STXD: Structural and Temporal Cross-Modal Distillation for Multi-View 3D Object Detection

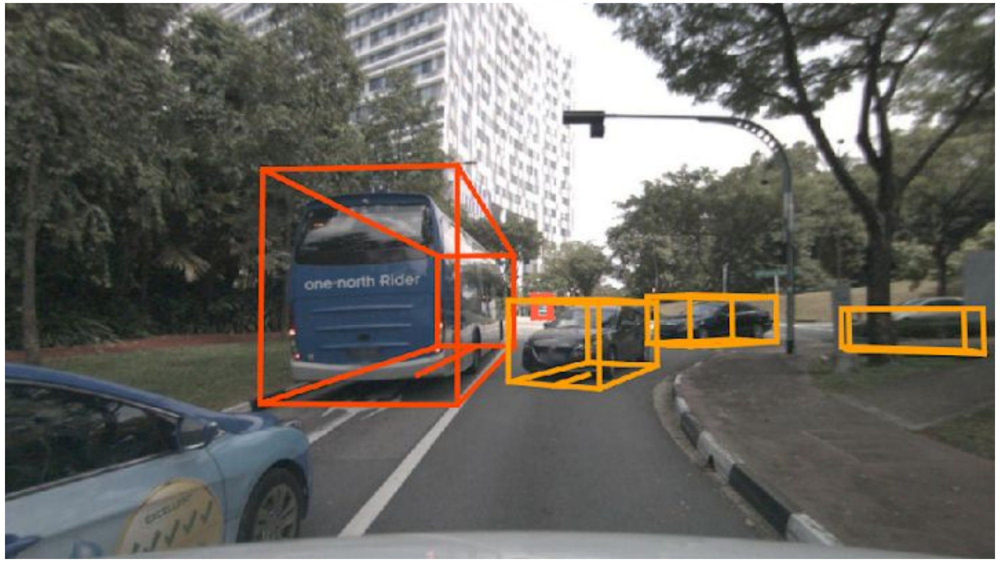
Sujin Jang*[1] Dae Ung Jo*[1] Sung Ju Hwang[2] Dongwook Lee[1] Daehyun Ji[1]

Samsung Advanced Institute of Technology (SAIT) [1]

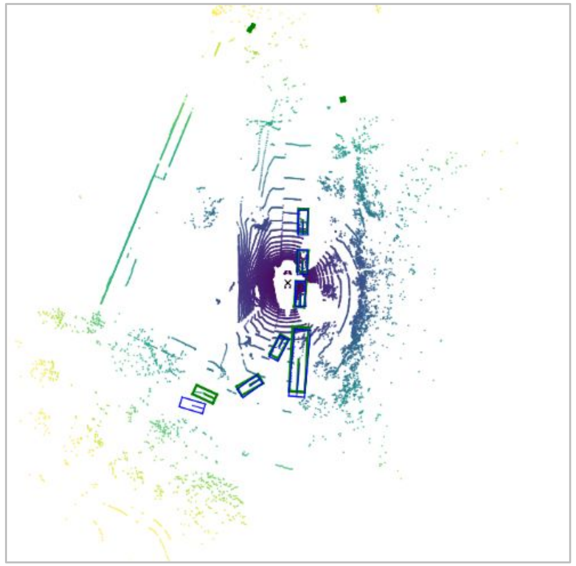
Korea Advanced Institute of Science and Technology (KAIST) [2]

3D Object Detection in Autonomous Driving

- Locate and classify various objects (e.g. car, truck, …) in 3D space
- From 3DOD, we can understand the surrounding environment
- Widely applied to various complex vision systems, such as autonomous driving



Multi-view camera



Bird’s-eye-view (BEV)

Cross-Modal Knowledge Distillation

Distillation

LiDAR-based

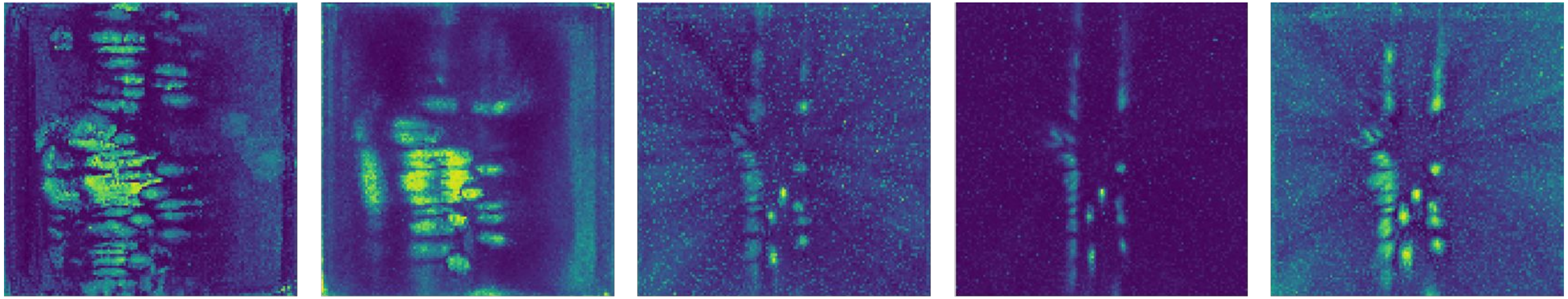
- ↑ Rich 3D information
- ↑ High performance in 3DOD
- ↓ Lack of color information
- ↓ Expensive

Camera-based

- ↑ Rich color information
- ↑ Low cost
- ↓ Lack of 3D information
- ↓ Low performance in 3DOD

Contributions

- We propose **STXD**, a cross-modal knowledge distillation framework from LiDAR to camera sensor for the multi-view 3D object detection
- **Correlation Regularizing Distillation (CD)** is introduced to prevent information collapse in student model arisen by modality gap between teacher and student sensors
- **Temporal Consistency Distillation (TD)** is introduced to leverage temporal knowledge embedded in features of previous frames



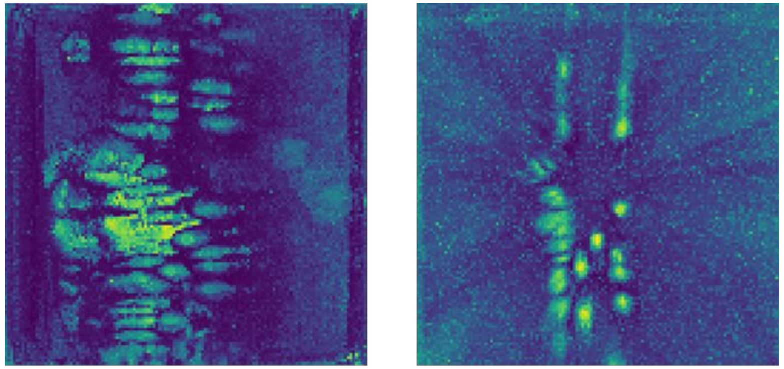
(a) LiDAR (b) Ours (c) MSE w/ GT (d) MSE (e) Camera

Correlation Regularizing Distillation (CD)

- In our empirical observation, CMKD causes **<u>information collapse</u>** in student features
- This is because features learned by different modalities are typically non-homogeneous.
- Thus, there exist distributional divergences between the feature spaces.

Method	d_{eff}	d^2_{eff}
MSE	3.810	14.519
MSE w/ GT	4.059	16.474
CD (Ours)	**4.389**	**19.267**
Teacher	5.757	33.137

d_{eff}: Effective dimension of student / teacher features



(a) LiDAR (e) Camera

Correlation Regularizing Distillation (CD)

- To mitigate information collapse problem, we introduce **<u>cross-correlation regularization</u>**
- Given LiDAR ($\mathbf{F}$) and camera ($\mathbf{G}$) BEV features,

$$\mathbf{C} = \hat{\mathbf{F}}^T \hat{\mathbf{G}} \in \mathbb{R}^{D \times D},$$

$$\mathcal{L}_{CD} := \underbrace{\sum_i (1 - \mathbf{C}(i,i))^2}_{\text{Maximize the similarity between aligned features}} + \lambda_c \underbrace{\sum_i \sum_{j \neq i} \mathbf{C}(i,j)^2}_{\text{Regularize cross-correlation along feature dimensions to prevent information collapse}}$$

Maximize the similarity between aligned features

Regularize cross-correlation along feature dimensions to **<u>prevent information collapse</u>**

Temporal Consistency Distillation (TD)

- To further enhance distillation quality, temporal information is also transferred

- To avoid spatial false matching across the time frames,
- Indirectly distills the teacher's information from past frames by a temporal similarity map

Temporal similarity map

$$\mathbf{T}^{(-k)} = \mathbf{F}^{(0)}\mathbf{F}^{(-k)^T} \in \mathbb{R}^{N\times N}, \quad k \in [1, K]$$
$$\mathbf{S}^{(-k)} = \mathbf{G}^{(0)}\mathbf{F}^{(-k)^T} \in \mathbb{R}^{N\times N}, \quad k \in [1, K]$$
$$\mathcal{L}_{TD} := \sum_k D_{KL}(\mathbf{S}^{(-k)} || \mathbf{T}^{(-k)})$$

Response-Level Distillation (RD)

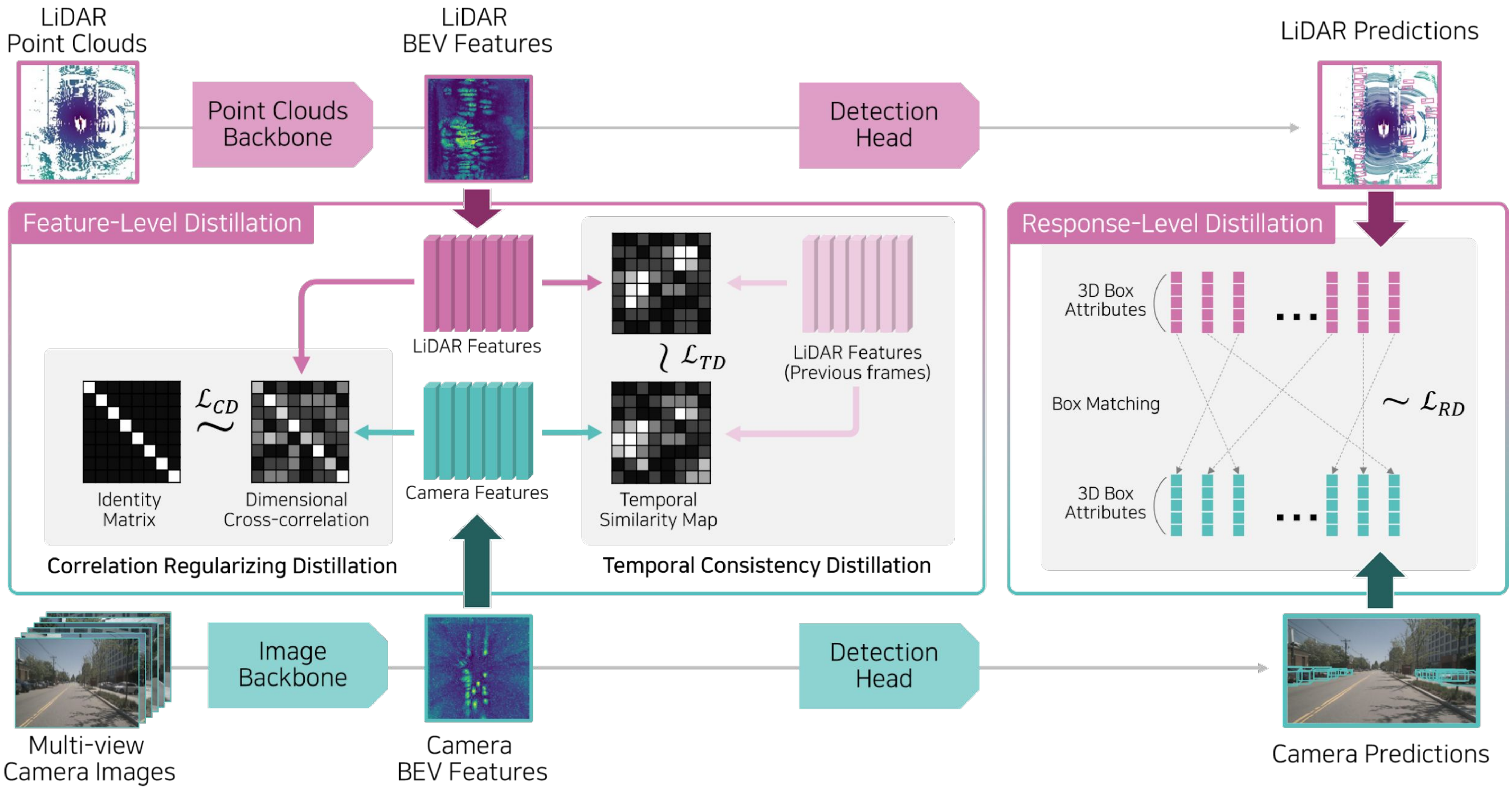
- Distillation is also performed on the predictions (responses)
- We applied a quality-based response-level distillation method

Quality of teacher's prediction

$$\boxed{q_i = \left(c_i^*\right)^{1-\gamma} \cdot \left(\mathrm{IoU}\left(\mathbf{b}_i^*, \mathbf{b}_i\right)\right)^{\gamma}}$$

$$\mathcal{L}_{RD} := \sum_j \boxed{q_{\pi(j)}} \cdot \left(\|\mathbf{b}_{\pi(j)} - \tilde{\mathbf{b}}_j\|_1 + D_{KL}(\mathbf{c}_{\pi(j)} || \tilde{\mathbf{c}}_j) \right)$$

Overall Framework



Evaluation on NuScenes Dataset

Method	Modality	NDS(%)	mAP(%)
BEVFormer† [34]	C	51.4	40.5
+BEVDistill [9]	L → C	52.4	41.7
+STXD (Ours)	L → C	**54.3** +2.9	**44.0** +3.5
UVTR-C [29]	C	44.1	36.2
+L2C [29]	L → C	45.0	37.2
+STXD (Ours)	L → C	**46.1** +2.0	**39.0** +2.8
UVTR-CS [29]	C	48.3	37.9
+L2CS [29]	L → C	48.8	39.2
+STXD (Ours)	L → C	**50.8** +2.5	**41.4** +3.5

Validation Set

Method	Modality	NDS(%)	mAP(%)
BEVFormer† [34]	C	52.6	42.4
+STXD (Ours)	L → C	**55.5** +2.9	**46.5** +3.9
BEVFormer‡ [34]	C	55.5	45.7
+STXD (Ours)	L → C	**58.3** +2.8	**49.7** +4.0
UVTR-C [29]	C	43.0	36.4
+L2C [29]	L → C	44.0	38.2
+STXD (Ours)	L → C	**45.8** +2.8	**40.2** +3.8
UVTR-CS [29]	C	48.6	39.0
+L2CS [29]	L → C	48.7	39.8
+STXD (Ours)	L → C	**51.8** +3.2	**43.5** +4.5

Test Set

Thank You for Your Attentions!

If you're interested, please visit our poster at

Poster Session 1
Tue 12 Dec 10:45 a.m. CST – 12:45 p.m. CST

@Great Hall & Hall B1+B2 #222