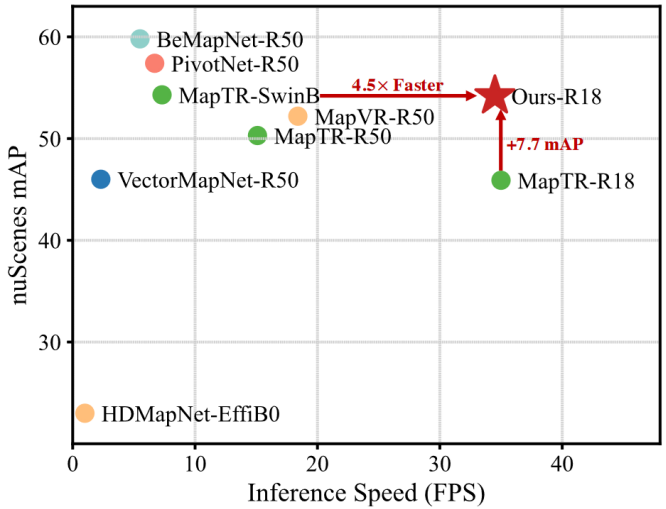
ECCV 2024

# MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation

Xiaoshuai Hao[1†], Ruikai Li[2†], Hui Zhang[1], Dingzhe Li[1], Rong Yin[3], Sangil Jung[4], Seung-In Park[4], ByungIn Yoo[4], Haimei Zhao[5§], and Jing Zhang[5§]

[1] Samsung R&D Institute China–Beijing
[2] State Key Lab of Intelligent Transportation System, Beihang University
[3] Institute of Information Engineering, Chinese Academy of Sciences
[4] Computer Vision TU, SAIT, SEC, Korea [5] The University of Sydney
{xshuai.hao, hui123.zhang, dingzhe.li, byungin.yoo}@samsung.com
ricky_developer@buaa.edu.cn {si14.park, sang-il.jung}@samsung.com
yinrong@iie.ac.cn hzha7798@uni.sydney.edu.au jingzhang.cv@gmail.com




- Problem/Objective
  - HD map construction
- Contribution/Key Idea
  - Camera-only(lightweight) based student model
  - Cross modality distillation을 위한 comprehensive 한 scheme
  - SOTA in HD map construction task

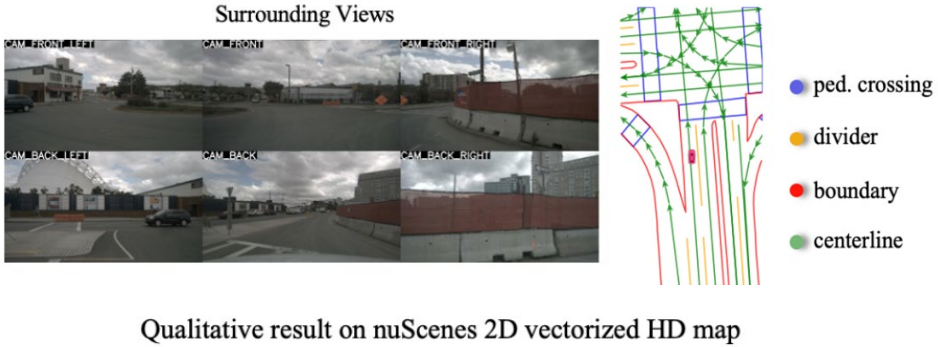
김범준

ECCV 2024

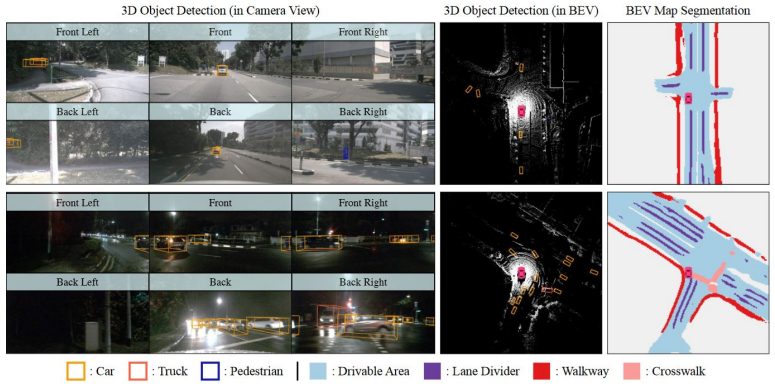
## MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation

*Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang*

- **HD map construction**
  - Driving scene과 navigation에 중요
  - BEV perception의 발전에 더욱 주목


<HD map construction>




<BEV map segmentation>

김범준

**MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation**

*Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang*

# Introduction

- Multi-view Camera
  - Lack of depth information
  - Cost-effective

    → KD(Knowledge Distillation) 을 사용하여 해결

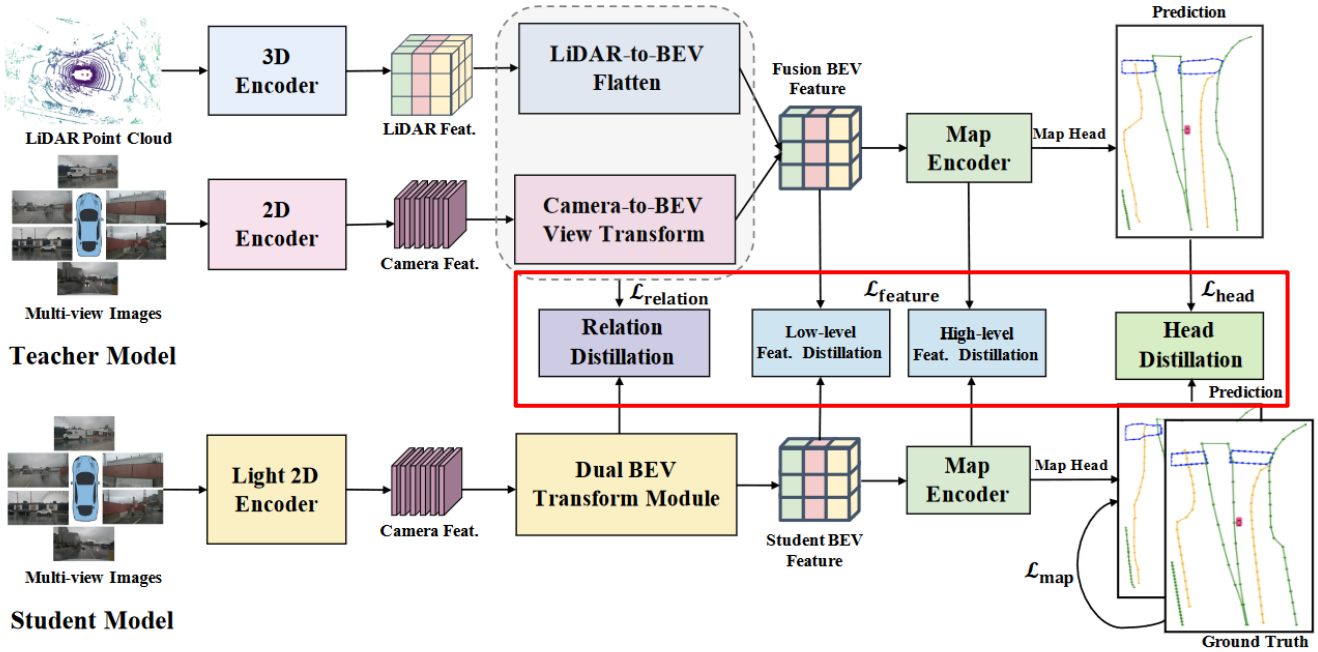
- 현존하는 BEV-based KD의 문제점
  - object detection task 기준으로 만들어진 모듈들
  - foreground object에 집중하기 위해 background의 adverse impact를 mitigate

    → 이런 task 간의 KD 방법 gap을 해결하기 위해
    - Relation distillation
    - Head Distillation

**MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation**

*Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang*

- **Method**

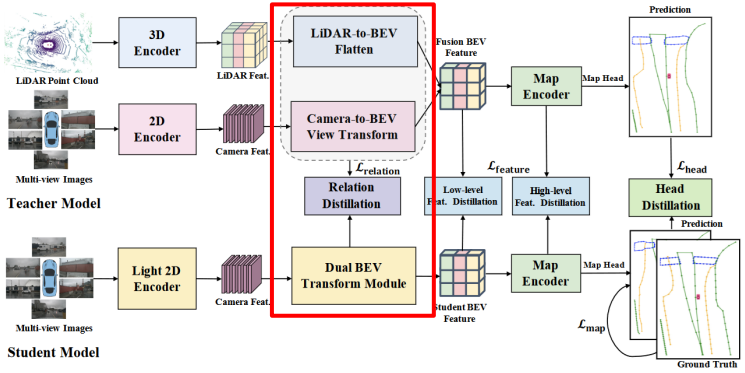



Teacher : MapTR / Student: MapTR(Camera Branch)

# MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation

*Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang*

ECCV 2024




deploy. On the base from MapTR, to mimic the multimodal fusion pipeline of the teacher model, we propose a Dual BEV Transform module to convert the multi-view features into two distinct BEV subspaces, whose effect will be verified

$$\mathbf{F}^T_{L_{bev}} \rightarrow A^T_{c2l} = \text{softmax}\left(\frac{\mathbf{Fp}^T_{C_{bev}} \text{Transpose}\left(\mathbf{Fp}^T_{L_{bev}}\right)}{\sqrt{D_k}}\right)$$

$$\mathbf{F}^T_{C_{bev}} \rightarrow A^T_{l2c} = \text{softmax}\left(\frac{\mathbf{Fp}^T_{L_{bev}} \text{Transpose}\left(\mathbf{Fp}^T_{C_{bev}}\right)}{\sqrt{D_k}}\right)$$

$$\mathbf{F}^S_{C_{sub1}} \rightarrow A^S_{c2l} = \text{softmax}\left(\frac{\mathbf{Fp}^S_{C_{sub1}} \text{Transpose}\left(\mathbf{Fp}^S_{C_{sub2}}\right)}{\sqrt{D_k}}\right)$$

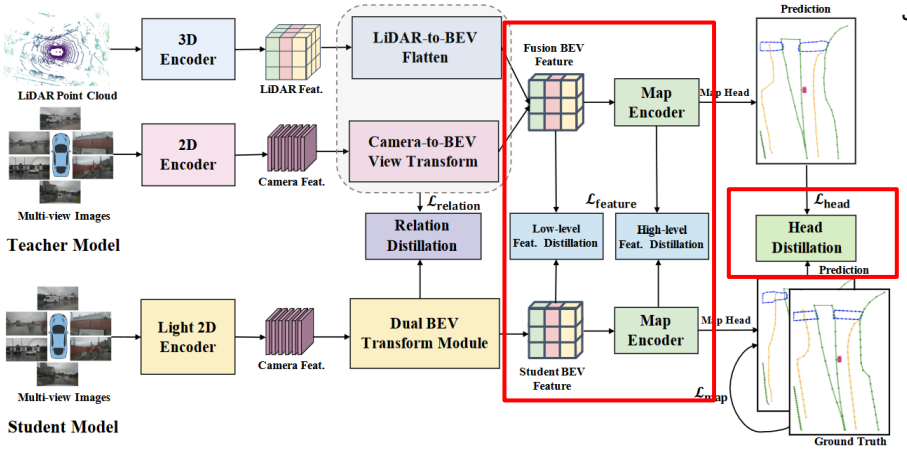
$$\mathbf{F}^S_{C_{sub2}} \rightarrow A^S_{l2c} = \text{softmax}\left(\frac{\mathbf{Fp}^S_{C_{sub2}} \text{Transpose}\left(\mathbf{Fp}^S_{C_{sub1}}\right)}{\sqrt{D_k}}\right)$$

$$\mathcal{L}_{relation} = D_{KL}(A^T_{c2l} || A^S_{c2l}) + D_{KL}(A^T_{l2c} || A^S_{l2c})$$

김범준

# MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation

*Xiaoshuai Hao, Ruikai Li, Hui Zhang, Jing Zhang*




$$\mathcal{L}_{low} = \mathrm{MSE}(\mathbf{F}^T_{fused}, \mathbf{F}^S_{fused})$$

$$\mathcal{L}_{high} = \mathrm{MSE}(\mathbf{F}^T_{high}, \mathbf{F}^S_{high})$$

$$\mathcal{L}_{feature} = \mathcal{L}_{low} + \mathcal{L}_{high}$$

$$\begin{aligned}\mathcal{L}_{head} &= \mathcal{L}_{cls} + \mathcal{L}_{point} \\ &= \mathcal{L}_{Focal}(\mathbf{F}^T_{cls}, \mathbf{F}^S_{cls}) + \mathcal{L}_{p2p}(\mathbf{F}^T_{point}, \mathbf{F}^S_{point})\end{aligned}$$

Manhattan distance

$$\mathcal{L} = \mathcal{L}_{map} + \lambda_1 \mathcal{L}_{relation} + \lambda_2 \mathcal{L}_{feature} + \lambda_3 \mathcal{L}_{head}$$

**MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation**

*Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang*

- **Experiments**

| Method | Student Modality | Teacher Modality | Backbone | Epochs | $AP_{ped.}$ | $AP_{div.}$ | $AP_{bou.}$ | mAP |
|---|---|---|---|---|---|---|---|---|
| HDMapNet [19] | C | – | Effi−B0 | 30 | 14.4 | 21.7 | 33.0 | 23.0 |
| VectorMapNet [25] | C | – | R50 | 110 | 36.1 | 47.3 | 39.3 | 40.9 |
| MapVR [47] | C | – | R50 | 24 | 47.7 | 54.4 | 51.4 | 51.2 |
| PivotNet [7] | C | – | R50 | 30 | 58.5 | 53.8 | 59.6 | 57.4 |
| BeMapNet [34] | C | – | R50 | 30 | 62.3 | 57.7 | 59.4 | 59.8 |
| MapTR [23] | C | – | R50 | 24 | 45.3 | 51.5 | 53.1 | 50.3 |
| MapTR [23] | L | – | Sec | 24 | 48.5 | 53.7 | 64.7 | 55.6 |
| MapTR [23] | C & L | – | R50 & Sec | 24 | 55.9 | 62.3 | 69.3 | 62.5 |
| MapTR [23] | C | – | R18 | 110 | 39.6 | 49.9 | 48.2 | 45.9 |
| BEV-LGKD† [18] | C | C | R18 | 110 | 42.2 | 47.6 | 49.7 | $46.5_{+0.6}$ |
| BEVDistill† [5] | C | L | R18 | 110 | 42.4 | 48.5 | 50.2 | $47.1_{+1.2}$ |
| UniDistill† [53] | C | C&L | R18 | 110 | 43.9 | 48.6 | 52.1 | $48.2_{+2.3}$ |
| MapDistill | C | C | R18 | 110 | 43.3 | 48.8 | 51.9 | $48.0_{+2.1}$ |
| MapDistill | C | L | R18 | 110 | 45.9 | 50.7 | 53.6 | $50.1_{+4.2}$ |
| **MapDistill** | **C** | **C & L** | **R18** | **110** | **49.2** | **54.5** | **57.1** | **$53.6_{+7.7}$** |

Camera-only

Baseline

KD

Tab. 1 shows that: (1) KD methods originally designed for BEV-based 3D object detection fail to achieve satisfying results due to task discrepancies between 3D object detection and HD map construction.

**MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation**

*Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang*

- **Experiments**

**Table 2:** Ablation study on the components in MapDistill.

| Setting | $\mathcal{L}_{relation}$ | $\mathcal{L}_{feature}$ | $\mathcal{L}_{head}$ | $\text{AP}_{ped.}$ | $\text{AP}_{div.}$ | $\text{AP}_{bou.}$ | mAP |
|---|---|---|---|---|---|---|---|
| Baseline | ✗ | ✗ | ✗ | 39.6 | 49.9 | 48.2 | 45.9 |
| a | ✓ | ✗ | ✗ | 44.1 | 49.7 | 52.4 | 48.8 |
| b | ✗ | ✓ | ✗ | 44.3 | 49.4 | 51.5 | 48.4 |
| c | ✗ | ✗ | ✓ | 44.2 | 50.1 | 52.7 | 49.0 |
| d | ✓ | ✓ | ✗ | 45.4 | 51.4 | 54.1 | 50.3 |
| e | ✗ | ✓ | ✓ | 46.3 | 51.8 | 54.3 | 50.8 |
| f | ✓ | ✗ | ✓ | 46.5 | 52.3 | 54.5 | 51.1 |
| g | ✓ | ✓ | ✓ | **49.2** | **54.5** | **57.1** | **53.6** |

# MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation

*Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang*

**(a)** Cross-modal relation distillation loss

| Method | $AP_{ped.}$ | $AP_{div.}$ | $AP_{bou.}$ | mAP |
|---|---|---|---|---|
| MapDistill (w/o $\mathcal{L}_{relation}$) | 46.3 | 51.8 | 54.3 | 50.8 |
| +Uni-modal Relation | 48.0 | 52.9 | 55.1 | 52.0 |
| +Hybrid Relation | 48.3 | 53.4 | 55.5 | 52.4 |
| **+Cross-modal Relation** | **49.2** | **54.5** | **57.1** | **53.6** |

**(b)** Dual-level feature distillation loss

| Method | $AP_{ped.}$ | $AP_{div.}$ | $AP_{bou.}$ | mAP |
|---|---|---|---|---|
| MapDistill (w/o $\mathcal{L}_{feature}$) | 46.5 | 52.3 | 54.5 | 51.1 |
| +Low-level (only) | 48.4 | 53.7 | 56.0 | 52.7 |
| +High-level (only) | 48.7 | 53.9 | 56.1 | 52.9 |
| **+Dual-level (ours)** | **49.2** | **54.5** | **57.1** | **53.6** |

**(c)** Map head distillation loss

| $\mathcal{L}_{cls}$ | $\mathcal{L}_{point}$ | $AP_{ped.}$ | $AP_{div.}$ | $AP_{bou.}$ | mAP |
|---|---|---|---|---|---|
| ✗ | ✗ | 45.4 | 51.4 | 54.1 | 50.3 |
| ✓ | ✗ | 47.3 | 52.8 | 55.3 | 51.8 |
| ✗ | ✓ | 47.1 | 53.0 | 55.6 | 51.9 |
| ✓ | ✓ | **49.2** | **54.5** | **57.1** | **53.6** |

**MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation**

*Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang*

**Table 4:** Ablation study of Dual BEV Transform Module.

| | subspace1 | subspace2 | $AP_{ped.}$ | $AP_{div.}$ | $AP_{bou.}$ | mAP |
|---|---|---|---|---|---|---|
| (a) | GKT | ✗ | 44.9 | 49.6 | 52.8 | 49.1 |
| (b) | LSS | LSS | 45.9 | 51.2 | 54.4 | 50.5 |
| | GKT | GKT | 46.7 | 51.6 | 54.5 | 50.9 |
| | Deform. | Deform. | 46.8 | 51.6 | 54.6 | 51.0 |
| (c) | GKT | Deform. | 47.1 | 53.2 | 56.2 | 52.1 |
| | Deform. | GKT | 47.3 | 53.4 | 56.1 | 52.3 |
| | LSS | Deform. | 48.9 | 53.9 | 56.2 | 53.0 |
| | Deform. | LSS | 48.7 | 53.8 | 55.9 | 52.8 |
| | LSS | GKT | 49.1 | 54.2 | 56.7 | 53.3 |
| | **GKT** | **LSS** | **49.2** | **54.5** | **57.1** | **53.6** |

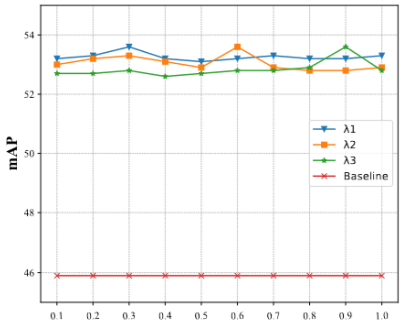



**Fig. 3:** Sensitivity of hyper-parameters.

**MapDistill: Boosting Efficient Camera-based HD Map Construction via Camera-LiDAR Fusion Model Distillation**

*Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang*

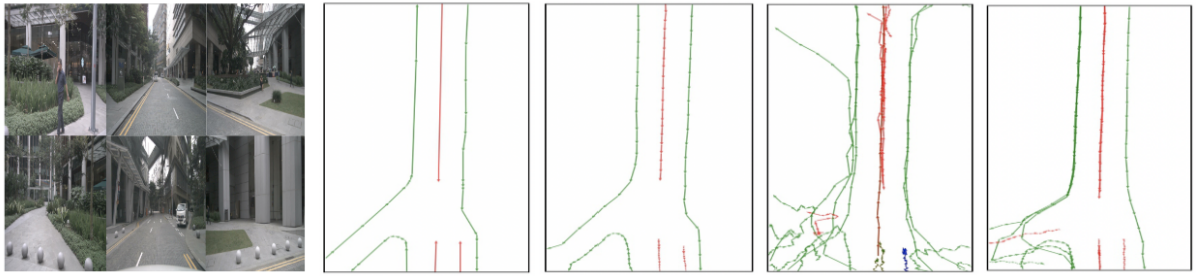



**Fig. 4:** Qualitative results on nuScenes val set. (a) Six camera inputs. (b) Ground-truth vectorized HD map. (c) Result of the camera-LiDAR-based teacher model. (d) Result of the camera-based student model without MapDistill (Baseline). (e) Result of the camera-based student model with MapDistill. MapDistill helps correct substantial errors in the Baseline's predictions and improves its accuracy.