# UniDistill: A Universal Cross-Modality Knowledge Distillation Framework for 3D Object Detection in Bird's-Eye View
# ——Supplementary Material——

Shengchao Zhou[1*] Weizhou Liu[1*] Chen Hu[1†] Shuchang Zhou[1] Chao Ma[2]

[1] MEGVII Technology

[2] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

{zhoushengchao,liuweizhou,huchen,zsc}@megvii.com, chaoma@sjtu.edu.cn

This supplementary material contains additional details of the main manuscript. Section 1 presents additional details of the models and training strategies. Section 2 complements more experiments not included in the main manuscript. Section 3 shows more visualization results to prove the effectiveness of UniDistill.

## 1. Details of Models and Training

To prove the effectiveness of UniDistill, we introduce BEVDet, CenterPoint and BEVFusion as the camera based, LiDAR based and LiDAR-camera based detectors. For BEVDet, the features of input images are firstly extracted by the backbone of ResNet-50 and then projected to BEV through LSS [4]. We set the projected features to be the low-level BEV features $\boldsymbol{F}^{\text{low}}_{\text{cam}}$. With respect to CenterPoint, the input LiDAR points are distributed to regular voxels and the features of each voxel are extracted by 3D convolution. Then, the features of voxels in the same column are concatenated and we set the result as $\boldsymbol{F}^{\text{low}}_{\text{ldr}}$. BEVFusion builds on the above detectors by concatenating $\boldsymbol{F}^{\text{low}}_{\text{cam}}$ with $\boldsymbol{F}^{\text{low}}_{\text{ldr}}$ and then processing it with a fully convolutional network (FCN). The output features are set to be $\boldsymbol{F}^{\text{low}}_{\text{fuse}}$. The following steps are the same for different detectors, where a FCN follows as an encoder to produce $\boldsymbol{F}^{\text{high}}$ and then a detection head of CenterPoint generates classification and regression heatmaps. These heatmaps are used to form $\boldsymbol{F}^{\text{resp}}$.

For all detectors, during training, the detection head will calculate a classification loss $\mathcal{L}_{\text{Cls}}$ and a regression loss $\mathcal{L}_{\text{Reg}}$ that are combined to form the detection loss $\mathcal{L}_{\text{Det}}$. In Section 4.2 of the main manuscript, to help the detectors perform better, we use auto-scaling to balance the scales between $\mathcal{L}_{\text{Cls}}$ and $\mathcal{L}_{\text{Reg}}$ but turn it off in Section 4.3 for efficiency.

The training of detectors is finished on 20 GeForce RTX 2080Ti GPUs. These GPUs are distributed on 5 machines, where each machine has 4 GPUs, so that we adopt distributed training. Because of the limited memory, each GPU is distributed with 1 training sample.

Table 1. Comparison between UniDistill and BEVDepth on testing dataset to show the advantages of knowledge distillation. "L" and "C" represent LiDAR and camera.

| Method | Teacher Modality | mAP ↑ | mASE ↓ | mAOE ↓ | NDS ↑ |
|---|---|---|---|---|---|
| Baseline | - | 26.4 | 26.6 | 55.8 | 36.1 |
| BEVDepth [2] | - | 28.4 | 26.3 | 55.3 | 37.7 |
| UniDistill | L | **28.9** | **25.9** | **51.4** | **38.4** |
| UniDistill | L+C | **29.6** | **25.7** | **49.2** | **39.3** |

## 2. Complementary Experiments

In this section, experiments not included in the main manuscript are complemented. In Section 2.1, UniDistill is compared with BEVDepth [2] and MonoDistill to show its advantages. In Section 2.2, the performance of UniDistill on Waymo is evaluated to show its generalization to different datasets. In Section 2.3, we replace the detection head with a TransFusion [1] based one and the backbone of BEVDet to Swin Transformer [3] to show the generalization to different architectures. In Section 2.4, more ablation studies about the adaptive layers and feature distillation are supplemented. In Section 2.5, the training time and memory usage of UniDistill are listed.

### 2.1. Comparison with BEVDepth and MonoDistill

To transfer the depth knowledge of LiDAR points to the camera based detector, which is BEVDet in our experiments, UniDistill introduces knowledge distillation for help. BEVDepth provides another approach for knowledge transfer by supervising the depth prediction of LSS in BEVDet with ground truth generated by projecting LiDAR points to the perspective view. Therefore, we compare Uni-

[*] Equal Contribution

[†] Corresponding Author

Table 2. Comparison between UniDistill and MonoDistill on nuScenes test dataset. “L” and “C” represent LiDAR and camera.

| Method | Modality | Teacher Modality | mAP ↑ | mASE ↓ | NDS ↑ |
|---|---|---|---|---|---|
| MonoDistill | C | L | 23.2 | 28.7 | 34.3 |
| UniDistill | | | **28.9** | **25.9** | **38.4** |

Table 3. Analysis to show the generalization of UniDistill to Waymo dataset. “L” and “C” represent LiDAR and camera.

| Method | Modality | Teacher Modality | mAPL ↑ | mAPH ↑ | mAP ↑ |
|---|---|---|---|---|---|
| UniDistill | L+C | - | 71.0 | 71.4 | 75.3 |
| | C | - | 22.3 | 33.0 | 34.5 |
| | C | L+C | **24.5** | **36.2** | **37.7** |

Distill with BEVDepth to show the advantages of knowledge distillation. We build BEVDepth based on BEVDet by combining the detection loss $\mathcal{L}_{\text{Det}}$ with another depth prediction loss $\mathcal{L}_{\text{Depth}}$ and train it with the full training dataset. The performance of BEVDepth on the testing dataset is in Table 1. From the results, UniDistill helps BEVDet obtain better performance than BEVDepth, showing the advantages of knowledge distillation.

MonoDistill is another knowledge distillation framework that transfers the knowledge from a LiDAR-based teacher to a camera-based student. It directly unifies the architecture of the teacher and student by training the teacher with LiDAR points projected to the perspective view. Therefore, we further compare UniDistill with MonoDistill and the results are listed in Table 2, showing the better performance of UniDistill for the modality combination (C, L).

## 2.2. Generalization to Waymo

In the main manuscript, all experiments are conducted on the nuScenes dataset. To show the generalization of UniDistill to different datasets, in this section, we further evaluate its performance on Waymo dataset. Specifically, UniDistill is first trained on the Waymo-mini dataset for 18 epochs and then tested on the whole validation set. The results in distillation path (2) are listed in Table 3, showing the effectiveness and generalization of UniDistill on Waymo.

## 2.3. Generalization to More Architectures

In the main manuscript, we set the detection head of all detectors to be the same as that of CenterPoint. Therefore, we substitute it with a TransFusion based one and re-evaluate UniDistill to show the generalization of UniDistill to other detection heads. The evaluation is conducted in distillation path (1) on the validation dataset and the modified detectors are trained on 1/2 training dataset for efficiency. Since the response distillation in UniDistill is not applicable to the TransFusion head, we only leverage the feature distillation and relation distillation for knowledge transfer. The results in Table 4 reveal that UniDistill also improves the performance of the student detector, showing its generalization to different detection heads.

Table 4. Performance analysis to show the generalization of UniDistill to TransFusion. “L” and “C” represent LiDAR and camera.

| Method | Modality | Teacher Modality | mAP ↑ | mASE ↓ | NDS ↑ |
|---|---|---|---|---|---|
| TransFusion [1] | L+C | - | 63.4 | 25.2 | 67.6 |
| | L | - | 58.5 | 27.2 | 63.4 |
| UniDistill | L | L+C | **60.9** | **25.9** | **65.9** |

Table 5. Analysis to show the generalization of UniDistill to Swin Transformer. “L” and “C” represent LiDAR and camera.

| Method | Modality | Teacher Modality | mAP ↑ | mASE ↓ | NDS ↑ |
|---|---|---|---|---|---|
| UniDistill | L+C | - | 63.3 | 24.7 | 69.0 |
| | C | - | 27.8 | 27.6 | 36.0 |
| | C | L+C | **32.5** | **25.7** | **39.7** |

In addition, we substitute the ResNet-50 in BEVDet with Swin Transformer to show the generalization of UniDistill to different backbones. For efficiency, the modified detectors are trained on 1/2 training dataset and then evaluated in distillation path (2) on the validation dataset. The results in Table 5 show that UniDistill improves the performance of student detector and generalizes to different backbones.

## 2.4. Additional Ablation Studies

In Section 4.3.4 of the main manuscript, we conduct experiments to show that when evaluating in distillation path (3), the adaptive layers can avoid the performance degradation of the student after knowledge distillation. Some experiments in distillation path (4) are further designed to show that when the teacher detector performs better than the student, adopting the adaptive layers will decrease the effectiveness of UniDistill. The results are listed in Table 6 and reveal that with the adaptive layers, the performance of the student slightly decreases. Therefore, when the teacher performs better than the student, there is no need to introduce the adaptive layers.

We also compare the detection loss $\mathcal{L}_{\text{Det}}$ with/without the adaptive layers and the baseline is the student without UniDistill. The results in Figure 1 show that with the adaptive layers, although the detection loss is lower than the baseline, it is always higher than that without the adaptive layers. We think the problem results from that the adaptive layers make it too free for the student to choose what to learn from the teacher. However, since the teacher detector is strong enough to instruct the student, directly aligning the features of the student with teacher can help the student learn better.

In addition, since most of the ablation studies are con-

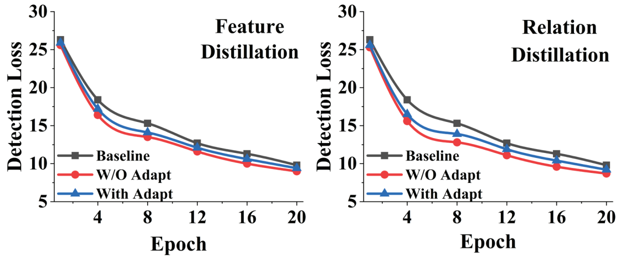

Figure 1. Illustration to show that the adaptive layers increase the detection loss when the teacher performs better than the student.

Table 6. Ablation study in path (4) to show that the adaptive layers decrease the effectiveness of feature distillation and relation distillation when the teacher detector performs better than student.

| Method | $\mathcal{L}_{\text{Fea}}$ | | | $\mathcal{L}_{\text{Rel}}$ | | |
|---|---|---|---|---|---|---|
| | mAP ↑ | mAVE ↓ | NDS ↑ | mAP ↑ | mAVE ↓ | NDS ↑ |
| Baseline | 20.3 | 95.2 | 33.1 | 20.3 | 95.2 | 33.1 |
| With Adapt | 20.7 | 91.4 | 33.9 | 21.2 | 89.6 | 34.2 |
| W/O Adapt | **21.1** | **88.5** | **34.3** | **21.7** | **84.5** | **35.0** |

ducted in path (4), we complement the ablation studies in path (1) to improve the reliability. As in Section 4.3.1 of the main manuscript, we compare the original feature distillation with two modified ones that align the low-level BEV features (1) completely or (2) inside a Gaussian-like mask. The results are listed in Table 7 and we can get the same conclusion that feature distillation performs better when selecting 9 crucial points for alignment.

## 2.5. Training Time and Memory Usage

In this section, the training time and memory usage of detectors with/without UniDistill are listed. The detectors are trained on 1 GeForce RTX 2080Ti GPU and the training batch size is 1. For the training time, we list the average time to calculate the training loss. With respect to memory usage, we report the max allocated memory during training. The results are illustrated in Table 8 and show that UniDistill will increase the training time and memory usage a lot. Therefore, we plan to introduce the block-wise distillation and other techniques to accelerate the training of UniDistill and decrease its memory usage.

# 3. More Visualization Results

In this section, we provide more visualization results to show the effectiveness of UniDistill. For the response features of one teacher detector and one student with/without UniDistill, we calculate the mean along the channel dimension and visualize them. The results of the LiDAR-camera based teacher and the camera based student are illustrated in Figure 2 and that of the LiDAR-camera based teacher and the LiDAR based student are in Figure 3. From the results, it is revealed that with UniDistill, the background areas are suppressed and the boundaries between objects are more clear. Therefore, there will be fewer false positive predictions and the detection performance is improved.

Table 7. Ablation study in path (1) to show that feature distillation performs better when selecting crucial points for alignment.

| Method | AP ↑ | | | | | NDS ↑ |
|---|---|---|---|---|---|---|
| | car | truck | ped | motor | mean | |
| Baseline | 82.8 | 52.0 | 76.4 | 54.2 | 53.5 | 63.9 |
| Complete | 82.4 | 52.1 | 77.4 | 56.8 | 54.3 | 64.2 |
| Gaussian | **84.7** | **54.3** | 76.1 | 53.4 | 54.7 | 64.8 |
| Crucial | 82.9 | 50.5 | **82.4** | **61.7** | **56.1** | **65.5** |

Table 8. Training time and memory usage of the detectors. "L" and "C" represent LiDAR and camera respectively.

| Modality | Teacher Modality | Training Time (s) | Memory Usage (GB) |
|---|---|---|---|
| L+C | - | 0.27 | 5.96 |
| C | - | 0.13 | 4.60 |
| C | L | 0.33 (+153%) | 5.07 (+0.47) |
| C | L+C | 0.40 (+207%) | 6.44 (+1.84) |
| L | - | 0.22 | 3.21 |
| L | C | 0.46 (+109%) | 4.51 (+1.30) |
| L | L+C | 0.53 (+140%) | 5.63 (+2.42) |

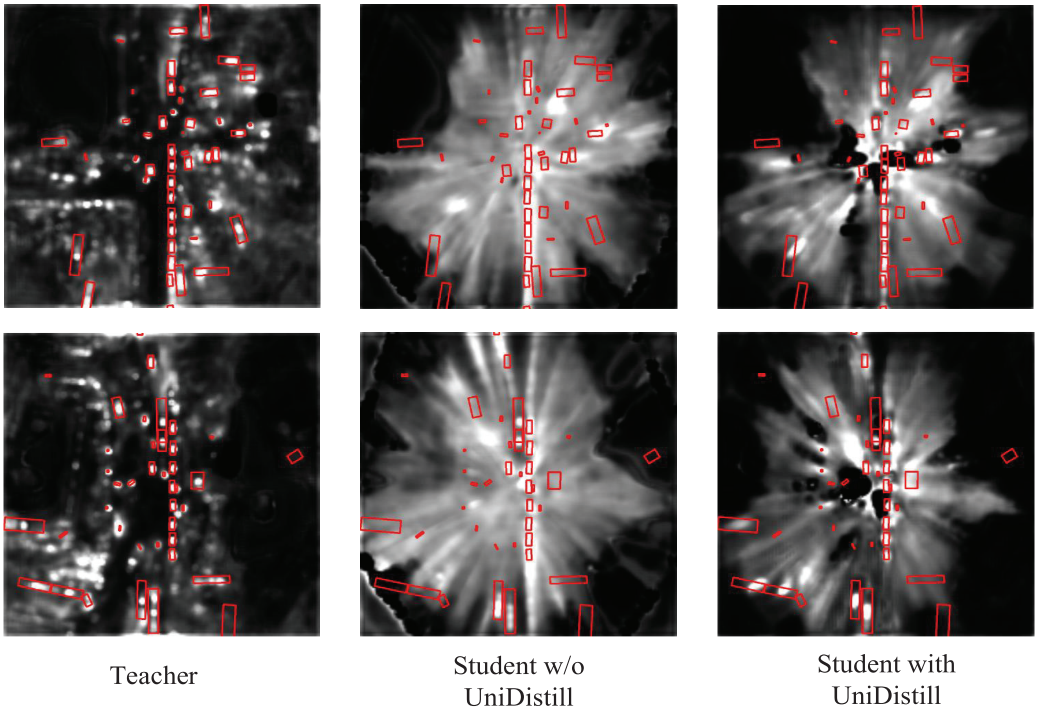

Figure 2. Visualization of the response features. The boxes in red are the ground truth bounding boxes. The teacher and student detectors are LiDAR-camera based and camera based respectively. The first and second rows represent the results of two scenes. With UniDistill, the background areas are suppressed and the object boundaries are more clear.

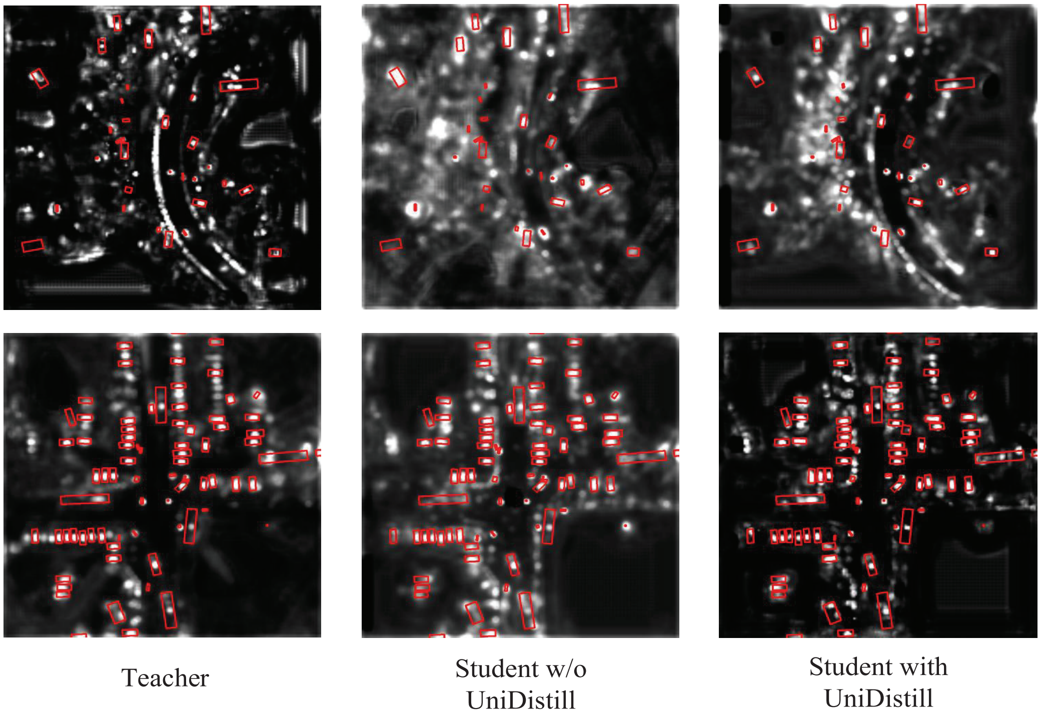

Figure 3. Visualization of the response features. The boxes in red are the ground truth bounding boxes. The teacher and student detectors are LiDAR-camera based and LiDAR based respectively. The first and second rows show the results of two scenes. With UniDistill, the background areas are suppressed and the object boundaries are more clear.

## References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022. 1, 2

[2] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 1

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[4] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210, 2020. 1