Supplementary

A. Reproducibility

All the source code, figures, and models will be available at https://github.com/yo0oenu/LGKD-SS

B. Linguistic knowledge extraction

In this section, we explain different language sets for text guidance used in teacher model training. We leverage a LLM to construct diverse text guidance sets. Specifically, the LLM is implemented with the following prompt instructions for the *Sentence* type: "All class in an urban driving scene are "Building", "Bicyclist", "Fence", "Pole", "Pedestrian", "Road", "Sky", "Sign Symbol", "Sidewalk", "Tree", "Car". Tell me about the general sentence that describe each class." For the *Keywords* type, the LLM is prompted as follows: "All class in an urban driving scene are "Building", "Bicyclist", "Fence", "Pole", "Pedestrian", "Road", "Sky", "Sign Symbol", "Sidewalk", "Tree", "Car". Tell me about the five general words that describe each class."

B.1. Label

- **Sky**: "Sky"
- **Building**: "Building"
- **Pole**: "Pole"
- **Road**: "Road"
- **Sidewalk**: "Sidewalk"
- **Tree**: "Tree"
- **SignSymbol**: "SignSymbol"
- **Fence**: "Fence"
- **Car**: "Car"
- **Pedestrian**: "Pedestrian"
- **Bicyclist**: "Bicyclist"

B.2. Sentence

We used the class names as the simplest form of text guidance.

- **Sky**: "The area above the earth, usually blue during the day with clouds."
- **Building**: "A tall structure made of concrete and glass that houses offices and apartments"
- **Pole**: "A vertical post used to support electrical wires or street lamps."
- **Road**: "A paved surface for vehicles to drive on, often marked with lanes."
- **Sidewalk**: "A paved path beside the road for people to walk safely."
- **Tree**: "A tall plant with a trunk, branches, and leaves providing shade."
- **SignSymbol**: "A traffic sign or used to guide or warn drivers and pedestrians."
- **Fence**: "A wooden or metal barrier that encloses a yard or property."
- **Car**: "A motor vehicle with four wheels used for transportation."
- **Pedestrian**: "A person walking on the sidewalk or crossing the street."
- **Bicyclist**: "A person wearing a helmet riding a bicycle on the street."

B.3. Words

- **Sky**: "Sky", "Cloud", "Blue", "Sun", "Daylight"
- **Building**: "Building", "Concrete", "Window", "Foundation", "Architecture"
- **Pole**: "Pole", "Wood", "Barrier", "Boundary", "Post"
- **Road**: "Road", "Drive", "Path", "Route", "Street"
- **Sidewalk**: "Sidewalk", "Pavement", "Path", "Walkway", "Walk"
- **Tree**: "Tree", "Branch", "Leaf", "Forest", "Wood"
- **SignSymbol**: "SignSymbol", "Traffic", "Board", "Sign", "Marker"
- **Fence**: "Fence", "Wood", "Barrier", "Boundary", "Post"
- **Car**: "Car", "Wheel", "Door", "Engine", "Tire"
- **Pedestrian**: "Pedestrian", "Walk", "Foot", "Step", "Crossing"
- **Bicyclist**: "Bicyclist", "Rider", "Wheels", "Cycling", "Path"

C. Experiments

C.1. Datasets

For the constrained area considering the limited field of view, We applied invalid black areas at the edge of the image plane considering limited field of view constraint for real-world environment. We applied zero-padding to the boundary regions of all input images while the ground truth contains class information for every pixel.

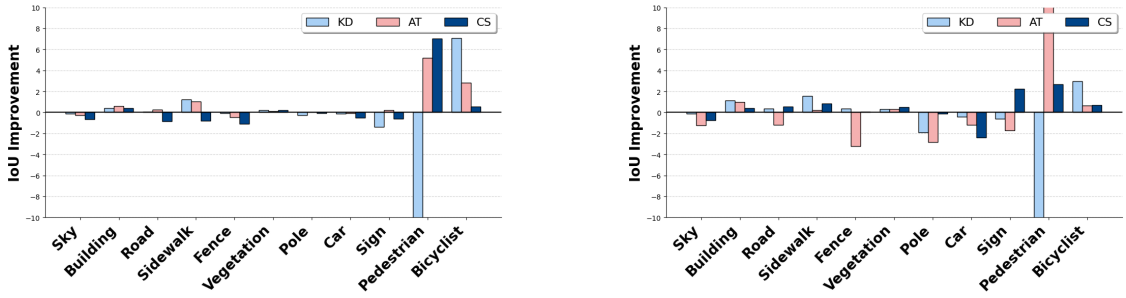
C.2. Implementation details

We explain more details for training teacher and student models, as well as baselines for KD.

The data augmentation as the teacher model training is implemented for both datasets. We trained the Student Model using AdamW with an initial learning rate of 6×10^{-5}weight decay of 0.01, and a poly scheduler with 1,500 warm-up iterations in all KD and Student only training.

We set hyperparameters for λ_{KD}, λ_{AT}, λ_{SP}, λ_{Proj}, and λ_{CS} to 0.01, 10000, 1.0, 0.01, and 0.1, considering their optimal performance.

All experiments are conducted on a desktop with an AMD Ryzen Threadripper PRO 5955WX CPU (16 cores, 32 threads, 3.6GHz), 256GB RAM, and an NVIDIA GeForce RTX 4090 GPU with 24 GB memory.



(a) Distillation with Image-Only Teacher

(b) Distillation with Sentence Teacher

Figure 8. Per-class IoU improvement over the student model on KITTI.

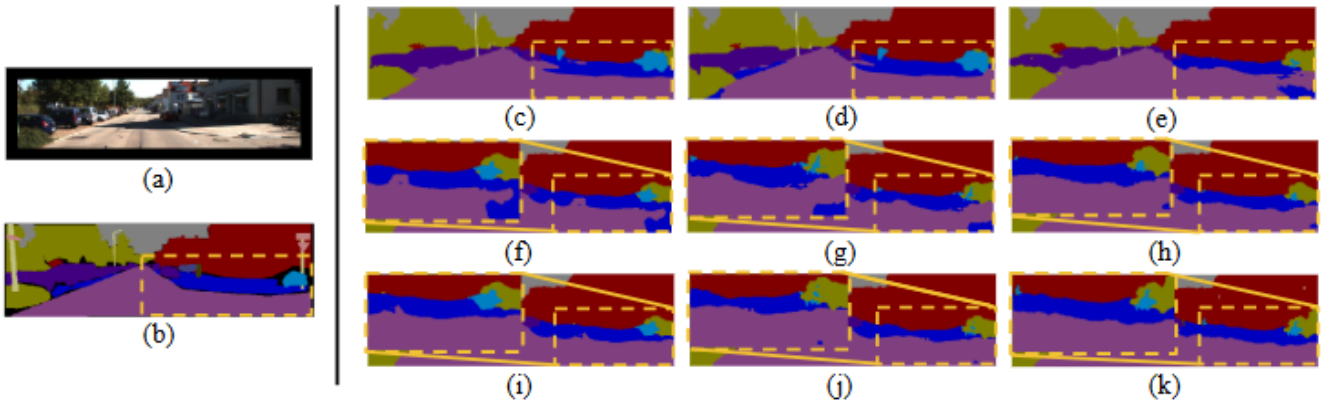


Figure 9. Results of semantic segmentation prediction on the KITTI dataset. (a) Input Image under limited FoV. (b) Ground Truth (GT). (c) Image-Only Teacher. (d) Sentence Teacher. (e) Learned from scratch (Student). (f) KD with Image-Only Teacher. (g) AT with Image-Only Teacher. (h) CS with Image-Only Teacher. (i) KD with Sentence Teacher. (j) AT with Sentence Teacher. (k) CS with Sentence Teacher

C.3. Distillation with language guided teacher

In this section, we provide additional results of per-class IoU improvement over the student that is learned from scratch, as shown in Figure 8. In Figure 8 (a), when distilling knowledge from the *Image-Only* Teacher, the AT method exhibits superior performance across most classes compared to CS. In contrast, when transferring knowledge from a teacher guided by linguistic information (Figure 8(b)), AT degrades the performance for most classes, whereas CS achieves larger IoU improvements than AT and also compared to when distilling from *Image-Only* Teacher.

Consistent with the CamVid results, when distilling knowledge from a teacher trained solely on visual information, distillation methods that directly match the teacher's spatial information, such as AT, demonstrate superior performance. However, when the teacher is guided by rich linguistic information, the semantic relationships between objects in the teacher's knowledge are enhanced. Consequently, AT that struggles to utilize this semantic information, suffers a performance degradation. This enhancement in semantic knowledge allows the performance of CS to improve compared to when distilling the *Image-Only* Teacher.

D. Additional analysis and ablations

D.1. Visualization

We plot additional visualizations including semantic segmentation results and feature relationship maps for qualitative evaluations.

D.1.1. Segmentation result

We visualize the semantic segmentation results of various models on KITTI in Figure 9. Although (c) provides cleaner segmentation results compared (d), (d) alone was able to predict the sidewalk underneath the car on the left, where all other models failed. This indirectly implies that the Text guidance facilitated this recognition by reflecting the relationship between "Road" and "Sidewalk".

For the student results, a common observation is that not only (e), but all students ((f), (g) and (h)), which were distilled from the Image-Only Teacher, show poor performance in the limited FoV condition within the bottom-right yellow dashed area. In contrast, students which were distilled from the Sentence Teacher ((i), (j) and (k)) show better prediction. Especially, (k) shows superior segmentation result compared to all other students.

These results imply that since the knowledge formed by the language guided vision teacher includes object-level semantics, it yields superior segmentation predictions in limited area compared to knowledge derived solely from vision information. Furthermore, to effectively leverage these rich semantic cues, similarity-based distillation strategies, such as CS are particularly effective.

D.1.2. Feature relationship

We also visualize channel similarity maps for CS and batch similarity maps for SP on KITTI dataset.

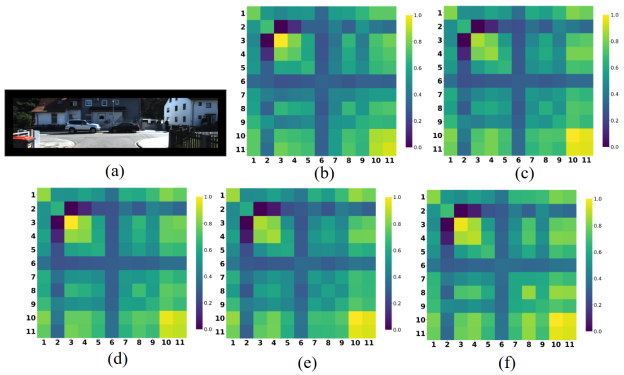


Figure 10. Visualization of channel similarity maps for CS on KITTI dataset. (a) Input Image. (b) Learned from scratch (Student). (c) Student from Image-Only Teacher. (d) Student from Label Teacher. (e) Student from Words Teacher. (f) Student from Sentence Teacher

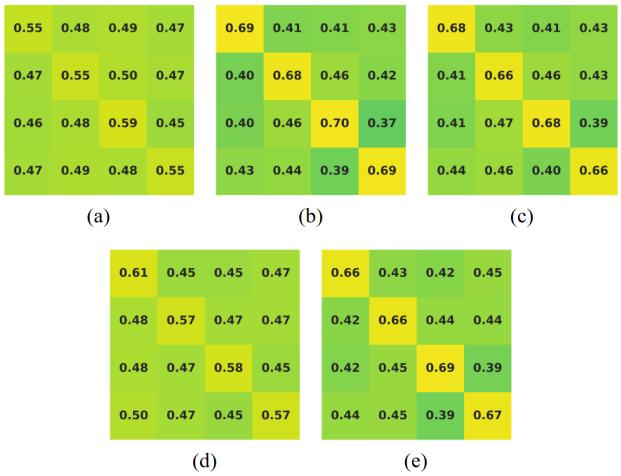


Figure 11. Visualization of batch similarity maps for SP on KITTI dataset. (a) Learned from scratch (Student). (b) Student from Image-Only Teacher. (c) Student from Label Teacher. (d) Student from Words Teacher. (e) Student from Sentence Teacher.

Figure 10 shows that (f) exhibits more stronger relationships for specific class pairs compared to others. For instance, the relationships between class 4 and 9 (sidewalk and sign) and class 8 and 11 (car and bicyclist) are reinforced in (f). This confirms that rich linguistic guidance enhances the relationships between objects for the text-guided teacher. Furthermore, it indicates that the CS distills these relationships into the Student, enabling it to learn semantic cues formed by linguistic information.

In Figure 11, Similar to the observations on CamVid, (a) shows indiscernible similarity patterns, whereas (b) exhibits the highest diagonal values compared to other models, demonstrating a relatively strict distinction between similar and dissimilar samples. (e), which achieved the best SP performance on the KITTI dataset by distilling knowledge from a teacher trained with rich linguistic guidance, outperforms (c) and (d), which relied on ambiguous text guidance. The results confirm that sentence-level text guidance establishes more precise semantic bridges between samples by semantically capturing detailed relationships within the visual information.

E. Discussion

In this paper, we explored the role of text guidance during knowledge distillation process with teacher models trained with text guidance in limited field of view environments. Since vision-language based teacher including diffusion module is adopted, the teacher takes a lot of time and computational resources that is explained in Section 4.2 and 5.3. To consider lightweight model generation, we adopt student model with Segformer. With this, teacher and student have dissimilar architecture styles and implementations, which make large gap of knowledge between the models. Also, because the combination has much different capacities, this causes hindrance in distillation process. As explained with the results in Table 3 and 4, along with extreme capacity differences (the capacity of the student is less than 1% compared to the one of the teacher), we confirmed that effective knowledge distillation is difficult with conventional logit matching methods alone, denoted as KD (standard KD). Also, the results on distilled students prominently showed that a better teacher does not generate a better student, as investigated in the prior study [8]. Thus, these findings present using representations well aligned with the teacher's knowledge is crucial in the KD process.

Experimental results showed that when the Teacher was trained primarily on visual information, AT, which directly transfers spatial information, and Proj, based on projection layers, were effective. Conversely, when the Teacher structured semantic relationships between visual elements through Text-Vision multimodal learning, particularly as linguistic information became richer, these relationships were reinforced, and most of the approaches showed performance degradation. In such cases, we confirmed that CS and SP methods leveraging semantic cue were more effec-

tive than other methods such as response-based or feature-based methods.

We believe that our observations can provide a new insight in generating a lightweight model through multimodal KD for autonomous driving and driving scene understanding.

As future work, we aim to develop an advanced KD methodology that can simultaneously transfer spatial information and semantic knowledge when distilling from Text-Image multimodal teachers to uni-modality students even under extreme architectural differences. With this, more superior and lightweight model can be generated, which can operate efficiently in autonomous driving environments with limited field of view.