Multimodal Material Segmentation
Supplementary Material

Yupeng Liang Ryosuke Wakaki Shohei Nobuhara Ko Nishino
Graduate School of Informatics, Kyoto University
https://vision.ist.i.kyoto-u.ac.jp/

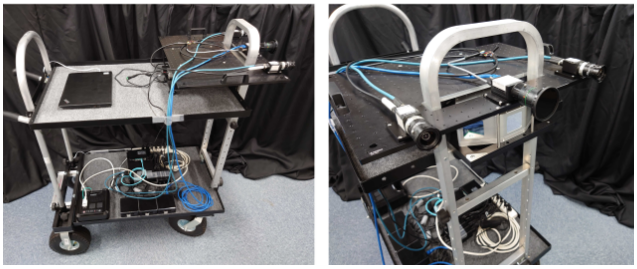

Figure 1. Our image capture system consists of a stereo pair of RGB-polarization cameras, an NIR camera between them, and a LiDAR below the three cameras, all mounted on a push cart roughly at the height of the hood of a car.

A. Image Capture System

Figure 1 shows the multi-imaging system we used to capture the MCubeS dataset.

B. Modified DDF and DRConv

We compare MCubeSNet with DDF and DRConv by modifying them to use semantic segmentation as the guidance field. Semantic segmentation is computed on the input image independent from the material segmentation. As such, the guided feature generation process of DRConv can safely be deleted; we delete the learnable guided mask branch and substitute the guided feature with our semantic guidance mask. Other parts, including the filter generator module, are the same. For DDF, we retain the spatial filter branch and channel filter branch. The original DDF generates filters for each pixel and can only be applied to a layer whose input channel size is equal to the output channel size. In contrast, we need region-wise filters rather than pixel-wise filters and the first layer of the decoder has different numbers of input and output channels. We introduce the semantic guidance mask into the spatial dynamic filters and compute the average of spatial filters in each semantic region as the new spatial filter. An output channel branch with a similar structure to the channel branch is also added to DDF. The output filter size is $oc \times 1 \times 1$, where o is the output channel size and c is the input channel size. The new spatial filters, channel filters, and output filters are multiplied to generate the convolution kernel for each region. Finally, we apply standard convolution in each semantic region.

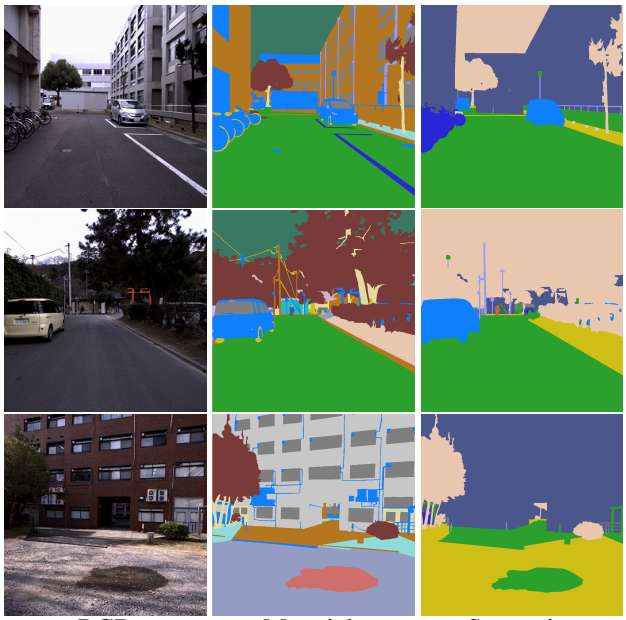

Figure 2. RGB Images (left) Material annotations (middle) and semantic annotations (right).

C. Details of Semantic Guidance

As we discussed in Section 4.1, we consolidate the semantic classes of CityScapes down to 10 classes. They are road, human, car, bicycle, building, wall, bridge, pole, terrain, and nature. When consolidating classes, we focus on the consistency of materials in each semantic class. For example, although cars and buses usually have different shapes, they are made of metal, rubber, glass and plastic. Thus, we treat cars and buses (including truck, caravan, trailer, and "on rails") as "car." Notice that glass rarely ap-

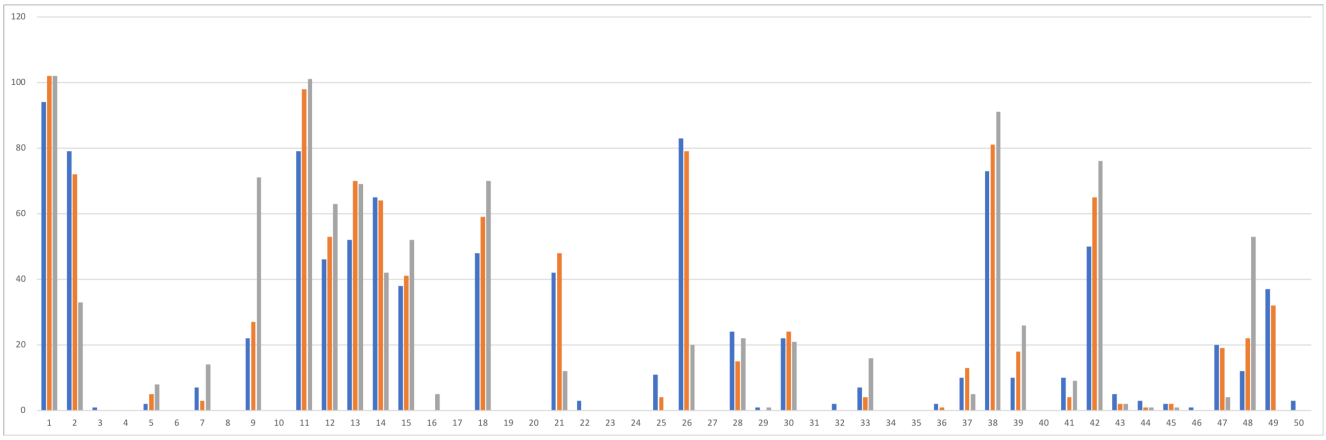

Figure 3. The frequency of the first 50 filters selected in "car" (blue), "bicycle" (orange) and "road" (gray) semantic regions.

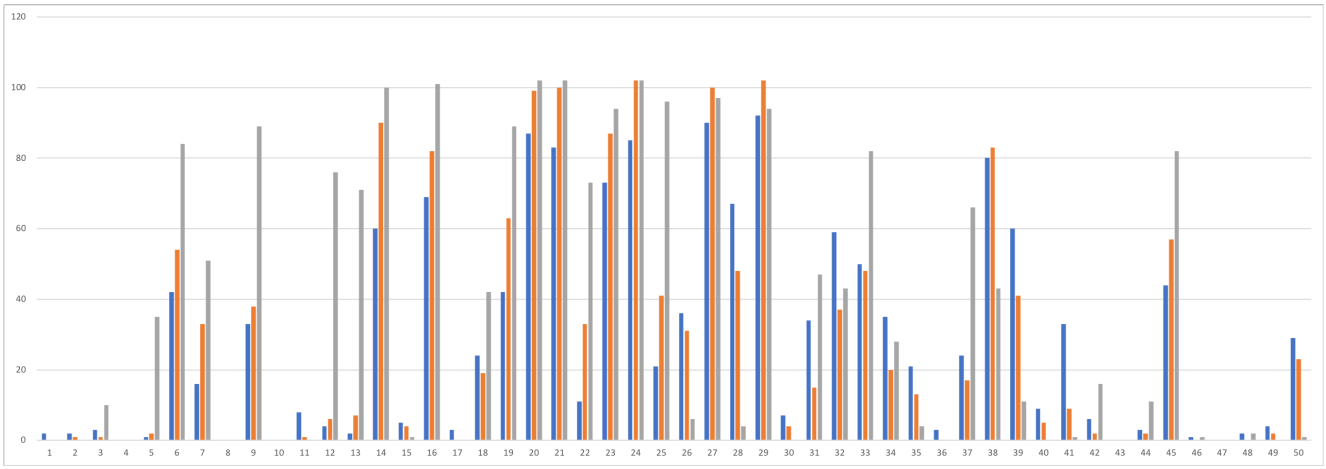

Figure 4. The frequency of the 51st-100th filters selected in "car" (blue), "bicycle" (orange), and "road" (gray) semantic regions.

pears on bicycles and bikes. Thus, we treat them as "bicycle" rather than "car." The "road" class includes road and sidewalk, mostly made from asphalt, concrete, and brick. Considering that glass is a critical component of "building" and rarely appears in bridges, tunnels, we treat these two objects as "bridge." The "Wall" class consists of wall, fence, and guard rail. We treat street lamps, traffic lights and all the poles nearby the street as "pole." All the ground that cannot be classified as "road" is "terrain." The "nature" class covers sky and trees. Figure 2 shows examples of semantic annotations and corresponding material annotations.

C.1. Selected Filters by RGFSConv

In our RGFSConv, we select filters based on their responses to each semantic region. We conduct additional experiments to verify that our RGFSConv selects different filters in each semantic class. We count the times that each filer is selected in each region in our test sets. RGFSConv selects 256 filters from 768 filters when the ratio λ is 3. Figure 3 and Fig. 4 show the frequency of the first 50 filters and 51st to 100th filters selected in "cars," "bicycle," and "road." For most filters, the times they are selected for "car" and "bicycle" are similar as both semantic classes consist of similar material components. The appearance of glass in "car" results in a small discrepancy between "car" and "bicycle." In contrast, the filters used for "road" is very different from "car" and "bicycle" as their materials are generally different. Besides, Figure 5 further validates the effectiveness of the filter selection scheme in RGFSConv. In a fraction of channels, the frequency of filter selection for "building," "terrain," and "nature" are the same. We think the appearance of "wood" as a material in these three semantic classes explains this. Again, different material compositions in these three classes cause large frequency differences in most filters.

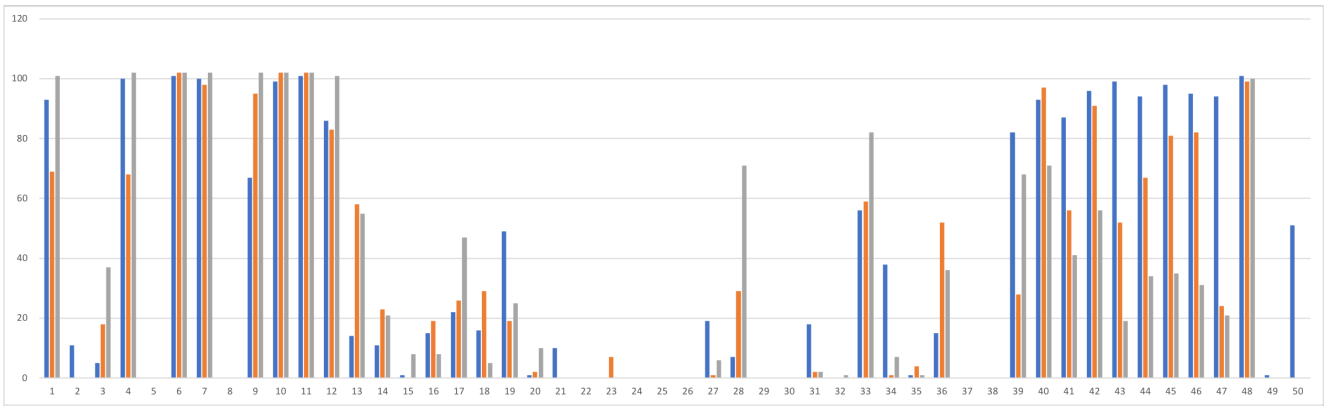

Figure 5. The frequency of the last 50 channels filters selected in "nature" (blue), "terrain" (orange), and "building" (gray) semantic regions.

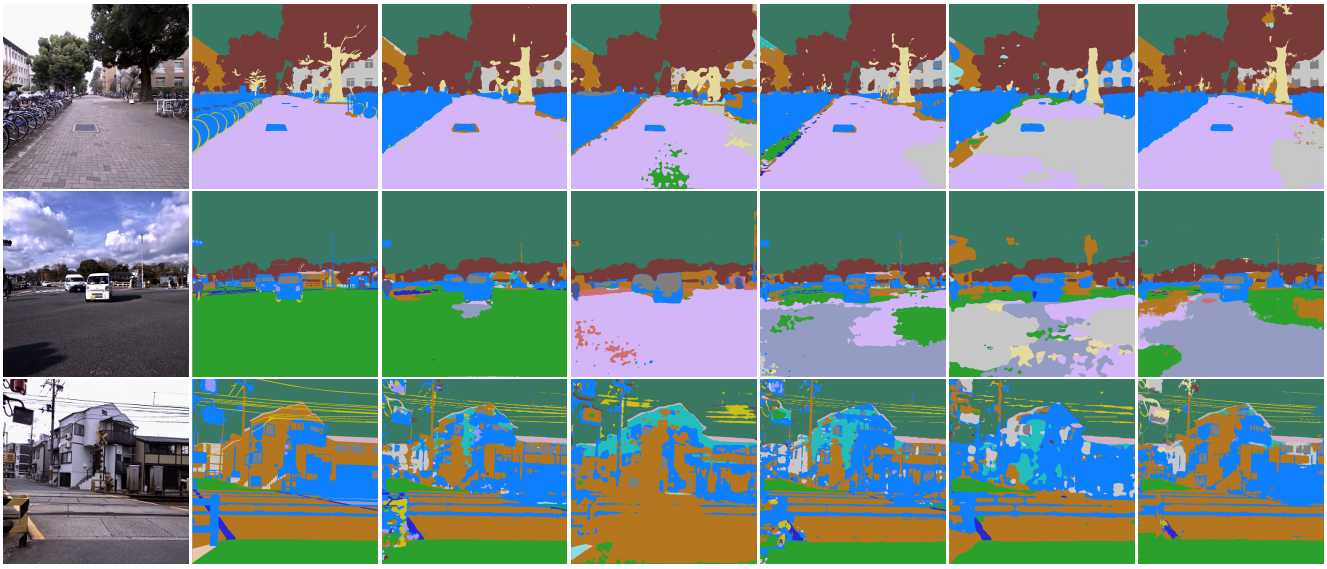

Figure 6. Material segmentation results of MCubeSNet, Modified-DRC, and Fusenet. MCubeSNet achieves highest accuracy by integrating multiple imaging modalities.

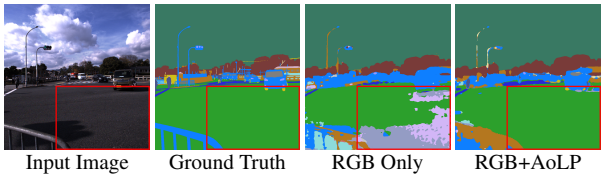

Figure 7. Contribution of AoLP for material segmentation. With the help of AoLP, MCubeSNet recognizes asphalt more accurately.

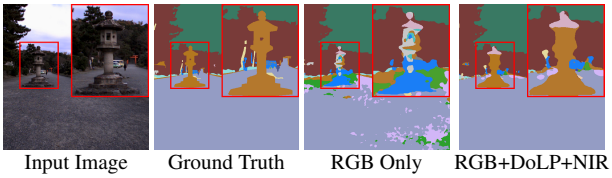

Figure 8. Contribution of DoLP and NIR for material segmentation. The combination of DoLP and NIR help discriminate between concrete and other materials.

C.2. More results

Figure 6 shows more results on the comparison of MCubeSNet, Modified-DRC, and Fusenet. Overall, our MCubeSNet achieves higher accuracy in both tiny material regions (rail track) and large material regions (road). Figure 7,Figure 8 and Figure 9 show more cases where combinations of different modalities achieve better result than using RGB images alone.

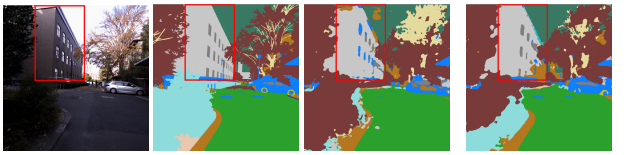

Figure 9. Contribution of polarization for material segmentation. Polarization behavior adds significant information to achieve higher accuracy especially for discerning metal and dielectrics.

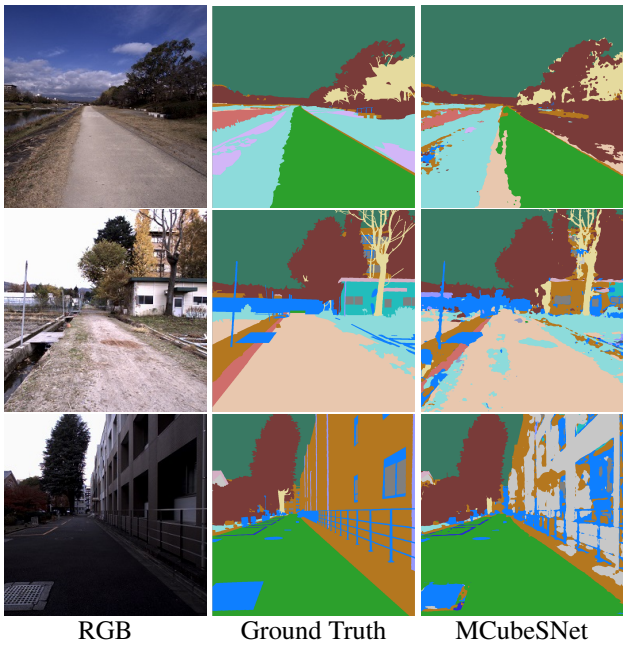

Figure 10. Some faliure cases of McubeSNet.

C.3. Limitation

Figure 10 demonstrates some failure cases of our MCubeSNet. Apparently, MCubeSNet gets confused when classifying leaf and grass. It wrongly segments the grass near the tree while successfully recognizing the grass on the left. Grass and leaf contain abundant water, which causes these two classes to appear similar in NIR images. The outer layer of buildings also confuse MCubeSNet. As shown in Fig. 10, it is difficult for MCubeSNet to distinguish concrete, brick, and plaster.