Multimodal Material Segmentation

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

Figure 1. We realize multimodal material segmentation, *i.e.*, per-pixel recognition of materials from multiple imaging modalities, by introducing a novel dataset and network. MultiModal Material Segmentation (MCubeS) dataset captures 42 different road scenes in 500 image sets each consisting of RGB, near-infrared, and polarization images. Each image is densely annotated with materials. We introduce MCubeSNet which learns to focus on the most informative combinations of imaging modalities for each material class through a newly derived region-guided filter selection (RGFS) layer.

Abstract

Recognition of materials from their visual appearance is essential for computer vision tasks, especially those that involve interaction with the real world. Material segmentation, i.e.*, dense per-pixel recognition of materials, remains challenging as, unlike objects, materials do not exhibit clearly discernible visual signatures in their regular RGB appearances. Different materials, however, do lead to different radiometric behaviors, which can often be captured with non-RGB imaging modalities. We realize multimodal material segmentation from RGB, polarization, and near-infrared images. We introduce the MCubeS dataset (from MultiModal Material Segmentation) which contains 500 sets of multimodal images capturing 42 street scenes. Ground truth material segmentation as well as semantic segmentation are annotated for every image and pixel. We also derive a novel deep neural network, MCubeSNet, which learns to focus on the most informative combinations of imaging modalities for each material class with a newly derived region-guided filter selection (RGFS) layer. We use semantic segmentation as a prior to "guide" this filter selection. To the best of our knowledge, our work is the first comprehensive study on truly multimodal material segmentation. We believe our work opens new avenues of practical use of material information in safety critical applications.*

1. Introduction

Thanks to the large strides made in object recognition research, computers can now tell what the object in an image is with sufficient accuracy. Telling what an object is, however, often insufficient to act in the real world. Our own visual system can not only tell a cup from a table, but also a paper cup from a ceramic one so that we can plan our grasp before touching it. If a computer could similarly tell what an object is made of, critical decisions can be made faster and more accurately. In particular, dense pixel-wise recognition of materials in an image becomes an essential task. We refer to this as material segmentation and distinguish it from classic "material recognition" which focuses on recognizing materials image-wise or for isolated objects.

Successful material segmentation would be particularly beneficial for road scene analysis. If an autonomous vehicle or an advanced driver assistance system (ADAS) can tell an asphalt road from a concrete one or a leaf on the road from dirt, it can execute safer control. Outdoor material segmentation, however, remains elusive mainly due to the rich variety of materials encountered in the real world and the lack of annotated data. Closest works only realize image-wise recognition of materials or are primarily of indoor architectural, professional photographs [1, 27]. It is also worth clarifying the distinction of material segmentation from stuff segmentation. "Stuff" is not a material but rather refers to

objects without discernible boundaries (*e.g*., a road, a representative "stuff," is composed of different materials such as asphalt, paint for markings, and metal for manholes).

The difficulty of material segmentation is exacerbated by the fact that materials lack well-defined visual features in regular RGB images. Unlike objects which largely exhibit different looks including shape contours and surface textures, different materials often result in similar appearance in regular color images. For instance, a ceramic cup and a plastic cup would have similar global shapes and local surface textures in an image. Some materials don't even have their own defined appearances. For instance, water does not have its own color and metal mostly mirror-reflects, both of which take on the appearance of their surroundings.

Where should we look for reliable visual cues to recognize materials? The surface composition of different materials not just in their spatial distribution but also in their subsurface structure give rise to characteristic radiometric behaviors. For instance, subtle differences in the mesoscopic surface structure change polarization of incident light and variation in subsurface composition result in different absorption of near-infrared (NIR) light. These radiometric features can potentially let us discern different materials robustly. Few works in the past have exploited different imaging modalities in isolation for only material recognition. Recent advances in imaging sensors, most notably the introduction of quad-Bayer CMOS, have brought the opportunity to leverage a variety of imaging modalities in a compact passive setup at low cost, making it particularly suitable for autonomous vehicles and mobile robots. We believe the time is ripe to systematically study what multimodal imaging can offer to material segmentation.

In this paper, we realize multimodal material segmentation, the recognition of per-pixel material categories from a set of images from the same vantage point but of different imaging modalities. In particular, we consider the combination of regular RGB, polarization, and near-infrared images at each instance. We build an imaging system consisting of a binocular stereo of quad-Bayer RGB polarization cameras, a monocular near-infrared camera, and a LiDAR to capture outdoor road scenes. We introduce a new dataset of multimodal material images which we refer to as the *MCubeS* dataset (from MultiModal Material Segmentation). The MCubeS dataset contains 500 sets of images of these imaging modalities taken at walking speed that emulates the vantage of an autonomous vehicle in 42 scenes and is fully annotated for each pixel. In addition to materials, we also annotate semantic segmentation labels. To our knowledge, MCubeS is the first of its kind and scale and opens new avenues of research on material segmentation.

We derive a novel deep architecture, which we refer to as MCubeSNet, for learning to accurately achieve multimodal material segmentation. We introduce region-guided filter selection (RGFS) to let MCubeSNet learn to focus on the most informative combinations of imaging modalities for each material class. We use object categories obtained with vanilla semantic segmentation as a prior to "guide" the filter selection. The network learns to select different convolution filters for each semantic class from a learned image-wise set of filters. This region-guided filter selection layer enables "dynamic" selection of filters, and thus combinations of imaging modalities, tailored to different potential materials underlying different semantic regions (*i.e*., object categories) without significant computational overhead.

To the best of our knowledge, our work is the first for multimodal material segmentation. In the absence of past methods, we experimentally validate the effectiveness of MCubeSNet by comparing its accuracy to state-of-the-art semantic segmentation methods. The experimental results, including ablation studies, clearly show that MCubeSNet can accurately and robustly recognize materials from multimodal data. The selected filters also reveal that characteristic radiometric properties of different materials are captured with unique combinations of imaging modalities. We believe our work makes an important step forward in material segmentation and opens new avenues of practical use of material information in safety critical applications. All data and code can be found on our project web page.

2. Related Work

Let us first review past works on material recognition and material segmentation including those that use non-RGB imaging modalities. We also review semantic segmentation works relevant to our method and as a baseline.

Material Recognition Early work on recognizing materials from images focused on image-wise recognition, *i.e*., determining a single material label for a whole image. We refer to this as material recognition to distinguish it from material segmentation. Dana *et al*. [5] introduced the CuReT dataset consisting of images of 61 different texture samples captured from over 205 different combinations of illumination and viewing conditions. The KTH-TIPS [7] dataset extended this work by introducing scale variation to 10 of the CuReT samples. These datasets were used to study texture modeling (*e.g*., Bidirectional Texture Function) and recognition. It is worth mentioning that material recognition is not the same as texture recognition as the cause of material appearance variation is not limited to just their spatial textures. Sharan *et al*. pioneered material recognition with the introduction of Flickr Materials Database (FMD) [13] which consists of 100 images each for 10 materials. Each image has only one material label with a spatial mask that isolates its region. Xue *et al*. introduced the Ground Terrain in Outdoor Scenes (GTOS) dataset which captures ground images from spatially and angularly different viewpoints for outdoor terrain (ground type) recognition [27]. This work

was extended to the GTOS-mobile dataset which 81 videos for 31 image-wide material classes [26].

Several works have introduced the use of r imaging for material recognition. Wolff *et al*. us ization to discern metallic and dielectric surfaces. *et al*. [18] used NIR in addition to RGB to distingu tile, linoleum, and wood. Mertz *et al*. [15] used flectivity to identify common materials. Erickson combined near-infrared spectroscopy and high-r texture images to classify materials of household Hu *et al*. [10] used passive millimeter-wave polariz classification of several metals and dielectric mate

Material Segmentation Our work concerns mat mentation, the pixel-wise recognition of materials world scene. In contrast to material recognition, segmentation requires densely annotated data an sification method that can combine both local visual features and global context. For this, deep neural networks trained on abundant data naturally becomes the primary approach. Bell *et al*. introduced the Materials in Context Database (MINC) [1] which consists of 2 million 64×64 image patches of 23 kinds of materials. They demonstrated material segmentation with a patch-based CNN followed by a CRF. The original images from which patches were extracted were sourced from Houszz and Flickr which are strongly biased towards professional architectural photographs (*e.g*., planned lighting and viewpoint).

Schwartz and Nishino introduced the Local Material Database [21] together with a canonical material category hierarchy as a three-level tree. Using this dataset, they derived a method for automatically discovering material attributes from local image-patches inside the object boundaries, and demonstrated its use for material recognition [19] which was later extended to simultaneous recognition of both material attributes (*e.g*., soft and fuzzy) with material categories (*e.g*., fabric) [21]. For material segmentation, they introduced a deep network architecture that combines fully convolutional per-pixel material classification with semantic segmentation and place recognition to integrate local material appearance and global contextual cues [20]. We significantly expand the horizon of material segmentation in terms of scenes, with self-driving scenarios in mind, as well as imaging modalities, and by deriving a novel deep material segmentation network that leverages the unique radiometric behaviors of different materials.

Semantic Segmentation Material segmentation is a challenging problem on its own, due to the reasons discussed in Sec. 1. The task, however, does bear a similarity to semantic segmentation which provides us with inspirations for network design. The introduction of fully convolutional networks which substitutes fully connected layers with convolution layers led to significant advancement in semantic segmentation [14]. Yu *et al*. introduced dilated convolution to overcome the limited receptive fields of ConvNets [28]. Deeplabv2 adopts atrous spatial pyramid pooling which consists of a group of dilated convolutions with varying dilation rates to extract multi-scale information [3]. Recent works have introduced various approaches to adapt the filters across the spatial domain by dynamically generating filters based on the image content [11]. Dynamic region-aware convolution introduced a guided feature branch to dynamically generate different filter sets to different image regions [2]. Decoupled Dynamic Filter (DDF) separates the spatial and channel components of these dynamic filters to reduce computational cost [30]. Our network is inspired by these recent dynamic filter designs, not just for spatial adaptability, but mainly to leverage the multimodal observations of different material appearance. Instead of dynamically generating filters, which is computational expensive, our network learns to select filters that find the best combination of imaging modalities adapted to the material underlying semantic regions from an overcomplete set of filters.

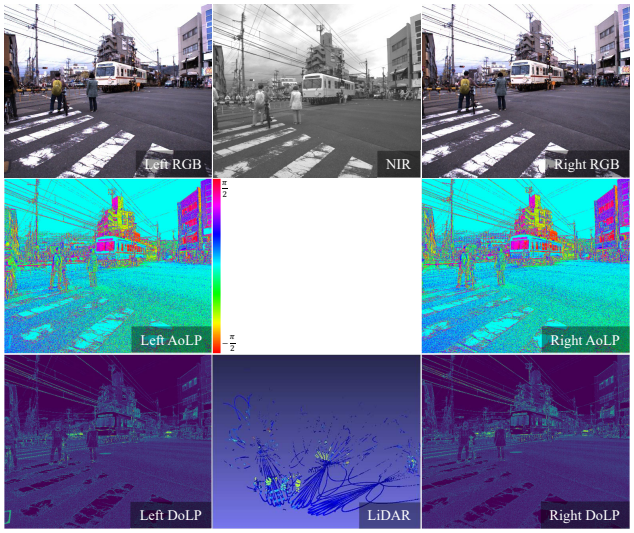


Figure 2. Example multimodal images from the MCubeS Dataset. Different imaging modalities capture characteristic radiometric behaviors of different materials.

3. MCubeS Dataset

We introduce the MultiModal Material Segmentation Dataset (MCubeS). MCubeS captures the visual appearance of various materials found in daily outdoor scenes from a viewpoint on a road, pavement, or sidewalk. At each viewpoint, we capture images with three fundamentally different imaging modalities, RGB, polarization, and near-infrared (NIR). The key challenges lie in systematic image capture with precise spatio-temporal alignment and annotation with pixel-wise material categories.

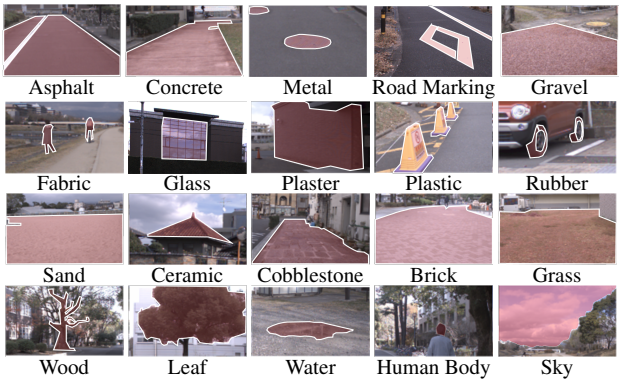


Figure 3. We annotate every pixel with 20 distinct material categories that were determined by thoroughly examining the MCubeS dataset. Here we show example regions for each class. The actual images are annotated densely for all pixels. Human body (skin and hair), leaf, grass, and sky are object and stuff names selected to represent their unique materials.

3.1. Imaging Modalities

Surfaces in the real-world made of different materials cause different subsurface compositions and surface structures at the mesoscopic scale. These differences give rise to different behaviors of incident light. Most notably, they alter the radiometric behavior of incident light in its reflection, refraction, and absorption. These differences can often be observed in their polarization properties, *i.e.*, different degree of linear polarization (DoLP) or angle of linear polarization (AoLP), and absorption of NIR light, respectively. For instance, surfaces made of metal specularly reflects light compared with wood which has a more balanced combination of specular and diffuse reflections, which result in differences in observed DoLP and AoLP. Water is transparent in the visible spectrum but absorbs light in the NIR range—water and water-containing surfaces (*i.e.*, foliage) exhibit different shades of intensity in an NIR image.

As shown in Fig. 2, MCubeS dataset captures these rich radiometric characteristics of different materials with a camera system consisting of a stereo pair of RGB-polarization (RGB-P) camera and a near-infrared (NIR) camera. The image capture system is also equipped with a LiDAR to assist label propagation (Sec. 3.3). The RGB-P camera is equipped with a quad-Bayer sensor which has a Bayer color filter array over 2×2 block of pixels with four on-chip polarizers of different angles.

Polarization Light is a transverse wave of electric and magnetic fields which are perpendicular to each other. Light with an electric field lying on a single plane is called linearly polarized. The angle of orientation of polarized light can be defined on the plane perpendicular to the transverse direction. Unpolarized light, in contrast, consists of light with electric fields oriented in all directions uniformly (*i.e.*, orientations form a circle). When the orientations of the electric fields of light is distributed as an ellipse on the plane perpendicular to the transverse direction, the light is partially linearly polarized. The major axis of this ellipse becomes the angle of polarization, and the ratio of the magnitudes of the major and minor axes is called the degree of polarization. The intensity of a partially polarized light with AoLP ϕ and DoLP ρ measured with a polarization filter angle ϕ_c is defined as $I(\phi_c) = \bar{I}(1+\rho\cos(2\phi_c - 2\phi))$, where $\bar{I}$ is the DC component [9]. The RGB-P camera captures four images corresponding to $\phi_c = 0$, $\frac{\pi}{4}$, $\frac{\pi}{2}$, and $\frac{3\pi}{4}$ which enables the calculation of ϕ, ρ, and $\bar{I}$ at each pixel in a single exposure.

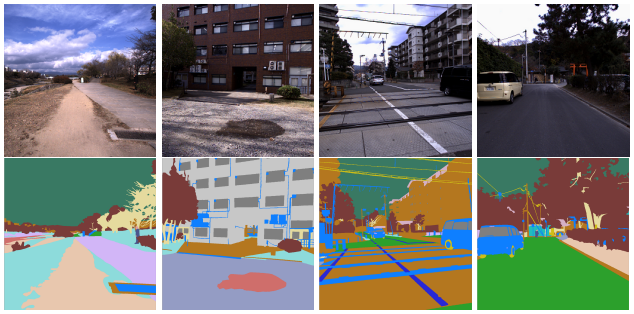
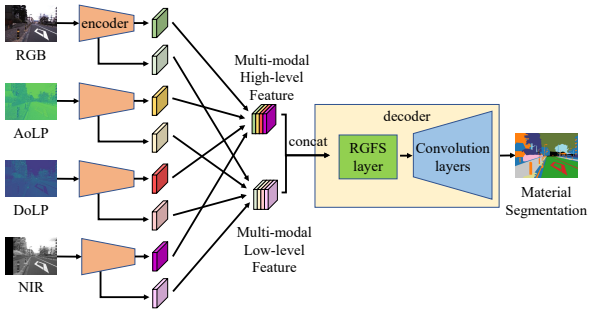


Figure 4. MCubeS dataset spans a wide range of road scenes (top row), including river sidewalks to railroad tracks, each densely annotated with materials (bottom row).

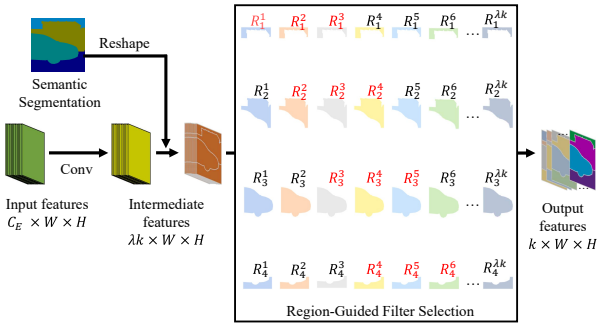
Near-infrared Light Light is a spectrum consisting of electromagnetic waves of different wavelengths. The radiometric behavior of light varies depending on the wavelength. Objects take on different colors as their reflection is wavelength-dependent. At the same time, the absorption and scattering properties of light transmitted into the subsurface or the volume of a medium varies depending on the wavelength. For instance, shorter wavelength light forward scatters more than light of longer wavelength. The absorption of light, in particular, dramatically changes outside the visual spectrum. Absorption of light through water, which is prevalent in our daily lives (not just puddles but also natural surfaces including leaves containing water), is particularly sensitive to wavelength. In the near-infrared range of 800nm to 1000nm, the absorption coefficient of light in water almost linearly increases from 0 to 1 [16]. That is, for a camera observing water or water-containing surfaces with a near-infrared filter of a wavelength in this range, different intensity encodes the depth or "wetness" of the surface (the deeper/wetter the darker) [22].

3.2. MCubeS Capture

Image Capture System We built a custom imaging system and mounted it on a cart to collect multimodal sequences of real-world scenes from a vantage point similar



(a) Overall network architecture.



(b) Region-guided filter selection layer.

Figure 5. (a) MCubeSNet extracts low-level and high-level feature maps for each imaging modality, separately, using DeepLab v3+ as the backbone. Polarization is represented by the AoLP and DoLP as independent modalities. (b) The encoder features are then input to a decoder that learns and selects different convolutional filters for different regions of a guidance field so that different materials integrate those imaging modalities in a way most relevant to identify them correctly. We refer to this as region-guided filter selection (FGFS) and use semantic segmentation as the guidance field.

to a car. The system is also equipped with a sparse LiDAR that will later be used to propagate annotations between different image modalities. Please see the image of the capture rig in the supplemental material. The imaging system consists of a pair of RGB-polarization (RGB-P) cameras (LUCID TRI050S-QC, 2/3-in sensor), one NIR camera (FLIR GS3-U3-41C6NIR-C, 1-in sensor), and one LiDAR (Livox Mid-100). A 6mm lens is attached to the RGB-P cameras, and an 8mm lens is attached to the NIR camera to make the field-of-views of the cameras approximately the same. The NIR camera is also equipped with a long-pass filter from 750 nm to cutoff light outside the near-infrared range.

Calibration Calibration of this multi-camera multimodal imaging system poses unique challenges due to the viewpoint differences, non-overlapping wavelength, and different modalities. We first calibrate the RGB-P camera pair against the NIR camera. Since the RGB and NIR images can resolve regular gray-scale patterns in their images, we use regular chessboard-based calibration [29] for estimating the intrinsic and extrinsic camera parameters.

We next calibrate the cameras to the LiDAR. LiDAR returns the reflectance intensity in addition to the distance to the target. This lets us model the LiDAR as a grayscale camera to run chessboard-based calibration by identifying the chess corners from the reflectance intensity image [25].

Data Collection Our MCubeS dataset consists of two distinct types of image sequences. The first type of data consists of sequences of images captured while the imaging system continuously moves forward. These sequences mimic the point of view of a moving vehicle. The second type of sequences is those captured at a single position while the imaging system pans. Some materials, like water, rarely appear in a road scene. The second type of data helps increase sample images of such materials.

MCubeS dataset consists of 42 scenes of the first type (continuous forward move) sampled at three frames per second. Average length of sequences are 309 seconds. The raw image sequences consist of 26650 image sets, from which 424 image sets at almost equal temporal spacing were annotated. We also capture 19 different scenes for the second type of data (fixed location panning). For each scene, eight image sets are captured to cover 360 degrees and a total of 76 image sets are annotated from these sequences. The total number of annotated image sets are 500.

3.3. Material Segmentation Annotation

Annotating each pixel of MCubeS poses significant challenges due to the different viewpoints. For this, we leverage the RGB-P stereo pair and LiDAR to propagate the labels across imaging modalities.

Material Classes We define 20 distinct materials by thoroughly examining the data we captured in the MCubeS dataset. Figure 3 shows examples of all materials. MCubeS scenes mainly consist of road scenes and sidewalks whose constituents vary from "stuff" like pavements and roads to objects including bicycles and manholes. Each of these can be made of different materials even for the same instance. For example, pavements can be made of asphalt, concrete, brick, cobblestone, gravel, and sand. We also treat road markings and manholes as made of different materials as they are of special interest in driving scenarios (*e.g.*, they can cause slips when wet). Vehicles are mainly made of metal, rubber, and glass. For people, clothes are made of fabric, while skin and hair are protein which are classes almost exclusive to people. For this, we use human body as the material label for parts other than their clothes. For natural objects, we mainly observe trees and grass as objects in the scene. Considering that wood can appear in sleepers on

Method	mIoU
MCubeSNet	**42.86%**
DeepLab v3+	38.13%
FuseNet	40.58%
TransFuser	37.66%
MMTM	39.71%
Modified-DRConv	34.63%
Modified-DDF	36.16%

(a)

Ratio	mIoU
1	38.13%
1.5	41.96%
2	42.85%
3	**42.86%**
4	39.54%
8	39.13%

(b)

Location	mIoU
First	**42.86%**
Second	39.65%
Both	33.08%

(c)

Table 1. Accuracy of (a) MCubeSNet and baseline methods, MCubeSNet with different RGFSConv (b) ratios and (c) locations.

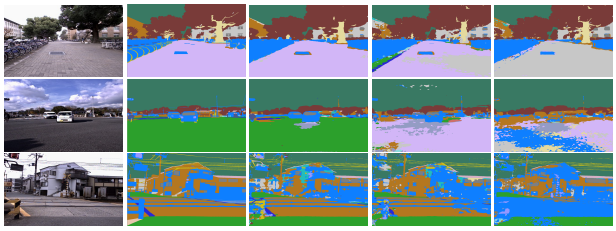


Figure 6. Material segmentation results of MCubeSNet, DeepLab v3+ with three additional encoders, and DDF. MCubeSNet successfully leverages multiple imaging modalities.

railroad tracks and fallen leaves may occupy some ground area, we use wood and leaf as material labels (*i.e.*, trees are made of wood and leaves). Another notable material we add is water, which we find in puddles, rivers, and ponds. The "Other" category includes ceramic, plaster, plastic, and sky, which are common but less significant in occurrences or for downstream tasks in a driving scenario.

Each image set of each sequence consists of 5 images: a stereo-pair of RGB and polarization images and an NIR image captured from three distinct viewpoints. As Fig. 4 shows, we densely annotate every image with per-pixel material classes. We annotate the left RGB image and propagate the per-pixel labels to other views of the same frame. We recover a dense depth image for the left RGB-P camera by integrating RGB stereo and LiDAR 3D points [23]. We use this dense depth image as a proxy to map the pixel-level annotations to the right RGB-P camera and the NIR camera. In this mapping, occluded pixels cannot find the corresponding pixels in the annotated left RGB image. We fill such holes with image inpainting [24].

4. MCubeSNet

We introduce a novel deep neural network that fully leverages the multimodal material appearance. We refer to this MultiModal Material Segmentation network as MCubeSNet. Inspired by recent advances in content-driven filtering, we introduce a novel guided convolution layer for integrating multimodal imaging data.

Architecture Figure 5a depicts the architecture of MCubeSNet. We chose DeepLab v3+ [4] as the backbone. The input consists of four images (left RGB and polarization image represented as AoLP and DoLP images, and near-infrared image). Since each of the different imaging modalities capture different radiometric aspects of different materials in the scene, we first extract image features using separate encoders for each image modality. For this, we use Resnet-101 with the ASPP module. As a result, each of the imaging modalities are encoded into low-level and high-level features. The high-level features are upsampled four times [4] and concatenated with the low-level features which is then input to the decoder.

4.1. Region-Guided Filter Selection

The decoder takes in the multimodal image features and produces a dense material segmentation. We fully leverage the multimodal material characteristics in this decoder by introducing a *region-guided filter selection convolution layer (RGFSConv)*.

Material occurrences are strongly correlated with object instances. For instance, gravel is often seen in roads and pavements but not as part of cars. Metal is often seen in cars and poles, but not as a road. We leverage this rich interdependence between objects and materials by using object categories as priors on material segmentation. Object categories are extracted by semantic segmentation on the input RGB image(s). We use DeepLab v3+ [4] for this, which is applied to the input RGB image outside MCubeSNet. We consolidate the semantic classes of CityScapes down to 10 classes. Please refer to the supplemental material for more details. We use this semantic segmentation as a guide field to apply different filters to different semantic regions.

The idea is to learn different convolutional filters to apply to each imaging modality feature (*i.e.*, channel) from the encoder so that different materials integrate those imaging modalities that are most relevant to identify them correctly. The semantic segmentation provides a region field that narrows down the possible materials. For instance, the network can learn to focus on learning filters that would help recognize metal, glass, and rubber for a car region. A naive implementation of this idea, however, incurs too much computational cost. For instance, a guided dynamic filter [2] with C input channels, O output channels, and m classes would require $m^2C(O+1)$ parameters in the feature generator module. At the same time, simply learning different filter sets for each semantic region would not model the underlying materials with consistent features.

Our key idea is to learn an overcomplete set of (regular) convolutional filters across the entire image, but learn to assign different sets of those filters to each semantic region. The filters are selected by picking the top k responsive filters for each semantic region, *i.e.*, picking the filters that have the highest average activation. This approach enables

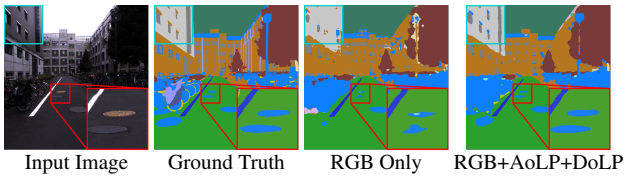


Figure 7. Contribution of polarization for material segmentation. Polarization behavior adds significant information to achieve higher accuracy especially for discerning metal and dielectrics.

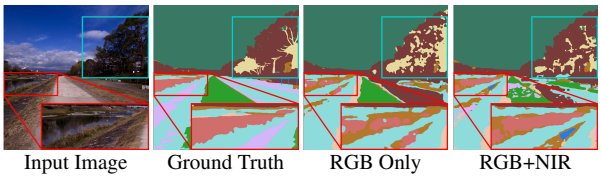


Figure 8. Contribution of NIR. NIR helps recognition of water (river) and wet surfaces (wood vs. leaves).

learning of material-specific filter sets that integrates different imaging modalities (channels of encoder output features) with different weightings guided by semantic regions, which we refer to as RGFSConv.

In the RGFSConv layer, a $C_{\mathrm{E}} \times W \times H$ feature map, which has C_{E} channels from the encoder output, first passes through vanilla convolutions to generate an intermediate feature map of size $\lambda k \times W \times H$. The ratio λ is a hyperparameter that defines the relative size of the pool of convolution filters the network can use for different semantic (in turn, material) regions. RGFSConv outputs feature maps of k channels by adaptively selecting different sets for different semantic regions. By matching k to the channel size of an existing layer in a backbone network, we can easily insert an RGFSConv layer to achieve this region-guided filter selection for any problem at hand. We believe RGFSConv would be effective in various multimodal imaging tasks.

The guide field, in our case the semantic segmentation, can be computed on the input independent from the network and fed into the RGFSConv layer after resizing its spatial dimensions to $W \times H$. For each semantic region, we compute the average response of each channel (r_j^m) by

$$r_j^m = \frac{1}{|D^m|} \sum_{(x,y) \in D^m} F_j(x,y)\,, \tag{1}$$

where F_j is the j-th channel of the intermediate feature, D^m represents the region(s) corresponding to the m-th semantic class, and $|D^m|$ denotes the number of pixels in D^m. The output feature map of the RGFSConv layer in each semantic region becomes

$$f^m = \mathrm{cat}(F_{j^*}^m)\,, \qquad j^* = \underset{j}{\overset{k}{\mathrm{argmax}}}(r_j^m)\,, \tag{2}$$

where j^* are the indices of the k largest average responses and cat is the concatenation operator.

5. Experiments

We thoroughly verify the effectiveness of MCubeSNet on our dataset by also comparing with baseline methods and through ablation studies. Please see the supplemental material for many more results.

Training Details We conduct all experiments using PyTorch on Quadro RTX A6000. Data augmentation including random horizontal flips and scale crops is applied. Training, validation, and test sets contain 302, 96 and 102 images, respectively. Polynomial decay from 0.05 with 0.9 as the power is used for the learning rate schedule for maximum epochs of 500. For the four ASPP modules and the decoder, the learning rate is increased by ten times. Stochastic gradient descent with momentum 0.9 and weight decay 5e-4 is used. We measure accuracy by the class mean Intersection-over-Union (mean IoU, mIoU). We use a pretrained DeepLab v3+ which we then fine-tune with the semantic labels of the training set of MCubeSNet as the semantic segmentation network for generating the guide fields for the RGFSConv layer of MCubeSNet.

Baseline Comparisons In the absence of other multimodal material segmentation methods, we compare our method with a state-of-the-art semantic segmentation method, namely DeepLab v3+ with four encoders, as the baseline method. Note that this "DeepLab v3+" performance also likely represents an upper bound on past RGB material segmentation networks (*e.g.*, [1] and [20]) given the progress in semantic segmentation research. We also compare with three multimodal semantic segmentation methods: FuseNet [8], Multi-Modal Fusion transformer [17], and MMTM [12]. For these methods, we modify their network structures to accept four imaging modalities. In FuseNet, we add two more encoders with the same structure as the original ones. The element-wise addition fusion mechanism in multiple layers are untouched. For the TransFuser and MMTM, we add the fusion modules after each block of ResNet-101. Additionally, we also test two dynamic filter methods DRConv [2] and DDF [30]. Similar to our method, we only substitute the first layer of the decoder with these two convolution methods. We also add the semantic guidance to them. Please refer the supplemental material for details.

Table 1(a) shows material segmentation results of our MCubeSNet and other models. MCubeSNet achieves 42.86% in mIoU and at least 2.28% gain over other methods. Compared to the baseline, RGFSConv improves the mIoU by 4.73%. Figure 6 shows the performance comparison of MCubeSNet, baseline, and Modified-DDF. Overall,

RGB	AoLP	DoLP	NIR	asphalt	concret	metal	road ma	fabric	glass	plaster	plastic	rubber	sand	gravel	ceramic	cobbles	brick	grass	wood	leaf	water	sky	mean
✓				75.8	32.3	36.1	53.7	0.0	23.1	0.8	5.2	3.1	61.9	53.6	6.3	38.1	25.7	53.6	27.0	70.2	13.1	95.1	33.7
✓	✓			83.3	42.3	43.0	58.4	8.8	27.3	0.6	9.8	12.0	55.5	57.7	18.1	64.6	36.6	56.5	34.8	71.8	6.8	95.0	39.1
✓		✓		75.2	40.2	37.8	53.9	4.2	32.3	1.9	14.3	11.3	59.7	21.8	11.6	28.9	29.1	54.6	29.4	71.4	9.6	94.3	34.1
✓			✓	82.4	41.5	47.0	**65.3**	15.2	**45.4**	0.5	14.1	**15.2**	59.9	47.3	20.6	39.9	27.7	**59.4**	38.0	**75.9**	**18.1**	**96.0**	40.5
✓		✓	✓	81.7	**45.2**	44.9	54.3	6.1	42.8	1.4	17.3	0.8	54.0	60.7	23.0	60.5	34.9	57.9	34.4	72.5	2.0	94.7	39.5
✓	✓		✓	83.0	43.9	**47.8**	57.9	10.4	40.3	0.7	17.7	13.0	57.5	52.8	20.7	65.0	38.2	58.2	38.1	75.1	8.2	95.3	41.2
✓	✓	✓		83.0	42.6	45.5	59.8	**17.0**	44.2	1.2	**18.6**	4.8	54.8	51.5	26.4	**67.6**	**41.9**	57.0	39.4	74.0	15.5	95.3	42.0
✓	✓	✓	✓	**85.7**	42.6	47.0	59.2	12.5	44.3	**3.0**	10.6	12.7	**66.8**	**67.1**	**27.8**	65.8	36.8	54.8	**39.4**	73.0	13.3	94.8	42.9

Table 2. Performance comparison of using different modalities in per-class IoUs and mIoU (%). The ratio λ is set to 3. When a certain modality is excluded, the encoder is fed with a zeroed-out image for that modality to ensure fair comparison. Best results are **highlighted**. We omit the human body class as its result is 0%.

our MCubeSNet achieves more accurate segmentation results. MCubeSNet, for instance, can discriminate the rail track from its surrounding concrete and overall performs better in recognizing the road. We tested class weight balancing on the network outputs which results in 45.95% mIoU (DeepLabv3+ achieves 41.32% mIoU). Please see the supplemental material for examples from all methods.

Ablation Studies on RGFSConv We analyze the effectiveness of RGFSConv by varying the ratio λ to 1, 1.5, 2, 3, 4, and 8. Notice that when the ratio is 1, our RGFSConv reduces to traditional convolution and the decoder has the same structure as the decoder of the original DeepLab v3+. Table 1(b) shows the results. The significant decrease in performance when changing the ratio from 3 to 1 verifies the effectiveness of our RGFSConv. When the ratio becomes larger than 3, MCubeSNet cannot maintain its accuracy. We believe this is because when λ is too large, the chance that each channel is selected in the output of RGFSConv drops and only a small fraction of parameters are updated in each iteration, which decreases accuracy.

We also explore the use of RGFSConv at different locations in the network: the first or second layer of the decoder or both. Table 1(c) shows that when the RGFSConv is closest to the encoder, MCubeSNet achieves highest performance. When we use two RGFSConvs simultaneously (Both), the network suffers from the inconsistency of channels between the two layers. These results indicate that RGFSConv finds optimal combinations of imaging modalities for different semantic regions (*i.e.*, potential materials).

Contributions of Imaging Modalities We investigate how each imaging modality contributes to the per-pixel recognition of different materials. For this, we train and test MCubeSNet with eight different combinations of imaging modalities. Table 2 shows these cases and their results. Clearly, using all modalities (AoLP, DoLP and NIR) leads to a 9.2% improvement in mIoU. The performance comparison in the four cases of RGB $\rightarrow$ RGB+NIR, RGB+AoLP $\rightarrow$ RGB+AoLP+NIR, RGB+DoLP $\rightarrow$ RGB+DoLP+NIR, and RGB+DoLP+AoLP $\rightarrow$ All modalities verify that NIR contributes significantly in recognizing gravel, ceramic, cobblestone, and brick. We notice the same increase in these classes after adding AoLP. Combining NIR and RGB also achieves the best result in water-related classes, such as grass, leaf, and water. This conclusion is consistent with the fact that water absorbs more NIR light than other materials. With the help of the two polarization modalities, MCubeSNet obtains 9.4%, 13.4%, and 20.9% performance improvement in segmentation of metal, plastic, and glass, respectively (RGB $\rightarrow$ RGB+AoLP+DoLP), which is also consistent with the polarization characteristics of metal and dielectrics. Overall, these results clearly demonstrate the importance of multimodal imaging for material segmentation, especially for outdoor scenes.

Limitations Please see the supplementary material for more results and analysis including failure cases. The combination of modalities we employed in this paper sometimes cannot resolve ambiguities of material appearance under different illumination and viewing conditions. We plan to quantify such uncertainties. We also plan to study how to combat the natural class imbalance of materials.

6. Conclusion

In this paper, we introduced a new dataset and novel method for multimodal material segmentation. The new MCubeS dataset consists of 500 sets of RGB, polarization, and NIR images of outdoor scenes captured from a vantage point similar to a car. MCubeSNet fully leverages these imaging modalities to accurately recognize per-pixel material categories. The experimental results clearly show the importance of multimodal imaging for outdoor material segmentation. We believe the dataset and network serve as an important platform for fully utilizing rich material information in safety critical applications.

Acknowledgement This work was in part supported by JSPS 20H05951, 21H04893, Sensetime, and JST JPMJCR20G7.

References

[1] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material Recognition in the Wild with the Materials in Context Database. In *CVPR*, pages 3479–3487, 2015. 1, 3, 7

[2] Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun. Dynamic Region-Aware Convolution. In *CVPR*, pages 8064–8073, 2021. 3, 6, 7

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE TPAMI*, 40(4):834–848, 2017. 3

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *ECCV*, pages 801–818, 2018. 6

[5] Kristin J Dana, Bram Van Ginneken, Shree K Nayar, and Jan J Koenderink. Reflectance and Texture of Real-World Surfaces. *ACM TOG*, 18(1):1–34, 1999. 2

[6] Zackory Erickson, Eliot Xing, Bharat Srirangam, Sonia Chernova, and Charles C Kemp. Multimodal Material Classification for Robots Using Spectroscopy and High Resolution Texture Imaging. In *IROS*, pages 10452–10459, 2020. 3

[7] Mario Fritz, Eric Hayman, Barbara Caputo, and Jan-Olof Eklundh. The KTH-TIPS Database. 2004. 2

[8] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN architecture. In *ACCV*, pages 213–228, 2016. 7

[9] Eugene Hecht. *Optics*. Pearson Education, 2016. 4

[10] Fei Hu, Yayun Cheng, Liangqi Gui, Liang Wu, Xinyi Zhang, Xiaohui Peng, and Jinlong Su. Polarization-Based Material Classification Technique Using Passive Millimeter-wave Polarimetric Imagery. *Applied Optics*, 55(31):8690–8697, 2016. 3

[11] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic Filter Networks. *NeurIPS*, 29:667–675, 2016. 3

[12] Hamid Joze, Reza Vaezi, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. MMTM: Multimodal Transfer Module for CNN Fusion. In *CVPR*, pages 13289–13299, 2020. 7

[13] Ce Liu, Lavanya Sharan, Edward H Adelson, and Ruth Rosenholtz. Exploring Features in a Bayesian Framework for Material Recognition. In *CVPR*, pages 239–246, 2010. 2

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *CVPR*, pages 3431–3440, 2015. 3

[15] Christoph Mertz, Sanjeev J Koppal, Solomon Sia, and Srinivasa Narasimhan. A Low-Power Structured Light Sensor for Outdoor Scene Reconstruction and Dominant Material Identification. In *CVPR*, pages 15–22, 2012. 3

[16] Satoshi Murai, Meng-Yu J. Kuo, Ryo Kawahara, Shohei Nobuhara, and Ko Nishino. Surface Normals and Shape from Water. In *ICCV*, 2019. 4

[17] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. In *CVPR*, pages 7077–7087, 2021. 7

[18] Neda Salamati, Clément Fredembach, and Sabine Süsstrunk. Material Classification Using Color and NIR Images. In *Color Imaging Conference*, pages 216–222, 2009. 3

[19] Gabriel Schwartz and Ko Nishino. Automatically discovering local visual material attributes. In *CVPR*, 2015. 3

[20] Gabriel Schwartz and Ko Nishino. Material Recognition from Local Appearance in Global Context. *arXiv preprint arXiv:1611.09394*, 2016. 3, 7

[21] Gabriel Schwartz and Ko Nishino. Recognizing Material Properties from Images. *IEEE TPAMI*, 42(8):1981–1995, 2020. 3

[22] Mihoko Shimano, Hiroki Okawa, Yuta Asano, Ryoma Bise, Ko Nishino, and Imari Sato. Wetness and Color from A Single Multispectral Image. In *CVPR*, 2017. 4

[23] Shreyas S Shivakumar, Kartik Mohta, Bernd Pfrommer, Vijay Kumar, and Camillo J Taylor. Real Time Dense Depth Estimation by Fusing Stereo with Sparse Depth Measurements. In *ICRA*, pages 6482–6488, 2019. 6

[24] Alexandru Telea. An Image Inpainting Technique Based on the Fast Marching Method. *Journal of Graphics Tools*, 9(1):23–34, 2004. 6

[25] Weimin Wang, Ken Sakurada, and Nobuo Kawaguchi. Reflectance Intensity Assisted Automatic and Accurate Extrinsic Calibration of 3d LiDAR and Panoramic Camera Using a Printed Chessboard. *Remote Sensing*, 9(8), 2017. 5

[26] Jia Xue, Hang Zhang, and Kristin Dana. Deep Texture Manifold for Ground Terrain Recognition. In *CVPR*, pages 558–567, 2018. 3

[27] Jia Xue, Hang Zhang, Kristin Dana, and Ko Nishino. Differential Angular Imaging for Material Recognition. In *CVPR*, pages 764–773, 2017. 1, 2

[28] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. *ICLR*, 2016. 3

[29] Zhengyou Zhang. A Flexible New Technique for Camera Calibration. *IEEE TPAMI*, 22(11):1330–1334, 2000. 5

[30] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled Dynamic Filter Networks. In *CVPR*, pages 6647–6656, 2021. 3, 7