RGB Road Scene Material Segmentation

Sudong Cai[0000−0002−5446−5618], Ryosuke Wakaki[0000−0003−3917−9012],
Shohei Nobuhara[0000−0002−3204−8696], and Ko Nishino[0000−0002−3534−3447]

Graduate School of Informatics, Kyoto University, Kyoto, Japan
https://vision.ist.i.kyoto-u.ac.jp/

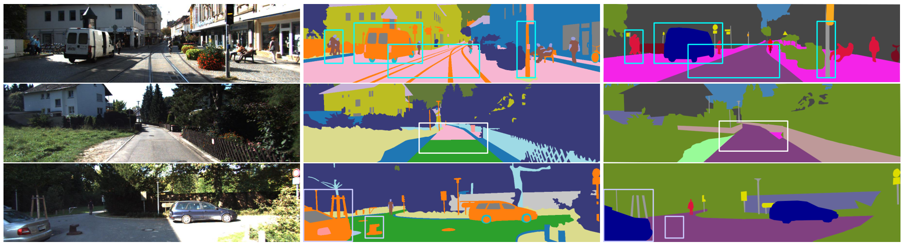
(a) Images (b) Materials (c) Semantics

Fig. 1. Materials vs. Semantics. Top: One semantic object may consist of multiple materials and different semantic objects may contain the same material. Middle: The same "Road" can be made of "asphalt," "concrete," or "brick." Bottom: A metal obstacle which is unclear in the semantic annotations, can cause hazard for driving.

Abstract. We address RGB road scene material segmentation, *i.e.*, per-pixel segmentation of materials in real-world driving views with pure RGB images, by building a new tailored benchmark dataset and model for it. Our new dataset, KITTI-Materials, based on the well-established KITTI dataset, consists of 1000 frames covering 24 different road scenes of urban/suburban landscapes, annotated with one of 20 material categories for every pixel in high quality. It is the first dataset tailored to RGB material segmentation in realistic driving scenes which allows us to train and test any RGB material segmentation model. Based on an analysis on KITTI-Materials, we identify the extraction and fusion of texture and context as the key to robust road scene material appearance. We introduce Road scene Material Segmentation Network (**RMSNet**), a new Transformer-based framework which will serve as a baseline for this challenging task. RMSNet encodes multi-scale hierarchical features with self-attention. We construct the decoder of RMSNet based on a novel lightweight self-attention model, which we refer to as **SAMixer**. SAMixer achieves adaptive fusion of informative texture and context cues across multiple feature levels. It also significantly accelerates self-attention for feature fusion with a balanced query-key similarity measure. We also introduce a built-in bottleneck of local statistics to achieve further efficiency and accuracy. Extensive experiments on KITTI-Materials validate the effectiveness of our RMSNet. We believe our work lays a solid foundation for further studies on RGB road scene material segmentation.

1 Introduction

Recognition of materials, what things are made of, in an image is critical for many computer vision applications. Materials inform the physical properties of the objects and regions in a scene which are otherwise inaccessible from just knowing the object categories. The way an action is planned for a paper cup as opposed to a ceramic cup would be different and gauging from sight would be advantageous. The importance of material recognition becomes even more significant for road scenes, particularly for self-driving vehicles to successfully navigate in daily environments. Despite its potential contributions to safety, and past work on object-level material recognition, little has been studied on regular color-image-level visual material understanding in road scenes.

Per-pixel material recognition (*i.e.*, material segmentation, in contrast to semantic segmentation) in a regular color image would be particularly informative for self-driving and driving assistance. Knowing that the asphalt-made road turns into gravel or brick would help an autonomous system to plan its speed, discerning a twig from a metal bar would help decide whether to avoid it, and telling a bronze statue from a live pedestrian would help anticipate its movements. Material segmentation, however, is not yet another semantic segmentation problem with a different set of labels. The challenge lies in the fact that the same object category can have different material categories, *e.g.*, a road can be made of asphalt, concrete, or even dirt and brick, yet they have the same shapes. The difficulty is exacerbated by the fact a single object can have multiple regions of different materials, *e.g.*, a bicycle made of metal, rubber, plastic, and leather. In contrast, objects can mostly be discerned with shape cues for category-level recognition, *i.e.*, semantic segmentation and object recognition.

In this paper, we address RGB road scene material segmentation by introducing a new benchmark dataset and a novel network that exploits the unique properties of material appearance and serve as a baseline model for this challenging task. We build a new dataset tailored to RGB road scene material segmentation by annotating images from the KITTI dataset [12]. We refer to this new dataset as the *KITTI-Materials* dataset. By building on a widely adopted road scene dataset, we are able to establish a dataset guaranteed to be relevant for autonomous driving research. KITTI-Materials consists of 1000 frames densely annotated with one of 20 material categories covering 24 different road scenes of common urban/suburban landscapes. The KITTI-Materials is the first tailored benchmark dataset for pure RGB road scene material segmentation which enables us to train and evaluate our ideas for the task and others to follow.

Figure 1 illustrates the key differences of RGB road scene material and semantic segmentation. A careful examination of the new dataset reveals that effective texture and context information extraction and fusion is essential for robust RGB road scene material recognition. The characteristic textures of materials provide vital visual cues for their identification. The appearance of material texture, however, changes dramatically with scale (*i.e.*, distance from viewpoint) and occlusion. We may incorporate structural dependencies of local texture features to arrive at representations robust to these scale and occlusion variations. Such structural context, however, is unreliable for material appearance in contrast to global shape cues often exploited for object recognition. For this, effective fusion of texture and context cues to produce discriminative joint representations becomes vital for discerning road scene materials.

We introduce Road scene Material Segmentation Network (**RMSNet**), a new Transformer-based material segmentation network which generates discriminative material appearance representations through joint texture-context learning with low computational cost. RMSNet adopts the efficient hierarchical encoder introduced by SegFormer [40] to extract features of local textures and long-range context from multi-level hierarchies. It then merges multi-level multi-scale features with a novel self-attention-based feature fusion model which we refer to as **SAMixer**. SAMixer introduces a new balanced query-key similarity (Q-K-Sim) measure with a container feature generated by aggregating all input feature maps. This results in a highly efficient self-attention mechanism with only $O(N + 1)$ complexity, where N denotes the number of input feature maps. SAMixer also uses a bottleneck local statistics encoding-decoding (BLSED) strategy for additional efficiency and accuracy.

We evaluate the effectiveness of RMSNet through extensive quantitative analysis and ablation studies on KITTI-Materials RGB RMS, and compare its accuracy with existing RGB material segmentation and road scene semantic segmentation methods. The results clearly demonstrate the effectiveness of RMSNet. We believe our work can contribute to richer visual understanding, particularly of road scenes, for safer driving. RMSNet will serve as a sound baseline model for this important task. We disseminate our project[1] to catapult this emerging avenue of research.

2 Related Work

Bell *et al.* [1] demonstrated material segmentation with a fully convolutional network cascaded with a fully-connected CRF [20, 34, 19], which is essentially semantic segmentation with material categories applied to mainly architectural photographs. Schwartz and Nishino introduced the use of material attributes as an intermediate representation for per-pixel material recognition without regard to shape features [29, 30, 32]. Later they introduced the integration of global contextual information in the form of semantic segmentation and place recognition and demonstrated its application to material segmentation on a material dataset consisting of local image patches sourced from COCO dataset and ImageNet [33].

Xue *et al.* [42] introduced the GTOS dataset, consisting of over 30k images of 40 ground surface material taken as top-down fronto-parallel images. They investigated the advantage of differential angular imaging for material recognition. Zhang *et al.* [48] proposed the Deep-TEN model by using "orderless" texture encoding [25]. More recently, Xue *et al.* [43] incorporated texture encoding showing superior results to DAIN [42]. These methods, however, focus on image-wise material recognition.

[1] https://github.com/kyotovision-public/RGB-Road-Scene-Material-Segmentation

Recently, Demir *et al.* [9] introduced the DeepGlobe dataset which consists of satellite images mainly for road and building extraction. Purri *et al.* [26] also proposed building material segmentation datasets from satellite images [2] and proposed reflectance residual encoding. Xue *et al.* [41] derived AngLNet which uses per-pixel angular luminance from multiple views. Material segmentation on road scenes is distinct from these bird-eye-view material segmentation as scale variation due to a dynamic perspective is inevitable.

Road scene semantic segmentation is a popular research field which provides us with inspirations for model design. Cordts *et al.* [8] introduced the Cityscapes dataset for scene understanding of urban driving environments. Many works have tackled road scene semantic segmentation using this dataset [50, 24, 4, 6, 44, 7, 52, 28, 27, 36, 40]. In contrast, road scene material segmentation with regular RGB images has not been intensively explored.

More related to our work, Liang *et al.* [22] introduced the MCubeS dataset, a multimodal material segmentation dataset consisting of RGB, NIR, and polarization images of city scenes, where the material categories are identical to our KITTI-Materials. Based on the dataset, they proposed the MCubeSNet modified on DeepLabv3+ [6] equipped with the RGFS layer to jointly apply various imaging modalities for improving material segmentation with the help of semantic segmentation annotations. In contrast, our KITTI-Materials dataset is tailored to pure RGB road scene material segmentation and comprises more images and scenes covering both city and suburban landscapes.

Vision transformers leverage multi-head self-attention (MSA) [35] to model long-range visual cues [38, 45, 18]. Full MSA to $2D$ spatial features, however, incurs excessive computational cost. Ramachandran *et al.* [27] modified the transformer to work on a fixed region and added positional biases. Wang *et al.* [36] introduced Stand-Alone Axial-MSA which processes feature maps along the height- and the width-axis separately to balance computational cost and accuracy for semantic segmentation. Zhang *et al.* [49] showed that co-occurrence of semantics including object categories exhibit long-range dependencies.

ViT [11] computes MSA within each non-overlapping image patch (*i.e.*, window) to achieve a speed-accuracy tradeoff for object recognition. PVT [37] introduced the first pyramid transformer architecture and demonstrated its potential for dense prediction tasks. Liu *et al.* [23] suggested applying MSA within fine-grained shifted windows and model cross-window connections to enhance local cues. Related models LeViT [13] and TNT [14] also improved window-MSA by infusing extra local details. Pure window-MSA, however, is computationally expensive for high-resolution features.

SegFormer [40] built a hierarchical transformer encoder with an efficient MSA in which the keys and values with reduced resolution were computed from condensed features with convolutions. It also introduced a lightweight All-MLP decoder and demonstrated its advantage over existing heavy decoders. Relevant idea was suggested in CvT [39], where Q-K-V projections were realized by convolutions. In contrast, RMSNet introduces a novel SAMixer model that fuses

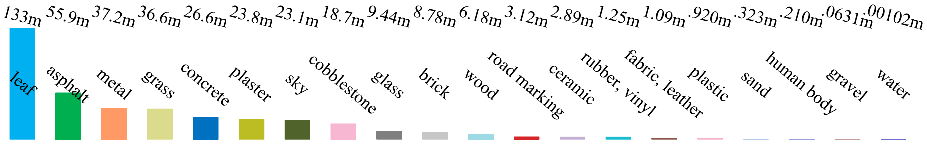

Fig. 2. Per-class pixel statistics (in "millions") of KITTI-Materials. Pixel labels show a clear long-tail distribution of material categories.

multi-level features of local textures and long-range contextual cues to generate robust representations for road scene material segmentation.

Past works have explored multi-scale feature learning in object recognition and semantic segmentation. Chen *et al.* [5] plugged a spatial attention layer into the bottom of a two-branch network to learn weights for features at different scales. SKNet [21] expanded SE-Net [17] to aggregate multi-scale features. The Deeplab family [3, 6, 44, 4] used atrous spatial pyramid pooling to learn scale-invariance through global statistics and a set of convolutions with different dilations. To enhance multi-level scale-aware feature learning, our RMSNet selectively activates meaningful features of local textures and non-local contextual interactions to form discriminative representations with SAMixer.

3 KITTI-Materials Dataset

We introduce the KITTI-Materials dataset, the first comprehensive RGB road scene material segmentation dataset. The images used in KITTI-Materials are sourced from the KITTI raw data [12]. It consists of 1000 images covering 24 different driving scenes including downtown, campus, residential area, highway, and other city/suburban landscapes captured from a car. In the 24 road scenes, there are 19 scenes consisting of 50 images sampled for every 5 consecutive frames, and other 5 scenes that contain 1, 5, 20, 15, and 9 images.

We annotated per-pixel material labels of 20 categories by professional paid annotators. All annotations are 1216 × 320 in resolution, and the raw images are center-cropped to this size beforehand. Figure 1 includes an example of road scene color images with their corresponding material annotations. Note that more visual examples can be found in the supplementary material.

Naturally reflecting the real world, our dataset has a very strong imbalance in the material categories, which is a significant challenge for accurate segmentation. Note that there is no manual selection applied to adjust the class-wise distribution of pixels in our dataset. Figure 2 shows pixel statistics with respect to each of the material categories. It shows that 16 material categories span $0.9 \times 10^6 - 1.4 \times 10^8$ pixels, *i.e.*, 99.84% of the total number of pixels. In contrast, 4 categories including "sand," "gravel," "water," and "human body," accounts for 0.083%, 0.016%, 0.00026%, and 0.054% of the overall pixels, respectively.

For evaluation on KITTI-materials, we define two training-test data splits (*i.e.*, Split-1 and -2), where the test set of Split-1 contains more scenes with highways and rural areas while Split-2 is biased to city scenes. Both splits contain 800

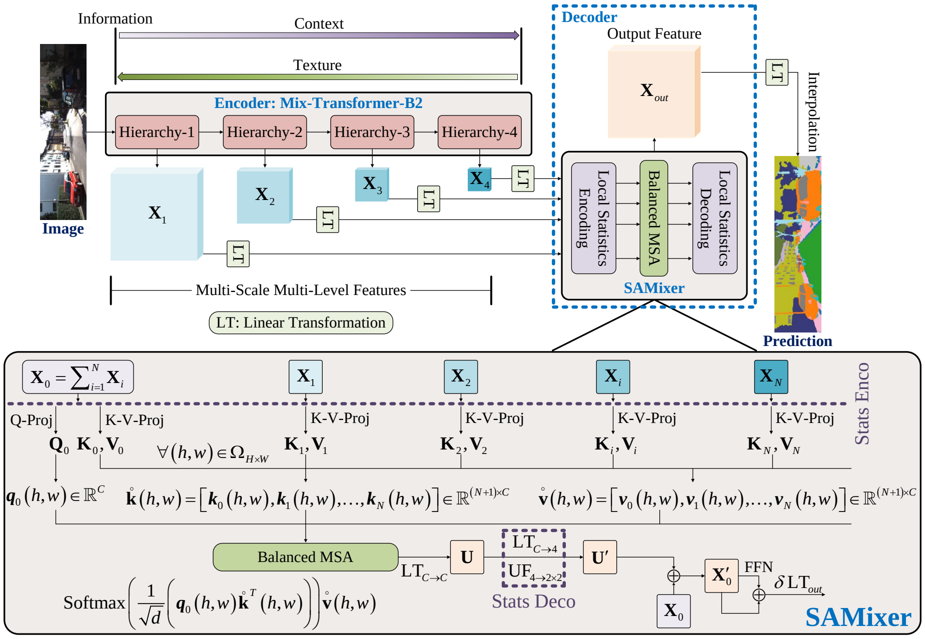
Information
Context
Texture
Decoder
Output Feature
X_{out}
LT
Interpolation
Encoder: Mix-Transformer-B2
Hierarchy-1
Hierarchy-2
Hierarchy-3
Hierarchy-4
Image
X_1
X_2
X_3
X_4
LT
Local Statistics Encoding
Balanced MSA
Local Statistics Decoding
SAMixer
Prediction
Multi-Scale Multi-Level Features
LT: Linear Transformation
$X_0 = \sum_{i=1}^{N} X_i$
X_1
X_2
X_i
X_N
Stats Enco
Q-Proj
K-V-Proj
Q_0 K_0, V_0
$\forall (h,w) \in \Omega_{H\times W}$
K_1, V_1
K_2, V_2
K_i, V_i
K_N, V_N
$q_0(h,w) \in \mathbb{R}^C$
$\mathring{k}(h,w) = [k_0(h,w), k_1(h,w), \ldots, k_N(h,w)] \in \mathbb{R}^{(N+1)\times C}$
$\mathring{v}(h,w) = [v_0(h,w), v_1(h,w), \ldots, v_N(h,w)] \in \mathbb{R}^{(N+1)\times C}$
Balanced MSA
$\mathrm{LT}_{C\to C}$
U
$\mathrm{LT}_{C\to 4}$
$\mathrm{UF}_{4\to 2\times 2}$
Stats Deco
U′
X_0'
FFN
$\delta\,\mathrm{LT}_{out}$
X_0
$\mathrm{Softmax}\left(\frac{1}{\sqrt{d}}\left(q_0(h,w)\mathring{k}^T(h,w)\right)\right)\mathring{v}(h,w)$
SAMixer

Fig. 3. Overview of RMSNet. "LT" denotes "Linear Transformation" layer with corresponding input and output channel-sizes. "Q-Proj" and "K-V-Proj" are "Query-Projection" and "Key- and Value-projection," respectively. "UF" means "Unfold" operation. After obtaining the output feature $\mathbf{X}_{out}$ of SAMixer, we employ a linear layer to generate the segmentation mask from $\mathbf{X}_{out}$ to achieve per-pixel material recognition.

images for training and 200 images for testing, but with different combinations of scenes. Both training and test sets of these two splits show very strong imbalance in the material categories. Further discussions and details including specific components of scenes, visual examples exhibiting their different characteristics of test sets, and per-class statistics are shown in the supplementary material.

4 RMSNet

We introduce RMSNet as a new baseline model for road scene material segmentation. RMSNet effectively fuses texture and contextual cues of material appearance with SAMixer. To the best of our knowledge, it is the first model to realize multi-level feature fusion with the MSA mechanism. Figure 3 depicts the overall architecture of RMSNet.

4.1 Mix-Transformer as Encoder

Hierarchical Feature Encoding. Our RMSNet adopts the middle-size hierarchical transformer encoder introduced by SegFormer [40], namely Mix-Transformer-B2 (MiT-B2), to extract a set of multi-level multi-scale feature maps, from 4 sequential learning stages (*i.e.*, hierarchies). Feature maps extracted from low to high hierarchy levels have high to low resolutions and contain gradually fewer local details of texture and more non-local context cues. For each hierarchy level, an overlapping patch merging layer with corresponding down-sampling ratio is employed to reduce the resolution of the input feature map. Specifically, given an input image with size $H_{in} \times W_{in} \times 3$, the encoder generates a set of hierarchical feature maps $\{\mathbf{X}_i\}$ with corresponding resolutions of $\{H_i \times W_i \times C_i\}$, where $i \in \{1, 2, 3, 4\}$ and C_i denotes the channel-size of $\mathbf{X}_i$. Note that we set $H_i \times W_i \times C_i = \frac{H_{in}}{2^{i+1}} \times \frac{W_{in}}{2^{i+1}} \times C_i$ by default.

Efficient MSA. To process high-resolution features efficiently, MiT-B2 employs *efficient MSA* which uses a $2D$ convolution $Conv_{R\times R}$ with kernel-size of $R \times R$ and stride of R to reduce the resolutions of key and value. Suppose that $\mathbf{Q}$, $\hat{\mathbf{K}}$, and $\hat{\mathbf{V}}$ denote the query, key, and value, efficient MSA can be calculated as

$$\text{Attention}\left(\mathbf{Q}, \hat{\mathbf{K}}, \hat{\mathbf{V}}\right) = \text{Softmax}\left(\frac{\mathbf{Q}\hat{\mathbf{K}}^T}{\sqrt{d}}\right)\hat{\mathbf{V}}\,, \tag{1}$$

where the query, the key, and value are transformed from a given feature map $\mathbf{X}_I \in \mathbb{R}^{H_I \times W_I \times C_I}$ and its condensed feature map $Conv_{R\times R}(\mathbf{X}_I) \in \mathbb{R}^{\frac{H_I}{R} \times \frac{W_I}{R} \times C_I}$. Note that feature maps $\mathbf{X}_I$ and $Conv_{R\times R}(\mathbf{X}_I)$ are reshaped to sizes $M_I \times C_I$ and $\frac{M_I}{R^2} \times C_I$ (*i.e.*, $M_I = H_I W_I$), respectively. Here, $d = \frac{C_I}{g}$, where g denotes the number of heads for an MSA computation. In this way, the computational complexity of an MSA can be controlled with the resolution reduction ratio R. For hierarchy-1 to =4, MiT-B2 assigns $R = 8, 4, 2, 1$, respectively.

Position-aware FFN. MiT-B2 inserts a 3×3 depth-wise convolution $\text{DWConv}_{3\times3}$ in each feed-forward network (FFN), at the top of the first linear layer, to enforce position awareness without additional positional encodings. With this modification, local details can be preserved without the sacrificing accuracy due to interpolation for matching resolutions. The position-aware FFN is defined as

$$\mathbf{X}'' = \text{LT}_2\left(\delta\left(\text{DWConv}_{3\times3}\left(\text{LT}_1\left(\mathbf{X}'\right)\right)\right)\right) + \mathbf{X}'\,, \tag{2}$$

where $\mathbf{X}'$ denotes the attended feature by the MSA layer and $\mathbf{X}''$ is the output feature of the FFN; δ denotes the assigned nonlinear activation (GELU[16] by default); LT_1 and LT_2 are the first and second linear layers, respectively.

4.2 SAMixer-based Decoder

Through careful examination of KITTI-Materials images and also in agreement to past works on material recognition (*e.g.*, [30, 29, 43, 31, 33, 46]), we find that efficient and faithful encoding of local texture patterns of different materials is critical for per-pixel material recognition. The appearance of textures, however, vary significantly with scale and occlusion. Structural dependencies and co-occurrences of local texture features may help extract a representation robust to this variability. Unlike semantic objects, however, materials often show more complicated spatial distributions (*i.e.*, more fragmented) and lack prominent shape cues. This makes fusion of local textures and long-range context cues even challenging. To realize effective fusion for Transformer-induced features, we propose a novel multi-level multi-scale feature fusion model based on MSA, which we refer to as **SAMixer**. Figure 3 depicts the diagram of the SAMixer-based decoder. SAMixer can efficiently fuse local and non-local features to generate robust representations for road scene materials.

Discussion on the Challenge of MSA-based Feature Fusion. MSA introduces informative context dependencies to deep learning representations. For multi-scale feature fusion, however, it inevitably causes excessive computational overhead. Suppose that $\chi = \left\{\mathbf{X}_i \in \mathbb{R}^{H_i \times W_i \times C_i} \mid i = 1, 2, \dots, N\right\}$ is a set of feature maps for fusion ($N = 4$ in our experiments). The fused feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ is generated by mixing all element feature maps $\mathbf{X}_i \in \chi$ at each aligned position $(h, w) \in \Omega_{H \times W}$, where $\Omega_{H \times W}$ denotes the spatial lattice of $\mathbf{X}$. Note that before fusion, each of the feature maps $\mathbf{X}_i$ of different sizes should be transformed and interpolated to the same size $H \times W \times C$ which we refer to as the *anchor size.*

With MSA feature fusion, each transformed and interpolated $\mathbf{X}_i$ with the anchor size is projected to $\mathbf{Q}_i$, $\mathbf{K}_i$, $\mathbf{V}_i$, where $\boldsymbol{q}_i\left(h, w\right)$, $\boldsymbol{k}_i\left(h, w\right)$, $\boldsymbol{v}_i\left(h, w\right)$ with the unified length C are corresponding feature vectors of $\mathbf{Q}_i$, $\mathbf{K}_i$, $\mathbf{V}_i$ at the given spatial position (w, h), respectively. To fuse each feature vector at an aligned position (w, h), MSA can be defined as

$$\text{Attention}\left(\mathbf{q}\left(h, w\right), \mathbf{k}\left(h, w\right), \mathbf{v}\left(h, w\right)\right) = \text{Softmax}\left(\frac{\mathbf{q}\left(h, w\right) \mathbf{k}^T\left(h, w\right)}{\sqrt{d}}\right) \mathbf{v}\left(h, w\right) , \tag{3}$$

where $\mathbf{q}\left(h, w\right), \mathbf{k}\left(h, w\right), \mathbf{v}\left(h, w\right) = \left[\boldsymbol{q}_i\left(h, w\right)\right], \left[\boldsymbol{k}_i\left(h, w\right)\right], \left[\boldsymbol{v}_i\left(h, w\right)\right] \in \mathbb{R}^{N \times C}$ are the query, key, and value of the position (h, w), respectively, formed by arranging the corresponding feature vectors along the row-axis. Let $\mathbf{F}\left(h, w\right) \in \mathbb{R}^{N \times C}$ denote the attended feature descriptor at position (h, w) computed by the MSA layer, each fused feature vector $\mathbf{X}(h, w) \in \mathbb{R}^C$ can be obtained by applying a simple average aggregation along the row-axis or a linear projection on $\mathbf{F}\left(h, w\right)$. In this way, the computational complexity is $O(N^2)$. General MSA can be excessively expensive in computational cost for feature fusion since coarse features of local texture patterns are usually of large sizes.

Proposed SAMixer. We construct SAMixer with MSA consuming only $O(N+1)$ computational complexity by deriving a new balanced query-key similarity (Q-K-Sim) measure in which a container feature is introduced by simply aggregating (*i.e.*, summing) all input features to trigger the MSA computation. SAMixer also introduces a new built-in bottleneck local encoding-decoding (BLSED) strategy to realize further efficiency and accuracy. Figure 3 depicts the SAMixer. In the following paragraphs, we present the two core components of SAMixer, *i.e.*, the balanced Q-K-Sim measure and BLSED strategy.

Balanced Query-Key Similarity Measure. Vanilla query-key similarity measure $\mathbf{q}\left(h, w\right) \mathbf{k}^T\left(h, w\right)$ defines a balanced (*i.e.*, symmetric) computation on the feature set χ for fusion. In contrast, for each query vector $\boldsymbol{q}_i\left(h,w\right)$ (where $i = 1,2,\ldots,N$), its corresponding decomposed group of query-key similarity measure $\boldsymbol{q}_i\left(h,w\right)\mathbf{k}^T\left(h,w\right)$ is imbalanced for different key vectors. That is, for $\forall i,j \in \left\{1,2,\ldots,N\right\}, i \neq j$, the query vector $\boldsymbol{q}_i\left(h,w\right)$ is likely closer to the corresponding key vector $\boldsymbol{k}_i\left(h,w\right)$ than a key vector $\boldsymbol{k}_j\left(h,w\right)$, because both $\boldsymbol{q}_i\left(h,w\right)$ and $\boldsymbol{k}_i\left(h,w\right)$ are generated from the same feature vector $\boldsymbol{x}_i\left(h,w\right) \in \mathbb{R}^C$. As a result, employing a decomposed group of query-key similarity measures independently, although efficient, leads to imbalanced feature fusion and the representational ability of the fused feature is limited.

Our goal is to calculate an efficient balanced MSA on the feature set χ by applying only one group of query-key similarity measures. As depicted in Figure 3, We achieve this by introducing a novel query-key similarity measure which we refer to as the *balanced query-key similarity measure*.

The core idea of this balanced Q-K-Sim measure is the new tailored element feature referred to as the *container feature* that enables balanced computation on a single group of query-key similarity measures. We generate this container feature $\mathbf{X}_0 \in \mathbb{R}^{H\times W\times C}$ by aggregating each of the features in χ with a simple summation (*i.e.*, $\mathbf{X}_0 = \sum_{i=1}^{N}\mathbf{X}_i$). Then, the feature set χ can be expanded into a new set $\mathring{\chi}$ comprising of $N+1$ feature elements by introducing $\mathbf{X}_0$.

Similarly, for $\forall\left(h,w\right) \in \Omega_{H\times W}$, we generate the key and value descriptors $\mathring{\mathbf{k}}\left(h,w\right), \mathring{\mathbf{v}}\left(h,w\right) = [\mathring{\boldsymbol{k}}_i\left(h,w\right)], [\mathring{\boldsymbol{v}}_i\left(h,w\right)] \in \mathbb{R}^{(N+1)\times C}$ from features in $\mathring{\chi}$ and the single query vector $\boldsymbol{q}_0\left(h,w\right) \in \mathbb{R}^C$ from the container feature vector $\boldsymbol{x}_0\left(h,w\right) \in \mathbb{R}^C$, respectively. With this, we can compute an efficient balanced MSA on $\mathring{\chi}$

$$\text{Attention}\left(\boldsymbol{q}_0\left(h,w\right), \mathring{\mathbf{k}}\left(h,w\right), \mathring{\mathbf{v}}\left(h,w\right)\right) = \text{Softmax}\left(\frac{\boldsymbol{q}_0\left(h,w\right)\mathring{\mathbf{k}}^T\left(h,w\right)}{\sqrt{d}}\right)\mathring{\mathbf{v}}\left(h,w\right) . \tag{4}$$

With the proposed balanced Q-K-Sim measure, we can preserve the balance of MSA while reducing the quadratic complexity of $O(N^2)$ to only $O(N+1)$. As a result, our model can effectively fuse high-resolution features to produce discriminative representations for road scene materials.

Bottleneck Local Statistics Encoding-Decoding Strategy. We achieve further efficiency and accuracy by introducing a lightweight embedded encoder-decoder strategy in the SAMixer. Figure 3 depicts the process of the proposed BLSED strategy. We first assign an anchor size $H\times W\times C$, where $H = \frac{H_1}{2^l}$ and $W = \frac{W_1}{2^l}$. Here, $l \in \mathbb{Z}^+$; H_1 and W_1 are the largest height and width of all the input features extracted by the hierarchical encoder. Note that we apply $l = 1$ such that $S = 2^l = 2$ by default in our experiments.

Before the MSA computation, we encode local statistics $\mathbf{U}_i \in \mathbb{R}^{H_i\times W_i\times C}$ from each of the input feature maps $\mathbf{X}_i$ in $\mathring{\chi}$ whose spatial resolutions $H_i \times W_i$ are higher than $H\times W$, respectively, by employing corresponding $2D$ convolutions $Conv_{S_i\times S_i}$ with kernel-size of $S_i \times S_i$ and stride of S_i ($S_i = \frac{H_i}{H}$ is divisible by 2). To reduce computational cost, each $Conv_{S_i\times S_i}$ is replaced by splicing a depthwise convolution with a linear layer. We interpolate all the feature maps to the anchor size whose spatial resolution is smaller than $H\times W$. We preserve the size of any feature naturally possessing spatial resolution of $H \times W$. Note that $\mathring{\chi}$ comprises the container feature $\mathbf{X}_0$, so index $i = 0, 1, \dots, N$. Particularly, since the anchor size is smaller than the highest resolution of features, we produce the container feature by $\mathbf{X}_0 = \sum_{i=1}^{N} \mathrm{UP}_i \left(\mathbf{X}_i\right)$, where UP_i denotes up-sampling via an bilinear interpolation with a scale factor of $\frac{H_1}{H_i}$ (equivalent to $\frac{W_1}{W_i}$). For $i = 1$, UP degrades to an identity mapping.

After the MSA computation, we decode the attended fused feature map $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$ to a high-resolution feature map $\mathbf{U}' \in \mathbb{R}^{H_1 \times W_1 \times C}$ with a channel-spatial decoupled combination scheme which significantly reduces the computational cost. We first generate the spatial mask $\mathbf{P} \in \mathbb{R}^{H_1 \times W_1 \times 1}$ from $\mathbf{U}$ by applying a linear transformation LT_{deco} with input channels of C and output channels of S^2. That is, for each feature vector $\boldsymbol{u}\left(h, w\right) \in \mathbb{R}^{C}$ of $\mathbf{U}$, LT_{deco} generates corresponding S^2 spatial feature units and unfolds these feature units into a spatial feature patch $\mathbf{p}\left(h, w\right) \in \mathbb{R}^{S \times S \times 1}$. Then, each feature vector $\boldsymbol{u}\left(h, w\right)$ and its corresponding spatial feature patch $\mathbf{p}\left(h, w\right)$ are combined by element-wise summation $\oplus$ to produce the feature patch $\mathbf{u}'\left(h, w\right) \in \mathbb{R}^{S \times S \times C}$ of $\mathbf{U}'$:

$$\mathbf{u}'\left(h, w\right) = \left(\mathbf{1}_{S \times S} \otimes \boldsymbol{u}\left(h, w\right)\right) \oplus \left(\mathbf{1}_{C} \otimes \mathbf{p}\left(h, w\right)\right) \,, \tag{5}$$

where $\otimes$ denotes Kronecker product; $\mathbf{1}_{S \times S}$ and $\mathbf{1}_{C}$ each denotes a ones matrix of the corresponding size. Then, we obtain $\mathbf{U}'$ by arranging each of the feature patches $\mathbf{u}'\left(h, w\right)$ according to their spatial position order.

With the proposed BLSED strategy, SAMixer achieves higher efficiency by operating on condensed feature maps.

Segmentation Mask Generation. The output feature $\mathbf{X}_{out} \in \mathbb{R}^{H_1 \times W_1 \times C}$ of SAMixer is generated by applying a linear layer LT_{out} (with a GELU activation δ) over the output of the FFN layer:

$$\mathbf{X}_{out} = \delta\left(\mathrm{LT}_{out}\left(\mathrm{LT}_2\left(\delta\left(\mathrm{LT}_1\left(\mathrm{DWConv}_{3\times3}\left(\mathbf{X}_0'\right)\right)\right)\right) + \mathbf{X}_0'\right)\right) \,, \tag{6}$$

where $\mathbf{X}_0' = \mathbf{X}_0 + \mathbf{U}'$, and LT_1 and LT_2 denote the first and second linear layers of the FFN, respectively. Unlike the FFN in MiT-B2, we employ a depth-wise convolution before LT_1, which increases the speed of the FFN. The segmentation mask is obtained by employing a linear layer with an output channel-size of the number of material classes (*i.e.*, 20) on $\mathbf{X}_{out}$.

5 Experiments and Discussions

We evaluate the effectiveness of our method on the KITTI-Materials dataset with detailed ablation studies and also thorough comparison with past material segmentation methods [1, 32], road scene semantic segmentation methods with CNN encoders [6, 44, 21, 50], related state-of-the-art transformers [11, 39, 40] and a gating-based dynamic network [51] that have been applied to semantic segmentation. Note that DeepLabv3+ [6] also represents Liang *et al.* [22] without semantic segmentation masks and using RGB only.

Table 1. Material segmentation results on KITTI-Materials dataset for our methods and other methods. “Trs” denotes “Transformer”; “DLv3+” denotes DeepLabv3+; Symbols “◇” and “⋆” denote “modified” and “our implementation,” respectively; Symbol “‡” denotes methods whose original code cannot support multi-GPU training/inference settings. ViT [11] and CvT [39] are applied with the All-mlp decoder [40].

Method	Backbone	Params	Fps↑	Split-1	Split-2
				mIoU(%)↑	
MINC‡ [1]	VGG16 [34]	134.34M	15.88	29.73	32.12
Matcontext‡ [32]	VGG16◇ [32]	25.42M	7.40	30.87	33.16
DLv3+ [6]	ResNet101 [15]	59.34M	14.60	41.35	46.09
	SK-ResNet101 [21]	60.47M	14.08	41.96	46.04
DeeperLab [44]	ResNet101 [15]	240.58M	11.29	42.56	47.12
PSPNet [50]	ResNet101 [15]	43.38M	14.22	31.92	37.11
DDF-DL [51]	DDFNet101 [51]	42.94M	12.66	41.55	46.41
ViT [11]	ViT-B/16 [11]	89.03M	13.69	40.02	46.06
CvT [39]	CvT-13 [39]	21.89M	18.02	41.72	47.54
SegFormer [40]	Mix-Trs-B2 [40]	27.36M	18.87	44.47	48.32
RMSNet **(ours)**	Mix-Trs-B2 [40]	31.53M	16.81	**46.82**	**50.34**

5.1 Implementation Details

Two different training-test data splits (denoted by *Split*-1 and -2, respectively) of KITTI-Materials with different characteristics are used for evaluation. The test set of “split-1” contains more scenes with highways and rural areas while “split-2” is biased to city scenes (see the supplemental material for details). Both splits consist of all 1000 images of KITTI-Materials where 800 images for training and 200 images for testing with different split rules. For all models, we apply the AdamW optimizer with a weight decay of 0.01 for 300 epochs including 10 epochs of linear warm-up. Following [40], we start the learning rate from 6×10^{-5} and 6×10^{-4} for encoders and decoders, respectively, with a cosine decay scheduler and a mini-batch of 16. We adopt standard image augmentation settings [6]. In the training phase, images are randomly center-cropped and then resized to 512×512 pixels, while in the testing, images are fixed to the original size (*i.e.*, 1216×320 pixels). To reduce the negative effect of extreme data imbalance, we calculated balancing weights based on class frequencies of materials and applied them to CE-losses of all models. Experiments are conducted on a computer with 4 × RTX A5000 GPUs. For fair comparisons, all encoders of our method and compared methods use ImageNet [10] pre-trained weights obtained from corresponding open-sourced projects or websites. All methods are evaluated in the raw image size without multi-scale averaging augmentation [47]. We use mean intersection of union (mIoU) to evaluate the performance of each model.

5.2 Experimental Results on KITTI-Materials

Based on the proposed KITTI-Materials dataset, we verify the effectiveness of our network designs by comparing with (1) existing general material segmenta-

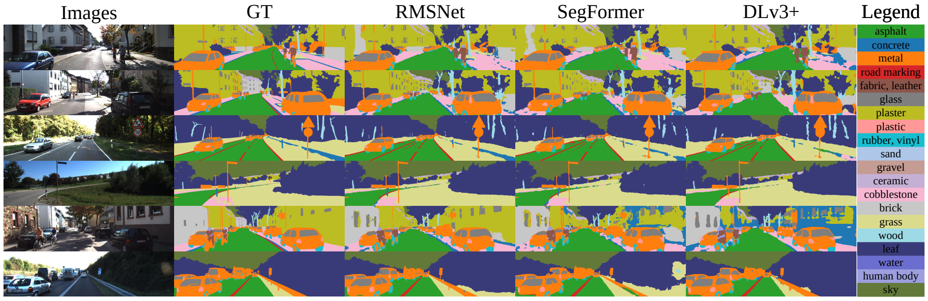

Fig. 4. Visual examples on KITTI-Materials. Compared with DeepLabv3+ [6] (denoted by "DLv3+") and SegFormer [40]. "GT" denotes "ground truth".

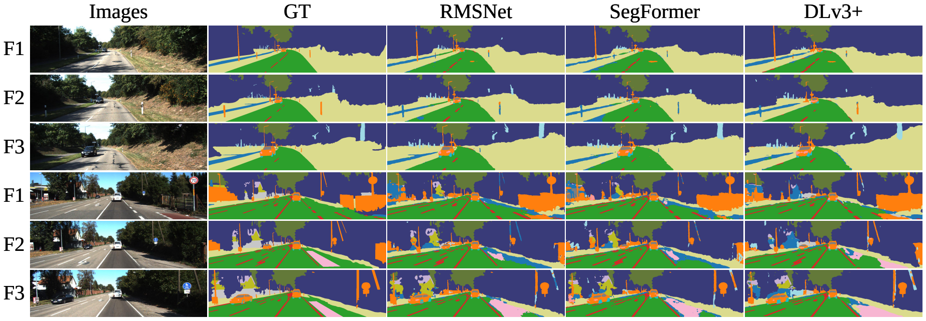

Fig. 5. Visual examples of moving cars of different scales. "F" denotes "frame."

tion methods for RGB images [1, 32]; (2) popular road scene semantic segmentation methods with CNN encoders [6, 44, 50, 15]; (3) enhanced DeepLabv3+ [6] with a multi-scale fusion method (*i.e.*, SKNet [21]) and a state-of-the-art (SOTA) gating-induced dynamic networks [51]; (4) related SOTA transformers [11, 39, 40] that have been validated on semantic segmentation, where SegFormer [40] is the closest model to our RMSNet.

As shown in Table 1, our RMSNet enjoys clear improvements over all compared methods for general material segmentation and road scene semantic segmentation in accuracy. Note that the major difference between RMSNet and SegFormer-B2 is the replacement of the All-MLP decoder with our SAMixer-based decoder. The results demonstrate the effectiveness of our network for road scene material segmentation. Compared with popular semantic segmentation frameworks [6, 44, 50] with CNN encoders [15, 21] and the SOTA gating-induced dynamic network [51], RMSNet yields significant gains in accuracy. Our RMSNet also shows further accuracy improvements over other compared SOTA Transformers [11, 39, 40] with competitive efficiency.

To demonstrate detailed performance differences on each material, we report the per-class comparative results (see the supplemental material) with visual examples shown in Figure 4, where we find that our network outperforms competing baseline DeepLabv3+ [6] and the SOTA SegFormer [40] by a clear margin on categories "fabric," "glass," "metal," "rubber," and "human body." These materials span a wide range of appearances as part of different semantic objects (*e.g.*, vehicles, bicycles, road markings, and pedestrians). Figure 5 demonstrates the significance of incorporating tailored texture-context feature fusion for road scene material segmentation through visual comparison between RMSNet and the compared methods. Our RMSNet with the SAMixer module achieves cleaner segmentation on windows, headlights, vehicle bodies, and wheels of moving cars of different scales (*i.e.*, different distances from viewpoint).

6 Ablation Study

Using split-1 of the KITTI-Materials dataset, we conduct targeted ablation studies on the proposed SAMixer and its core ingredients, *i.e.*, the *balanced query-key similarity measure* and *bottleneck local encoding-decoding strategy*, to verify their effectiveness in adding efficiency and accuracy.

6.1 Balanced Query-Key Similarity Measure

Here we independently discuss and validate the proposed balanced Q-K-Sim measure. We build an abridged SAMixer (denoted by "SAM-a") by removing the BLSED strategy, and conduct a comparative experiment by introducing three targeted control groups of feature fusion models built on (1) the raw MSA mechanism (denoted by "MSA-raw"); (2) a series of imbalanced partial MSA mechanisms where the queries are only transformed from one of the four feature maps for fusion (denoted by "MSA-ib1" to "-ib4"); (3) the All-MLP module of SegFormer [40] (denoted by "SegF"). We use SAM-a and these control groups to replace the original SAMixer to build abridged RMSNets and compare them on our KITTI-Materials dataset. To prevent excessive computational overhead, all methods SAM-a, MSA-raw, and "MSA-ib1" to "-ib4" employ the resolution reduction strategy suggested in the Mix-Transformer [40] with reduction ratio $R = 2$. We also add DeepLabv3+ [6] (denoted by "DLv3+") as a baseline method. Table 2(a) and (b) show that SAM-a and MSA-raw introduce similar accuracy gains while SAM-a consumes far fewer computational costs; MSA-ib1 to -ib4 lead to close/less accuracy gains to the SegFormer, demonstrating the effectiveness of the balanced Q-K-Sim measure. This confirms the effectiveness of our proposed multi-scale feature fusion model.

6.2 Bottleneck Local Statistics Encoding-Decoding Strategy

Effectiveness. We propose this tailored built-in strategy to achieve further efficiency and accuracy. To evaluate its effectiveness, we compare the original SAMixer (denoted by "SAM") with two targeted control groups (1) the abridged SAMixer without the BLSED strategy (*i.e.*, "SAM-a" introduced in Section 6.1);

Table 2. Ablation studies on (a) and (b) the balanced Q-K-Sim measure, (c) effectiveness and (d) resolution reduction ratio setting of the BLSED strategy. "DLv3+", "SegF," and "SAM" denote "DeepLabv3+ [6]," "SegFormer [40]," and our "SAMixer," respectively. "SAM-a" denotes the abridged SAMixer without BLSED strategy.

(a)

Method	Fps↑	mIoU↑
DLv3+	14.60	41.35
SegF	18.87	44.47
MSA-raw	10.98	**45.51**
SAM-a	15.39	45.33

(b)

Method	Fps↑	mIoU↑
MSA-ib1	17.47	**44.17**
MSA-ib2	17.24	43.08
MSA-ib3	17.55	43.94
MSA-ib4	17.71	43.29

(c)

Method	Fps↑	mIoU↑
DLv3+	14.60	41.35
SegF	18.87	44.47
SAM-a	15.39	45.33
SAM	16.81	**46.82**

(d)

Ratio	Fps↑	mIoU↑
DLv3+	14.60	41.35
W/o	15.39	45.33
2	16.81	**46.82**
4	15.75	45.74

(2) the SegFormer All-MLP module (denoted by "SegF"). The comparative results are reported in Table 2(c). Our RMSNet with SAMixer outperforms the abridged SAMixer SAM-a in both accuracy and efficiency. RMSNet also improves the SOTA SegFormer by a clear margin on accuracy. The results verify the effectiveness of our network design.

Reduction Ratio Setting. We assign an unified resolution reduction ratio (denoted by "Ratio") as the stride and kernel-size for corresponding depth-wise convolutions to control the encoding process of the local statistics of feature maps for fusion. We conduct this ablation study to evaluate the effectiveness of the BLSED strategy with different Ratio settings. Table 2(d) reports the comparative results with reduction ratios of "W/o", 2, and 4, where "W/o" denotes removing BLSED strategy (*i.e.*, "SAM-a" introduced in Section 6.1). Based on the results, we set "ratio=2" by default, since it reaches high accuracy with competitive efficiency, compared with other counterparts.

7 Conclusion

We address RGB road scene material segmentation by constructing a new benchmark dataset, KITTI-Materials, and by deriving a new network that effectively fuses texture and contextual cues for accurate per-pixel material recognition. The network, *i.e.*, RMSNet, achieves this with a newly derive SAMixer module built on a balanced Q-K-Sim measure and a BLSED strategy. Experimental validations and ablation studies on KITTI-Materials dataset confirm the effectiveness of our proposed designs. We believe our data and model can contribute to further studies on leveraging rich visual material information for road scene understanding and will serve as a sound baseline to tackle this challenging task. As a limitation, RMSNet falls short in extracting very high-resolution features of abundant local textures with adequate computational efficiency, which we plan to address in future work. We hope dissemination of our data and code will catalyze further studies on this important and challenging visual task.

Acknowledgements This work was in part supported by JSPS 20H05951, 21H04893, JST JPMJCR20G7, and SenseTime Japan.

References

1. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material Recognition in the Wild with the Materials in Context Database. In: Proc. CVPR. pp. 3479–3487 (2015)
2. Brown, M., Goldberg, H., Foster, K., Leichtman, A., Wang, S., Hagstrom, S., Bosch, M., Almes, S.: Large-Scale Public Lidar and Satellite Image Data Set for Urban Semantic Labeling. In: Laser Radar Technology and Applications XXIII. vol. 10636 (2018)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. TPAMI **40**(4), 834–848 (2018)
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1706.05587 (2017)
5. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to Scale: Scale-Aware Semantic Image Segmentation. In: Proc. CVPR. pp. 3640–3649 (2016)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Proc. ECCV. pp. 833–851 (2018)
7. Choi, S., Kim, J.T., Choo, J.: Cars Can't Fly Up in the Sky: Improving Urban-Scene Segmentation via Height-Driven Attention networks. In: Proc. CVPR. pp. 9373–9383 (2020)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: Proc. CVPR. pp. 3213–3223 (2016)
9. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raska, R.: Deepglobe 2018: A Challenge to Parse the Earth Through Satellite Images. In: Proc. CVPR Workshop. pp. 172–17209 (2018)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A Large-Scale Hierarchical Image Database. In: Proc. CVPR. pp. 248–255 (2009)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: Proc. ICLR (2021)
12. Geiger, A., Lenz, P., Urtasun, R.: Are We Ready for Autonomous Driving? The Kitti Vision Benchmark Suite. In: Proc. CVPR (2012)
13. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: A Vision Transformer in Convnet's Clothing for Faster Inference. In: Proc. ICCV (2021)
14. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in Transformer. In: Proc. NeurIPS (2021)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proc. CVPR. pp. 770–778 (2016)
16. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. Proc. CVPR (2018)
18. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNet: Criss-Cross Attention for Semantic Segmentation. In: Proc. ICCV. pp. 603–612 (2019)
19. Kraehenbuehl, P., Koltun, V.: Parameter Learning and Convergent Inference for Dense Random Fields. In: Proc. ICML. pp. 513–521 (2013)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Proc. NeurIPS. pp. 1097–1105 (2012)

21. Li, X., Wang, W., Hu, X., Yang, J.: Selective Kernel Networks. In: Proc. CVPR. pp. 510–519 (2019)
22. Liang, Y., Wakaki, R., Nobuhara, S., Nishino, K.: Multimodal Material Segmentation. In: Proc. CVPR (2022)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: Proc. ICCV (2021)
24. Neuhold, G., Ollmann, T., Bulo, S.R., Kontschieder, P.: The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In: Proc. ICCV. pp. 5000–5009 (2017)
25. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification. In: Proc. ECCV. vol. 6314, pp. 143–156 (2010)
26. Purri, M., Xue, J., Dana, K.J., Leotta, M.J., Lipsa, D., Li, Z., Xu, B., Shan, J.: Material Segmentation of Multi-View Satellite Imagery. arXiv preprint arXiv:1904.08537 (2019)
27. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-Alone Self-Attention in Vision Models. In: Proc. NeurIPS. vol. 32, pp. 68–80 (2019)
28. Reda, F.A., Liu, G., Shih, K.J., Kirby, R., Barker, J., Tarjan, D., Tao, A., Catanzaro, B.: SDC-Net: Video Prediction Using Spatially-Displaced Convolution. In: Proc. ECCV. pp. 747–763 (2018)
29. Schwartz, G., Nishino, K.: Visual Material Traits: Recognizing Per-Pixel Material Context. In: IEEE Color and Photometry in Computer Vision Workshop (2013)
30. Schwartz, G., Nishino, K.: Automatically Discovering Local Visual Material Attributes. In: Proc. CVPR (2015)
31. Schwartz, G., Nishino, K.: Integrating Local Material Recognition with Large-Scale Perceptual Attribute Discovery. arXiv preprint arXiv:1604.01345 (2016)
32. Schwartz, G., Nishino, K.: Material Recognition from Local Appearance in Global Context. arXiv preprint arXiv:1611.09394 (2016)
33. Schwartz, G., Nishino, K.: Recognizing Material Properties from Images. TPAMI **42**(8), 1981–1995 (2020)
34. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions. In: Proc. CVPR. pp. 1–9 (2015)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All You Need. In: Proc. NeurIPS. vol. 30, pp. 5998–6008 (2017)
36. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A.L., Chen, L.C.: Axial-Deeplab: Stand-Alone Axial-Attention for Panoptic Segmentation. In: Proc. ECCV. pp. 108–126 (2020)
37. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In: Proc. ICCV (2021)
38. Wang, X., Girshick, R., Gupta, A., He, K.: Non-Local Neural Networks. In: Proc. CVPR. pp. 7794–7803 (2018)
39. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: CvT: Introducing Convolutions to Vision Transformers. In: Proc. ICCV. pp. 22–31 (2021)
40. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: Proc. NeurIPS (2021)

41. Xue, J., Purri, M., Dana, K.: Angular luminance for material segmentation. In: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium (2020)
42. Xue, J., Zhang, H., Dana, K., Nishino, K.: Differential Angular Imaging for Material Recognition. In: Proc. CVPR. pp. 6940–6949 (2017)
43. Xue, J., Zhang, H., Nishino, K., Dana, K.: Differential Viewpoints for Ground Terrain Material Recognition. TPAMI pp. 1–1 (2020)
44. Yang, T.J., Collins, M.D., Zhu, Y., Hwang, J.J., Liu, T., Zhang, X., Sze, V., Papandreou, G., Chen, L.C.: Deeperlab: Single-Shot Image Parser. arXiv preprint arXiv:1902.05093 (2019)
45. Zhang, H., Goodfellow, I.J., Metaxas, D.N., Odena, A.: Self-Attention Generative Adversarial Networks. In: Proc. ICML. pp. 7354–7363 (2018)
46. Zhang, H., Dana, K., Nishino, K.: Reflectance Hashing for Material Recognition. In: Proc. CVPR. pp. 3071–3080 (2015)
47. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context Encoding for Semantic Segmentation. In: Proc. CVPR. pp. 7151–7160 (2018)
48. Zhang, H., Xue, J., Dana, K.: Deep Ten: Texture Encoding Network. In: Proc. CVPR. pp. 2896–2905 (2017)
49. Zhang, H., Zhang, H., Wang, C., Xie, J.: Co-Occurrent Features in Semantic Segmentation. In: Proc. CVPR. pp. 548–557 (2019)
50. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid Scene Parsing Network. In: Proc. CVPR (2017)
51. Zhou, J., Jampani, V., Pi, Z., Liu, Q., Yang, M.H.: Decoupled Dynamic Filter Networks. In: Proc. CVPR (2021)
52. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving Semantic Segmentation via Video Propagation and Label Relaxation. In: Proc. CVPR. pp. 8856–8865 (2019)