RawSeg: Grid Spatial and Spectral Attended Semantic Segmentation Based on Raw Bayer Images

Guoyu Lu
https://sites.google.com/view/guoyulu/home

Intelligent Vision and Sensing Lab
University of Georgia
Athens, USA

Abstract

Semantic segmentation methods are typically designed for RGB color images, which are interpolated from raw Bayer images. While RGB images provide abundant color information and are easily understood by humans, they also add extra storage and computational burden for neural networks. On the other hand, raw Bayer images preserve primitive color information with a single channel, potentially increasing segmentation accuracy while significantly decreasing storage and computation time. In this paper, we propose RawSeg-Net to segment single-channel raw Bayer images directly. Different from RGB images that already contain neighboring context information during ISP color interpolation, each pixel in raw Bayer images does not contain any context clues. Based on Bayer pattern properties, RawSeg-Net assigns dynamic attention on Bayer images' spectral frequency and spatial locations to mitigate classification confusion, and proposes a re-sampling strategy to capture both global and local contextual information.

1 Introduction

Scene segmentation is a fundamental and challenging topic in computer vision with a wide range of applications, such as in autonomous driving, augmented reality, medical imaging, etc [1] [26] [11]. The vast majority of current semantic segmentation algorithms take the 3-channel color images after image signal processor (ISP) pipelines as inputs. To output RGB color images, the ISP pipeline will consume extra time from raw Bayer images and may damage or lose primitive pixel information captured by the raw camera sensor due to the operations like demosaicing, exposure adjustment, and many other middle processes in ISP [10]. Raw Bayer images contain all the necessary color and intensity gradient information in a single channel, making them an efficient source for RGB images. They save up to 67% of image storage

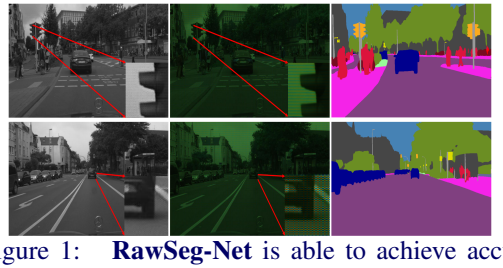

Figure 1: **RawSeg-Net** is able to achieve accurate scene segmentation from 8-bit (single channel) Bayer pattern image (**left**). The middle image is the result of the left image overlaid with color Bayer pattern for better observation. Our method precisely segments moving objects (pedestrians, cars) and object boundaries (buildings, roads) (**right**).

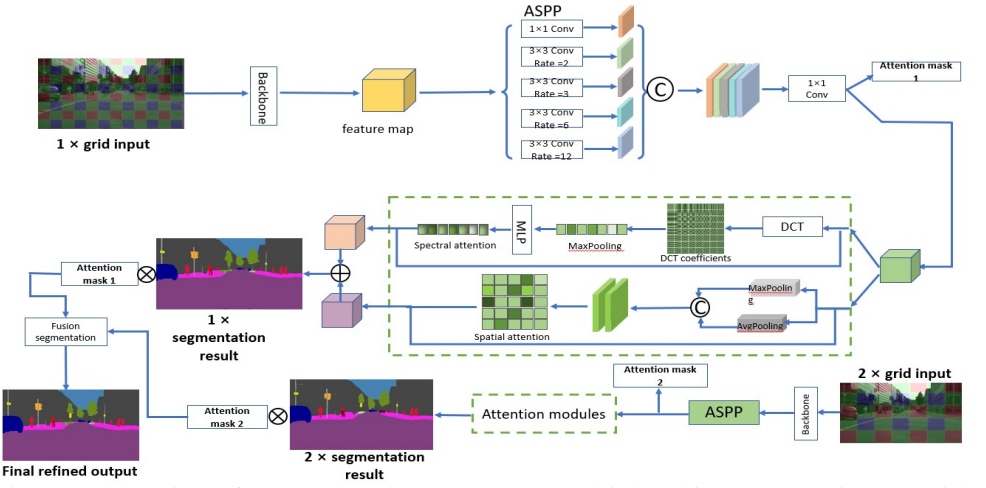

Figure 2: Overview of our proposed RawSeg-Net. Multiple 8-bit raw Bayer images with different grid sizes are input to the backbone to extract low-level features. We further deploy ASPP module to extract the contextual information, introduce spatial coordinate attention to focus on each grid coordinate, and utilize spectral frequency attention to focus on each split spectrum. By concatenating different weighted feature maps and fusing the class maps from different grid sizes, the raw Bayer image can be accurately segmented.

space and can potentially increase the image processing pipeline by eliminating the ISP process, which is a significant time-consuming step. By processing single-channel images, the computation burden and neural network complexity can also be reduced. Therefore, raw Bayer images have several advantages, including completeness and accuracy of color information, efficient storage, fast processing speed, and reduced network complexity. The widely used Bayer pattern, arranged in a repeated 2×2 matrix grid containing one red component, one blue component, and two green components, is typically used to generate raw Bayer images.Despite the numerous benefits of raw Bayer images over RGB images, there is currently a shortage of segmentation algorithms specifically designed for Bayer patterns.

In this work, we demonstrate the usability of raw Bayer images on scene segmentation tasks and propose a semantic segmentation network designated for raw Bayer image RawSeg-Net in order to accurately segment raw Bayer images, as Fig. 1. Unlike RGB color images that maintain neighboring contextual information during ISP color interpolation, raw Bayer images' pixels miss the context clues from neighboring locations from spectral and spatial perspectives. Therefore, to effectively utilize Bayer pattern, we explore a spatial coordinate attention mechanism to accurately allocate attention weights to each specific pixel by aggregating diverse feature maps and spectral frequency attention to capture different light wavelengths and high frequency details contained in the raw Bayer image. As scene images are commonly composed of objects of various sizes (e.g., building as large structures and traffic signs as fine structures), we compose the grids into different sizes (e.g., one composed grid maintains 4 small grids with the same color) to segment the image and fuse segmentation outputs with different grid sizes to precisely segment the images with objects in various scales, benefiting from the Bayer grid pattern. With convolution kennels dedicated to Bayer patterns, RawSeg-Net can capture spatial and spectral features at various grid sizes to realize precise segmentation based on raw images. Our method is detailed in Fig. 2.

To sum up, this paper makes several significant contributions. 1) We demonstrate that

single-channel raw Bayer images are highly suitable for image segmentation tasks, offering advantages such as reduced storage requirements, faster image processing, and less complex neural networks. 2) We propose novel spatial coordinate attention and spectral frequency attention mechanisms designed specifically for Bayer images, allowing for highly accurate semantic segmentation. 3) We introduce a fusion strategy that leverages different grid sizes of the Bayer pattern to effectively segment objects of varying scales.

2 Related work

Semantic Segmentation: Benefiting from the successful usages of deep Convolutional Neural Networks (CNNs) [21] [25] [15] [3] [16], semantic segmentation has achieved significant improvement towards understanding a complex scene. Fully Convolutional Networks (FCN) [21] first applied a fully convolutional network in semantic segmentation tasks. Following FCN, extensive research was proposed based on the FCN architecture, such as UNet [25], SegNet [2], PSPNet [15] and DeepLab-based [3] [4] works. Recently, PSANet [37] proposed a point-wise attention network to learn attention for each feature map position for scene parsing. HRNet [32] started from a high resolution convolution stem and gradually added high-to-low resolution blocks. In addition to CNN features based on color and texture information for segmentation, depth information is also applied to support segmentation tasks [22, 23]. Large models, like SAM [13], are also proposed for segmentation tasks. Existing semantics segmentation schemes are mainly designed for RGB color images without focusing on raw Bayer images, which are the source of RGB images.

Context Attention: Contextual information is critical in various vision-based tasks such as semantic segmentation. An increasing number of works have explored contextual dependencies and context-weighted information, especially attention mechanisms. Different strategies are proposed to explore long-term attention dependencies [31] [28] [33] [6]. Wang et al. [33] presented a self-attention module with non-local operations to capture long-range dependencies in spatial-temporal dimensions to process videos and images. DANet [6] applied a dual-attention strategy to combine information from the input images and the final feature maps. Different from attention mechanisms commonly applied to RGB color images, this paper focuses on the affluent contextual relationships contained in the Bayer patterns to better capture the shape and spectral information explicitly existing in raw Bayer images.

Bayer Pattern: Most of the works using Bayer Color Filter Array (CFA) are designed for image demosaicing, which is to interpolate the vacant red, green and blue values in the raw Bayer pattern images to restore 3-channel RGB color images [17] [35] [24] [20]. Various clues have been investigated to interpolate RGB color information, such as color difference [5], edge direction [14] and image reconstruction [27]. Deep learning approaches have also been applied in image demosaicing [30] [29] [19]. In particular, Liu et al. [19] proposed a self-guidance network to use an initially estimated green channel as guidance to recover all missing values in the input image. Another typical application for Bayer images is image restoration. Bayer images have also bee applied to object detection tasks [7]. Zhou et al. [39] proposed to restore images from the raw Bayer domain. However, Bayer images have rarely been applied to image segmentation tasks, mainly because Bayer images are not convenient for human eyes to observe.

3 Raw Bayer Image Segmentation Framework

RawSeg-Net is specially designed for raw Bayer images to map to pixel-level class annotations. The introduction of dynamic attention mechanisms on the Bayer pattern helps coordinate and split spectral wavelengths under multiple Bayer grid sizes.

3.1 Raw Bayer Pattern

Most commercial digital cameras have a single CCD/CMOS sensor that captures the intensity of light, but not its color wavelength. To produce color information, the sensor is overlaid with a Bayer "color filter array" (CFA), which filters the captured pixels and produces different spectral channels. This results in a raw Bayer image I^{bayer}, which is an image mosaic. To recover the full RGB color S from the separate spectral channels S^R, S^G and S^B, where $S = S^R \bigcup S^G \bigcup S^B$, S^B and S^R each occupy a quarter of all pixels, and S^G occupies half of all image pixels arranged in a quincunx lattice. Fig. 3 shows the zoom-in details of a 20×20 region in the captured raw Bayer image by a single CCD sensor equipped with a Bayer pattern filter, as well as a rendering illustration where each sample point is plotted with Bayer color.

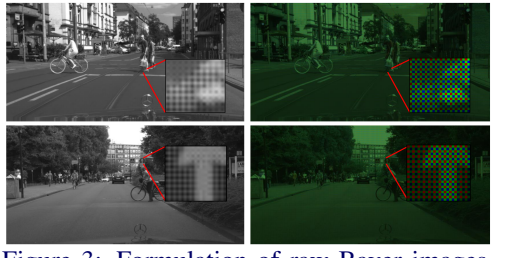

Figure 3: Formulation of raw Bayer images. For each sample image (row), left is the zoom-in 20×20 pixels' region of the raw Bayer image; right is the raw Bayer image overlaid with color Bayer pattern for better observation.

Demosaicing methods are typically used to interpolate missing color information and recover the full RGB color image from a raw Bayer image. However, in scenes with high contrast and constantly changing colors or objects, demosaicing may result in the loss of details and introduce color artifacts like bleeding and zippering. Furthermore, post-processing stages such as demosaicing can be computationally expensive, which makes raw Bayer images a more cost-effective option for end-to-end semantic segmentation. In contrast to RGB images, raw Bayer images preserve the most primitive color information, making them ideal for semantic segmentation. The Bayer CFA used in typical post-processing steps is illustrated in Fig. 4.

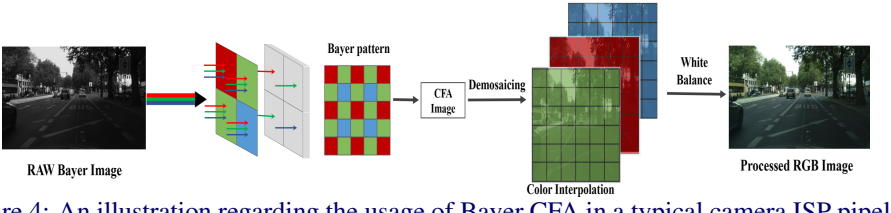

Figure 4: An illustration regarding the usage of Bayer CFA in a typical camera ISP pipeline.

3.2 Spectral Frequency Attention Block

Effective utilization of color information is crucial for various computer vision tasks, including segmentation and detection, as it provides a wider spectral perception field with multiple color channels. To strengthen the features that encode spectral information and reduce the impact of ineffective features, it is essential to recalibrate them. This is particularly relevant for Bayer images that contain only a single color channel per pixel, as all three RGB channels have already been interpolated from neighboring pixels during the ISP process, which encodes contextual information in the image. However, this contextual information is not encoded in the raw Bayer images. To learn spectral at-

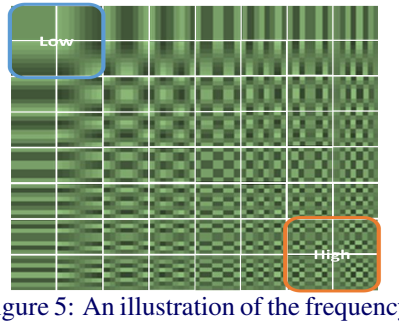

Figure 5: An illustration of the frequency components from the Discrete Cosine Transform (DCT) of feature maps.

tention from the context information, we propose to decompose the image into frequency spectra using Discrete Cosine Transform (DCT), which has been largely used in image and video compression applications. The DCT representation expresses an image as a sum of sinusoids at varying magnitudes and frequencies. Given an input image x of size $M \times N$, the 2D DCT spectrum $B \in R^{M \times N}$ is obtained as:

$$B_{pq} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} a_{p,q} \cos \frac{\pi(2m+1)p}{2M} \cos \frac{\pi(2n+1)q}{2N} \tag{1}$$

where $a_{0,0}$ is $\frac{1}{\sqrt{MN}}$, which corresponds to the lowest frequency component in the left top regions of Fig. 5. $a_{p,q}$ is $\frac{2}{\sqrt{MN}}$ for all other frequency components of the 2D DCT. $B_{M-1,N-1}$ corresponds to DCT coefficients of the highest frequency in bottom right regions of Fig. 5.

Given the input feature map $X \in R^{C \times H \times W}$, DCT coefficients $A \in R^{F \times C \times H \times W}$ are computed for the selected F frequency components. Reshaping X to $1 \times C \times H \times W$, conducting element-wise multiplication with A, and summarizing the output across the spatial coordinate, the embedded frequency matrix will be $D \in R^{C \times J}$, $\mathrm{J} = \mathrm{H} \times \mathrm{W}$. The embedding is then forwarded to choose the maximum frequency response per channel via max pooling. The final weighted feature map output Y is generated by a fully connected layer and sigmoid activation, as shown in Fig. 6.

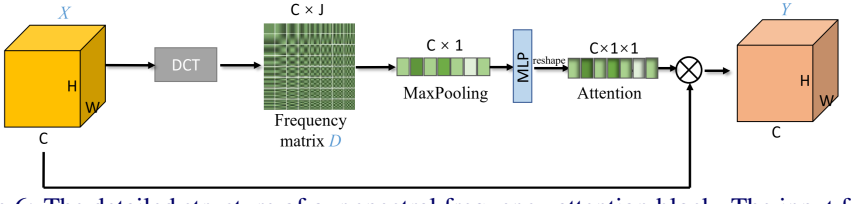

Figure 6: The detailed structure of our spectral frequency attention block. The input feature map is first decomposed to spectral frequency domain. The DCT coefficients are then formulated to one channel attention map via a MaxPooling and MLP. The output feature map weighted by the attention map through element-wise multiplication shares the same dimension as the input.

3.3 Spatial Attention Module

To extract smooth and continuous segmentation boundaries, a larger spatial perception field covering locations with salient and continuous color information is necessary for raw Bayer images where neighboring pixels do not have continuous color changes. To address this, we introduce a spatial attention module, as shown in Fig. 7. The raw Bayer image is composed of grids of light-sensitive cells, and the spatial attention block enhances a wide range of contextual information into the local Bayer point. To explore spatial attention in raw Bayer images, we process the input feature $X' \in R^{C \times H \times W}$ separately with global average pooling (GAP) and global max pooling (GMP) along the feature channels, and aggregate the results for concatenation. This process is expressed as:

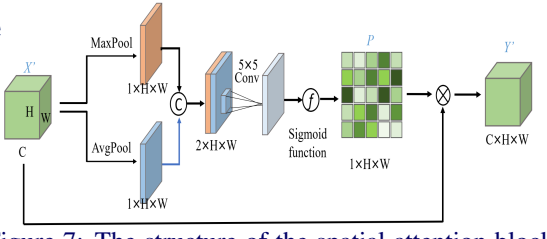

Figure 7: The structure of the spatial attention block, where " $\otimes$" denotes the element-wise Handamard multiplication. Given an input feature map X', the spatial attention block learns a spatial attention map P and generates the spatial attention guided feature Y'.

$$Pool\_block(X') = Concat\left\{GAP(X'), GMP(X')\right\} \tag{2}$$

where the output of the Pool_block (X') is in a tensor of shape $2 \times H \times W$. The output is then followed by a 5×5 convolutional layer and a batch normalization layer. The output is then passed through a sigmoid activation layer $(\cdot)$ to generate a $1 \times H \times W$ attention map. The final weighted attention output is element-wisely multiplied with the original input feature map as:

$$Y' = \sigma\left(Conv\left(Pool\_block(X')\right)\right) \otimes X' \tag{3}$$

where the output feature map Y' shares the same dimension as the input feature map X' as $C \times H \times W$. With the introduced spectral frequency attention (SFA) and spatial coordinate attention (SCA) blocks, the extracted features from the backbone are able to adaptively emphasize on both Bayer spectrum and coordinate. The SFA and SCA are concatenated together to several convolution layers to generate the final pixel-level estimation map, as shown in Fig. 8.

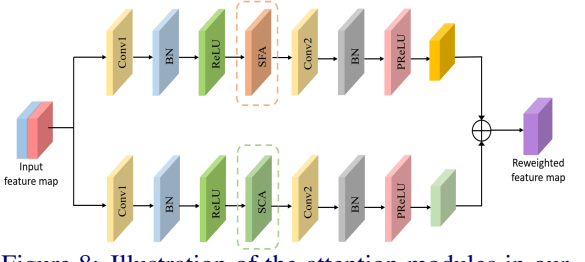

Figure 8: Illustration of the attention modules in our entire pipeline.

The overall objective function for training the RawSeg-Net is a combination of the normal cross-entropy loss (between the estimated segmentation output y and the ground truth label y) and the RMI loss [38] (between the estimated probability of segmentation labels and the probability of the ground truth labels) as:

$$L_{seg}(y,\tilde{y}) = \lambda_{ce} L_{ce}(y,\tilde{y}) - \lambda_{rmi} I_l(\mathbf{Y},\tilde{\mathbf{Y}}) \tag{4}$$

where L_{ce} is the per-pixel cross-entropy loss, and $I_l(\mathbf{Y},\tilde{\mathbf{Y}})$ denotes the lower bound of the mutual information of estimated and the ground truth variables. $L_{\text{seg}}(y,\tilde{y})$ is designed to simultaneously minimize the dissimilarity and maximize the lower bound of the mutual information to enable the estimated segmentation map to achieve high-order consistency with the ground truth segmentation map.

3.4 Multi-grid Re-sampling Strategy

Considering the raw Bayer image is composed of pixel level light-sensitive cells, a more focused strategy for processing multi-level grids is essential for the final segmentation output. In contrast with [56] [8] [9] that use different sampling operations or pyramid pooling to obtain a multiscale representation, we instead opt for re-sampling the original grid size to a larger size Bayer grid. More specifically, we group pixels in each original 2×2 Bayer pattern together and assign the same color to the newly composed larger Bayer grid so that the entire Bayer image is composed of Bayer patterns with larger size. Therefore, the individual pixel-wise segmentation estimations are combined with dynamically learned scale-aware weights followed by a pixel-wise summation for generating the final refined segmentation. The re-sampling strategy of the raw Bayer image and the dynamic weight guided output refinement steps are depicted in Fig. 9.

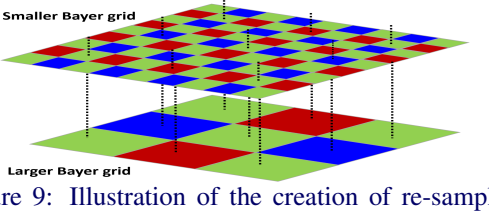

Figure 9: Illustration of the creation of re-sampled Bayer grids while the image size remains the same. A larger Bayer grid composes of small Bayer pixels without reducing original dimension.

With this strategy, we observe that the final refined output performs better than the estimations with original Bayer input in the large structures such as on building boundaries and pedestrian road, and performs better than the estimations with the re-sampling Bayer input in the fine structures such as lamp poles and traffic signs. With the re-sampling and dynamic weighting strategy, the refined output can benefit from both global and local contextual information, as demonstrated in Fig. 10.

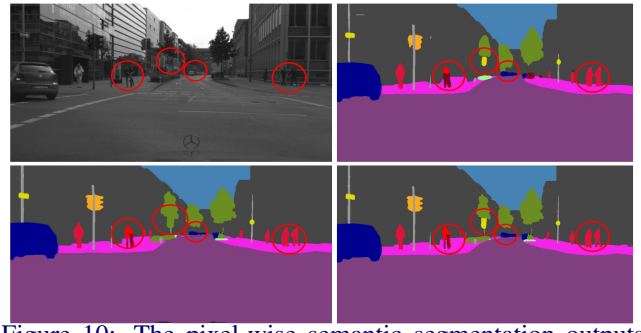

Figure 10: The pixel-wise semantic segmentation outputs with different Bayer grid sizes. First row: input raw Bayer image; segmentation output from large Bayer grid. Second row: segmentation output from small Bayer grid; final refined segmentation output.

The final loss will be a combination of L_{seg} from the raw Bayer image and the re-sampling Bayer image. Assuming the dynamic attention mask and the pixel-wise segmentation for the input Bayer image are M and $Y_{\text{seg}\,1}$, the corresponding attention mask and the segmentation output for the re-sampled Bayer image are $1-M$ and $Y_{\text{seg}\,2}$, the refined segmentation output can be formulated as:

$$Y_{seg\_final} = M \cdot Y_{seg_1} + (1-M) \cdot Y_{seg_2} \tag{5}$$

Therefore, the final refined objective function L_{final} is based on Eq. 4 that can be updated as:

$$L_{final} = L_{seg}\left(y, \tilde{Y}_{s_{seg_1}}\right) + L_{seg}\left(y, \tilde{Y}_{seg_2}\right) + L_{seg}\left(y, \tilde{Y}_{seg\_final}\right) \tag{6}$$

4 Experiments

4.1 Datasets and Configuration

We evaluate our proposed framework on three datasets: Cityscapes, Mapillary, and a dataset we collected using a NIKON-D3500 digital camera. Cityscapes is a high-resolution dataset consisting of around 5,000 images with pixel-level segmentation annotations for 19 classes, including road, sidewalk, building, person, car, and more. We used reverse-engineered RGB color images published by the dataset and converted them to 8-bit Bayer images for training and evaluation. The training, validation, and testing sets are split into 2,975, 500, and 1,525 images. We used the most common RGGB Bayer pattern to extract one channel for each 3-channel pixel in the order of the Bayer pattern.

	Train	Validation	Test
Images	2000	500	1000
Masks	14016	3350	7010

Table 1: Split of our collected raw Bayer image dataset.

We also collected a real raw Bayer image dataset using a NIKON-D3500 camera. The dataset has the same class categories as Cityscapes, and its partition details are shown in Table 1. Fig. 11 displays some samples of collected raw Bayer images and their corresponding pixel-wise label annotations. Finally, we evaluated our method on Mapillary and our collected dataset with and without re-training.

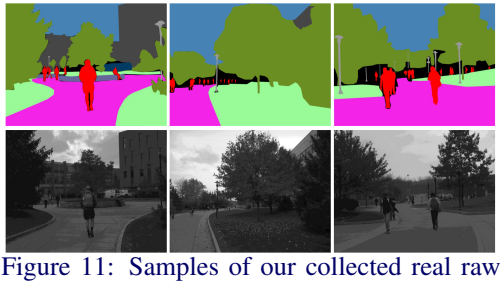

Figure 11: Samples of our collected real raw Bayer image dataset.

Method	MIoU	FPS	road	sidewalk	build	wall	fence	pole	tlight	tsign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bike
RefineNet [18]	0.735	1.8	0.872	0.737	0.826	0.593	0.481	0.524	0.706	0.750	0.803	0.636	0.920	0.796	0.642	0.900	0.616	0.725	0.610	0.583	0.609
Deeplab-v3 [4]	0.757	2.3	0.919	0.754	0.855	0.732	0.475	0.554	0.760	0.752	0.828	0.462	0.914	0.846	0.589	0.865	0.540	0.627	0.538	0.682	0.646
Deeplab-v3+ [5]	0.763	3.6	0.921	0.752	0.867	0.735	0.478	0.571	0.761	0.769	0.826	0.487	0.919	0.846	0.603	0.879	0.533	0.636	0.541	0.685	0.651
DANet [8]	0.721	1.1	0.875	0.810	0.838	0.584	0.526	0.519	0.618	0.534	0.816	0.355	0.919	0.850	0.568	0.848	0.628	0.700	0.615	0.570	0.552
PSANet [36]	0.767	2.9	0.894	0.756	0.862	0.655	0.547	0.630	0.730	0.743	0.824	0.479	0.917	0.881	0.646	0.860	0.581	0.717	0.681	0.618	0.668
HRNetV2-W48 [32]	0.774	5.6	0.878	0.760	0.851	0.751	0.569	0.531	0.694	0.425	0.812	0.314	0.927	0.832	0.536	0.917	0.746	0.823	0.765	0.659	0.532
DNLNet [34]	0.788	2.4	0.890	0.810	0.839	0.574	0.570	0.576	0.740	0.635	0.809	0.649	0.918	0.815	0.565	0.843	0.679	0.714	0.780	0.503	0.613
Our	0.803	8.7	0.947	0.851	0.896	0.636	0.623	0.661	0.792	0.825	0.905	0.726	0.946	0.859	0.749	0.939	0.768	0.822	0.827	0.704	0.735

Table 2: Per-category results on the validation set of the simulated Cityscapes Bayer image data. All the models are trained with only on the fine annotated data of the train set. Our network achieves 80.3% accuracy in mIoU, which outperforms all other compared approaches [18] [4] [12] [6] [37] [32] [34] on 16 out of 19 categories.

We use a ResNet-50 based network (configured with a stride of 2 and convolution kernel of 2×2 for adapting Bayer pattern) and ASPP module as the backbone for extracting features in the network. The learning rate with warm-up steps of 10 and the poly learning rate policy [36] for decaying the initial learning rate by multiplying $\left(1 - \frac{\text{iter}}{\text{total iter}}\right)^{0.9}$ are adopted to help the training stage converge efficiently. The optimizer of stochastic gradient descent (SGD) with a batch size of 4 is utilized and the initial learning rate is set to be $5e-3$. Limited by the GPU memory, we resize images to 1024×512 for all experiments. Data augmentation with random horizontal flip, color transforms in brightness, contrast, hue, and saturation is applied. Additionally, we re-train the state-of-the-arts [18] [4] [12] [6] [37] [32] [34] on the same Bayer image datasets for fair comparisons.

4.2 Visual and Quantitative Analysis

We report the mean Intersection over Union (IoU) of each specific category on the simulated Cityscape dataset in Table 2 to compare the proposed method with recent state-of-the-art approaches. Our method outperforms [6] [37] significantly on large objects such as road and sky while improving the accuracy on relatively tiny objects like pole and bike by a large margin compared to [32] and [34]. This performance enhancement is mainly attributed to the design of spectral frequency attention, spatial coordinate attention, and multi-grid resampling strategy that consider both global context information and local pattern shapes. Furthermore, our method achieves high accuracy while significantly increasing the time performance, with a speed of 8.7 fps compared to 1.8-5.6 fps for other methods.

We also present a qualitative evaluation of the segmentation results in Fig. 12, which shows that our method outperforms other state-of-the-art methods in terms of maintaining accurate segmentation boundaries and preserving object shapes, particularly in comparison with [34] on road and persons. Furthermore, even without re-training on the real collected dataset, our method produces significantly better segmentation results than other methods, demonstrating the suitability and generalizability of the proposed framework in real-world applications.

	Method	Backbone	mIoU(%)
Mapillary Dataset (without train)	Deeplab-V3[4].	ResNet-50	55.8
	HRNetV2-W48[32].	ResNet-50	56.9
	DNLNet[34].	ResNet-50	58.2
	Ours	ResNet-50	59.3
	Ours	ResNet-101	60.1
Our Collected Dataset (without train)	Deeplab-V3[4].	ResNet-50	29.7
	HRNetV2-W48[32].	ResNet-50	33.6
	DNLNet[34].	ResNet-50	31.2
	Ours	ResNet-50	42.7
	Ours	ResNet-101	44.9

Table 3: Results on Mapillary and our collected datasets.

We widely validate the performance of the proposed network and test its generalization on different scenes by evaluating it on the Mapillary dataset in Table 3. Our method achieves about 3.5% mIoU improvement compared to [4] using the same ResNet-50 backbone structure and 4.3% improvement using the ResNet-101 structure. The improvement is even more significant on the collected dataset, with our method achieving the top performance of 42.7% in mIoU, which is 43.7% higher than [4] and 36.9% higher than [34].

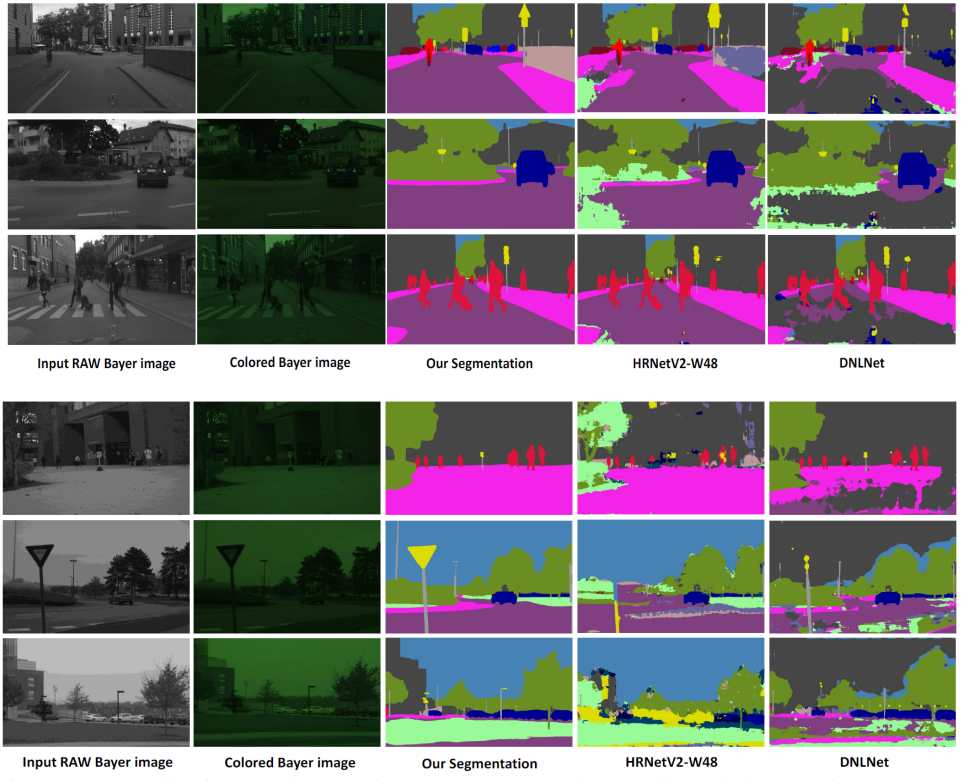

Figure 12: Qualitative results on Cityscapes (top) and our collected dataset (bottom). For each dataset, from left to right: input raw image; input image with Bayer pattern overlaid for better observation; our result; result from HRNetV2-W48 [32] and DNLNet [34].

4.3 Ablation Study

Method	Backbone	mIoU (%)	δ(%)
[12]	ResNet-50	76.3	–
Our Naive	ResNet-50	75.2	–
Ours	ResNet-50	80.3	–
[12]	ResNet-101	77.8	1.5 ↑
Our Naive	ResNet-101	77.0	1.8 ↑
Ours	ResNet-101	81.6	1.3 ↑

Table 4: Ablation study on different backbone settings.

Ablation study on network structures: we investigate the impact of different backbone settings on segmentation performance. As shown in Table 4, we first evaluate a naive structure without any proposed components, achieving mIoU scores of 75.2% and 77.0% with ResNet-50 and ResNet-101 backbones, respectively. After incorporating the introduced components, our method achieves a significant 5.1% and 4.6% improvement over the naive structure. While changing from ResNet-50 to ResNet-101 backbone only brings a 1.5% improvement for the compared method [12], the proposed method gains a 1.3% increase, indicating the improvement mainly comes from components specifically designed for Bayer images rather than deeper network structures.

Ablation study on network inputs: Table 5 compares the accuracy and computation cost of our proposed method with different types of inputs. Our method achieves higher mIoU performance on raw Bayer images compared to grayscale images, despite both having 8-bit channels. This indicates that the proposed network effectively learns spatial, spectral, and shape information from the Bayer pattern.

The segmentation accuracy is comparable when taking RGB images (81.0%) and Bayer images (80.3%) as inputs. However, the network size significantly decreases with Bayer images (26.5M parameters) compared to RGB images (39.8M parameters), saving not only image storage but also network operation space.

Method	Type	mIoU (%)	δ(%)	Params (M)
RefineNet [13]	Gray	73.6	–	108.7
	RGB	74.2	0.6 ↑	121.4
	Bayer	73.5	0.1 ↓	108.7
Our Naive	Gray	75.6	–	27.7
	RGB	76.8	1.2 ↑	39.2
	Bayer	75.2	0.4 ↓	27.7
Ours	Gray	79.7	–	26.5
	RGB	81.0	1.3 ↑	39.8
	Bayer	80.3	0.6 ↑	26.5

Table 5: Ablation study on segmentation performance and network parameters using different input image types.

Ablation study on key components: Table 6 presents a detailed analysis of each key component of our network design. The results show that the spatial coordinate and spectral frequency attention block together contribute to a 3.1% improvement in mIoU compared to the naive implementation without any proposed component. A multi-grid training strategy further achieves a 1.3% gain following the attention blocks. The incorporation of RMI loss in the objective function further raises the mIoU to 80.3%, which outperforms the naive implementation by almost 5.1%.

$L_{spectral}$	$L_{spatial}$	$L_{resampling}$	L_{rmi}	mIoU(%)	δ(%)
–	–	–	–	75.2	–
✓	–	–	–	77.9	2.7 ↑
✓	✓	–	–	78.3	3.1 ↑
✓	✓	✓	–	79.6	4.4 ↑
✓	✓	✓	✓	**80.3**	**5.1** ↑

Table 6: Ablation study on the effect of each component and loss.

The effects of different modules are also illustrated in Fig. 13, which demonstrates that with the introduced attention modules and re-sampling strategy, some misclassified categories such as trucks, poles, and traffic signs can be corrected, and object boundaries and details such as cars and persons are clearer.

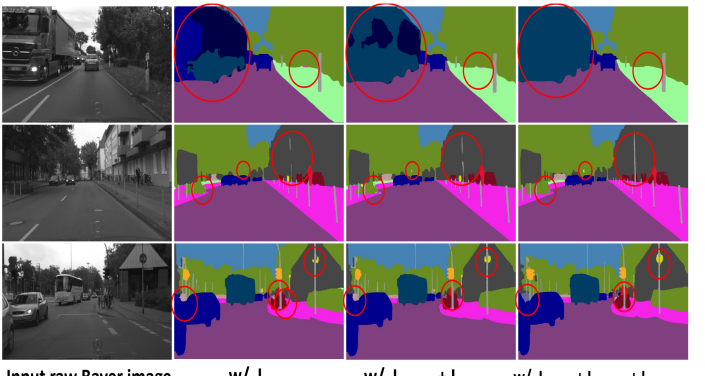

Figure 13: Visual analysis of the effects of different components and losses. Salient regions are highlighted for easy comparison.

5 Conclusion

We proposed RawSeg-Net, an end-to-end semantic segmentation network designed to segment raw Bayer images, enabling the elimination of the ISP process in image generation. Our approach uses 8-bit raw Bayer images, leading to large storage reductions and computational efficiency improvements. By introducing Bayer spectral frequency and spatial coordinate attention, as well as a multi-grid re-sampling strategy, we improved segmentation accuracy by combining local and global context information, offering a promising solution for efficient and accurate semantic segmentation of raw Bayer images.

Ack: This paper is supported by NSF Awards No. 2334624, 2334690, and 2334246.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets deep learning for car instance segmentation in urban scenes. In *British machine vision conference*, volume 1, page 2, 2017.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[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 transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[5] K-H Chung and Y-H Chan. Color demosaicing using variance of color differences. *IEEE transactions on image processing*, 15(10):2944–2955, 2006.

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[7] Lu Guoyu. Object detection based on raw bayer images. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[8] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3562–3572, 2019.

[9] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7519–7528, 2019.

[10] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pająk, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (ToG)*, 33(6):1–13, 2014.

[11] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32:582–596, 2019.

[12] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[14] Wonjae Lee, Seongjoo Lee, and Jaeseok Kim. Cost-efffective color filter array demosaicing using spatial correlation. *IEEE Transactions on Consumer Electronics*, 52(2): 547–554, 2006.

[15] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[16] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 435–452. Springer, 2020.

[17] Xin Li, Bahadir Gunturk, and Lei Zhang. Image demosaicing: A systematic survey. In *Visual Communications and Image Processing 2008*, volume 6822, pages 489–503. SPIE, 2008.

[18] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[19] Lin Liu, Xu Jia, Jianzhuang Liu, and Qi Tian. Joint demosaicing and denoising with self guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2240–2249, 2020.

[20] Shumin Liu, Jiajia Chen, Yuan Xun, Xiaojin Zhao, and Chip-Hong Chang. A new polarization image demosaicking algorithm by exploiting inter-channel correlations with guided filtering. *IEEE Transactions on Image Processing*, 29:7076–7089, 2020.

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[22] Guoyu Lu, Li Ren, Jeffrey Caplan, and Chandra Kambhamettu. Stromule branch tip detection based on accurate cell image segmentation. In *IEEE International Conference on Image Processing (ICIP)*, pages 3300–3304, 2017.

[23] Yawen Lu, Michel Sarkis, and Guoyu Lu. Multi-task learning for single image depth estimation and segmentation based on unsupervised network. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 10788–10794, 2020.

[24] Daniele Menon and Giancarlo Calvagno. Color image demosaicking: An overview. *Signal Processing: Image Communication*, 26(8-9):518–533, 2011.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[26] Abhinav Sagar and RajKumar Soundrapandiyan. Semantic segmentation with multi scale spatial attention for self driving cars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2650–2656, 2021.

[27] Ling Shao and Amin Ur Rehman. Image demosaicing using content and colour-correlation analysis. *Signal processing*, 103:84–91, 2014.

[28] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[29] Daniel Stanley Tan, Wei-Yang Chen, and Kai-Lung Hua. Deepdemosaicking: Adaptive image demosaicking via multiple deep fully convolutional networks. *IEEE Transactions on Image Processing*, 27(5):2408–2419, 2018.

[30] Runjie Tan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Color image demosaicking via deep residual learning. In *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, volume 2, page 6, 2017.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[32] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[34] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 191–207. Springer, 2020.

[35] Chao Zhang, Yan Li, Jue Wang, and Pengwei Hao. Universal demosaicking of color filter arrays. *IEEE Transactions on Image Processing*, 25(11):5173–5186, 2016.

[36] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[37] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 267–283, 2018.

[38] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[39] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9179–9188, 2021.