Frequency-Adaptive Dilated Convolution for Semantic Segmentation
Supplementary Material

Linwei Chen[1] Lin Gu[2,3] Dezhi Zheng[1] Ying Fu[1*]
[1]Beijing Institute of Technology [2]RIKEN [3]The University of Tokyo
chenlinwei@bit.edu.cn lin.gu@riken.jp zhengdezhi@bit.edu.cn fuying@bit.edu.cn

This supplementary material provides more details and results that are not included in the main paper due to space limitations. The contents are organized as follows.

- Section A introduces the scaling property of Fourier Transforms.
- Section B provides additional frequency analysis for dilated convolution.
- Section C offers sampling analysis for dilated convolution.
- Section D describes theoretical analysis of AdaKern.
- Section E presents a more in-depth ablation study.
- Section F includes additional experiment details.
- Section G provides more detailed results of real-time semantic segmentation.
- Section H reports more quantitative results on various tasks.
- Section I showcases more visualized results.

A. The scaling property of Fourier Transforms

In this section, we analyze how the dilation manipulate the frequency response of convolution. In the discrete two-dimensional case, considering a zero-padded normal convolution kernel $\mathbf{W}$ and its dilated version $\mathbf{W}'$. Increasing the dilation rate from 1 to D is equivalent to expanding the convolution kernel through zero-insertion by a factor of D. We have $\mathbf{W}'(Dm, Dn) = \mathbf{W}(m, n)$, where D are dilation rate, *i.e.*, scaling factors in the horizontal and vertical directions. Before applying the Fourier transform to the convolution weight to obtain its frequency response, it is a common practice to use zero-padding, enlarging the small size of the convolution weight. Notice that the size after zero-padding is significantly larger than the kernel size [13]. The Fourier transform of the convolution kernel $\mathbf{W}(m, n)$ denoted by $\mathbf{W}_F(u, v)$ can be expressed as:

$$\begin{aligned}\mathbf{W}_F(u,v) &= \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\mathbf{W}(m,n)e^{-j2\pi(um+vn)} \\ &= \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\mathbf{W}'(Dm,Dn)e^{-j2\pi(um+vn)}\end{aligned} \tag{1}$$

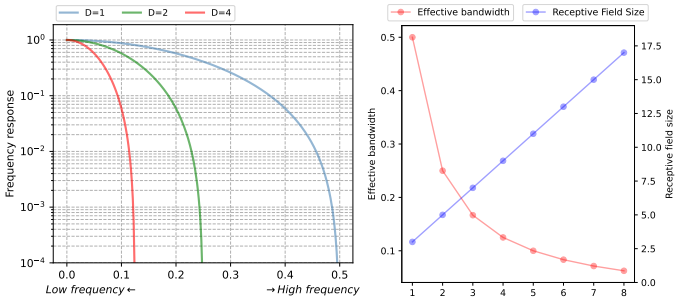


Figure 1. Left: 'D' is dilation. Higher dilation scales the kernel's frequency response curve to lower frequencies, narrowing its effective bandwidth. Normalized frequency $[0, 0.5]$ is used for simplicity, $\times 2\pi$ yields normalized angular frequency. Right: Receptive field expands with dilation rate, enhancing spatial coverage. However, higher dilation reduces the kernel's effective bandwidth, limiting its ability to capture frequency information.

Where the normalized frequencies in the height and width dimensions are given by $|u|$ and $|v|$. After shifting the low frequency to the center, u takes values from the set $\{-\frac{M}{2}, -\frac{M+1}{2}, \ldots, \frac{M-1}{2}\}$, and v takes values from $\{-\frac{N}{2}, -\frac{N+1}{2}, \ldots, \frac{N-1}{2}\}$. To establish the relationship between $\mathbf{W}_F(u, v)$ and $\mathbf{W}'_F(u, v)$, we introduce two new variables $p = Dm$ and $q = Dn$, such that $m = \frac{1}{D}p$ and $n = \frac{1}{D}q$. Using these new variables, the Fourier transform expression can be rewritten as:

$$\begin{aligned}\mathbf{W}_F(u,v) &= \frac{1}{MN}\sum_{p=0}^{M-1}\sum_{q=0}^{N-1}\mathbf{W}'(p,q)e^{-j2\pi\left(\frac{u}{D}p+\frac{v}{D}q\right)} \\ &= \mathbf{W}'_F\left(\frac{u}{D},\frac{v}{D}\right)\end{aligned} \tag{2}$$

$\mathbf{W}'_F$ is the Fourier transform of the convolution kernel $\mathbf{W}'$. Thus, the frequency response of the standard kernel $\mathbf{W}_F(u, v)$ is scaled to $\frac{1}{D}$ lower frequency in its dilated version $\mathbf{W}'_F\left(\frac{u}{D}, \frac{v}{D}\right)$.

We present a representative response curve of the convolution kernel in the left of Figure 1. With dilation increasing

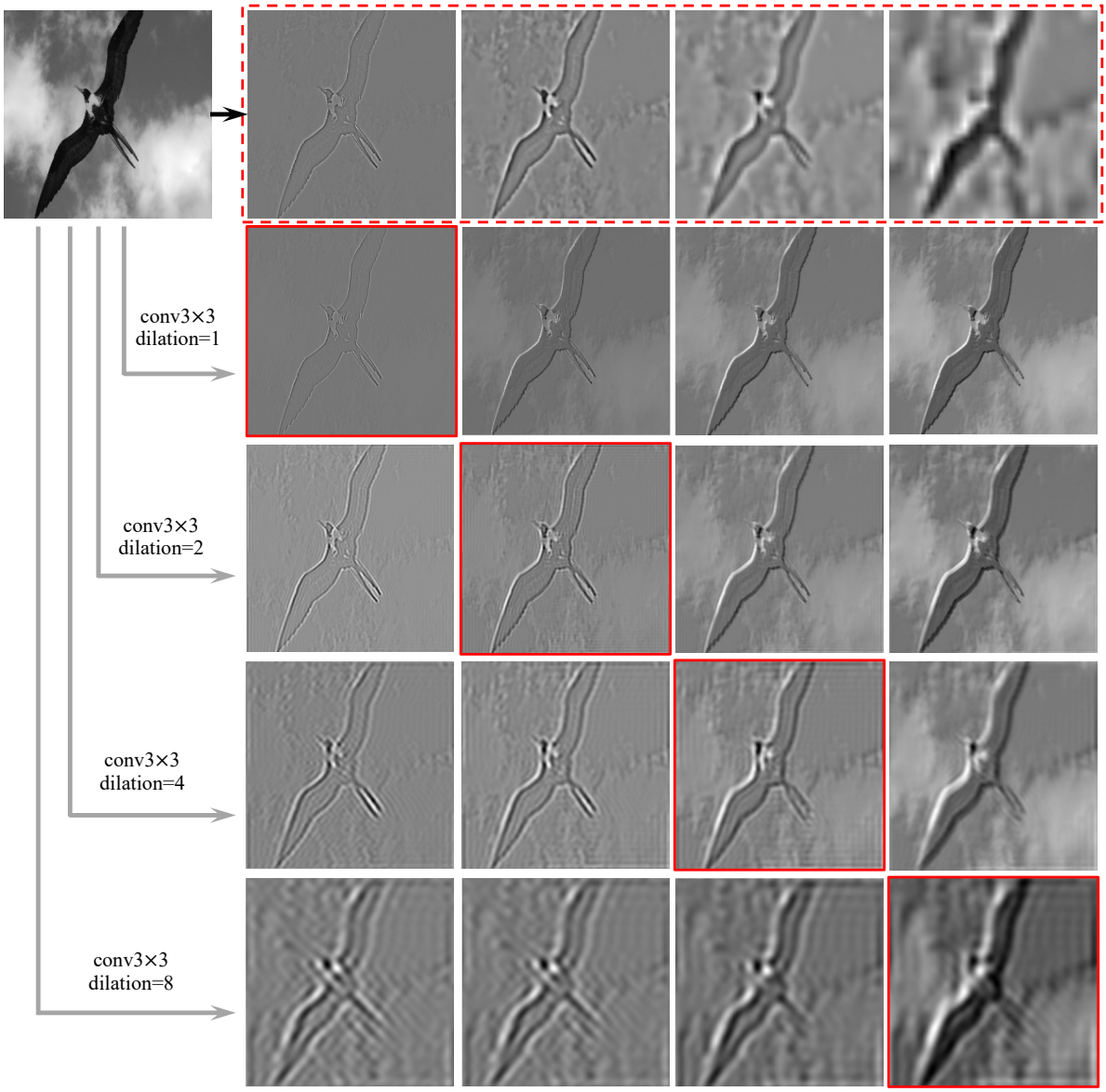
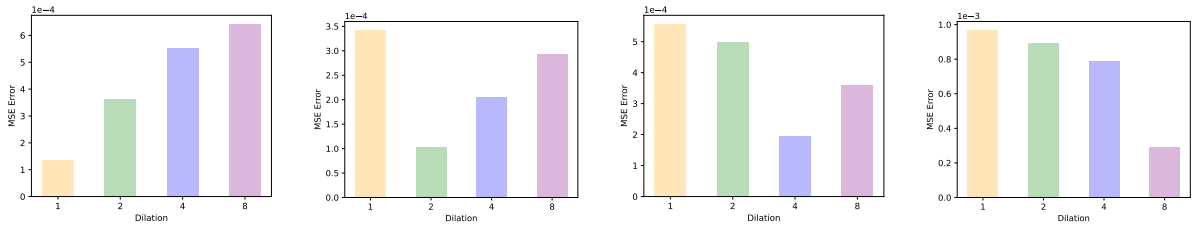


Figure 2. In the top row, we follow [20, 21] to decompose the image into different spatial frequency band, ranging from high to low frequency. The image is filtered within one-octave-wide (doubling of frequency) spatial frequency bands centered at $f_{center} \in \{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ normalized frequency (*i.e.*, $f_{center} \in \{128.0, 64.0, 32.0, 16.0\}$ cycles/image for 256×256 images). We employ one layer of 3×3 convolution with different dilation to learn to extract four corresponding spatial frequency bands from the image. Frequencies higher than $\frac{1}{2D}$ are removed to avoid aliasing artifacts. The red box indicates the lowest Mean Squared Error (MSE) for the corresponding frequency band among different dilation rates. The corresponding MSE error results are shown in Figure 3. We find that convolutions with higher dilation rates excel in extracting lower frequencies, while convolutions with lower dilation rates are better at extracting higher frequencies. This observation also supports our AdaDR, which assigns a small dilation rate for high-frequency areas full of details and a large dilation rate for low-frequency smooth areas.

from 1 to D, the response curves are scaled to $\frac{1}{D}$, reducing the effective bandwidth to its $\frac{1}{D}$. Additionally, the receptive field of the convolution is positively correlated with dilation. We illustrate the relationship between dilation, bandwidth, and receptive field in the right of Figure 1. Our work is motivated by this observation and aims to balance the bandwidth and receptive field of dilated convolution.

B. Frequency Analysis for Dilated Convolution

Inspired by spatial-frequency channel analysis [20, 21], we evaluate convolution with varying dilation rates to learn and



(a) (b) (c) (d)

Figure 3. The Mean Squared Error (MSE) values in (a)-(d) represent the difference between the top-row image decomposed into various spatial frequency bands in Figure 2 and the corresponding extracted images. The extraction is conducted on the original image using a single layer of convolution with different dilation rates.

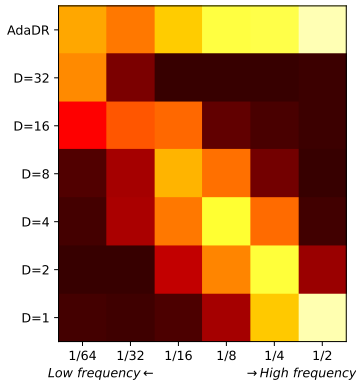


Figure 4. Intensity increases in color correspond to reduced MSE errors in the extraction of corresponding frequencies. Convolutions with higher dilation rates tend to excel at extracting lower frequencies. AdaDR can extract different frequencies from the original image by adjusting its dilation rate.

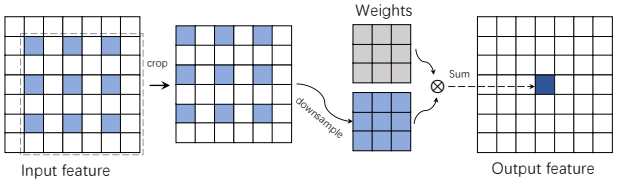


Figure 5. Illustration of dilated convolution. The computation of dilated convolution can be broken down into several steps [23]. First, crop the corresponding feature map, then, downsample it by $D\times$ (dilation rate), and finally, pixel-wise multiply and sum. Thus, dilated convolution is actually operated at a sampling rate of $\frac{1}{D}$.

extract distinct frequency components from images.

As shown in Figure 2, in the top row, we follow [20, 21] to decompose the image into different spatial frequencies, ranging from high to low frequency. The image is filtered within one-octave-wide (doubling of frequency) spatial frequency bands centered at $f_{center} \in \{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ normalized frequency. We employ one layer of 3×3 convolution with different dilation to learn to extract four corresponding spatial frequency bands from the image. Frequencies higher than $\frac{1}{2D}$ are removed to avoid aliasing artifacts. The red box indicates the lowest Mean Squared Error (MSE) for the corresponding frequency band among different dilation rates. The MSE error results are shown in Figure 3. We find that convolutions with higher dilation rates excel in extracting lower frequencies, while convolutions with lower dilation rates are better at extracting higher frequencies.

This observation also supports our AdaDR, which assigns a small dilation rate for high-frequency areas full of details and a large dilation rate for low-frequency smooth areas. By dynamically adjusting dilation rates spatially variantly, AdaDR demonstrates the capability to extract a broad spectrum of frequencies, as shown in Figure 4. This highlights its effectiveness in capturing diverse frequency information.

C. Sampling Analysis for Dilated Convolution

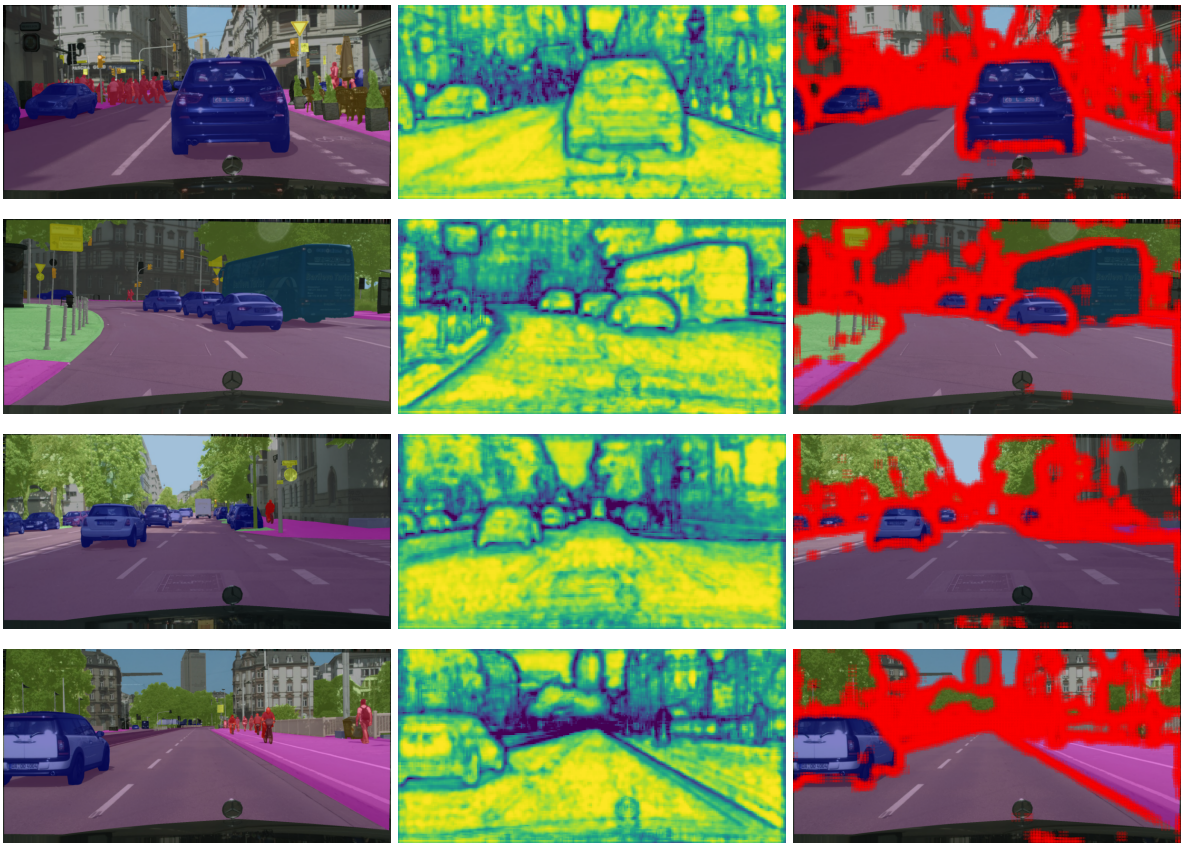
As shown in Figure 5, the computation of dilated convolution can be broken down into several steps [23]. First, crop the corresponding feature map, then downsample it by $D\times$ (dilation rate), and finally, pixel-wise multiply and sum. Thus, dilated convolution is actually operated at a sampling rate of $\frac{1}{D}$.

D. Theoretical Analysis of AdaKern

For a static convolutional kernel, AdaKern decompose its weights $\mathbf{W}$ as follows

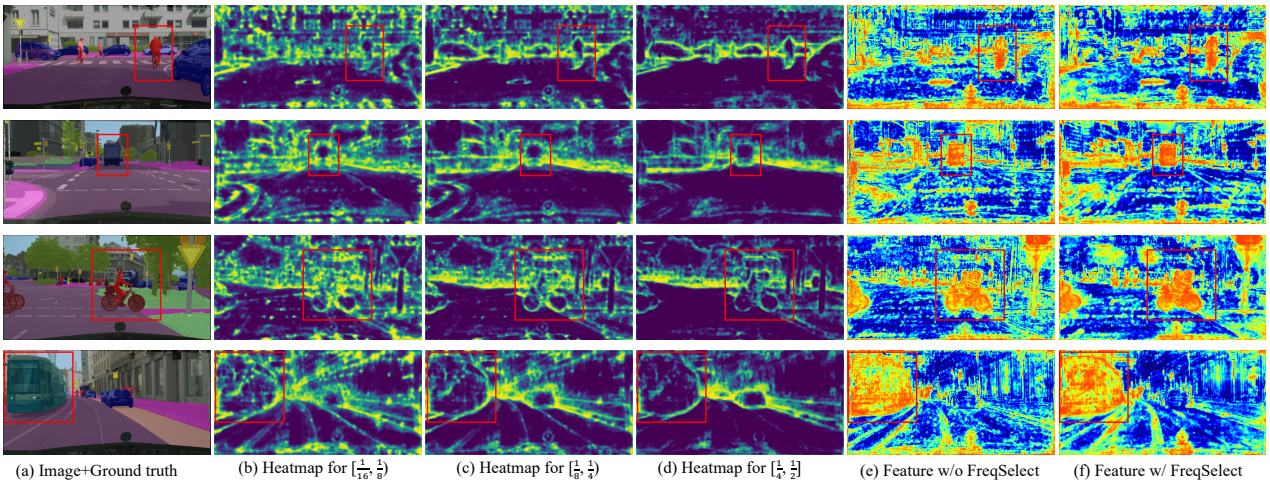
$$\mathbf{W} = \bar{\mathbf{W}} + \hat{\mathbf{W}}. \tag{3}$$

Here, $\bar{\mathbf{W}} = \frac{1}{K\times K}\sum_{i=1}^{K\times K}\mathbf{W}_i$ represents the kernel-wise averaged $\mathbf{W}$. It functions as a low-pass $K \times K$ mean filter, followed by a 1×1 convolution with parameters defined by $\bar{\mathbf{W}}$. Through Fourier analysis, $\bar{\mathbf{W}}$ is proportional to the lowest frequency component of 2D DFT. Specifically, the Fourier transform of static convolution kernel $\mathbf{W}$ denoted



(a) Image & Ground truth (b) Dilation map (c) Sampling coordinates

Figure 6. Visualized results for AdaDR. AdaDR learns to spatially adaptively assign dilation rates based on the local feature frequency. We show the sampling coordinates of convolution with adaptive dilation. For better visualization, we only display the sampling coordinates of convolution with a low dilation rate (≤ 2) in (c), which mainly distribute along object boundaries.



(a) Image+Ground truth (b) Heatmap for $[\frac{1}{16}, \frac{1}{8})$ (c) Heatmap for $[\frac{1}{8}, \frac{1}{4})$ (d) Heatmap for $[\frac{1}{4}, \frac{1}{2}]$ (e) Feature w/o FreqSelect (f) Feature w/ FreqSelect

Figure 7. Visualized results for FreqSelect. Brighter colors indicate a higher value. FreqSelect selectively suppresses high frequencies in areas that do not contribute to accurate predictions, such as the background and the center of objects. Notably, FreqSelect exhibits a tendency to assign a higher attention weight to object boundaries, especially in higher frequency bands. Consequently, features enhanced by FreqSelect showcase a reduction in high frequencies within the background, resulting in clearer object boundaries.

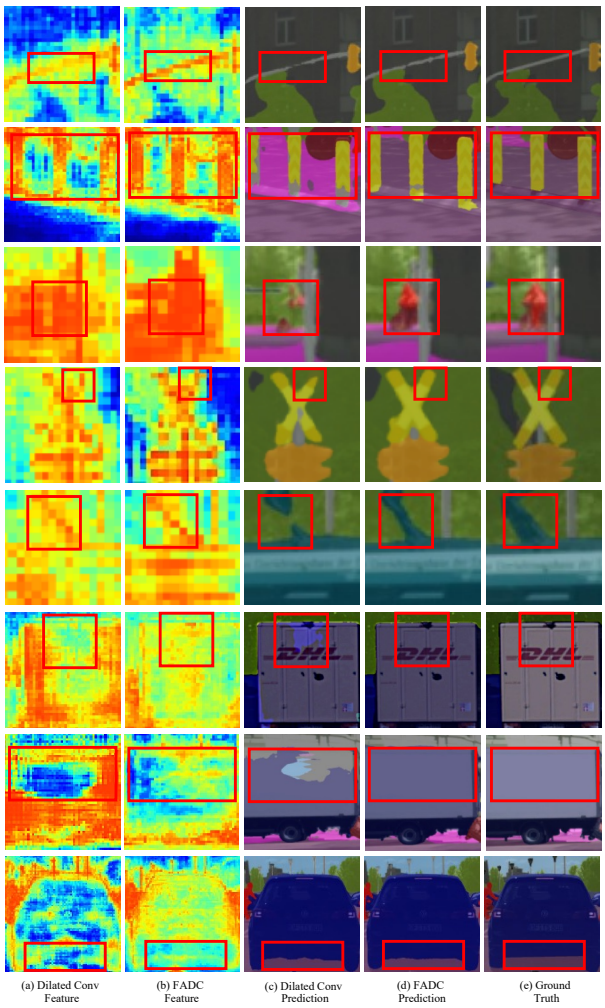


Figure 8. Feature and prediction visualization on Cityscape [3]. Aliasing artifacts are evident in (a), leading to the loss of details in the representation of a thin pole and truck boundary, resulting in inferior predictions in (c). In contrast, our proposed FADC method in (b) exhibits an accurate and uniform response to both thin poles and large trucks, contributing to consistently accurate predictions in (d).

by $\mathbf{W}_F(u,v)$ can be expressed as

$$\mathbf{W}_F(u,v) = \frac{1}{K \times K} \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} \mathbf{W}(m,n) e^{-j2\pi(um+vn)}, \tag{4}$$

We set $(u,v) = (0,0)$ to obtain the lowest frequency component, which is equal to $\bar{\mathbf{W}}$

$$\begin{aligned} \mathbf{W}_F(0,0) &= \frac{1}{K \times K} \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} \mathbf{W}(m,n) e^{0} \\ &= \bar{\mathbf{W}} \end{aligned} \tag{5}$$

The term $\hat{\mathbf{W}} = \mathbf{W} - \bar{\mathbf{W}}$ denotes the residual part, indicating the frequency components excluding the lowest, capturing local differences and extracting the high-frequency components. In this way, AdaKern decomposes the static convolution kernel into low and high-frequency parts, and reweights them using dynamic weights λ_l, λ_h

$$\mathbf{W}' = \lambda_l \bar{\mathbf{W}} + \lambda_h \hat{\mathbf{W}}. \tag{6}$$

Table 1. Ablation study on ADE20K [29] dataset. We adopt the widely used ResNet-50 with UPerNet as the segmentation model and replace the standard convolutional layer (marked as Conv.) with the proposed FADC, encompassed with three plug-in strategies, *i.e.*, AdaDR, AdaKern, and FreqSelect.

Method	AdaDR	AdaKern	FreqSelect	mIoU
Conv.				40.8
FADC	✓			43.6
FADC	✓	✓		43.8
FADC	✓		✓	44.1
FADC	✓	✓	✓	**44.4**

Table 2. Ablation study for AdaDR on the ADE20K dataset [29].

Percentage for training	12.5%	25%	50%
mIoU	43.5	44.4	44.1

Table 3. Ablation study for FreqSelect on the ADE20K dataset [29]. The phrase "Fixed weight for lowest" indicates setting a fixed selection weight of 1.0 for the lowest frequency band.

Frequency Band	Fixed weight for lowest	mIoU
2: $[0,\frac{1}{4}),[\frac{1}{4},\frac{1}{2}]$	✓	44.0
3: $[0,\frac{1}{8}),[\frac{1}{8},\frac{1}{4}),[\frac{1}{4},\frac{1}{2}]$	✓	44.2
4: $[0,\frac{1}{16}),[0,\frac{1}{8}),[\frac{1}{8},\frac{1}{4}),[\frac{1}{4},\frac{1}{2}]$	✓	**44.4**
5: $[0,\frac{1}{32}),[0,\frac{1}{16}),[0,\frac{1}{8}),[\frac{1}{8},\frac{1}{4}),[\frac{1}{4},\frac{1}{2}]$	✓	44.1
4: $[0,\frac{1}{16}),[0,\frac{1}{8}),[\frac{1}{8},\frac{1}{4}),[\frac{1}{4},\frac{1}{2}]$		44.0

E. Ablation study

In this section, we conduct an extensive ablation study on the ADE20K dataset [29] to evaluate the effectiveness of our proposed Frequency-Aware Dilated Convolution (FADC) compared to the standard convolutional layer in the widely used ResNet-50 [6] architecture with UPerNet [24] as the segmentation model. We replace the standard convolutional layer in the last three stages with the proposed FADC in the ResNet-50 [6], all models are trained with the same setting described in Section F.

AdaDR. The ablation study results are summarized in Table 1. AdaDR dynamically adjusts dilation rates based on local feature frequency, enhancing the receptive field size as well as keep appropriate frequency bandwidth to capture frequency information. The application of FADC, with

Table 4. Comparison of speed and accuracy on Cityscapes [3]. The models pre-trained by other segmentation datasets are marked with †.

Model	Val	Test	#FPS	GPU	Resolution	#GFLOPs	#Params
DF2-Seg1 [10]	75.9	74.8	67.2	GTX 1080Ti	1536x768	-	-
DF2-Seg2 [10]	76.9	75.3	56.3	GTX 1080Ti	1536x768	-	-
SwiftNetRN-18 [16]	75.5	75.4	39.9	GTX 1080Ti	2048×1024	104.0	11.8M
SwiftNetRN-18 ens [16]	-	76.5	18.4	GTX 1080Ti	2048×1024	218.0	24.7M
CABiNet [8]	76.6	75.9	76.5	RTX 2080Ti	2048×1024	12.0	2.64M
BiSeNet(Res18) [27]	74.8	74.7	65.5	GTX 1080Ti	1536×768	55.3	49M
BiSeNetV2-L [26]	75.8	75.3	47.3	GTX 1080Ti	1024×512	118.5	-
STDC1-Seg75 [4]	74.5	75.3	74.8	RTX 3090	1536×768	-	-
STDC2-Seg75 [4]	77.0	76.8	58.2	RTX 3090	1536×768	-	-
PP-LiteSeg-T2 [17]	76.0	74.9	96.0	RTX 3090	1536×768	-	-
PP-LiteSeg-B2 [17]	78.2	77.5	68.2	RTX 3090	1536×768	-	-
HyperSeg-M [14]	76.2	75.8	59.1	RTX 3090	1024×512	7.5	10.1
HyperSeg-S [14]	78.2	78.1	45.7	RTX 3090	1536×768	17.0	10.2
SFNet(DF2) [9]	-	77.8	87.6	RTX 3090	2048×1024	-	10.53M
SFNet(ResNet-18) [9]	-	78.9	30.4	RTX 3090	2048x1024	247.0	12.87M
SFNet(ResNet-18)† [9]	-	80.4	30.4	RTX 3090	2048×1024	247.0	12.87M
DDRNet-23-S [7]	77.8	77.4	108.1	RTX 3090	2048×1024	36.3	5.7M
DDRNet-23 [7]	79.5	79.4	51.4	RTX 3090	2048×1024	143.1	20.1M
DDRNet-39 [7]	-	80.4	30.8	RTX 3090	2048×1024	281.2	32.3M
PIDNet-S [25]	78.8	78.6	93.2	RTX 3090	2048×1024	47.6	7.6M
PIDNet-M [25]	80.1	80.1	39.8	RTX 3090	2048×1024	197.4	34.4M
PIDNet-L [25]	80.9	80.6	31.1	RTX 3090	2048×1024	275.8	36.9M
PIDNet-M +FADC (Ours)	**81.0**	**80.6**	37.7	RTX 3090	2048×1024	198.4	34.6M

AdaDR incorporated alone, demonstrates a noticeable improvement, mIoU from 40.8 to 43.6.

For optimizing the spatial variant dilation rate assigned to coordinate p, we choose to optimize $\hat{\mathbf{D}}(p)$ directly. That is, by increasing the dilation rate at position p where the high-frequency power $\mathrm{HP}(p)$ is low to encourage a large receptive field, and suppressing the dilation rate where $\mathrm{HP}(p)$ is high to reduce the loss of frequency information. To formalize this optimization, we express it as follows

$$\theta = \max_{\theta} \left(\sum_{p \in \mathrm{HP}^-} \hat{\mathbf{D}}(p) - \sum_{p \in \mathrm{HP}^+} \hat{\mathbf{D}}(p) \right). \tag{7}$$

Here, HP^+ and HP^- represent pixels with the highest/lowest high-frequency power, *i.e.*, the brighter/darker areas in Figure 2(b), respectively. We empirically set the weight of this optimization target to 0.01 to balance it during training with task loss (*e.g.*, pixel-wise cross-entropy for segmentation). As shown in Table 2, we use 25% pixel with the highest/lowest high-frequency power for training.

AdaKern. The AdaKern operates on convolution kernel weights, decomposing them into low-frequency and high-frequency components. This manipulation optimizes the frequency response curve, enabling dynamic adjustments on a per-channel basis. The subsequent inclusion of AdaKern further enhances the performance achieved by AdaDR alone, resulting in an mIoU of 43.8.

FreqSelect. High frequencies require a lower dilation with high bandwidth to capture adequately. Considering the convolution's inclination to amplify high-frequency components, which negatively impacts the average dilation rate predicted by AdaDR, there is a resultant decrease in receptive field size. FreqSelect strategically intervenes by spatially reweighting high frequencies. It selectively suppresses high frequencies in areas that do not contribute to accurate predictions, such as the background and the center of objects. This prompts FADC to learn higher dilation rates, thereby enlarging the receptive field. The subsequent integration of FreqSelect contributes to a further performance boost, yielding an mIoU of 44.1.

We also conduct an ablation study on the number of frequency bands in FreqSelect. Specifically, we decompose the frequency in an octave-wise manner [21] into different frequency bands. As illustrated in Table 3, we observe an increase in segmentation accuracy with the number of frequency bands, from 2 to 4, and achieve the best results with 4 frequency bands, namely, $[0, \frac{1}{16})$, $[\frac{1}{16}, \frac{1}{8})$, $[\frac{1}{8}, \frac{1}{4})$, and $[\frac{1}{4}, \frac{1}{2}]$. Additionally, we notice that setting a fixed selection weight for the lowest frequency band, instead of dynamically predicting the selection weight, leads to better im-

Table 5. Various task results on the COCO dataset.

Task Model	**Object Detection** Faster RCNN	**Instance Segmentation** Mask RCNN	**Panoptic Segmentation** PanopticFPN
Standard Conv	AP: 37.4	AP: 34.7	PQ: 40.7
FADC	AP: **40.5 (+3.1)**	AP: **37.2 (+2.5)**	PQ: **42.8 (+2.1)**

provements. The reason for this is that the lowest frequency band only requires low effective bandwidth and does not impact the optimization of the dilation rate of AdaDR. Dynamically reweighting the lowest frequency band may result in a faded response of the object. Thus, we assign a fixed weight of 1.0 to the lowest frequency band.

Furthermore, the introduction of FreqSelect in conjunction with AdaDR and AdaKern culminates in the highest mIoU of 44.4, underscoring the synergistic impact of the three strategies. This comprehensive analysis establishes the effectiveness of our proposed FADC, with each constituent strategy playing a pivotal role in enhancing semantic segmentation performance.

F. Experiments Settings

Datasets and Metrics. We evaluate our methods on several challenging semantic segmentation datasets, including Cityscapes [3] and ADE20K [29]. Cityscapes [3] comprises 19 semantic categories designed for semantic segmentation tasks, featuring 5,000 finely annotated images with dimensions of 2048×1024 pixels. The training, validation, and test sets consist of 2,975, 500, and 1,525 samples, respectively. We only utilize the training set for learning. ADE20K [29] is a challenging dataset encompassing 150 semantic classes, distributed across 20,210, 2,000, and 3,352 images in the training, validation, and test sets.

Additionally, we leverage the COCO [11] dataset to evaluate our methods on object detection and instance segmentation tasks. We employ the mean Intersection over Union (mIoU) for semantic segmentation and Average Precision (AP) for object detection/instance segmentation as our evaluation metrics.

Implement details. For Mask2Former [2], PIDNet [25] on the Cityscapes [3]. Our training protocols are the same as the original paper [2, 25]. Specifically, we adopt the poly strategy to update the learning rate and random cropping, random horizontal flipping, and random scaling in the range of [0.5, 2.0] for data augmentation. The number of training epochs, the initial learning rate, weight decay, cropped size, batch size, and optimizer for Mask2Former and PIDNet are [90k, 1e-4, 5e-2, 512×1024, 16, AdamW], [120k, 1e-2, 5e-4, 1024×1024, 12, SGD]. For PSPNet [28] and DeepLab [1], we adopt the settings of [40K, 1e-2, 5e-4,, 512×1024, 8, SGD] is adopted. On the ADE20K [29] dataset, we use ResNet [6] with HorNet [18] utilizing the UperNet [24] framework. HorFPN is employed for HorNet [18]. All models undergo training for 160k iterations using the AdamW [12] optimizer, with a batch size of 16, following the same settings as HorNet [18].

On the COCO [11] dataset, we adhere to common practices [5, 18, 22] and train object detection and instance segmentation models for 12 (1× schedule) or 36 (3× schedule) epochs.

G. Real-Time Semantic Segmentation

We provide more detailed results in Table 4. Equipped with FADC, our PIDNet-M achieves a mIoU of 81.0 at a frame rate of 37.7 frames per second (FPS), surpassing the performance of the heavier PIDNet-L while maintaining a faster speed (37.7 *vs.* 31.1), thereby establishing a new state-of-the-art. This demonstrates the efficiency of the proposed method.

H. More Results on Various Tasks

In Table 5, our FADC method demonstrates consistent improvements across multiple computer vision tasks. Specifically, compared to the baseline models using standard convolution, our approach yields notable enhancements in object detection, instance segmentation, and panoptic segmentation performance metrics. We observe an increase of +3.1 in box Average Precision (AP), indicating improved accuracy in localizing objects within bounding boxes. Moreover, our method achieves a +2.5 improvement in mask AP, showcasing enhanced precision in delineating object boundaries. Additionally, there is a noteworthy enhancement of +2.1 in Panoptic Quality (PQ), reflecting improved overall performance in jointly handling instance and semantic segmentation tasks. These results underscore the effectiveness of our FADC approach in advancing the state-of-the-art across various visual recognition tasks.

I. More Visualized Results

In this section, we provide a more visual demonstration of the effectiveness of the proposed method.

Visualization for AdaDR. As shown in Figure 6, AdaDR learns to spatially adaptively assign dilation rates based on local feature frequency. We depict the sampling coordinates of convolution with adaptive dilation. For better visualization, we only display the sampling coordinates of convolution with a low dilation rate ($\leq$ 2) in (c), which mainly distribute along object boundaries. In other words, AdaDR

learns to assign denser sampling coordinates for higher frequency areas, following the Shannon-Nyquist sampling theorem [15, 19].

Analysis for FreqSelect. As shown in Figure 7, brighter colors indicate a higher value. FreqSelect selectively suppresses high frequencies in areas that do not contribute to accurate predictions, such as the background and the center of objects. Notably, FreqSelect exhibits a tendency to assign a higher attention weight to object boundaries, especially in higher frequency bands. Consequently, features enhanced by FreqSelect showcase a reduction in high frequencies within the background, resulting in clearer object boundaries.

Visualized feature and prediction. We present more representative visualization results in Figure 8. In Figure 8(a), dilated convolution is shown to fail in accurately extracting high-frequency information, such as the fine details of thin poles depicted in Figure 8(c). In contrast, our FADC accurately captures these details in Figure 8(b), resulting in superior predictions, as shown in Figure 8(d).

It is evident that dilated convolution struggles to respond uniformly to large trucks due to an insufficient receptive field to extract local information. On the other hand, FADC uniformly responds to large trucks, leading to more consistent and accurate segmentation predictions. These visualizations serve to illustrate the effectiveness of our proposed FADC in addressing the limitations of dilated convolution.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of European Conference on Computer Vision*, pages 801–818, 2018. 7

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 7

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 5, 6, 7

[4] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 9716–9725, 2021. 6

[5] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. 2022. 7

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5, 7

[7] Yuanduo Hong, Huihui Pan, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv preprint arXiv:2101.06085*, 2021. 6

[8] Saumya Kumaar, Ye Lyu, Francesco Nex, and Michael Ying Yang. Cabinet: Efficient context aggregation network for low-latency semantic segmentation. In *IEEE International Conference on Robotics and Automation*, pages 13517–13524. IEEE, 2021. 6

[9] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *Proceedings of European Conference on Computer Vision*, pages 775–793. Springer, 2020. 6

[10] Xin Li, Yiming Zhou, Zheng Pan, and Jiashi Feng. Partial order pruning: for best speed/accuracy trade-off in neural architecture search. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 9145–9153, 2019. 6

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[13] Michael Mathieu, Mikael Henaff, and Yann LeCun. Fast training of convolutional networks through ffts. In *Proceedings of International Conference on Learning Representations*, pages 1–9, 2014. 1

[14] Yuval Nirkin, Lior Wolf, and Tal Hassner. Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4061–4070, 2021. 6

[15] Harry Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928. 8

[16] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 12607–12616, 2019. 6

[17] Juncai Peng, Yi Liu, Shiyu Tang, Yuying Hao, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen, Zhiliang Yu, Yuning Du, et al. Pp-liteseg: A superior real-time semantic segmentation model. *arXiv preprint arXiv:2204.02681*, 2022. 6

[18] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser Nam Lim, and Jiwen Lu. Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Proceedings of Advances in Neural Information Processing Systems*, 35:10353–10366, 2022. 7

[19] Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949. 8

[20] Joshua A Solomon and Denis G Pelli. The visual filter mediating letter identification. *Nature*, 369(6479):395–397, 1994. 2, 3

[21] Ajay Subramanian, Elena Sizikova, Najib J Majaj, and De-

nis G Pelli. Spatial-frequency channels, shape bias, and adversarial robustness. pages 1–10, 2023. 2, 3, 6

[22] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023. 7

[23] Zhengyang Wang and Shuiwang Ji. Smoothed dilated convolutions for improved dense prediction. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2486–2495, 2018. 3

[24] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of European Conference on Computer Vision*, pages 418–434, 2018. 5, 7

[25] Jiacong Xu, Zixiang Xiong, and Shankar P Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 19529–19539, 2023. 6, 7

[26] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, 2021. 6

[27] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of European Conference on Computer Vision*, pages 325–341, 2018. 6

[28] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2881–2890, 2017. 7

[29] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 5, 7