Frequency-Adaptive Dilated Convolution for Semantic Segmentation

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

Abstract

Dilated convolution, which expands the receptive field by inserting gaps between its consecutive elements, is widely employed in computer vision. In this study, we propose three strategies to improve individual phases of dilated convolution from the perspective of spectrum analysis. Departing from the conventional practice of fixing a global dilation rate as a hyperparameter, we introduce Frequency-Adaptive Dilated Convolution (FADC), which dynamically adjusts dilation rates spatially based on local frequency components. Subsequently, we design two plug-in modules to directly enhance effective bandwidth and receptive field size. The Adaptive Kernel (AdaKern) module decomposes convolution weights into low-frequency and high-frequency components, dynamically adjusting the ratio between these components on a per-channel basis. By increasing the high-frequency part of convolution weights, AdaKern captures more high-frequency components, thereby improving effective bandwidth. The Frequency Selection (FreqSelect) module optimally balances high- and low-frequency components in feature representations through spatially variant reweighting. It suppresses high frequencies in the background to encourage FADC to learn a larger dilation, thereby increasing the receptive field for an expanded scope. Extensive experiments on segmentation and object detection consistently validate the efficacy of our approach. The code is made publicly available at https://github.com/Linwei-Chen/FADC.

1. Introduction

Dilated convolution inserts gaps between the filter values at a dilation rate (D) to expand the receptive field without significantly increasing computational load. This technique is widely used in computer vision tasks, such as semantic segmentation [9, 78] and object detection [55].

While effective in expanding the receptive field size with a large dilation rate, it comes at the expense of high-frequency component response [78]. Increasing the dilation rate from 1 to D is equivalent to expanding the convolution kernel through zero-insertion by a factor of D. According to the scaling property of Fourier Transforms [51, 56], both the frequency response curve and the bandwidth of the convolution kernel will be scaled to $\frac{1}{D}$. As illustrated in Figure 1, the bandwidth of the red curve for $D = 4$ is only a quarter of the one for $D = 1$ in blue. The reduced bandwidth significantly limits the layer's ability to process high-frequency components. For instance, gridding artifacts occur when a feature map has higher-frequency content than the sampling rate of the dilated convolution [67, 78].

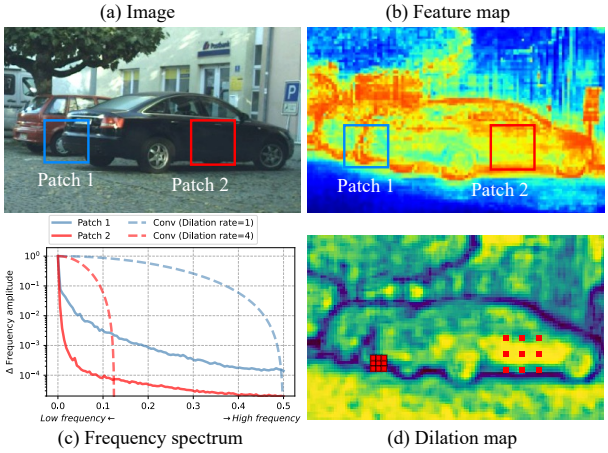


Figure 1. For the input image (a), its extracted features (b) exhibit spatial variation. Patch 1 contains more high-frequency information, whereas the frequency of patch 2 is predominantly concentrated in low frequency (c). Consequently, assigning a small dilation rate for patch 1 is essential to preserve a high effective bandwidth, while a larger dilation rate, with limited effective bandwidth, is sufficient for patch 2, benefiting the achievement of a larger receptive field (d).

In this paper, we introduce Frequency-Adaptive Dilated Convolution (FADC) to enhance dilated convolution through the lens of spectrum analysis. As illustrated in Figure 2, FADC comprises three key strategies, *i.e.*, Adaptive Dilation Rate (AdaDR), Adaptive Kernel (AdaKern), and Frequency Selection (FreqSelect), aimed at enhancing the individual phases of vanilla dilated convolution. AdaDR

*Corresponding Author

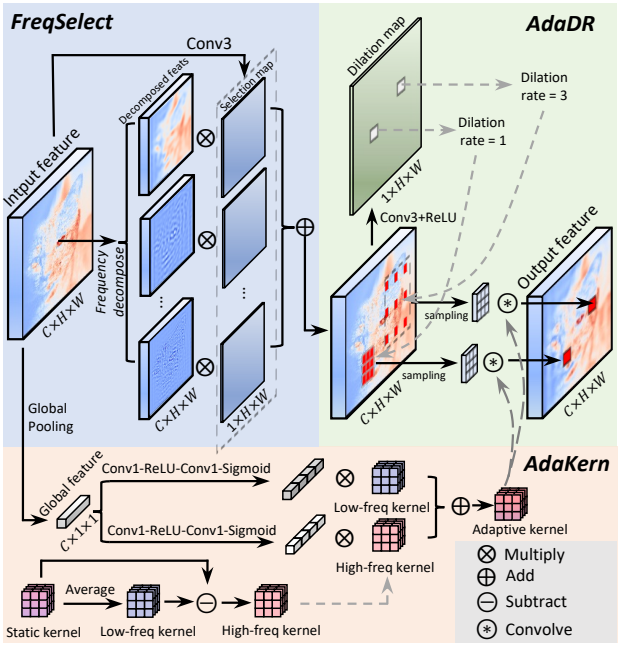


Figure 2. Overview of the proposed Frequency Adaptive Dilated Convolution (FADC). It comprises three strategies: Adaptive Dilation Rate (AdaDR), Adaptive Kernel (AdaKern), and Frequency Selection (FreqSelect).

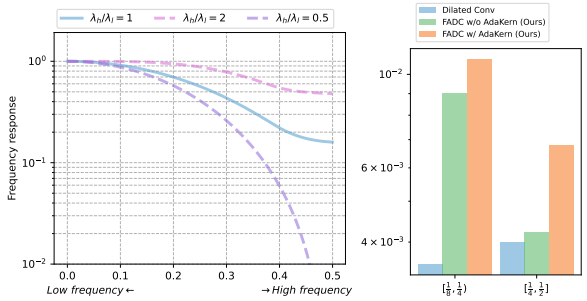


Figure 3. Analysis of AdaKern. Left: representative frequency response of a convolution kernel, λ_l, λ_h represent dynamic weights for decomposed low/high-frequency kernels. Right: the FADC with AdaKern increases the proportion of high-frequency band power within $[\frac{1}{8}, \frac{1}{4})$ and $[\frac{1}{4}, \frac{1}{2}]$ in the extracted feature, indicating an increase in effective bandwidth.

spatially adjusts the dilation rate, AdaKern operates on the convolution kernel weights, while FreqSelect directly balances the frequency power of the input feature to encourage the expansion of the receptive field.

Unlike the conventional approach of globally fixing the dilation rate, our AdaDR dynamically assigns dilation rates locally based on the spectrum. For instance, in patch 1 of Figure 1(a), where the car boundaries exhibit much high-frequency component (indicated by the blue solid line), AdaDR applies a small dilation rate ($D = 1$) with a broad effective bandwidth (represented by the blue dot curve). Conversely, for the car door in patch 2, where the frequency power is predominantly concentrated in the low-frequency domain, AdaDR increases the dilation rate D to 4, as a reduced bandwidth can still encompass a substantial amount of frequency power. The dilation map for these two patches is illustrated in Figure 1(d). In comparison to the fixed dilation rates (*e.g.*, D =1, 2, 4 in [41, 78]), our AdaDR enhances the theoretical average receptive field size of Figure 1 from ∼440 to ∼1000 pixels.

AdaKern is a plug-in module that manipulates the convolution kernel to optimize the frequency response curve in Figure 3 and enhances the effective bandwidth. As shown in Figure 3, this module decomposes convolution weights into low-frequency and high-frequency components. This allows us to dynamically manipulate both components on a per-channel basis. For example, increasing the weight of the high-frequency kernel (marked in red at the bottom of Figure 2) leads to a stronger response for high-frequency components, which in turn increases the effective bandwidth as shown in the left of Figure 3, curve of $\lambda_h/\lambda_l = 2$.

The FreqSelect increases the receptive field size by balancing high- and low-frequency components in the feature before feeding into dilated convolution. Since convolution tends to amplify high-frequency components [48], features after dilated convolution often exhibit a higher proportion of high-frequency components. To capture these added high-frequency components, a small dilation rate D will be favored for its large effective bandwidth, at the cost of compromised receptive field size. By suppressing the high-frequency power on the input features, our FreqSelect module is able to increase the respective field size. Specifically, As shown in Figure.2, FreqSelect decomposes the feature map into 4 frequency channels from low to high. Then, we spatially reweights each channel with a selection map to balance frequency power, enabling FADC to effectively learn a larger receptive field.

Our experimental results in semantic segmentation show that our proposed method consistently brings improvements, thus validating the effectiveness of our approach. In particular, when our proposed method is applied with PIDNet, it achieves the optimal balance between inference speed and accuracy on Cityscapes, resulting in an 81.0 mIoU at 37.7 FPS. Moreover, our proposed strategy can also be integrated into deformable convolution and dilated attention, resulting in a consistent boost in performance for both segmentation and object detection tasks. Our contributions can be summarized as follows:

- We conduct an in-depth exploration of dilated convolution using frequency analysis, reframing the assignment of dilation as a trade-off problem that involves balancing effective bandwidth and receptive field.
- We introduced Frequency-Adaptive Dilated Convolution (FADC). It adopts the Adaptive Dilation Rate (AdaDR), Adaptive Kernel (AdaKern), and Frequency Selection (FreqSelect) strategies. AdaDR dynamically adjusts dilation rates in a spatially variant manner to achieve a

balance between effective bandwidth and receptive field. AdaKern adaptively adjusts the kernel to fully utilize the bandwidth, and FreqSelect learns a frequency-balanced feature to encourage a large receptive field.
- We validate our approach through comprehensive experiments in the segmentation task, consistently demonstrating its effectiveness. Furthermore, the proposed AdaKern and FreqSelect also prove to be effective when integrated with deformable convolution and dilated attention in object detection and segmentation tasks.

2. Related work

Content-Adaptive Networks. As deep learning technology advances [**? ? ? ? ? ? ? ?**], the effectiveness of content-adaptive characteristics has been demonstrated by various works [13, 20, 57, 59, 65, 83]. One content-adaptive strategy involves weight adjustments, which are widely employed. Recent vision transformers [15, 22, 42] incorporate attention mechanisms to predict input-adaptive attention values. These models have achieved significant success with large receptive, but suffer from heavy computation.

In addition to weight adjustments, [1, 13, 30, 68, 74, 84] modify the sampling grid of the convolution kernel that is closely related to our work. Deformable convolution [13, 68, 84] is employed in various computer vision tasks, including object detection. It introduces $K \times K \times 2$ asymmetrical offsets for every position in the sampling grid, causing the extracted features to exhibit spatial deviations. In object detection tasks, estimated boxes are corrected through regression to mitigate these deviations. However, in position-sensitive tasks such as semantic segmentation, where strong consistency in density and features at each location is crucial, features with spatial deviations can lead to incorrect learning. In contrast, the proposed frequency-adaptive dilated convolution only requires one value as the dilation rate for each position. This approach necessitates fewer additional standard convolutions for computing sampling coordinates, making it lightweight. Moreover, it eliminates spatial deviations, thereby reducing the risk of erroneous learning and benefiting position-sensitive tasks.

Adaptive Dilated Convolution [1, 30, 74] also discards the use of globally fixed dilation. [30] formulates the dilation of each point in the kernel as learned fixed weights, while [1, 74] empirically adjust the dilation rate based on the assumption that dilation values are linked to inter-layer patterns between convolution layers or the object scale. In contrast to [1, 30, 74], which rely on intuitive assumptions, our proposed method is motivated by quantitative frequency analysis. Moreover, they overlook the aliasing artifacts that occur when the feature frequency exceeds the sampling rate, exposing them to a potential risk of degradation.

Aliasing Artifacts in Neural Networks. The issue of aliasing artifacts in neural networks is gaining increasing attention within the computer vision community. Several studies have analyzed the aliasing artifacts resulting from insufficient sampling during downsampling in neural networks [27, 32, 64, 80, 85]. Others have broadened their focus to include anti-aliasing techniques in various applications, such as vision transformers [52], tiny object detection [45], and image generation in generative adversarial networks (GANs) [29]. Regarding aliasing artifacts in dilated convolution, commonly referred to as the gridding artifact, they occur when a feature map contains higher-frequency content than the sampling rate of the dilated convolution [78]. Previous works either empirically applied learned convolution to acquire low-pass filters for anti-aliasing [78], employed dilated convolution with multiple dilation rates [61, 67], or used a fully connected layer to smooth dilated convolutions [69]. However, these methods are primarily empirically designed, involving stacking more layers, and do not explicitly handle the issue from a frequency perspective. In contrast, our proposed method avoids gridding artifacts by dynamically adjusting the dilation rate based on local frequency. Additionally, FreqSelect contributes by suppressing high frequencies in the background or object center. This approach offers a more principled and effective solution to address aliasing artifacts.

Frequency domain learning. Traditional signal processing has long relied on frequency-domain analysis as a fundamental tool [2, 50]. Notably, these well-established methods have recently found applications in deep learning, playing pivotal roles. In this context, they are employed to examine the optimization strategies [75] and generalization capabilities [66] of Deep Neural Networks (DNNs). Moreover, these frequency-domain techniques have been seamlessly integrated into DNN architectures. This integration has facilitated the learning of non-local features [11, 19, 28, 35, 54] or domain-generalizable representations [36]. Recent studies [48, 79] demonstrate that capturing balanced representations of both high- and low-frequency components can enhance model performance. Therefore, our method provides a frequency view for dilated convolution and improves its capability to capture different frequency information.

3. Frequency Adaptive Dilated Convolution

The overview of the proposed FADC is illustrated in Figure 2. In this section, we begin by introducing the AdaDR strategy, outlining how we balance bandwidth and receptive field. Subsequently, we delve into the details of the AdaKern and FreqSelect strategies, designed to fully leverage bandwidth and promote a large receptive field.

3.1. Adaptive Dilation Rate

Dilated Convolution. The widely-used dilated convolution can be formulated as follows:

$$\mathbf{Y}(p) = \sum_{i=1}^{K\times K} \mathbf{W}_i \mathbf{X}(p + \Delta p_i \times D), \quad (1)$$

where $\mathbf{Y}(p)$ represents the pixel value at position p in the output feature map, K is the kernel size, $\mathbf{W}_i$ denotes the weight parameters for the kernel, and $\mathbf{X}(p + \Delta p_i)$ represents the pixel value at the position corresponding to p offset by Δp_i in the input feature map. The variable Δp_i represents the i-th location of the pre-defined grid sampling $(-1,-1), (-1,0), (-1,+1), \ldots, (+1,+1)$. By receptive field can be enlarged by increasing the dilation rate D.

Frequency analysis. Previous works have observed that an increased dilation leads to the degradation of frequency information capture [67, 69, 78]. Specifically, increasing the dilation rate from 1 to D scales up the convolution kernel by a factor of D, following the scaling property of Fourier Transforms. Consequently, the response frequency of the convolution kernel is reduced to $\frac{1}{D}$, resulting in a shift in the frequency response from high frequency to lower frequency [51, 56], as depicted in Figure 1. Moreover, dilated convolution effectively operates at a sampling rate of $\frac{1}{D}$, making it unable to capture frequencies above the Nyquist frequency, *i.e.*, half the sampling rate $\frac{1}{2D}$.

Specifically, we first transform the feature map $\mathbf{X} \in \mathbb{R}^{H\times W}$ into the frequency domain using the Discrete Fourier Transform (DFT), $\mathbf{X}_F = \mathcal{F}(\mathbf{X})$, it can be represented as

$$\mathbf{X}_F(u,v) = \frac{1}{HW} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \mathbf{X}(h,w) e^{-2\pi j(uh+vw)}, \quad (2)$$

where $\mathbf{X}_F \in \mathbb{R}^{H\times W}$ represents the output array of complex numbers from the DFT. H and W denote its height and width. h, w indicates the coordinates of feature map $\mathbf{X}$. 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 $\{-\frac{H}{2}, -\frac{H+1}{2}, \ldots, \frac{H-1}{2}\}$, and v takes values from $\{-\frac{W}{2}, -\frac{W+1}{2}, \ldots, \frac{W-1}{2}\}$. Consequently, the set of high frequencies larger than the Nyquist frequency $\mathcal{H}_D^+ = \{(u,v) \mid |k| > \frac{1}{2D} \text{ or } |l| > \frac{1}{2D}\}$ is unable to be accurately captured, limiting its bandwidth.

Adaptive dilation rate. Building on the above analysis, the selection of the dilation rate can be viewed as a trade-off between a large receptive field and effective bandwidth. Considering that the input feature map is spatially variant, the optimal dilation for each pixel can be different. Thus, we introduce the strategy of Adaptive Dilation Rate (AdaDR) to achieve better balancing. It assigns each pixel a different dilation rate

$$\mathbf{Y}(p) = \sum_{i=1}^{K\times K} \mathbf{W}_i \mathbf{X}(p + \Delta p_i \times \hat{\mathbf{D}}(p)). \quad (3)$$

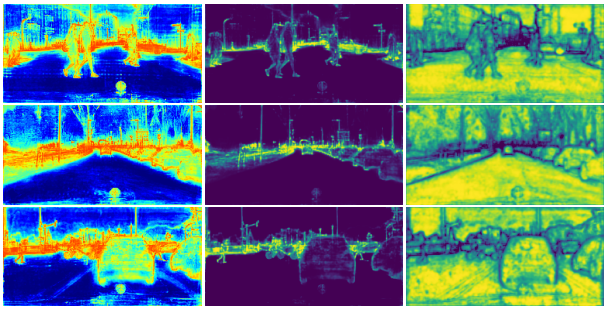
$\hat{\mathbf{D}}(p)$ can be predicted by a convolutional layer with parameters θ. Particularly, we incorporate a ReLU layer to ensure the non-negativity of the dilations, and we also adopt the modulation mechanism [84]. It aims to maximize the receptive field and minimize the lost frequency information for each pixel. For a local feature centered at p with a window size of s, we term it as $\mathbf{X}^{(p,s)}$. Its receptive field $\mathrm{RF}(p) = (K-1) \times \hat{\mathbf{D}}(p) + 1$ is positively related to $\hat{\mathbf{D}}(p)$. The frequencies in a set $\mathcal{H}^+_{\hat{\mathbf{D}}(\mathbf{p})}$ are unable to be captured accurately. Thus, the lost frequency information can be measured by calculating the high-frequency power $\mathrm{HP}(p) = \sum_{\mathcal{H}^+_{\hat{\mathbf{D}}(p)}} |\mathbf{X}_F^{(p,s)}(u,v)|^2$. Therefore, the optimization of θ can be written as

$$\theta = \max_\theta \left(\sum \mathrm{RF}(p) - \sum \mathrm{HP}(p) \right). \quad (4)$$

However, direct optimization can be impractical due to the discrete nature of the frequency set $\mathcal{H}^+_{\hat{\mathbf{D}}(p)}$, and the fact that the calculation of HP is non-differentiable. Consequently, we choose to optimize $\hat{\mathbf{D}}(p)$ directly, *i.e.*, by increasing the dilation rate at position p where the $\mathrm{HP}(p)$ is low to encourage 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

$$\theta = \max_\theta \left(\sum_{p\in \mathrm{HP}^-} \hat{\mathbf{D}}(p) - \sum_{p\in \mathrm{HP}^+} \hat{\mathbf{D}}(p) \right). \quad (5)$$

Here, HP^+ and HP^- represent pixels with the highest/lowest (*e.g.*, 25%) high-frequency power, *i.e.*, the brighter/darker areas in Figure 4(b), respectively.



(a) Feature (b) High frequency power map (c) Dilation map

Figure 4. Feature Visualization. Pixels with brighter colors on the high-frequency power map indicate a higher level of high-frequency components. Brighter colors on the dilation map indicate higher dilation rates. We observe that FADC learns to assign lower dilation rates to high-frequency areas, such as the boundaries of objects, and higher dilation rates for low-frequency areas, such as the center of objects and the background.

3.2. Adaptive Kernel

AdaDR achieves a delicate equilibrium between effective bandwidth and receptive field through the individual assignment of dilation rates to each pixel, optimizing both factors collectively. The effective bandwidth, intimately connected to the convolutional kernel's weight, plays a pivotal

role. Traditional convolutional kernels learn to capture features across diverse frequency bands, which are crucial for comprehending intricate visual patterns, but they become static once trained. To further enhance effective bandwidth, we decompose convolutional kernel parameters into low-frequency and high-frequency components before introducing dynamic weighting to adjust the frequency response. This process only adds minor additional parameters and computational overhead. For a static convolutional kernel, its weights $\mathbf{W}$ can be decomposed as follows

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

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 $\bar{\mathbf{W}}$. As discussed in [62], higher mean values are more likely to attenuate the high-frequency components. The term $\hat{\mathbf{W}} = \mathbf{W} - \bar{\mathbf{W}}$ denotes the residual part, capturing local differences and extracting the high-frequency components. After decomposition, our AdaKern dynamical adjust both high and low frequency components and can be formally represented as

$$\mathbf{W}' = \lambda_l\bar{\mathbf{W}} + \lambda_h\hat{\mathbf{W}}, \tag{7}$$

where λ_l, λ_h are the dynamic weights for each channel, which is predicted by a simple and lightweight global pooling + convolution layers. Dynamically adjusting the ratio of $\frac{\lambda_l}{\lambda_h}$ based on the input context, allows the network to focus on specific frequency bands and adapt to the complexity of visual patterns in the feature. This dynamic frequency-adaptive approach enhances the network's ability to capture both low-frequency context and high-frequency local details. This, in turn, increases the effective bandwidth, leading to improved performance in segmentation tasks that require diverse feature extraction across different frequencies.

3.3. Frequency Selection

As indicated by prior studies [48], conventional convolution often functions as a high-pass filter. Consequently, the resulting features tend to manifest a higher proportion of high-frequency components. This inclination leads to the adoption of smaller overall dilation rates to preserve a high effective bandwidth, unfortunately compromising the size of the receptive field. FreqSelect is devised to enhance the receptive field by balancing high- and low-frequency components in the feature representations.

Specifically, FreqSelect initially decomposes features into different frequency bands by applying distinct masks in the Fourier domain:

$$\mathbf{X}_b = \mathcal{F}^{-1}(\mathcal{M}_b\mathbf{X}_F), \tag{8}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fast Fourier Transform. $\mathcal{M}_b$ is a binary mask designed to extract the corresponding frequency:

$$\mathcal{M}_b(u,v) = \begin{cases} 1 & \text{if } \phi_b \le \max(|u|,|v|) < \phi_{b+1} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Here, ϕ_b, ϕ_{b+1} are from $B+1$ predefined frequency thresholds $\{0, \phi_1, \phi_2, ..., \phi_{B-1}, \frac{1}{2}\}$. Subsequently, FreqSelect dynamically reweights the frequency components in different frequency bands spatially. This is formulated as:

$$\hat{\mathbf{X}}(i,j) = \sum_{b=0}^{B-1}\mathbf{A}_b(i,j)\mathbf{X}_b(i,j), \tag{10}$$

where $\hat{\mathbf{X}}(i,j)$ is the learned frequency-balanced feature after FreqSelect, and $\mathbf{A}_b \in \mathbb{R}^{H\times W}$ denotes the selection map for the b-th frequency band. Specifically, we decompose the frequency in an octave-wise [60] manner into four frequency bands, *i.e.*, $[0, \frac{1}{16})$, $[\frac{1}{16}, \frac{1}{8})$, $[\frac{1}{8}, \frac{1}{4})$, and $[\frac{1}{4}, \frac{1}{2}]$.

4. Experiments

4.1. Experiments Settings

Datasets and Metrics. We evaluate our methods on several challenging semantic segmentation datasets, including Cityscapes [12] and ADE20K [82]. We employ the mean Intersection over Union (mIoU) for semantic segmentation [4, 8, 18, 39, 44] and Average Precision (AP) for object detection/instance segmentation [5–7, 23, 26] as our evaluation metrics.

Implement details. Mask2Former [10], PIDNet [72], ResNet/HorNet+UPerNet, we keep the same setting with the original paper [10, 53, 72]. On the COCO [37] dataset, we adhere to common practices [21, 53, 68] and train object detection and instance segmentation models for 12 ($1\times$ schedule) or 36 ($3\times$ schedule) epochs. In the case of Dilated-ResNet, we substitute the dilated convolution at stage-3∼4 with the proposed FADC. For PIDNet, the convolution at the bottleneck is replaced with the proposed FADC. For ResNet, we replace the convolution at stage-2∼4 with the proposed FADC.

4.2. Main Results

In this section, we initially assess the effectiveness of the proposed method through standard semantic segmentation benchmarks. Subsequently, we report results on real-time semantic segmentation. Finally, we seamlessly integrate the proposed method into pertinent deformable convolutions (DCNv2 [84]) and advanced frameworks, such as DCN3-based InternImage [68], along with incorporating dilated attention mechanisms as exemplified by DiNAT [21].

Standard Semantic Segmentation. As shown in Table 1, we compared the proposed FADC with Dilated Convolution [78], Deformable Convolution (DCNv2)[84], and Adaptive Dilated Convolution (ADC)[74]. On the widely used Cityscapes dataset [12], when equipped with our FADC, the results for PSPNet, DeepLabV3+, and

Table 1. Results are reported on Cityscapes validation set [12].

Method *Backbone: Dilated-ResNet-50 [78]*	#Params	#FLOPS	mIoU
PSPNet [81]	49.0M	1427.5G	77.8
PSPNet [81] + DCNv2 [84]	+0.7M	+24.5G	79.7
PSPNet [81] + FADC (Ours)	+0.5M	+9.2G	**80.4**
DeepLabV3+ [9]	43.6G	1410.9G	79.2
DeepLabV3+ [9] + DCNv2 [84]	+0.7M	+24.5G	79.9
DeepLabV3+ [9] + FADC (Ours)	+0.5M	+9.2G	**80.3**
Backbone: Dilated-ResNet-101 [78]			
DeepLabV3+ [9] + ADC [74]	62.8M	2032.3G	80.7
DeepLabV3+ [9] + FADC (Ours)	63.9M	2067.0G	**81.5**
Backbone: ResNet-50 [24]			
Mask2Former [10]	44.0M	-	79.4
Mask2Former [10] + DCNv2 [84]	+0.9M	+7.7G	80.4
Mask2Former [10] + FADC (Ours)	+0.5M	+4.3G	**80.6**

Table 2. Quantitative comparisons on semantic segmentation tasks with UPerNet [71] on the ADE20K validation set.

Method	#Params	#FLOPS	mIoU	
			SS	MS
ResNet-50 [24]	66M	947G	40.7	41.8
ResNet-101 [24]	85M	1029G	42.9	44.0
ResNet-50-FADC (Ours)	67M	949G	**44.4**	**45.5**
Swin-B [42]	121M	1188G	48.1	49.7
NAT-B [22]	123M	1137G	48.5	49.7
ConvNeXt-B [43]	122M	1170G	49.1	49.9
ConvNeXt-B-dcls [30]	122M	1170G	49.3	-
DAT-B [70]	121M	1212G	49.4	50.6
DiNAT-B [21]	123M	1137G	49.6	50.4
Focal-B [73]	126M	1354G	49.0	50.5
InternImage-B [68]	128M	1185G	50.8	51.3
HorNet-B [53]	126M	1171G	50.5	50.9
HorNet-B-FADC (Ours)	128M	1176G	**51.1**	**51.5**

Mask2Former show improvements of +2.6, +1.1, and +1.2 mIoU, respectively. These enhancements outperform DCNv2 by 0.7, 0.4, and 0.2 mIoU with fewer additional computations and parameters. FADC also outperforms ADC, which adopts an adaptive dilation strategy, by 0.8 mIoU. Furthermore, as demonstrated in Table 2 using the more challenging ADE20K dataset, FADC significantly enhances the mIoU of ResNet-50 with UPerNet by 3.7, surpassing even its heavier counterpart, ResNet-101 (44.4 *vs.* 42.9). When applied with larger HorNet-B, it leads to +0.6 gains and outperforms recent state-of-the-art methods, including Swin, ConvNeXt, RepLKNet-31L, InternImage, and DiNAT. Notably, HorNet-B-FADC exhibits superior performance and improvement (51.1 *vs.* 49.3, and +0.6 *vs.* +0.2) compared to ConvNeXt-B-dcls [30], which applies learning dilation spacing.

Real-time Semantic Segmentation. Real-time semantic segmentation is crucial for applications such as autonomous vehicles [17] and robot surgery [58]. We further evaluate the proposed method for real-time semantic segmentation on the Cityscapes dataset [12] as shown in Table 3.

Table 3. Comparison on Cityscapes [12]. † indicates the models pre-trained by extra datasets. We follow [72] to test our method on a single RTX 3090 with a resolution of 1024×2048.

Model	#Params	#FLOPs	#FPS	Val	Test
DF2-Seg1 [34]	-	-	67.2	75.9	74.8
DF2-Seg2 [34]	-	-	56.3	76.9	75.3
SwiftNetRN-18 [47]	11.8M	104.0G	39.9	75.5	75.4
SwiftNetRN-18 ens [47]	24.7M	218.0G	18.4	-	76.5
CABiNet [31]	2.64M	12.0G	76.5	76.6	75.9
BiSeNet(Res18)[77]	49M	55.3G	65.5	74.8	74.7
BiSeNetV2-L[76]	-	118.5G	47.3	75.8	75.3
STDC1-Seg75 [16]	-	-	74.8	74.5	75.3
STDC2-Seg75 [16]	-	-	58.2	77.0	76.8
PP-LiteSeg-T2 [49]	-	-	96.0	76.0	74.9
PP-LiteSeg-B2 [49]	-	-	68.2	78.2	77.5
HyperSeg-M [46]	10.1M	7.5G	59.1	76.2	75.8
HyperSeg-S [46]	10.2M	17.0G	45.7	78.2	78.1
SFNet(DF2)[33]	10.53M	-	87.6	-	77.8
SFNet(ResNet-18)[33]	12.87M	247.0G	30.4	-	78.9
SFNet(ResNet-18)†[33]	12.87M	247.0G	30.4	-	80.4
DDRNet-23-S[25]	5.7M	36.3G	108.1	77.8	77.4
DDRNet-23 [25]	20.1M	143.1G	51.4	79.5	79.4
DDRNet-39 [25]	32.3M	281.2G	30.8	-	80.4
PIDNet-S [72]	7.6M	47.6G	93.2	78.8	78.6
PIDNet-M [72]	34.4M	197.4G	39.8	80.1	80.1
PIDNet-L [72]	36.9M	275.8G	31.1	80.9	80.6
PIDNet-M-FADC (Ours)	34.6M	198.4G	37.7	**81.0**	**80.6**

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.

Integration with DCNv2, InternImage, and DiNAT. There exists a set of potent techniques for adjusting the sampling coordinates of convolution or attention, akin to dilated convolution. Examples include DCNv2 [84], InternImage [68] (a DCNv3-based model), and DiNAT [21]. DCNv2 and InternImage can be conceptualized as dynamically assigning a dilation rate to each point of the kernel. Conversely, DiNAT adjusts the sampling coordinates for calculating attention in a manner analogous to dilated convolution, thereby encountering similar challenges associated with dilation convolution. Here, we combine the proposed AdaKern and FreqSelect with DCNv2, InternImage (DCNv3-based model), and DiNAT to assess their effectiveness. Table 4 illustrates the impact of this integration. DCNv2 has previously demonstrated notable success in object detection tasks, and our proposed AdaKern and FreqSelect contribute a further enhancement of 0.9 in box AP. Furthermore, FreqSelect enhances the performance of InternImage by 0.8 on the ADE20K dataset, and DiNAT by 0.6 in mask AP on COCO [38]. These results serve as compelling evidence of the efficacy of our approach.

Table 4. Combining the proposed AdaKern and FreqSelect strategies with DCNv2 [84] on the object detection task. All models are trained with a 1× schedule on the COCO dataset [38].

Model: Faster-RCNN [55]	Param	FLOPs	AP^{box}	AP^{box}_{50}	AP^{box}_{75}
ResNet-50 [24]	41.7M	207.1G	37.4	58.1	40.4
ResNet-50 [24] + DCNv2 [84]	+0.9M	+3.9G	41.3	62.8	45.1
+AdaKern+FreqSelect (Ours)	+1.0M	+4.6G	**42.2**	**63.5**	**46.2**

Table 5. Combining the proposed FreqSelect strategies with InternImage [68] on the ADE20K validation set.

Method	#Params	#FLOPS	mIoU	
			SS	MS
Swin-T [42]	60M	945G	44.5	45.8
ConvNeXt-T [43]	60M	939G	46.0	46.7
SLAK-T [40]	65M	936G	47.6	-
InternImage-T [68]	59M	944G	47.9	48.1
+ FreqSelect (Ours)	60M	948G	**48.7**	**48.9**

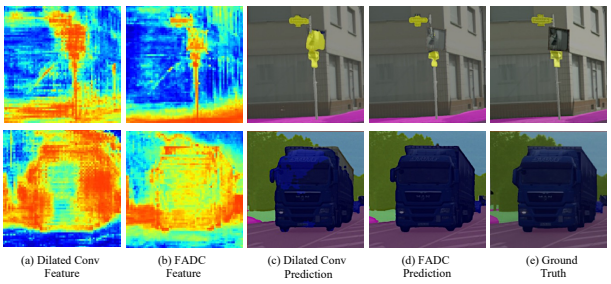


Figure 5. Visualized results on Cityscape [12]. Aliasing artifacts are evident in (a), resulting in the loss of details in the representation of thin poles and truck boundaries, which leads to inferior predictions in (c). In contrast, our proposed FADC method in (b) demonstrates an accurate and uniform response to both thin poles and large trucks, thereby contributing to consistently accurate predictions in (d).

Visualized Results. We present representative visualization results in Figure 5. The top row demonstrates that dilated convolution fails to accurately extract high-frequency information, such as the fine details of thin poles. In contrast, our proposed Frequency-Adaptive Dilated Convolution (FADC) accurately captures these details, resulting in superior predictions. In the bottom row, 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.

5. Analysis and Disccusion

We utilize dilated ResNet-50 [78] as the baseline model and conduct a thorough analysis of the proposed FADC. Additional analyses are available in the supplementary material.

Table 6. Object detection and instance segmentation performance on the COCO dataset [38] with the Mask R-CNN detector [23]. All models are trained with a 3× schedule [21, 68].

Model	Params	FLOPs	AP^{box}	AP^{mask}
ConvNeXt-S [43]	348G	70M	47.9	42.9
RegionViT-B+ [3]	307G	93M	48.3	43.5
NAT-S [22]	330G	70M	48.4	43.2
ConvNeXt-B [43]	486G	108M	48.5	43.5
Swin-S [42]	359G	69M	48.5	43.3
Internlmage-S [68]	340G	69M	49.7	44.5
DAT-S [70]	378G	69M	49.0	44.0
DiNAT-S [21]	330G	70M	49.3	43.9
+FreqSelect (Ours)	331G	71M	**49.8**	**44.5**

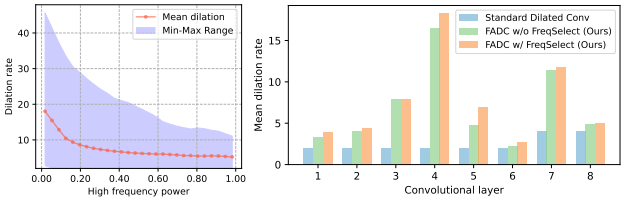


Figure 6. Left: The curve illustrates the relationship between normalized high-frequency power and the predicted dilation rate. FADC, incorporating AdaDR, assigns lower dilation rates to areas with more high-frequency components, such as the boundaries of cars. Right: Mean dilation rates at stages 3 and 4 in ResNet.

Analysis of AdaDR. As depicted in Figure 6, AdaDR learns to predict a small dilation rate for areas with high frequencies, such as the boundaries of cars, bicycles, and persons (refer to Figure 4(c)), to maintain a high bandwidth for capturing high frequency fine details. Conversely, it assigns a larger dilation rate for smoother areas with a lower level of high frequency to expand the receptive field. Furthermore, in comparison to deformable convolution [13, 84], AdaDR avoids spatial deviation illustrated in Figure 7, preventing incorrect learning and benefits position-sensitive tasks.

Analysis of AdaKern. Through adaptive adjustment of the ratio between high-frequency and low-frequency components in the static kernel based on input feature, AdaKern modulates the frequency response of the convolution kernel, empowering FADC to extract more high-frequency detailed information. As depicted in the right of Figure 3, we perform a statistical analysis of the frequency power in the feature map. In comparison to dilated convolution, FADC extracts a greater amount of high-frequency information, crucial for capturing segmentation details, and using AdaKern further amplifies this capability.

Analysis of FreqSelect. We conduct a statistical analysis of the average weights generated by FreqSelect for different frequency bands, as presented in Table 8. FreqSelect predicted a lower average weight for the higher frequency band, consistent with the inverse power law [63]. Upon visualizing the heatmap in Figure 8, we noted that FreqSelect

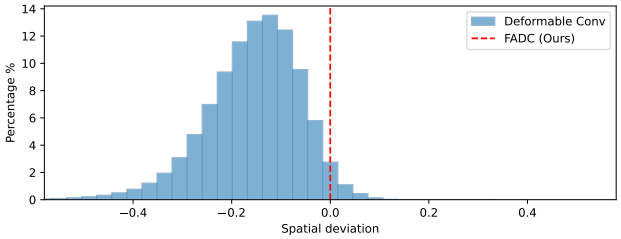


Figure 7. Spatial deviation analysis. We illustrate a histogram of the spatial deviation between the center of the predicted sampling coordinate and corresponding pixel coordinates.

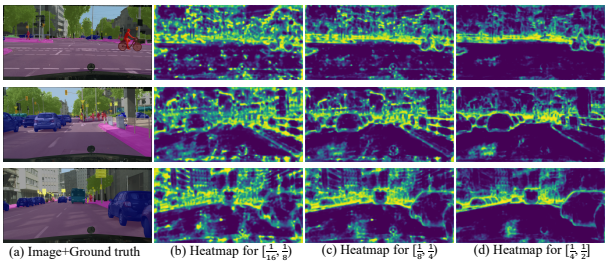


Figure 8. Visualization for Freqselect. Brighter colors indicate a higher attention value.

Table 7. Theoretical receptive field analysis using dilated ResNet-50 [78]. FADC considerably improves the receptive field of the entire model, and FreqSelect further enhances it.

Conv. Type	Standard Dilated Conv	FADC (w/o FreqSelect)	FADC (w/ FreqSelect)
Receptive Field	441	1007	**1100**

Table 8. Average attention weights of different frequency bands. Statistics are collected on the Cityscapes validation set [12].

Frequency band ($\times 2\pi$)	$[0, \frac{1}{16})$	$[\frac{1}{16}, \frac{1}{8})$	$[\frac{1}{8}, \frac{1}{4})$	$[\frac{1}{4}, \frac{1}{2}]$
Average selection weight	1.0	0.66	0.50	0.34

tends to assign a higher attention weight to object boundaries. This is more obvious for higher frequency bands. It selectively suppresses high frequencies in areas that do not contribute to accurate predictions, such as the background and the center of objects. This encourages FADC to learn higher dilation rates, thereby enlarging the receptive field.

Receptive Field. The importance of a large receptive field in scene understanding tasks has been emphasized [14, 15]. Adopting the AdaDR strategy, FADC can employ a higher overall dilation rate to expand the receptive field, surpassing the widely used dilated ResNet [78] with a global fixed dilation rate, as indicated in Table 7. Figure 8 visually demonstrates how FreqSelect contributes to an increased average dilation rate of FADC. By selectively weighting frequencies in the feature map, FreqSelect further encourages a higher dilation rate, ultimately resulting in an elevated receptive field, as shown in Table 7.

Bandwidth. Measuring the bandwidth of a complex model is not straightforward [57], instead, we directly assess the frequency information in extracted features. In Figure 3, in comparison with dilated convolution, FADC increases the power in the high-frequency band of $[\frac{1}{8}, \frac{1}{4})$ and $[\frac{1}{4}, \frac{1}{2}]$. AdaKern further enhances the power in the frequency band $[\frac{1}{4}, \frac{1}{2}]$. This indicates the extraction of more high-frequency information, demonstrating an improved bandwidth.

Aliasing Artifacts. As outlined in [67, 78], aliasing artifacts, commonly referred to as gridding artifacts, manifest when the frequency content of a feature map exceeds the sampling rate of dilated convolution, as depicted in Figure 5. To elaborate, these artifacts occur when the frequency within the feature map surpasses the effective bandwidth of dilated convolution. Previous studies have attempted to address this issue empirically by incorporating additional convolutional layers to learn a low-pass filter for artifact removal [61, 67] or by employing multiple dilation rates to increase the sampling rate [61, 67]. In contrast to these approaches, our proposed method mitigates gridding artifacts by dynamically adjusting the dilation rate based on local frequency. Furthermore, FreqSelect contributes to this by suppressing high frequencies in areas that do not contribute to accurate predictions in the background or object center.

6. Conclusion

In this work, we review dilated convolution from a frequency perspective and introduce FADC to improve individual phases with three key strategies: AdaDR, AdaKern, and FreqSelect. Diverging from the conventional approach of employing a fixed global dilation rate, AdaDR dynamically adjusts dilation rates based on local frequency components, enhancing spatial adaptability. AdaKern dynamically adjusts the ratio between low-frequency and high-frequency components in convolution weights on a per-channel basis, capturing more high-frequency information and improving overall effective bandwidth. FreqSelect balances high- and low-frequency components through spatially variant reweighting, encouraging FADC to learn larger dilations and, consequently, expanding the receptive field. In the future, we aim to extend our quantitative frequency analysis to deformable/dilated attention. Additionally, since FADC are demonstrated to be seamlessly replace standard convolution layers in the existing architectures, we are going to design specific architecture for FADC.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62331006, 62171038, and 62088101), the R&D Program of Beijing Municipal Education Commission (KZ202211417048), the Fundamental Research Funds for the Central Universities, and JST Moonshot R&D Grant Number JPMJMS2011, Japan.

References

[1] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4594–4603, 2020. 3

[2] Gregory A Baxes. *Digital image processing: principles and applications*. John Wiley & Sons, Inc., 1994. 3

[3] Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. Regionvit: Regional-to-local attention for vision transformers. In *Proceedings of International Conference on Learning Representations*, pages 1–15, 2021. 7

[4] Linwei Chen, Zheng Fang, and Ying Fu. Consistency-aware map generation at multiple zoom levels using aerial image. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:5953–5966, 2022. 5

[5] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 5

[6] Linwei Chen, Ying Fu, Shaodi You, and Hongzhe Liu. Efficient hybrid supervision for instance segmentation in aerial images. *Remote Sensing*, 13(2):252, 2021.

[7] Linwei Chen, Ying Fu, Shaodi You, and Hongzhe Liu. Hybrid supervised instance segmentation by learning label noise suppression. *Neurocomputing*, 496:131–146, 2022. 5

[8] Linwei Chen, Lin Gu, and Ying Fu. When semantic segmentation meets frequency aliasing. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[9] 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. 1, 6

[10] 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. 5, 6

[11] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pages 4479–4488, 2020. 3

[12] 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, 8

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision*, pages 764–773, 2017. 3, 7

[14] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 11963–11975, 2022. 8

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations*, pages 1–12, 2020. 3, 8

[16] 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

[17] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020. 6

[18] Ying Fu, Zheng Fang, Linwei Chen, Tao Song, and Defu Lin. Level-aware consistent multilevel map translation from satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2022. 5

[19] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. In *Proceedings of International Conference on Learning Representations*, pages 1–12, 2022. 3

[20] Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, and Jingdong Wang. On the connection between local attention and dynamic depth-wise convolution. In *International Conference on Learning Representations*, pages 1–14, 2021. 3

[21] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. 2022. 5, 6, 7

[22] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 6185–6194, 2023. 3, 6, 7

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 5, 7

[24] 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. 6, 7

[25] 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

[26] Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 1–15, 2021. 5

[27] Md Tahmid Hossain, Shyh Wei Teng, Guojun Lu, Mohammad Arifur Rahman, and Ferdous Sohel. Anti-aliasing deep image classifiers using novel depth adaptive blurring and activation function. *Neurocomputing*, 536:164–174, 2023. 3

[28] Zhipeng Huang, Zhizheng Zhang, Cuiling Lan, Zheng-Jun Zha, Yan Lu, and Baining Guo. Adaptive frequency filters as efficient global token mixers. In *Proceedings of IEEE*

International Conference on Computer Vision, pages 1–11, 2023. 3

[29] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Proceedings of Advances in Neural Information Processing Systems*, 34:852–863, 2021. 3

[30] Ismail Khalfaoui-Hassani, Thomas Pellegrini, and Timothée Masquelier. Dilated convolution with learnable spacings. In *Proceedings of International Conference on Learning Representations*, pages 1–13, 2023. 3, 6

[31] 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

[32] Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. Wavecnet: Wavelet integrated cnns to suppress aliasing effect for noise-robust image classification. *IEEE Transaction on Image Process.*, 30:7074–7089, 2021. 3

[33] 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

[34] 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

[35] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *Proceedings of International Conference on Learning Representations*, pages 1–12, 2021. 3

[36] Shiqi Lin, Zhizheng Zhang, Zhipeng Huang, Yan Lu, Cuiling Lan, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, Amey Parulkar, et al. Deep frequency filtering for domain generalization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 11797–11807, 2023. 3

[37] 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. 5

[38] 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 *Proceedings of European Conference on Computer Vision*, pages 740–755, 2014. 6, 7

[39] Songlin Liu, Linwei Chen, Li Zhang, Jun Hu, and Ying Fu. A large-scale climate-aware satellite image dataset for domain adaptive land-cover semantic segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205:98–114, 2023. 5

[40] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Tommi Kärkkäinen, Mykola Pechenizkiy, Decebal Constantin Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. In *Proceedings of International Conference on Learning Representations*, 2023. 7

[41] Sun-Ao Liu, Hongtao Xie, Hai Xu, Yongdong Zhang, and Qi Tian. Partial class activation attention for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 16836–16845, 2022. 2

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of IEEE International Conference on Computer Vision*, pages 10012–10022, 2021. 3, 6, 7

[43] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 6, 7

[44] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE International Conference on Computer Vision*, pages 3431–3440, 2015. 5

[45] Jinlai Ning and Michael Spratling. The importance of anti-aliasing in tiny object detection. In *Asian Conference on Machine Learning*, 2023. 3

[46] 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

[47] 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

[48] Namuk Park and Songkuk Kim. How do vision transformers work? In *Proceedings of International Conference on Learning Representations*, pages 1–14, 2021. 2, 3, 5

[49] 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

[50] Ioannis Pitas. *Digital image processing algorithms and applications*. John Wiley & Sons, 2000. 3

[51] John G Proakis and Dimitris G Manolakis. *Digital signal processing: Pearson new international edition*. Pearson Higher Ed, 2013. 1, 4

[52] Shengju Qian, Hao Shao, Yi Zhu, Mu Li, and Jiaya Jia. Blending anti-aliasing into vision transformer. *Proceedings of Advances in Neural Information Processing Systems*, 34:5416–5429, 2021. 3

[53] 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. 5, 6

[54] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Proceedings of Advances in Neural Information Processing*

Systems, volume 34, pages 980–993, 2021. 3

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 91–99, 2015. 1, 7

[56] Richard A Roberts and Clifford T Mullis. *Digital signal processing*. Addison-Wesley Longman Publishing Co., Inc., 1987. 1, 4

[57] David W Romero, Robert-Jan Bruintjes, Jakub M Tomczak, Erik J Bekkers, Mark Hoogendoorn, and Jan C van Gemert. Flexconv: Continuous kernel convolutions with differentiable kernel sizes. In *Proceedings of International Conference on Learning Representations*, pages 1–14, 2021. 3, 8

[58] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *IEEE international conference on machine learning and applications*, pages 624–628. IEEE, 2018. 6

[59] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019. 3

[60] Ajay Subramanian, Elena Sizikova, Najib J Majaj, and Denis G Pelli. Spatial-frequency channels, shape bias, and adversarial robustness. pages 1–10, 2023. 5

[61] Naoya Takahashi and Yuki Mitsufuji. Densely connected multi-dilated convolutional networks for dense prediction tasks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 993–1002, 2021. 3, 8

[62] Ling Tang, Wen Shen, Zhanpeng Zhou, Yuefeng Chen, and Quanshi Zhang. Defects of convolutional decoder networks in frequency representation. *arXiv preprint arXiv:2210.09020*, 2022. 5

[63] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391, 2003. 7

[64] Cristina Vasconcelos, Hugo Larochelle, Vincent Dumoulin, Rob Romijnders, Nicolas Le Roux, and Ross Goroshin. Impact of aliasing on generalization in deep convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision*, pages 10529–10538, 2021. 3

[65] Thomas Verelst and Tinne Tuytelaars. Dynamic convolutions: Exploiting spatial sparsity for faster inference. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2320–2329, 2020. 3

[66] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 3

[67] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, pages 1451–1460. Ieee, 2018. 1, 3, 4, 8

[68] 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. 3, 5, 6, 7

[69] 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, 4

[70] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 6, 7

[71] 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. 6

[72] 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. 5, 6

[73] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Proceedings of Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. 6

[74] Jie Yao, Dongdong Wang, Hao Hu, Weiwei Xing, and Liqiang Wang. Adcnn: Towards learning adaptive dilation for convolutional neural networks. *Pattern Recognition*, 123:108369, 2022. 3, 5, 6

[75] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Proceedings of Advances in Neural Information Processing Systems*, volume 32, 2019. 3

[76] 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

[77] 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

[78] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 472–480, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[79] Guhnoo Yun, Juhan Yoo, Kijung Kim, Jeongho Lee, and Dong Hwan Kim. Spanet: Frequency-balancing token mixer using spectral pooling aggregation modulation. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1–16, 2023. 3

[80] Richard Zhang. Making convolutional networks shift-invariant again. In *Proceedings of International Conference on Machine Learning*, pages 7324–7334, 2019. 3

[81] 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. 6

[82] 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

[83] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled dynamic filter networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 6647–6656, 2021. 3

[84] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 3, 4, 5, 6, 7

[85] Xueyan Zou, Fanyi Xiao, Zhiding Yu, and Yong Jae Lee. Delving deeper into anti-aliasing in convnets. In *Proceedings of the British Machine Vision Conference*, pages 1–13, 2020. 3