WHEN SEMANTIC SEGMENTATION MEETS FREQUENCY ALIASING (SUPPLEMENT MATERIAL)

Linwei Chen[1], **Lin Gu**[2,3] **& Ying Fu**[1] *
[1]Beijing Institute of Technology, Beijing, China
[2]RIKEN AIP, Tokyo, Japan
[3]The University of Tokyo, Tokyo, Japan
chenlinwei@bit.edu.cn lin.gu@riken.jp 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 provides more analysis of aliasing degradation.
- Section B compares the aliasing score with confidence-based hard pixel type identification.
- Section C discusses the equivalent sampling rate when the kernel size and stride differ in height and width dimensions.
- Section D provides more implementation details of training.
- Section E includes an additional ablation study for the Frequency Mixing (FreqMix) module.
- Section F provides visualization of the frequency response of existing blur filters and demonstrates the advantage of the proposed de-aliasing filter.
- Section G provides additional experimental results combined with segmentation boundary refinement methods.
- Section H provides an analysis of the orthogonality of downsampling filters.
- Section I provides a detailed visual analysis, illustrating how aliasing degrades features and leads to three types of errors, and how DAF and FreqMix effectively address aliasing degradation.
- Section K analyzes the feature map in the frequency domain.
- Section L discusses the aliasing for the transformer-based architecture.

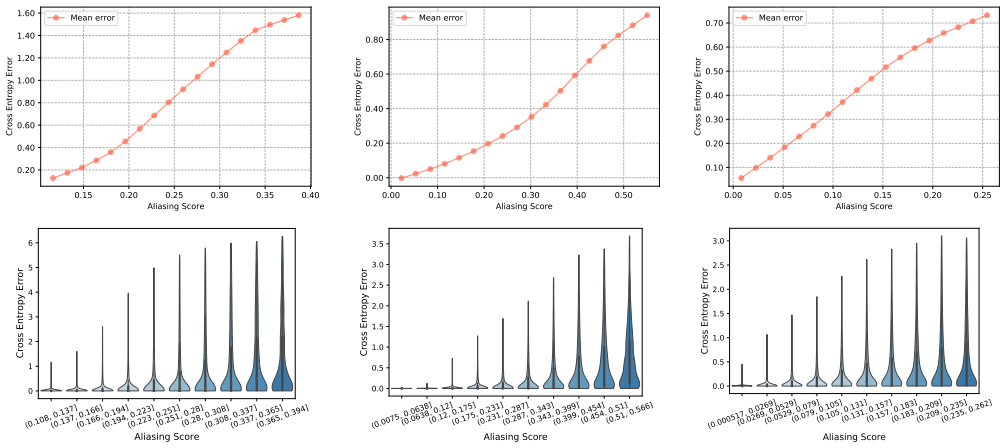
A ALIASING DEGRADATION

We have chosen ResNet (He et al., 2016), Swin Transformer (Liu et al., 2021), and ConvNeXt (Liu et al., 2022) to perform a quantitative analysis of the correlation between aliasing scores and errors. Our analysis has been concentrated on the results obtained at object boundaries, where the majority of challenging pixels are found, as mentioned in (Li et al., 2017; Gu et al., 2020). ResNet is a well-established and widely used backbone, whereas Swin Transformer and ConvNeXt represent the latest transformer-based and CNN-based backbones, respectively. Despite the differences in their model structures, our findings, illustrated in Figure 1, reveal a consistent pattern: boundary pixels with higher aliasing scores tend to demonstrate higher cross-entropy errors, indicating that they are more prone to misclassification. These results highlight the pervasive issue of aliasing-induced degradation within modern deep neural networks. This observation underscores the need for comprehensive solutions to mitigate and address this problem, especially in the context of computer vision and other applications where accurate pixel-level information is crucial.

B COMPARING WITH THE PREDICTION CONFIDENCE METHOD

Previous research efforts (Li et al., 2017; Gu et al., 2020) have identified hard pixels based on prediction confidence. Here, we compare the prediction confidence method with the aliasing-based

*Corresponding Author



(a) ResNet-50 (b) Swin-T (c) ConvNeXt-T

Figure 1: Illustration depicting the relationship between aliasing scores and hard pixels at boundaries. These results demonstrate the degradation caused by aliasing affecting three different widely-used models.

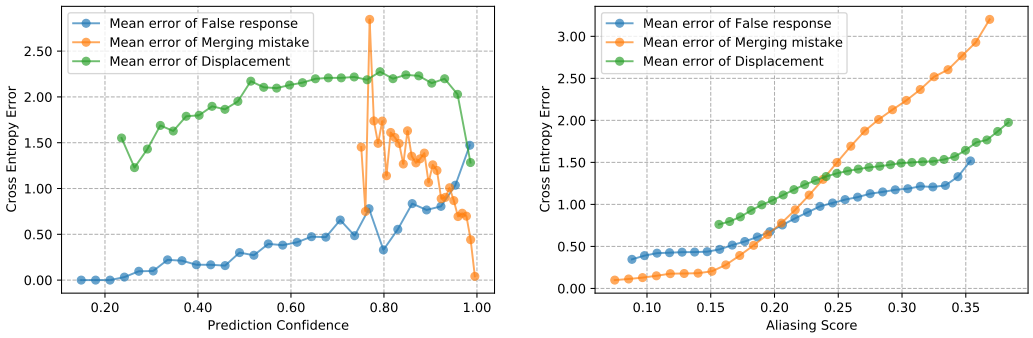


Figure 2: Left: Statistically correlated curve of cross-entropy error for three types of hard pixels with respect to prediction confidence (Li et al., 2017). Right: Distribution of the three types of hard pixels.

method for hard pixel identification. As illustrated in Figure 2, aliasing scores exhibit a more consistent trend than confidence-based results, especially when considering the effectiveness of distinguishing three types of errors: displacement errors, false responses, and merging mistakes. This finding underscores the potential of aliasing for categorizing errors effectively.

Notably, these errors display distinct characteristics when analyzed in the context of aliasing, and their importance varies across different scenarios. For instance, displacement errors tend to cluster in regions with high aliasing scores. In critical applications such as robotic surgery or radiation therapy, even a slight displacement of just two or three pixels from vital organs like the brainstem or the main artery can result in catastrophic consequences. Conversely, false responses and merging mistakes, which carry greater significance in autonomous driving scenarios, are more commonly found in areas with relatively low aliasing scores.

This analysis not only provides insights into distinguishing between the three error types but also offers valuable guidance for designing error-correcting methods tailored to specific scenarios.

C EQUIVALENT SAMPLING RATE

In the main paper, we provide the calculation of the equivalent sampling rate when the downsampling kernel and stride are the same in the height and width dimensions, which is the most common situation in existing modern deep neural networks (He et al., 2016; Liu et al., 2021; 2022; Wang et al., 2023). Here, we discuss the equivalent sampling rate when the kernel size and stride in the height and width dimensions are different.

Table 1: Ablation study for Frequency Mixing (FreqMix) module.

Method	mIoU↑	BIoU↑	BAcc↑	FErr↓	MErr↓	DErr↓	#FLOPs	#Params
FreqMix w/ decomposition	79.6	58.6	74.7	**23.5**	52.9	26.3	423.69G	39.64M
FreqMix w/o decomposition	**79.7**	**58.8**	**74.9**	24.0	**52.8**	**26.1**	**298.67G**	**31.54M**

Similarly, we consider both the kernel size and feature size (channel, height, and width), rather than just the downsampling stride. We introduce a simple equation for the equivalent sampling rate in the height and width dimensions (ESR^H, ESR^W) to calculate the actual sampling rate as follows:

$$\begin{aligned} \text{ESR}^H &= \min(K^H_{\text{down}}, \sqrt{\frac{C^{\text{out}}}{C^{\text{in}}}}) \times \frac{H^{\text{out}}}{H^{\text{in}}}, \\ \text{ESR}^W &= \min(K^W_{\text{down}}, \sqrt{\frac{C^{\text{out}}}{C^{\text{in}}}}) \times \frac{W^{\text{out}}}{W^{\text{in}}}, \end{aligned} \tag{1}$$

where C, H, and W are the size of the channel, height, and width. "in" and "out" indicate the input and output features. $K^H_{\text{down}}, K^W_{\text{down}}$ present the downsampling kernel size in height and width dimensions. $\frac{H^{\text{out}}}{H^{\text{in}}}$, $\frac{W^{\text{out}}}{W^{\text{in}}}$ are equal to downsampling stride in height and width dimension, which aligns with (Grabinski et al., 2022). $\min(K^H_{\text{down}}, \sqrt{\frac{C^{\text{out}}}{C^{\text{in}}}})$, $\min(K^W_{\text{down}}, \sqrt{\frac{C^{\text{out}}}{C^{\text{in}}}})$ indicates the influence of the downsampling kernel size K^{down} and channel expansion. Notice that we make the common assumption that the impact of channel expansion works for both height and width dimensions, using the square root to calculate the impact for both dimensions.

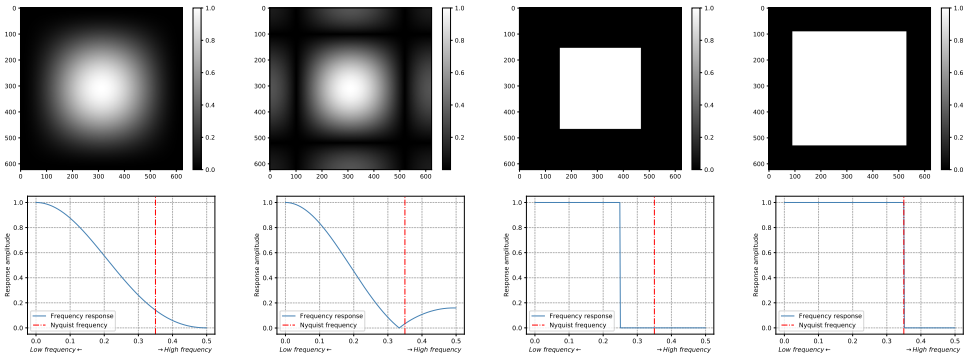
D More Implement Details

For the Cityscapes dataset, we employ a crop size of 768×768, a batch size of 8, and a total of 80K iterations. For the PASCAL VOC dataset, we utilize a crop size of 512×512, a batch size of 16, and a total of 40K iterations. On the ADE20K dataset, we adopt a crop size of 512×512, a batch size of 16, and a total of 80K iterations. We also set the channel of the feature pyramid to 128. We employ stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 5e-4. The initial learning rate is set at 0.01. During training, we adjust the learning rate using the common 'poly' learning rate policy, which reduces the initial learning rate by multiplying $(1 - \frac{\text{iter}}{\text{max\_iter}})^{0.9}$. We apply standard data augmentation techniques, including random horizontal flipping and random resizing within the range of 0.5 to 2.

For PointRend (Kirillov et al., 2020) and Mask2Former (Cheng et al., 2022) on the LIS dataset, we adopt the same training settings as described in the paper (Chen et al., 2023). We use random flip as data augmentation and train with a batch size of 8, a learning rate of 1e-2 for 12 epochs, with a learning rate dropping by 10× at 8 and 11 epochs, respectively. To make the model quickly adapt to low-light settings, we use COCO pre-trained model as initialization following (Chen et al., 2023).

E Ablation Study

In this section, we present an additional ablation study focused on the Frequency Mixing (FreqMix) module. In the main paper, we describe our approach to predicting weighting values. Specifically, we decompose the prediction of three-dimensional frequency weighting values $A^{\downarrow}$, $A^{\uparrow} \in \mathbb{R}^{C\times H\times W}$, into channel-wise components $A^{\downarrow}_1$, $A^{\uparrow}_2 \in \mathbb{R}^{H\times W}$ and spatial-wise components $A^{\downarrow}_1$, $A^{\uparrow}_2 \in \mathbb{R}^{H\times W}$. Where $A^{\downarrow}$ represents the weighting values for frequencies below the Nyquist frequency, and $A^{\uparrow}$ represents those above Nyquist frequency. As shown in Table 1, the prediction of three-dimensional frequency weighting values can be computationally intensive for prediction. It directly introduces an additional 7.2 million parameters and requires an extra 125.1 GFLOPs of computational cost. Interestingly, despite this increase in complexity, the overall results are largely similar to the decomposition solution. This highlights the efficiency of our proposed method.



(a) Blur (b) AdaBlur (c) FLC (d) Ours

Figure 3: Visualization of the frequency response of existing blur filters, including Blur (Zhang, 2019), AdaBlur (Zou et al., 2020), FLC (Grabinski et al., 2022), and our proposed methods. The top row shows the frequency response in two dimensions. We shift the low frequency to the center, and the four corners indicate high frequency, with brighter areas representing higher response. The bottom row displays the frequency response in one dimension, where the left side represents lower frequency, and the right side represents higher frequency. The red line indicates the Nyquist frequency.

Table 2: Combination with boundary refinement methods on the Cityscapes (Cordts et al., 2016) validation set. Results are reported from the original paper (Yuan et al., 2020).

Method	DeepLabv3 [arXiv2017] (Chen et al., 2017)	+GUM [BMVC2018] (Mazzini, 2018)	+DenseCRF [NeurIPS] (Krähenbühl & Koltun, 2011)	+SegFix [ECCV2020] (Yuan et al., 2020)	+SegFix+Ours (Ours)
mIoU	79.5	79.8	79.7	80.5	**81.1**

F VISUALIZATION OF VARIOUS BLUR FILTERS

For better comparison, we also illustrate the visualization of the frequency response of existing blur filters, including Blur (Zhang, 2019), AdaBlur (Zou et al., 2020), FLC (Grabinski et al., 2022), and our proposed methods in Figure 3. The top row shows the frequency response in two dimensions. We shift the low frequency to the center, and the four corners indicate high frequency, with brighter areas representing higher response. The bottom row displays the frequency response in one dimension, where the left side represents lower frequency, and the right side represents higher frequency. The red line indicates the Nyquist frequency. We observe that Blur (Zhang, 2019) and AdaBlur (Zou et al., 2020) cannot entirely eliminate frequencies higher than the Nyquist frequency. Conversely, due to an underestimation of the Nyquist frequency, FLC (Grabinski et al., 2022) excessively removes frequencies below the Nyquist frequency, resulting in information loss. In contrast, our proposed de-aliasing filter effectively and precisely removes the frequency power above the Nyquist frequency, which explains its effectiveness.

G COMBINATION WITH BOUNDARY REFINEMENT METHODS

In this section, we integrate our proposed method with SegFix (Yuan et al., 2020), a previously effective approach for semantic segmentation boundary refinement. SegFix significantly improves segmentation results by refining predictions, particularly at boundaries where most hard pixels occur. Our proposed method operates independently from SegFix, enhancing models by optimizing intermediate features, precisely removing frequencies leading to aliasing (via the de-aliasing filter), and adjusting frequencies using the encoder block (Frequency Mixing Module).

The synergy between these techniques is evident in the results presented in Table 2, where our method enhances SegFix by an additional 0.6 mIoU. This improvement highlights the effectiveness of our approach in addressing complex segmentation challenges.

Table 3: Orthogonality analysis for downsampling filters in ResNet (He et al., 2016), Swin Transformer (Liu et al., 2021), ConvNeXt (Liu et al., 2022), HorNet (Rao et al., 2022), and DiNAT (Hassani & Shi, 2022). Their weights are obtained by training on ImageNet. A higher absolute cosine similarity value indicates greater similarity in filter weights, suggesting a lower degree of orthogonality (0.0 = totally orthogonal, 1.0 = identical filters).

Model	ResNet-18	ConvNeXt-T	Swin-T	HorNet-T	DiNAT-L
Abs. CosSim.	0.067	0.072	0.06	0.069	0.046

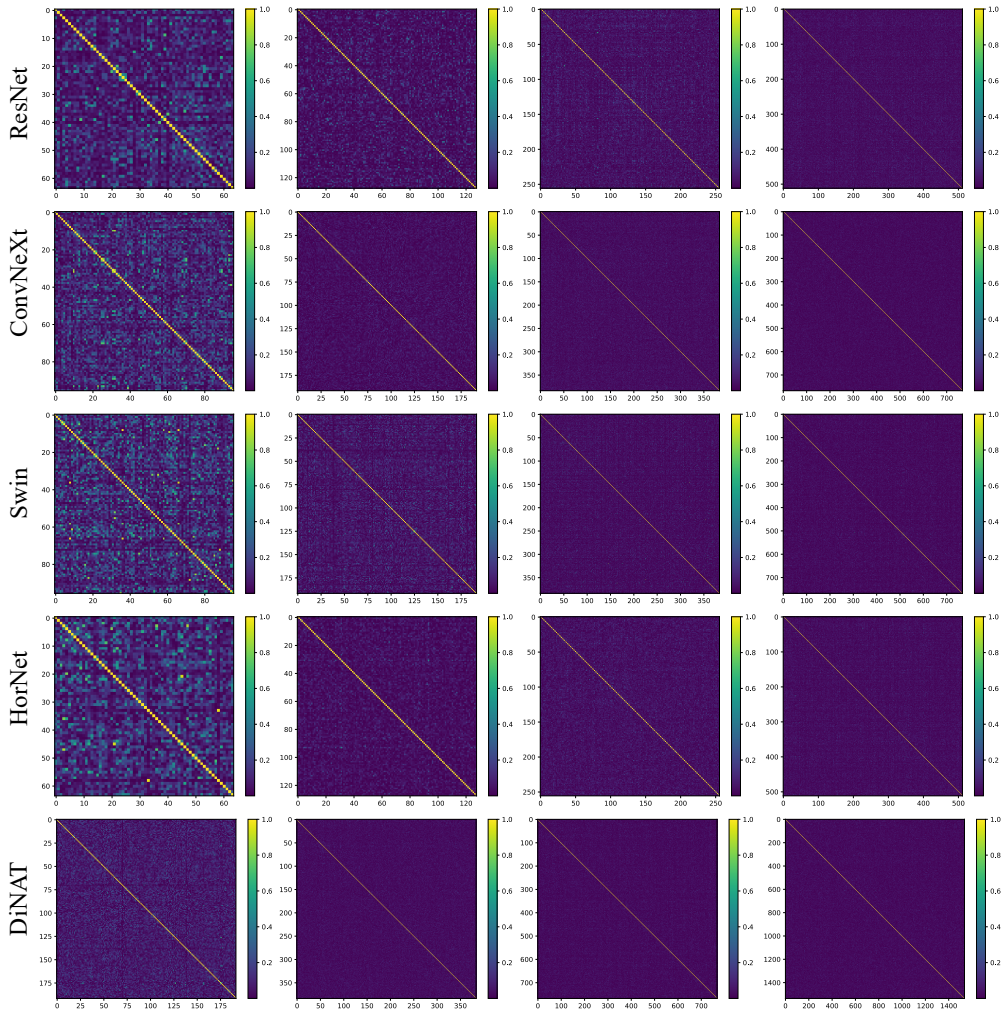


Figure 4: Orthogonality degree analysis of filters for downsampling in ResNet (He et al., 2016), Swin Transformer (Liu et al., 2021), ConvNeXtcitep2022convnet, HorNet(Rao et al., 2022), and DiNAT (Hassani & Shi, 2022). We illustrate the absolute cosine similarity matrix, where each element indicates the absolute cosine similarity between different filters. A brighter color indicates a higher similarity in filter weights, suggesting a lower degree of orthogonality. We observe the matrix showing dark colors, with bright colors only appearing along the diagonal (self-to-self), indicating that the filters are essentially orthogonal.

H Downsampling Filters Orthogonality Analysis

The introduced calculation of the equivalent sampling rate is based on the assumption that downsampling filters are orthogonal. In this section, we quantitatively analyze the orthogonality degree of downsampling filters in widely used models, including ResNet (He et al., 2016), Swin Transformer (Liu et al., 2021), ConvNeXt (Liu et al., 2022), HorNet (Rao et al., 2022), and DiNAT (Hassani & Shi, 2022). Their weights are obtained by training on ImageNet.

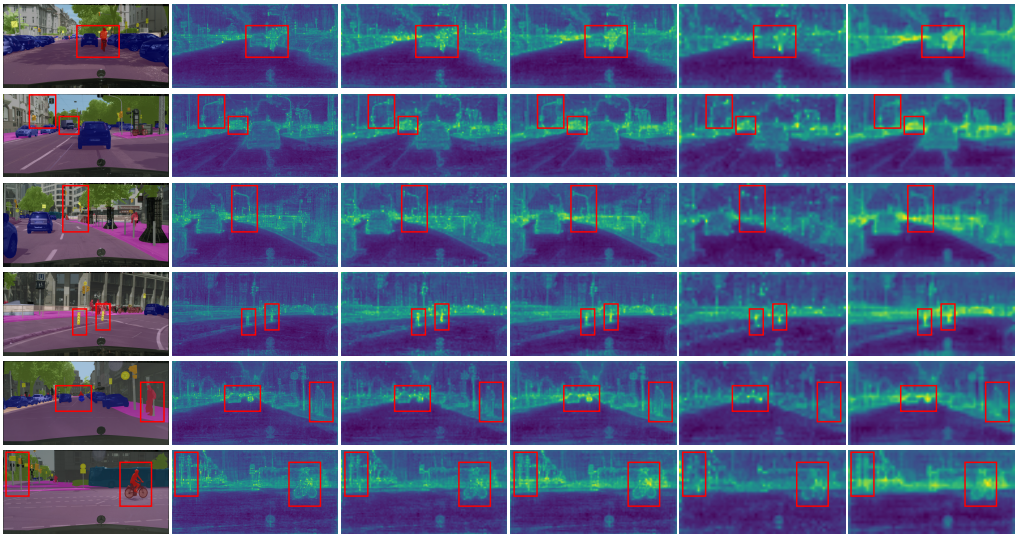


Figure 5: Visualization for DAF. We mark some high aliasing score areas in the feature with a red box. Without DAF in (c) and (e), the downsampling of features exhibits a severe "jagged" phenomenon (Zou et al., 2020; Qian et al., 2021), resulting in the degraded representation of object boundaries. The response of some objects is faded or lost in the 4× scale in (e). By directly removing the high frequency leads to aliasing degradation, the proposed DAF can largely relieve the "jagged" phenomenon (Zou et al., 2020; Qian et al., 2021), making the boundaries more clear in the (d) and (f). Furthermore, DAF largely preserves the object responses, as shown in (f), compared to (e).

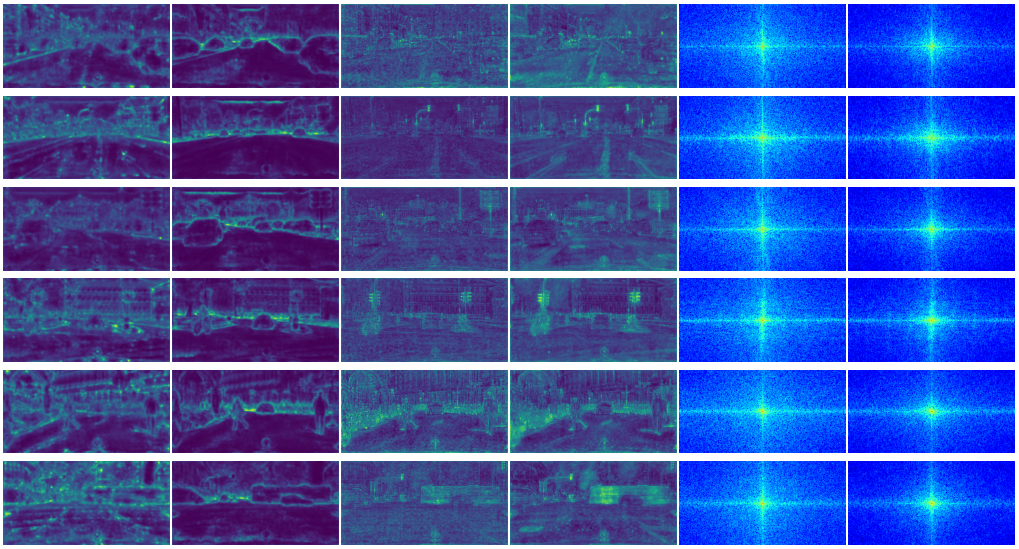


Figure 6: Visualization for FreqMix. In (a) and (b), the heatmap shows a brighter color for object boundaries and a darker color for object centers and backgrounds, especially for high-frequency components, where a brighter color indicates a high value. Thus, FreqMix not only reduces the overall high-frequency content responsible for aliasing degradation in (f) but also preserves the high frequency of object boundaries. This preservation is crucial for making the boundaries clear in (d), ensuring accurate segmentation, and lowering the occurrence of three types of errors.

As shown in Table 3 and Figure 4, we use the absolute cosine similarity as the quantitative measurement of the orthogonality degree. A higher absolute cosine similarity value indicates greater

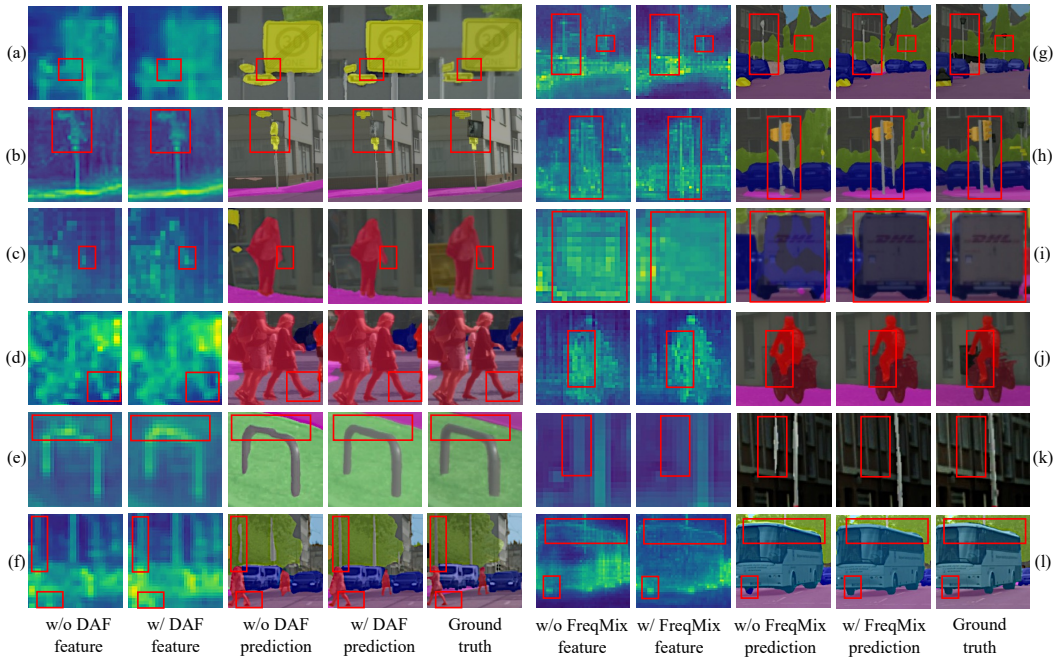


Figure 7: Visualization of how the proposed DAF and FreqMix address aliasing degradation and improve the feature and final segmentation. We randomly select image patches with a high aliasing score from Cityscapes (Cordts et al., 2016) dataset validation set. Zoom in for better view.

similarity in filter weights, suggesting a lower degree of orthogonality. As depicted in Table 3 and Figure 4, the quantitative measurements indicate that the downsampling filters are predominantly orthogonal (with an average absolute cosine similarity ranging from 0.046 to 0.072), thereby supporting the introduced equivalent sampling rate in Section C. The proposed equivalent sampling rate is designed as a heuristic for selecting the cutoff frequency, and we think that exploring how to finely adjust the equivalent sampling rate based on the orthogonality of the filter is a very interesting and important problem that is worth further investigation.

I Visualized Analysis for DAF and FreqMix

To investigate how the proposed DAF and FreqMix address aliasing degradation and enhance deep neural networks, we visualize the deep features in the model and randomly select some examples in Figures 5 and 6. Furthermore, in Figures 7, we visualize three types of errors and demonstrate how the proposed DAF and FreqMix alleviate these errors: 1) false responses, 2) merging mistakes, and 3) displacements.

I.1 Visualization for De-aliasing filter (DAF)

As depicted in Figures 5(c) and (e), we indicate some areas with a high aliasing score in the feature with the red box, they exhibit a severe "jagged" phenomenon (Zou et al., 2020; Qian et al., 2021), leading to the degraded representation of object boundaries. Moreover, in comparison with the original feature in Figure 5(b), the response of some objects is lost in the $4\times$ downsampling in Figures 5(e). It is noteworthy that widely used state-of-the-art models, such as Swin Transformer (Liu et al., 2021) and ConvNeXt (Liu et al., 2022), adopt a total downsampling stride of $32\times$, potentially leading to an even more severe loss of response.

By directly removing the high frequency leads to aliasing degradation, the proposed DAF can largely relieve the "jagged" phenomenon (Zou et al., 2020; Qian et al., 2021), making the boundaries more clear in the Figures 5(d) and (f). Furthermore, DAF largely preserves the object responses, as shown in Figure 5(f), compared to Figure 5(e), resulting in a more accurate segmentation prediction and a lower occurrence of the three types of errors shown in the left column Figure 7.

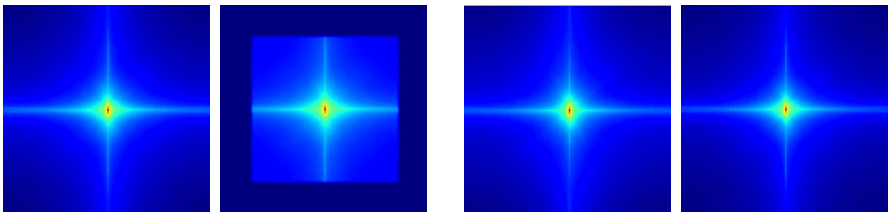


Figure 8: Visualization of the averaged frequency distribution of features. The center indicates low frequency, and the corners indicate high frequency. A brighter color indicates more corresponding frequency components. The DAF directly removes the frequency above the Nyquist frequency, while FreqMix suppresses the high frequency.

I.2 Visualization for Frequency Mixing Module (FreqMix)

FreqMix improves the model by decomposing features into low frequency and high frequency using a Nyquist frequency threshold and dynamically selecting them in a spatial-variant manner. We visualize the heatmap for selecting low frequency and high frequency in Figures 6(a) and (b), where a brighter color indicates a high value. The heatmap shows a brighter color for object boundaries and a darker color for object centers and backgrounds, especially for high-frequency components. Thus, FreqMix not only reduces the overall high-frequency content (see Figure 6(e) and (f)), responsible for aliasing degradation, but also preserves the high frequency of object boundaries. This preservation is crucial for making the boundaries clear (see Figure 6(c) and (d)), ensuring accurate segmentation and a lower occurrence of three types of errors. Further visualization in the right column Figure 7 verifies that FreqMix reduces the occurrence of the three types of errors.

I.3 Visualization of Three Types of Errors

As illustrated in Figure 7, we randomly selected image patches with a high aliasing score from the Cityscapes (Cordts et al., 2016) dataset validation set.

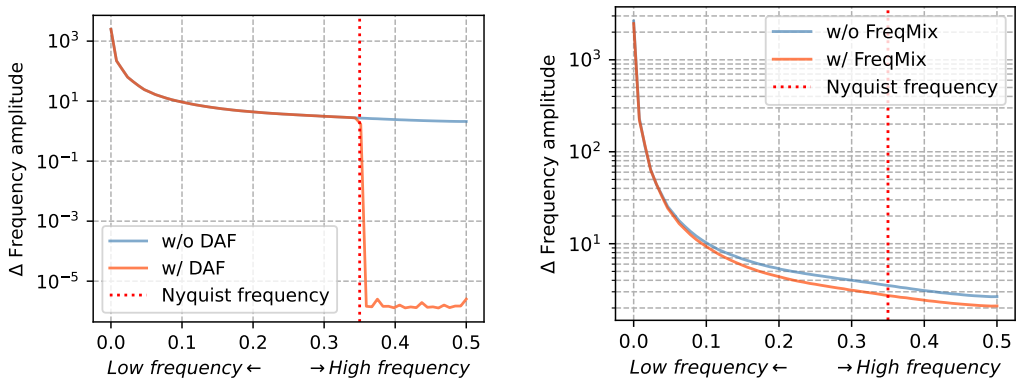
We observed that aliasing leads to a "jagged" phenomenon (Zou et al., 2020; Qian et al., 2021), disrupting object shapes and boundaries, resulting in false responses (Figures 7(b) and (l)) and displacement errors (Figures 7(c), (d), and (e)). DAF and FreqMix relieve this phenomenon and improve the feature representation thus resulting in lower false responses and displacement errors.

When high-frequency information is aliased, it transforms into false low-frequency information. For example, when two objects are close to each other, their high-frequency boundaries can be aliased to lower frequency during downsampling, causing the two objects to appear connected and their boundaries to be merged. This leads to merging errors (Figures 7(a), (g), and (j)). DAF/FreqMix solve this by removing/suppressing these high frequencies during downsampling/encoder block, leading to lower merging errors.

Moreover, high-frequency components in the object center or background can result in false responses (Figures 7(i) and (k)). FreqMix addresses this issue by suppressing the high frequency in the object center or background while preserving the high frequency at the boundaries.

J Feature Frequency Analysis

We present a feature frequency analysis in Figures 8 and 9. The DAF directly eliminates frequencies above the Nyquist frequency, while the FreqMix suppresses high-frequency components, alleviating aliasing degradation.



(a) DAF

(b) FreqMix

Figure 9: Visualization of the frequency distribution of features.

K Feature Frequency Analysis

We present a feature frequency analysis in Figures 8 and 9. The DAF directly eliminates frequencies above the Nyquist frequency, while the FreqMix suppresses high-frequency components, alleviating aliasing degradation.

L Discussion about aliasing in a transformer-based architecture

As for recent transformer-based architectures, aliasing remains a concern. Taking the renowned Vision Transformer (ViT) as an example, ViT (Dosovitskiy et al., 2020) tokenizes images by splitting them into non-overlapping patches, which are then fed into transformer blocks. The tokenization and self-attention operations performed on these discontinuous patch embeddings can be viewed as downsampling operations, introducing a potential side effect of aliasing. It is essential to note that this downsampling operation is virtually unavoidable due to the spatial redundancy nature of the image (He et al., 2022) and huge computational costs without downsampling (increasing by 256× without downsampling in ViT).

Several existing studies have acknowledged this concern. A straightforward solution to alleviate aliasing is to increase the sampling rate. Similar trends are observed in vision transformers, where the use of overlapped tokens (Yuan et al., 2021) and smaller patch sizes (Caron et al., 2021) contributes to improved performance. However, escalating sampling rates incur quadratic computational costs. Consequently, I hypothesize that integrating appropriate anti-aliasing filters into the 'attending' process could offer a viable solution. In fact, existing work has empirically explored blending anti-aliasing filters into the vision transformer, reporting observed improvements Qian et al. (2021).

In conclusion, aliasing persists as a potential concern in transformer-based architectures, and prior studies have endeavored to address it empirically by enhancing the sampling rate or integrating anti-aliasing filters into the attention mechanism. Our work represents a step forward in quantitatively assessing and addressing aliasing in contemporary models, supported by theoretical foundations. This issue still presents ample opportunities for further investigation and exploration.

References

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of IEEE International Conference on Computer Vision*, pp. 9650–9660, 2021.

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

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.

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*, pp. 1290–1299, 2022.

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*, pp. 3213–3223, 2016.

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*, pp. 1–12, 2020.

Julia Grabinski, Steffen Jung, Janis Keuper, and Margret Keuper. Frequencylowcut pooling-plug and play against catastrophic overfitting. In *Proceedings of European Conference on Computer Vision*, pp. 36–57, 2022.

Zhangxuan Gu, Li Niu, Haohua Zhao, and Liqing Zhang. Hard pixel mining for depth privileged semantic segmentation. *IEEE Transactions on Multimedia*, 23:3738–3751, 2020.

Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. 2022. URL https://arxiv.org/abs/2209.15001.

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*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9799–9808, 2020.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NeurIPS*, 24:1–14, 2011.

Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3193–3202, 2017.

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*, pp. 10012–10022, 2021.

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*, pp. 11976–11986, 2022.

Davide Mazzini. Guided upsampling network for real-time semantic segmentation. In *Proceedings of the British Machine Vision Conference*, pp. 1–12, 2018.

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.

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.

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*, pp. 14408–14419, 2023.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of IEEE International Conference on Computer Vision*, pp. 558–567, 2021.

Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 489–506. Springer, 2020.

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

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*, pp. 1–13, 2020.