國立台灣科技大學
資 訊 工 程 系

碩士學位論文

基於金字塔特徵分支的 SwinIR 圖像去模糊增強

Enhancing SwinIR for Image Deblurring via Pyramid Feature Branching

研 究 生： 呂承奕
學　　號： M11115062

指導教授 :吳怡樂

中 華 民 國 一 一 四 年 六 月 六 日

基於金字塔特徵分支的 SwinIR 圖像去模糊增強

國立台灣科技大學資訊工程系

學生：呂承奕

指導教授：吳怡樂

中華民國一一四年六月

論文摘要

影像去模糊是電腦視覺中一項基礎且關鍵的低階任務，目的是從模糊的輸入中還原出清晰影像，其模糊通常來自於相機晃動或物體運動。我們探討將原本應用於超解析度、影像去噪與壓縮雜訊還原任務的 SwinIR 架構，延伸至影像去模糊任務的潛力。雖然 SwinIR 並非爲去模糊任務所設計，我們仍嘗試以其作爲基線模型，並在第二個殘差 Swin Transformer 區塊中插入金字塔分支模組進行擴展。此層級化特徵整合設計能促進跨尺度資訊流動，藉此更有效捕捉模糊特徵。在 GoPro 與 HIDE 資料集上的實驗結果顯示，該方法相較原始 SwinIR 有約 +0.1 至 +0.2 dB PSNR 的小幅提升。然而，所提出的方法與目前最先進（SOTA）的影像去模糊模型相比，仍有約 2 dB PSNR 的差距。我們分析了 SwinIR 在去模糊任務中的侷限性，並探討這些發現的意涵。

Enhancing SwinIR for Image Deblurring via Pyramid Feature Branching

Department of Computer Science and Information Engineering

National Taiwan University of science and technology

Student：Cheng-Yi Lu

Advisor：Yi-Leh Wu

June, 2025

Abstract

Image deblurring is a fundamental low-level vision task aimed at recovering sharp images from blurred inputs, which often result from camera shake or object motion. We explore the potential of adapting the SwinIR architecture, originally proposed for superresolution, image denoising, and compression artifact reduction, to the task of image deblurring. Although SwinIR was not originally designed for image deblurring, we employed it as a baseline and extended it with a pyramid-style branch module inserted into the second residual Swin Transformer block. This hierarchical feature integration allows for cross-scale information flow, with the aim of better capturing blur patterns. Experimental results on the GoPro and HIDE datasets show a small improvement (around +0.1 to +0.2 dB PSNR) over the baseline SwinIR. However, the proposed method still lags behind state-of-the-art (SOTA) image deblurring models by approximately 2 dB PSNR. We analyze the limitations of SwinIR in the context of deblurring and discuss the implications of our findings.

Contents

論文摘要 i

Abstract in English ii

Contents iii

List of Figures v

List of Tables vi

1 Introduction 1

1.1 Research Background 1

1.2 Research Motivation 1

2 Related Work 3

2.1 Swin Transformer 3

2.2 SwinIR for Image Restoration 5

2.3 Pyramid Feature Integration 5

3 Proposed Method 7

3.1 Baseline Architecture 7

3.2 Proposed Architecture 8

4 Experiments 10

4.1 Dataset 10

4.1.1 GoPro 10

4.1.2 HIDE 10

4.1.3 RealBlur-R 11

4.2 Evaluation Metrics and Training Details 11

4.2.1 Evaluation Metrics 11

4.2.2 Training Details 12

4.3 Overall Performance Comparison 13
4.4 The Impact of Pyramid Branch 15
5 Conclusions and Future Work 17
References 18
Appendix:The overall performance comparison with other SOTA methods 19

List of Figures

2.1 Swin Transformer block with W-MSA and SW-MSA. . . . 4

2.2 W-MSA and SW-MSA in Swin Transformer 4

2.3 The architecture of a pyramid module 6

3.1 The architecture of SwinIR 7

3.2 The architecture of our proposed model 8

3.3 The architecture of Pyramid Branch we proposed 9

4.1 The result for SwinIR and PBSwinIR 13

4.2 Overall PSNR performance on SwinIR and PBSwinIR . . 14

4.3 Overall SSIM performance on SwinIR and PBSwinIR . . . 14

4.4 The impact of pyramid branch in different layers 16

A.1 Overall PSNR performance on SwinIR and PBSwinIR . . 19

A.2 Overall PSNR performance on SwinIR and PBSwinIR . . 20

List of Tables

4.1 Comparison of commonly used datasets for image deblurring. 11

4.2 Key hyperparameters used in our SwinIR-based deblurring model. 12

4.3 Overall performance(PSNR/SSIM) on GoPro and HIDE datasets. 13

4.4 PSNR on GoPro dataset when inserting Pyramid Branch into different SwinIR layers. 15

A.1 Overall performance(PSNR/SSIM) on GoPro and HIDE datasets. 19

Chapter 1 Introduction

1.1 Research Background

Image deblurring remains a challenging task in low-level vision due to the complex, spatially variant blur patterns present in real-world images. Although numerous CNN- and transformer-based architectures have been proposed for this task, most of them rely on encoder-decoder structures or multiscale designs to handle blur effectively.

1.2 Research Motivation

SwinIR [1], based on Swin Transformer [2], has shown strong performance in tasks such as image super-resolution, image denoising, and JPEG artifact reduction. Its window-based attention mechanism enables efficient computation on high-resolution images. However, its applicability to deblurring remains unexplored. In this work, we first used the original SwinIR as a baseline for single-image deblurring, and added a pyramid branch module to enhance multiscale feature extraction.

The main contributions are summarized as follows:

- We propose a PyramidBranch module to enhance multi-scale feature extraction. The design incorporates hierarchical feature processing via downsampling and upsampling paths, allowing the network to better handle spatially varying blur.

- We integrate the PyramidBranch into the SwinIR backbone with minimal architectural changes, maintaining the overall SwinIR pipeline while improving its ability to capture diverse contextual information.
- Experimental results on benchmark datasets (GoPro and HIDE) show that the proposed model achieves consistent improvement over the baseline SwinIR in terms of both PSNR and visual quality.

Chapter 2 Related Work

2.1 Swin Transformer

The Swin Transformer [2] is a hierarchical vision transformer architecture that enables linear computational complexity with respect to image size by computing self-attention within local windows. It also incorporates a shifted windowing mechanism to enhance cross-window connections. These properties make the Swin Transformer highly suitable for dense prediction tasks, including image restoration.

The core module of the Swin Transformer is the *Window-based Multi-head Self-Attention (W-MSA)*, which divides the image into non-overlapping windows and performs self-attention independently in each. The *Shifted Window Multi-head Self-Attention (SW-MSA)* is then used to connect information across windows in the next layer. This alternation between W-MSA and SW-MSA allows efficient and effective information propagation without full global attention. The architecture is shown in Figure 2.1 and Figure 2.2.

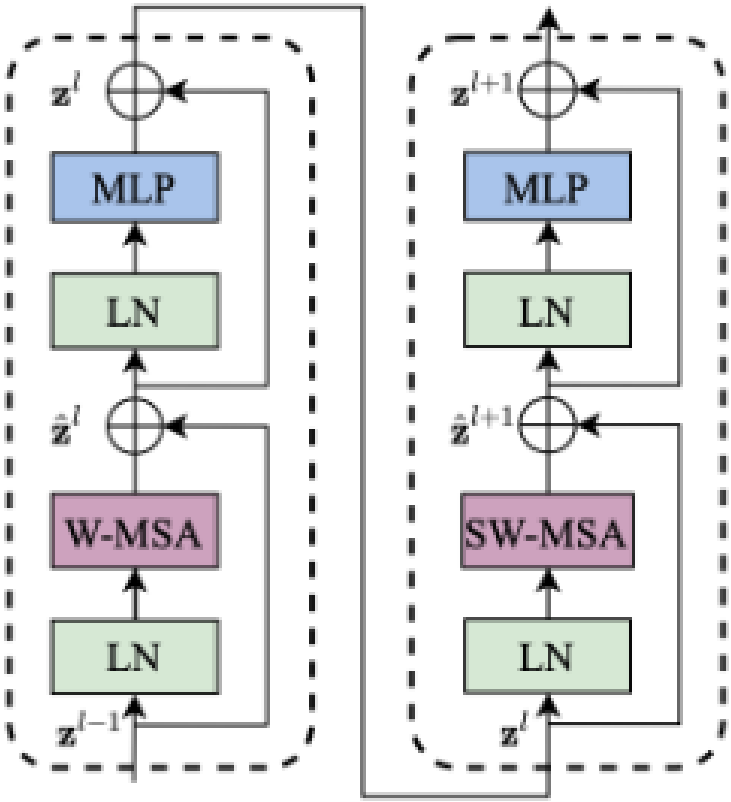

Figure 2.1: Swin Transformer block with W-MSA and SW-MSA.

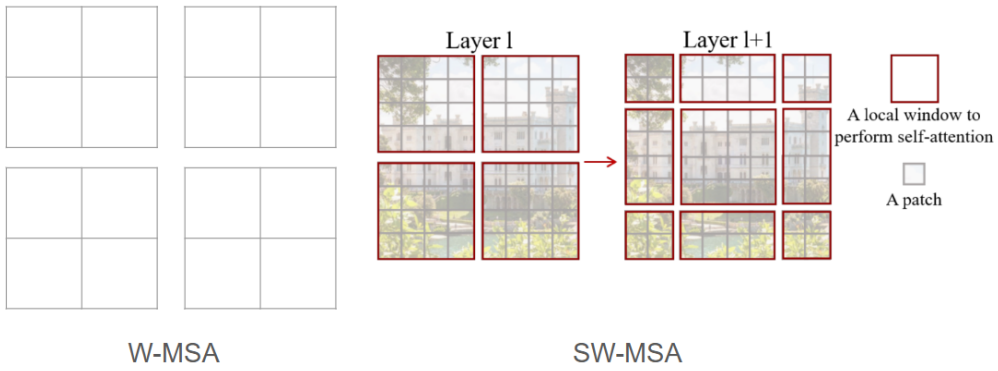

Figure 2.2: W-MSA and SW-MSA in Swin Transformer

The Swin Transformer forms the building blocks of various high-performance vision models, including SwinIR. Its ability to scale hierarchically and capture both local and global information plays a key role in learning complex

patterns such as motion blur and texture degradation.

2.2 SwinIR for Image Restoration

SwinIR (Swin Transformer for Image Restoration) [1] is a Transformer-based architecture specifically designed for low-level vision tasks such as super-resolution, denoising, and deblurring. It consists of three main stages: shallow feature extraction, deep feature extraction, and image reconstruction.

In the deep feature extraction stage, SwinIR stacks multiple *Residual Swin Transformer Blocks (RSTBs)* to hierarchically process feature maps. Each RSTB is composed of Swin Transformer layers followed by convolutional fusion and residual connection, enabling both non-linear representation learning and feature reuse.

SwinIR has demonstrated competitive results on various image restoration benchmarks. However, the original architecture processes features at a single scale within each RSTB, which limits its ability to model scale-variant degradation, such as spatially varying blur commonly found in real-world deblurring tasks.

2.3 Pyramid Feature Integration

Pyramid feature architectures have been widely used in tasks involving scale variance, such as deraining [3], defocus deblurring [4], and motion

deblurring [5,6]. These networks typically downsample features to aggregate global context, apply processing (e.g., attention or convolutions), and upsample for fusion. The architecture is shown in Figure 2.3. Motivated by their success, we integrate a lightweight pyramid-style module into the SwinIR backbone to enhance multiscale representation for deblurring tasks.

In particular, multiscale transformers such as MIMO-UNet [7] and FPN [8] have incorporated cross-scale attention or parallel branches to capture hierarchical features. Inspired by these, we propose a pyramid branch within SwinIR to enhance its scale-modeling capacity.

Figure 2.3: The architecture of a pyramid module

Chapter 3 Proposed Method

Our proposed model is based on SwinIR [1] . In this section, we introduce our model architecture and implementation details.

3.1 Baseline Architecture

SwinIR is a transformer-based image restoration model built upon the Swin Transformer backbone, adapted with task-specific modules such as residual Swin blocks and upsampling layers to better handle pixel-level prediction tasks.The complete architecture is shown in Figure 3.1.

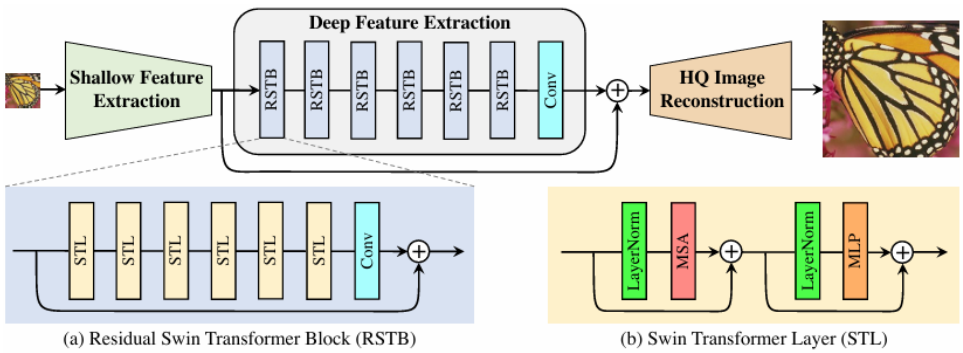

Figure 3.1: The architecture of SwinIR

The backbone of this model is the Residual Swin Transformer Block(RSTB). It consists of several Swin Transformer Layers (STL), each composed of a window-based multi-head self-attention (W-MSA) and a feed-forward network (FFN).The STL captures long-range dependencies within local windows, while the FFN processes each feature independently.The RSTB also utilizes a residual connection, allowing the input to bypass the transformation and directly combine with the output.

3.2 Proposed Architecture

In this section, we propose the PyramidBranch-SwinIR (PBSwinIR) architecture, which builds upon the original SwinIR design for image restoration tasks. To enhance its multiscale representation ability, we insert a lightweight pyramid-style module PyramidBranch into the second stage of the SwinIR feature extraction hierarchy. This module performs shallow multiscale fusion by downsampling and upsampling features locally, allowing the model to encode cross-scale context without the complexity of full pyramid networks like FPN [9].

Specifically, the PyramidBranch module first downsamples the input features using a strided convolution, then upsamples them via a transposed convolution to restore the original resolution. The resulting features are normalized through a LayerNorm layer before being processed by window-based self-attention, which captures localized spatial dependencies in non-overlapping windows. Finally, the attended features are merged and added back to the original input via a residual connection. This shallow multi-scale processing enables the module to capture coarse-to-fine contextual cues with minimal computational overhead.

As a result, PBSwinIR achieves improved performance in deblurring tasks while maintaining architectural efficiency. The complete model is illustrated in Figure 3.2.

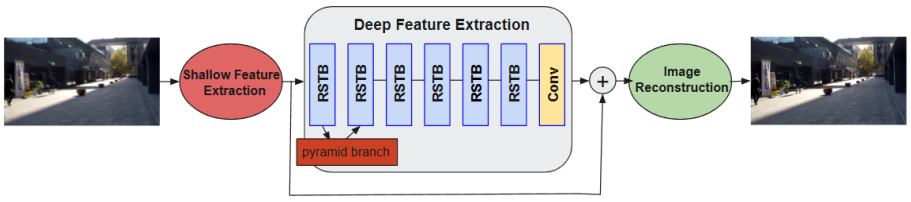

Figure 3.2: The architecture of our proposed model

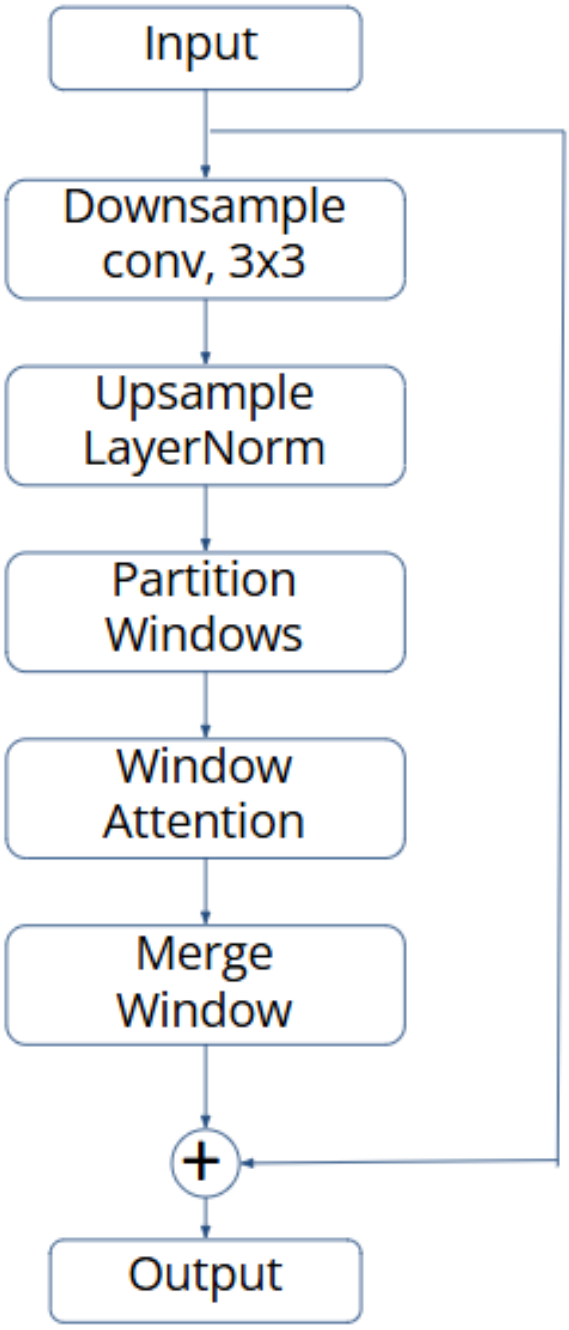

Figure 3.3: The architecture of Pyramid Branch we proposed

Chapter 4 Experiments

4.1 Dataset

In this paper, we use the GoPro [10] dataset and HIDE [11] dataset, which usually are used for image deblurring tasks.

4.1.1 GoPro

The GoPro dataset is widely used for evaluating image deblurring algorithms. It contains 3,214 image pairs with a resolution of 1280×720, split into 2,103 training images and 1,111 test images. Each pair includes a realistic blurry image and its corresponding sharp ground truth, captured using a high-speed camera. Specifically, the blurry images are synthesized by averaging consecutive frames recorded at 240 fps from 21 different scenes, simulating motion blur in real-world scenarios.

4.1.2 HIDE

The HIDE (Human-aware Deblurring) dataset is designed to benchmark deblurring methods in more complex, human-centric environments. It contains 2,025 image pairs featuring humans in dynamic scenes with varying poses and motion. The images are also captured using a high-speed camera, and the blurry frames are generated through temporal averaging.

4.1.3 RealBlur-R

The RealBlur dataset is designed to provide real-world blurry and sharp image pairs captured using DSLR cameras, without synthetic processing. It consists of two subsets, RealBlur-J and RealBlur-R, collected using two different camera models. Each subset contains around 4,000 pairs of images with spatial resolution of 720p or higher. The dataset focuses on realistic camera shake and object motion blur, which are more challenging and diverse than synthetically generated blur. The details of the three datasets is shown in Table 4.1

Table 4.1: Comparison of commonly used datasets for image deblurring.

Dataset	**GoPro** [10]	**HIDE** [11]	**RealBlur-R** [12]
Blur Type	Synthetic (frame averaging)	Synthetic (frame averaging)	Real-world(camera shake)
Data Source	High-speed video (240fps)	High-speed video with humans	DSLR handheld photos
Scene Content	General outdoor scenes	Human-centric scenes	Indoor and outdoor scenes
Image Pairs	3,214	2,025	8,422
Image Resolution	1280×720	1280×720	Varies (mostly high-res)
Motion Complexity	Moderate	High (human motion)	High (real blur, unpredictable)
Usage	Training + Testing	Testing	Testing

4.2 Evaluation Metrics and Training Details

4.2.1 Evaluation Metrics

We evaluate the predicted deblurred image by PSNR and SSIM as the evaluation indicators.

4.2.2 Training Details

In our experiments, the hyperparameter config is based on SwinIR. We uses the GetForce RTX 4070 for training. The training parameters of the experiments are shown in Table 4.2.

Table 4.2: Key hyperparameters used in our SwinIR-based deblurring model.

Model Architecture	
Image Patch Size	96×96
Window Size	8
Embedding Dimension	180
Depths	[6, 7, 6, 6, 6, 6]
Number of Attention Heads	[6, 6, 6, 6, 6, 6]
MLP Ratio	2
Training Configuration	
Loss Function	Charbonnier ($\epsilon = 1 \times 10^{-6}$)
Optimizer	Adam
Activation Function	GELU
Learning Rate	2×10^{-4}
Weight Decay	0
Batch Size	4

4.3 Overall Performance Comparison

In this paper, our experiments are based on the SwinIR framework. We propose the PBSwinIR architecture by integrating a Pyramid-style Branch module into the second stage of the network. This design enhances the network's ability to capture multi-scale contextual information while maintaining architectural efficiency. Compared to the original SwinIR, our model achieves slightly better results across multiple benchmarks, demonstrating the effectiveness of our proposed modification. However, the performance of PBSwinIR is still below the latest state-of-the-art(SOTA) deblurring methods. The detailed comparison is presented in Table A.1.

Table 4.3: Overall performance(PSNR/ SSIM) on GoPro and HIDE datasets.

Method	**GoPro**	**HIDE**	**RealBlur-R**
SwinIR (baseline)	30.96/0.9110	30.14/0.9071	33.76/0.9302
SwinIR + Pyramid Branch (ours)	31.13/0.9126	30.26/0.9094	33.98/0.9411

Figure 4.1: The result for SwinIR and PBSwinIR

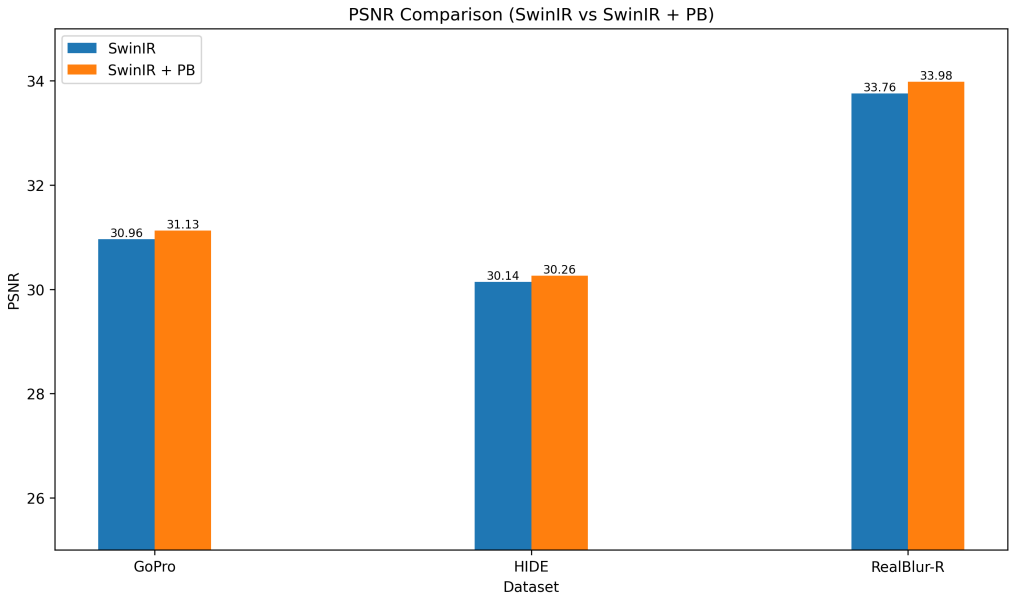

Figure 4.2: Overall PSNR performance on SwinIR and PBSwinIR

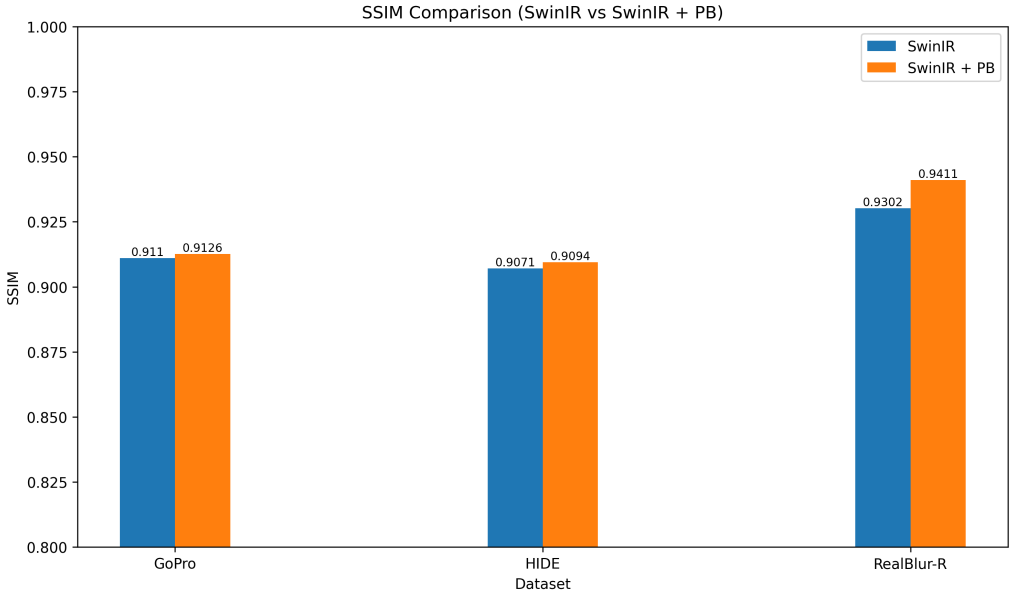

Figure 4.3: Overall SSIM performance on SwinIR and PBSwinIR

4.4 The Impact of Pyramid Branch

In this section, we explore the impact of integrating a pyramid attention branch into different layers of the SwinIR backbone. The proposed Pyramid Branch leverages a downsample-attend-upsample strategy and is inserted within each RSTB block of a specific layer. We maintain the original depth setting of SwinIR as depths = [6, 7, 6, 6, 6, 6] when inserting the module into the second layer, where the depth is slightly increased due to the added computations. Experimental results demonstrate that inserting the Pyramid Branch in the second layer yields the best PSNR performance on the GoPro dataset. This may be attributed to the layer ' s intermediate position, which allows enriched multi-scale features to be effectively fused by subsequent layers. Table 4.4 summarizes the PSNR performance for each configuration.

Table 4.4: PSNR on GoPro dataset when inserting Pyramid Branch into different SwinIR layers.

Insertion Layer	**PSNR (dB)**
SwinIR (baseline)	30.96
+ Pyramid Branch @ Layer 1	31.08
+ Pyramid Branch @ Layer 2	**31.13**
+ Pyramid Branch @ Layer 3	31.06
+ Pyramid Branch @ Layer 4	31.03
+ Pyramid Branch @ Layer 5	30.97
+ Pyramid Branch @ Layer 6	30.95

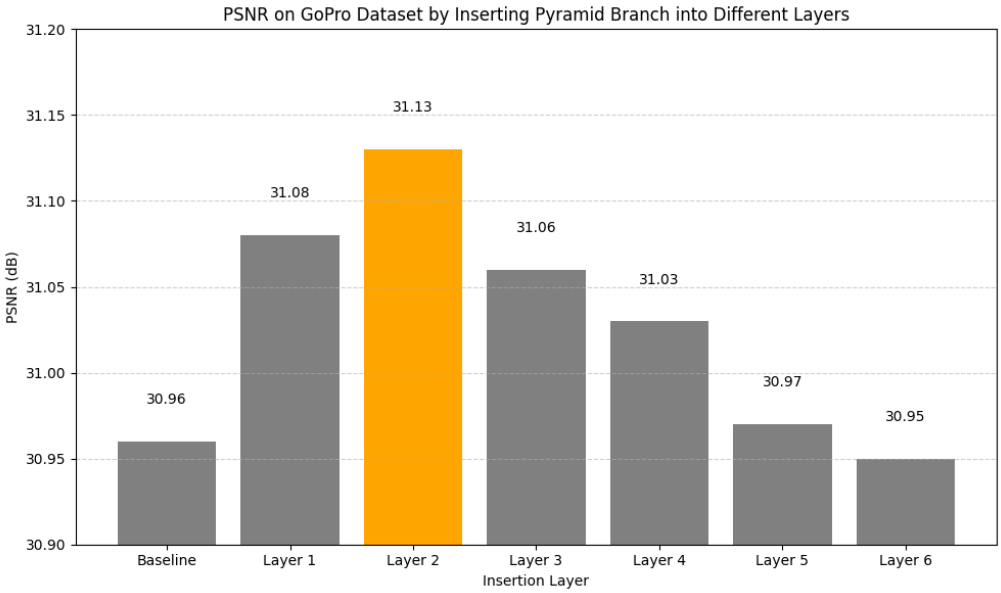

Figure 4.4: The impact of pyramid branch in different layers

Chapter 5 Conclusions and Future Work

We investigated the use of SwinIR for image deblurring and enhanced it with a pyramid-style branch for multiscale feature modeling. Although our approach yields small improvements, it remains significantly behind SOTA deblurring models. The result shows that while SwinIR ' s design is promising for high-resolution image restoration, it still needs substantial adaptation to handle the challenges of image deblurring effectively.

In the future, we plan to investigate more efficient feature interaction mechanisms within the PyramidBranch module and explore improved training strategies for handling extremely blurred regions. These directions may help close the gap between our model and state-of-the-art performance.

References

[1] J. Liang, J. Cao, K. Sun, Y. Zhang, L. V. G. Wang, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2021.

[2] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

[3] X. Li, J. Wu, Z. Lin, and H. Liu, "Progressive image deraining networks: A better and simpler baseline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3937–3946, 2019.

[4] A. Abuolaim and M. Brown, "Defocus deblurring using dual-pixel data," in *European Conference on Computer Vision (ECCV)*, pp. 748–765, Springer, 2020.

[5] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1671–1681, 2019.

[6] H. Zhang, Y. Xu, K. Zhang, W. Zuo, and L. Zhang, "Deblurring via stacked filter-responding convolutional networks," in *European Conference on Computer Vision (ECCV)*, pp. 720–736, Springer, 2020.

[7] S. Cho and S. Cho, "Rethinking coarse-to-fine approach in single image deblurring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[8] T.-Y. Lin, P. Dollaŕ, R. B. Girshick, K. He, B. Hariharan, and S. Bhattacharyya, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, 2017.

[9] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3883–3891, 2017.

[11] S. Shen, J. Chen, Y. Liu, and X. Tao, "Human-aware motion deblurring," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5572–5581, 2019.

[12] J. Rim, S. Kim, S. Lee, and S. J. Kim, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *European Conference on Computer Vision (ECCV)*, pp. 184–201, Springer, 2020.

Appendix:The overall performance comparison with other SOTA methods

Table A.1: Overall performance(PSNR/ SSIM) on GoPro and HIDE datasets.

Method	Params(M)	GoPro	HIDE	RealBlur-R
SwinIR (baseline)	11.5	30.96/0.9110	30.14/0.9071	33.76/0.9302
SwinIR + Pyramid Branch (ours)	12.7	31.13/0.9126	30.26/0.9094	33.98/0.9411
MAXIM	22	32.86/–	32.83/0.9556	35.78/–
CAPTNet	–	33.74/0.967	31.86/0.949	–
FFTformer	16.6	34.21/0.969	31.62/0.9455	–
Restormer	26.13	32.92/0.961	31.22/0.942	36.19/0.957

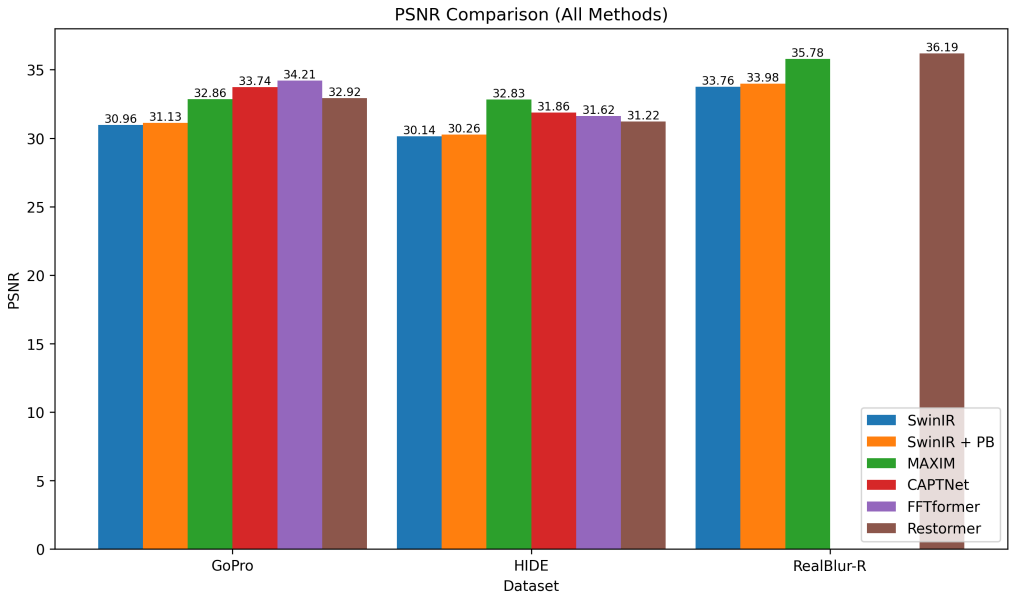

Figure A.1: Overall PSNR performance on SwinIR and PBSwinIR

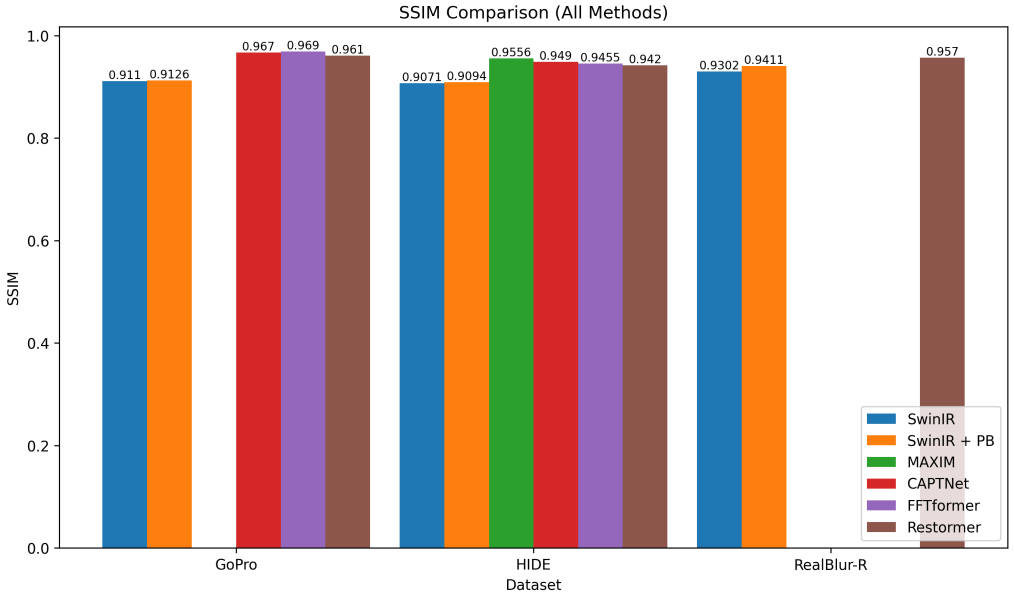

Figure A.2: Overall PSNR performance on SwinIR and PBSwinIR