PrecipFormer: Efficient Transformer for Precipitation Downscaling

Rohit Kumar[1], Tanishq Sharma[1], Vedanshi Vaghela[2], Sanjeev K. Jha[2], Akshay Agarwal[1]
[1]Department of Data Science and Engineering, [2]Earth and Environmental Sciences
Indian Institute of Science Education and Research (IISER), Bhopal
{rohitkumar20, tanishq22, vedanshi23, sanjeevj, akagarwal}@iiserb.ac.in

Abstract

Precipitation downscaling, which enhances the spatial resolution of gridded precipitation data, remains a critical challenge in climate modeling and hydrological applications. While Vision Transformers (ViTs) have demonstrated remarkable success in various computer vision tasks through their ability to capture long-range dependencies, their application to precipitation downscaling remains largely unexplored due to computational constraints and the challenge of effectively modeling both local and global precipitation patterns. This paper introduces PrecipFormer, a computationally efficient transformer architecture specifically designed for precipitation downscaling. Our model builds upon the Low-to-High Multi-Level Vision Transformer (LMLT) mechanism, enabling parallel processing of features at multiple spatial scales while significantly reducing computational overhead. We enhance the architecture with a Convolutional Block Attention Module (CBAM) in the shallow feature extractor to adaptively focus on critical spatial regions. Through extensive experiments, we demonstrate that the proposed PrecipFormer achieves superior performance compared to state-of-the-art baselines.

1. Introduction

Precipitation downscaling is a critical component in climate studies and weather forecasting, enabling the translation of coarse-resolution climate data into finer spatial scales necessary for localized impact assessments and decision-making [1]. Accurate downscaling improves the resolution of precipitation estimates, which is essential for applications such as flood prediction, agricultural planning, and infrastructure development [17]. However, precipitation downscaling poses significant challenges due to the inherently sporadic and highly localized nature of precipitation events, which are often unexpected and vary both spatially and temporally [3]. Statistical and dynamical downscaling approaches have been utilized to address these concerns. A regional climate model incorporates detailed terrain and land cover information in the dynamical downscaling approach. The high computational cost limits its application to a limited region of interest [19]. Although computationally less expensive, the statistical downscaling approach [9] has a limitation in that the empirical relationships between the coarse-scale and fine-scale climate variables are assumed to hold under future scenarios. In recent years, deep learning techniques, particularly convolutional neural networks (CNNs), have shown promise in addressing these challenges by effectively modeling spatial dependencies and capturing non-linearities in high-dimensional data [6]. Attention mechanisms have further advanced the capabilities of deep learning models by enabling the selective focus on relevant spatial features, thus improving the performance of various computer vision tasks. The convolutional block attention module (CBAM) [22] is one such mechanism that integrates channel and spatial attention to enhance feature representation in CNNs. While CBAM has demonstrated effectiveness in improving model performance, it still inherits the computational burdens associated with convolutional operations, which can be prohibitive for large-scale climate datasets. Vision transformers (ViTs) [7] has emerged as a powerful alternative to CNNs, leverag-

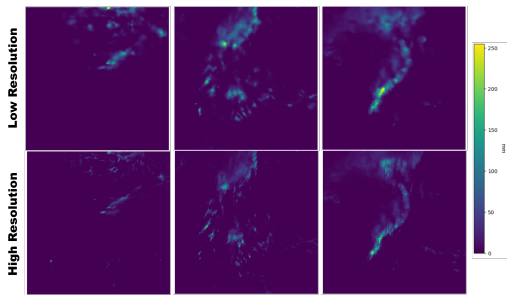

Figure 1. Visualization of input-output pairs from the RainNet dataset. Each pair shows a low-resolution input (12km) and its corresponding high-resolution ground truth (4km) precipitation map, demonstrating the significant detail captured at higher spatial resolutions.

ing self-attention mechanisms to model long-range dependencies and capture global context within images. Unlike CNNs, which rely on local receptive fields, ViTs partition images into non-overlapping patches and apply transformer architectures to process these patches, enabling more efficient and scalable spatial feature extraction [7]. Recent studies have successfully applied ViTs to various remote sensing tasks [15], including land cover classification and object detection [23], showcasing their versatility and superior performance compared to traditional CNN-based approaches. Despite the advancements brought by ViTs, their application to precipitation downscaling remains largely under-explored. Existing studies primarily focus on land cover and general weather prediction, with limited attention to the unique challenges posed by precipitation events. Precipitation events are characterized by their unexpected onset and highly localized distribution, necessitating models that can dynamically adapt to varying spatial scales and temporal patterns [3].

This gap highlights the need for specialized attention mechanisms that can efficiently capture the localized and transient nature of precipitation data. Most existing deep learning approaches for precipitation downscaling rely on synthetically generated training pairs, where low-resolution data is created by applying predefined downsampling operations (e.g., bicubic interpolation, Gaussian smoothing) to high-resolution observations [8]. Although this approach simplifies data collection, it does not effectively capture the realistic noise patterns and complex relationships found in actual multi-scale precipitation measurements. As a result, models may perform well on artificially downsampled data but struggle when applied to real-world scenarios. To address this limitation, Chen et al. [4] introduced the RainNet dataset, a large-scale and annotated dataset for precipitation downscaling. RainNet provides realistic simulations of both low and high-resolution precipitation maps in a geographical region that experiences moderate to heavy rainfall. Figure 1 shows a few samples from the RainNet dataset, depicting details preserved in high-resolution samples as compared to low-resolution samples, and localized, complex patterns of precipitation in high-resolution maps.

We introduce PrecipFormer, a novel approach that leverages self-attention-based spatial mechanisms from vision transformers to enhance precipitation downscaling. Our method addresses the computational inefficiencies of traditional attention modules by implementing a window-based self-attention mechanism, inspired by the ViT architecture, which reduces the computational burden while maintaining high accuracy in capturing localized precipitation events. *We hypothesize that efficient spatial attention mechanisms can significantly improve downscaling outcomes by better modeling the spatial dependencies and unexpected nature of precipitation phenomena.* More importantly, we also leverage the RainNet dataset [4], enabling the development and evaluation of more robust downscaling models. *To validate our hypothesis, we conduct comprehensive experiments comparing PrecipFormer with several baseline models, including those enhanced with the CBAM attention module.* Our results demonstrate that PrecipFormer consistently outperforms these baselines, achieving superior accuracy in downscaling "real world" precipitation data. Furthermore, our approach exhibits enhanced computational efficiency, making it suitable for large-scale climate applications. The primary contributions of our work are:

- We propose PrecipFormer, a computationally efficient transformer-based architecture for precipitation downscaling that integrates LMLT's parallel multi-scale processing [10] with CBAM enhancement. Unlike previous approaches that focus solely on performance, our model achieves competitive accuracy while significantly reducing computational requirements.

- Extensive experiments are conducted on the RainNet dataset, which provides realistic pairs of low-resolution (12km) and high-resolution (4km) precipitation maps, rather than relying on synthetic downsampling techniques commonly used in previous studies. This ensures our evaluation better reflects real-world application scenarios.

- Through comprehensive ablation studies and computational analysis, we provide detailed insights into the effectiveness of different architectural components and their impact on both performance and efficiency. The analysis demonstrates that PrecipFormer achieves up to 77.8% reduction in FLOPs while maintaining competitive performance compared to state-of-the-art (SOTA) models.

The remainder of this paper is structured as follows: Section 2 reviews related work on precipitation downscaling, attention mechanisms, and Vision Transformers. Section 3 details the methodology of the proposed PrecipFormer, including the architecture and implementation of the window-based self-attention mechanism [24]. Section 4 first presents the description of the dataset used, evaluation metrics, and baseline models used to assess the model performance. In the end, a comprehensive discussion of the results is given, highlighting the superiority of PrecipFormer over baseline models. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2. Related Works

Recent advances in deep learning have revolutionized both image super-resolution and precipitation downscaling. While these fields have evolved independently, they

share common challenges in reconstructing high-resolution outputs while preserving structural details. Understanding these parallel developments provides a crucial context for our work. We first review key developments in image super-resolution that inform our architectural choices, followed by their specific applications to precipitation downscaling.

2.1. Generic Image Super-Resolution Algorithms

The application of deep learning to image super-resolution (SR) has evolved significantly since the seminal work SRCNN [6], which first demonstrated the potential of CNNs for learning end-to-end mappings between low and high-resolution images. Subsequent advancements brought architectural innovations, notably SRGAN [12], which introduced adversarial training and perceptual loss to generate more realistic high-resolution images. This approach marked a significant shift from purely reconstruction-based objectives to perceptual quality enhancement.

The emergence of Vision Transformers (ViTs) [7] has further advanced the field of image super-resolution. SwinIR [14] effectively adapted the hierarchical Swin transformer architecture for SR tasks, leveraging its ability to model long-range dependencies through shifted window-based self-attention. While transformers excel at capturing global context due to their larger receptive fields, they often struggle with fine-grained local feature extraction and incur significant computational overhead. To address these limitations, low-to-high multi-level vision transformer (LMLT) [10] introduced an innovative approach that processes features at multiple scales in parallel through separate attention heads. This parallel multi-scale processing enables efficient capture of both local details and global context while significantly reducing computational requirements compared to sequential transformer blocks. Each head operates at a different spatial scale, with progressive feature integration from lower to higher resolutions, effectively balancing the model's ability to capture both fine details and long-range dependencies.

2.2. Image Super-Resolution Algorithms in Precipitation Downscaling

The success of SRCNN inspired its adaptation to climate science through DeepSD [20], which pioneered the application of deep learning to precipitation downscaling. DeepSD demonstrated significant improvements over traditional statistical and dynamical downscaling methods, establishing the potential of deep learning approaches for this domain. Building on this foundation, more sophisticated architectures are adapted for precipitation downscaling, with Super Resolution Deep Residual Network (SRDRN) [21] showing particularly strong performance due to its ability to learn complex spatial transformations through residual learning. The incorporation of generative modeling further advanced the field, with Kumar et al. [11] applying SRGAN to precipitation downscaling, enabling better preservation of extreme weather patterns and spatial coherence. Subsequently, attention mechanisms are integrated into downscaling models, Chiang et al. [5] incorporating CBAM to enhance the model's ability to focus on relevant spatial regions, given the localized nature of precipitation patterns.

Recent advances in computer vision have brought new possibilities to precipitation downscaling. Diffusion models, which have shown remarkable success in image generation, have been adapted for precipitation downscaling [18], offering improved uncertainty quantification and physical consistency. While vision transformers have demonstrated impressive results in various remote sensing tasks [15], their application to precipitation downscaling remains relatively unexplored, despite their potential for capturing long-range dependencies crucial for atmospheric processes. These approaches have highlighted key challenges specific to precipitation downscaling, including the need to preserve physical constraints, handle extreme events, and capture multi-scale atmospheric phenomena. This has led to an increasing focus on developing architectures that can effectively balance computational efficiency with the ability to model complex spatial patterns across different scales.

3. Proposed Precipitation Downscaling Algorithm

Precipitation downscaling presents unique challenges that require carefully designed architectural choices. Traditional transformer architectures, while powerful, often struggle with computational efficiency and with the preservation of fine-grained precipitation patterns. Our proposed PrecipFormer addresses these challenges through three key innovations: *efficient multi-scale feature processing, enhanced spatial attention, and optimized high-resolution reconstruction.* In this section, we detail these components and their integration into a cohesive architecture, followed by our implementation details of the proposed architecture.

3.1. PrecipFormer

Our proposed PrecipFormer architecture builds upon SwinIR [14] and integrates the multi-level transformer concept from LMLT [10] for efficient precipitation downscaling. As shown in Figure 2, the model has three main components: (a) shallow feature extraction enhanced with CBAM, (b) deep feature extraction utilizing parallel multi-level transformer blocks, and (c) high-resolution reconstruction. Given a low-resolution precipitation map x, our goal is to generate a super-resolved precipitation map $\hat{y}$ with upscaling factor s (for our experiments *s=3*) while maintaining computational efficiency.

Shallow Feature Extraction: The shallow feature extraction module begins with a 3×3 convolutional layer to

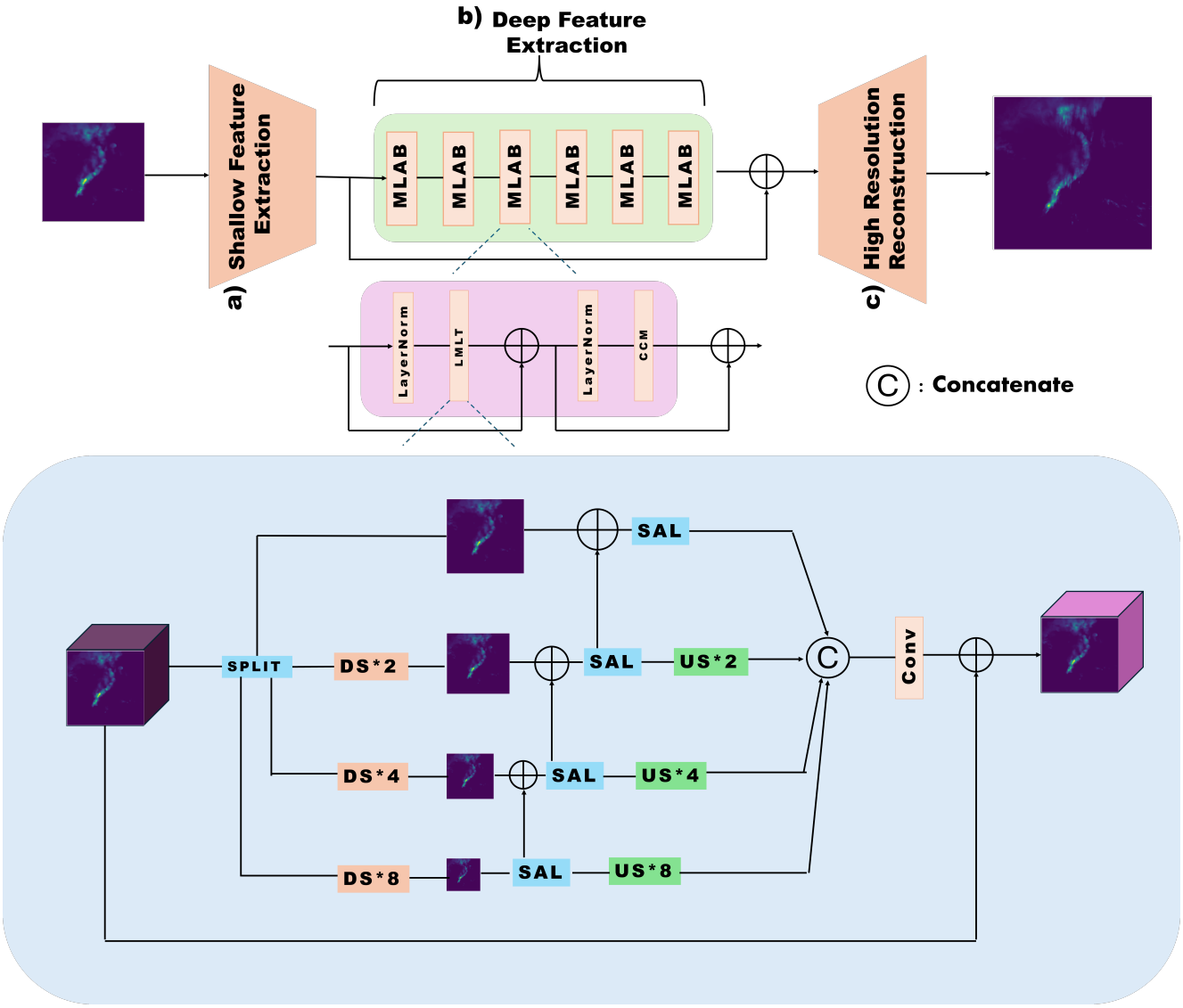

Figure 2. PrecipFormer architecture overview. (a) Shallow feature extraction with CBAM enhancement for initial feature refinement. (b) Deep feature extraction with multiple Multi-Level Attention Blocks, each containing parallel multi-scale self-attention processing. (c) High-resolution reconstruction module using pixel shuffle upsampling. Self-Attention Layer (SAL) operating on non-overlapping windows. CCM: 1×1 channel-wise convolution for feature dimension adjustment. DS/US: Downsample/Upsample operations for multi-scale processing.

extract initial features from the input precipitation map. Unlike SwinIR, we enhance this stage by incorporating CBAM [22] after the convolutional layer. The CBAM module sequentially applies channel and spatial attention mechanisms to adaptively refine features. *This enhancement is particularly beneficial for precipitation downscaling, as it helps the model focus on relevant spatial regions given the sparse and localized nature of precipitation patterns.* The channel attention emphasizes informative features along the channel dimension, while spatial attention helps identify regions of significant precipitation activity.

Deep Feature Extraction: The deep feature extraction module contains multiple sequential blocks similar to [14], which we refer to as multi-level attention blocks (MLAB). Within each block, we replace sequential attention calculation with parallel multi-level transformer architecture. Inside one such block, the input features first undergo layer normalization (LN) [13] and are then divided into H heads ($H = 4$ in our implementation), the division is done using channel-wise split, with each head processing features at progressively reduced spatial scales. The first head maintains the original spatial dimensions, while subsequent

heads process features at $1/2$, $1/4$, and $1/8$ of the original resolution through average pooling operations. A residual connection is employed around each self-attention block, followed by a channel-wise 1×1 convolution for feature dimension adjustment.

A key innovation we leverage from [10] is the progressive integration of features from lower-resolution heads to higher-resolution heads. Features from lower-resolution heads are upsampled through bilinear interpolation and integrated with higher-resolution features through element-wise addition. Each integration path includes a residual connection to maintain gradient flow. The final features from all heads are concatenated along the channel dimension and processed through a 1×1 convolution to adjust the channel dimension. This multi-scale processing effectively captures precipitation patterns at different scales while maintaining computational efficiency.

High-Resolution Reconstruction: The high-resolution reconstruction stage processes the concatenated multi-scale features through a 3×3 convolutional layer followed by pixel shuffle [16] upsampling to achieve the target resolution. The use of pixel-shuffle upsampling helps avoid checkerboard artifacts that can occur with transposed convolution operations.

Training Objective Following recent advances in transformer-based super-resolution networks, we employ the Charbonnier loss [2] for training:

$$\mathcal{L}(\hat{y}, y) = \sqrt{\|\hat{y} - y\|^2 + \epsilon^2} \tag{1}$$

where ϵ is set to $1e^{-3}$. The Charbonnier loss, a differentiable variant of L1 loss, offers better stability in handling outliers during training. This is particularly important for precipitation downscaling, where accurate reconstruction of extreme precipitation events is crucial.

3.2. Implementation Details

We implement PrecipFormer in PyTorch and train it using the Adam optimizer with an initial learning rate of $1e^{-4}$. The learning rate is adjusted using a cosine annealing scheduler to ensure stable convergence. The model is trained with a batch size of 16, which we assert provides realistic pairs of low-resolution (12km) and high-resolution (4km) precipitation maps, i.e., the upscaling factor s is 3. All experiments are conducted on a single NVIDIA RTX A6000 GPU.

4. Results and Analysis

Evaluating precipitation downscaling models requires careful consideration of both quantitative metrics and qualitative analysis, as different metrics capture distinct aspects of model performance. We first provide a detailed description of the dataset and associated protocol used in this study, followed by the description of metrics used to measure the performance of the proposed and baseline models. Later, detailed performance comparisons and ablation studies have been presented. We conclude with an analysis of computational efficiency, demonstrating how PrecipFormer achieves state-of-the-art performance while significantly reducing computational requirements.

4.1. Dataset

The RainNet dataset [4] is a comprehensive resource, offering around $62,500$ image pairs representing rainfall data at 4 km and 12 km spatial resolutions, with an hourly temporal resolution. The RainNet dataset is especially valuable for downscaling tasks, where the objective is to transform coarse-resolution data (such as 12 km images) into high-resolution data (such as 4 km images). This process, known as downscaling, is crucial in climate and weather prediction, as it allows for more detailed and accurate forecasting.

For the aforementioned dataset, we took the sub-crops of each image to 128×128 pixels for low resolution (LR) (which covers approximately 2,633,565 km sq. area) and 384×384 for high resolution (HR) (which covers approximately the same area) on the south-western part of the eastern coast of the US. i.e approximately latitude: from 25°N to 41°N and longitude: from 105°W to 89°W for the rainy season months i.e July through November. As a part of the training strategy, we apply temporal split to data for training, testing, and validation. We used the 70% for training, and the 15% each for validation and evaluation of the proposed and existing models.

4.2. State-of-the-art Baseline Models

To evaluate the effectiveness of our proposed PrecipFormer, we compare it against three representative super-resolution models that span different architectural paradigms. **SRCNN** [6] is one of the pioneering CNN-based super-resolution models that demonstrates the potential of deep learning for super-resolution tasks. The model employs three convolutional layers that simulate the sparse-coding-based super-resolution pipeline: patch extraction, non-linear mapping, and reconstruction. Despite its simple architecture, SRCNN established important baselines for learning-based super-resolution approaches. **SRResNet** [12] leverages multiple residual blocks with skip connections, enabling the training of deeper networks while maintaining stable gradient flow. The residual architecture allows the model to learn residual information between low and high-resolution images more effectively than sequential CNN architectures. **SwinIR** [14] represents the state-of-the-art in transformer-based super-resolution. It employs a hierarchical Swin Transformer as its backbone, which computes self-attention within shifted windows to reduce computational complexity while maintaining the abil-

ity to model long-range dependencies. The model demonstrates the potential of transformer architectures in capturing global context for super-resolution tasks. We implement all baseline models following their original architectures but train them specifically for precipitation downscaling on the RainNet dataset. This ensures a fair comparison while evaluating their effectiveness in capturing the unique characteristics of precipitation patterns. For consistent evaluation, all models are trained with the same batch size, optimization strategy, and number of epochs as our proposed model.

4.3. Evaluation Metrics

The performance of the proposed and existing algorithms are evaluated using multiple metrics, which are described below:

- The mean absolute error (MAE) measures the average magnitude of errors between predicted and ground truth precipitation values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

  where y_i and $\hat{y}_i$ are the ground truth and predicted values respectively.

- The structural similarity index (SSIM) assesses the perceptual quality of downscaled precipitation maps by comparing luminance, contrast, and structure:

$$\text{SSIM} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3}$$

  where μ and σ represent mean and standard deviation respectively, and c_1, c_2 are constants to avoid division by zero.

- The continuous ranked probability score (CRPS) evaluates probabilistic forecasts by measuring the integrated squared difference between the cumulative distribution functions of predictions and observations:

$$\text{CRPS} = \int_{-\infty}^{\infty}(F(y) - H(y - y_o))^2 dy \tag{4}$$

  where, F(y) is the predicted cumulative distribution and H is the Heaviside step function centered on the observation y_o.

4.4. Experimental Analysis

As shown in Table 1, PrecipFormer demonstrates consistent improvements across all evaluation metrics. For example, the proposed model achieves the lowest MAE of 0.595, showing substantial improvements of 12.2% and 15.9% over SRCNN (0.678) and SRResNet (0.708) respectively. Compared to the transformer-based SwinIR (0.611), PrecipFormer achieves a modest but meaningful improvement of 2.6% in MAE, indicating enhanced capability in accurate precipitation value reconstruction. The trend suggests that transformer-based architectures, particularly our enhanced design, are better suited for preserving the spatial patterns and structural details crucial for precipitation fields. Furthermore, the progressive reduction in CRPS from SRCNN (0.811) through SRResNet (0.791) and SwinIR (0.783) to our model (0.772) showcases the evolution of architectures in handling precipitation predictions.

Table 1. Quantitative evaluation on the RainNet test set (9,635 samples). Our PrecipFormer achieves superior performance across all metrics compared to state-of-the-art methods. The best results are shown in **bold**.

Model	MAE ↓	SSIM ↑	CRPS ↓
SRCNN	0.678	0.958	0.811
SRResNet	0.708	0.958	0.791
SwinIR	0.611	0.961	0.783
PrecipFormer (Ours)	**0.595**	**0.964**	**0.772**

The consistent performance improvements across all metrics, though incremental compared to SwinIR, are particularly noteworthy given our model's significantly reduced computational complexity (as detailed in our efficiency analysis). This demonstrates that PrecipFormer successfully achieves its goal of balancing performance and efficiency through the integration of parallel multi-scale processing and CBAM enhancement. Also, from the visual comparison of these models shown in Figure 3, we can observe that the CNN-based models, SRCNN, and SRResNet consistently overestimate the precipitation values compared to transformer-based models, SwinIR, and PrecipFormer.

4.5. Ablation Studies

To validate the effectiveness of our design choices and understand the contribution of each key component, we conduct ablation experiments by creating variants of our model. Table 2 presents the quantitative results of this study.

Impact of Multi-scale Parallel Attention: To evaluate the effectiveness of multi-scale feature processing in parallel attention layers, we create an ablated version of precipFormer which processes features at a single scale. The results show that removing multi-scale processing leads to slight performance degradation across metrics, with MAE increasing from 0.595 to 0.614 (3.1% drop) and CRPS increasing from 0.772 to 0.788. The degradation validates our hypothesis that processing features at multiple scales help better capture precipitation patterns at different resolutions. The parallel processing of features at multiple scales

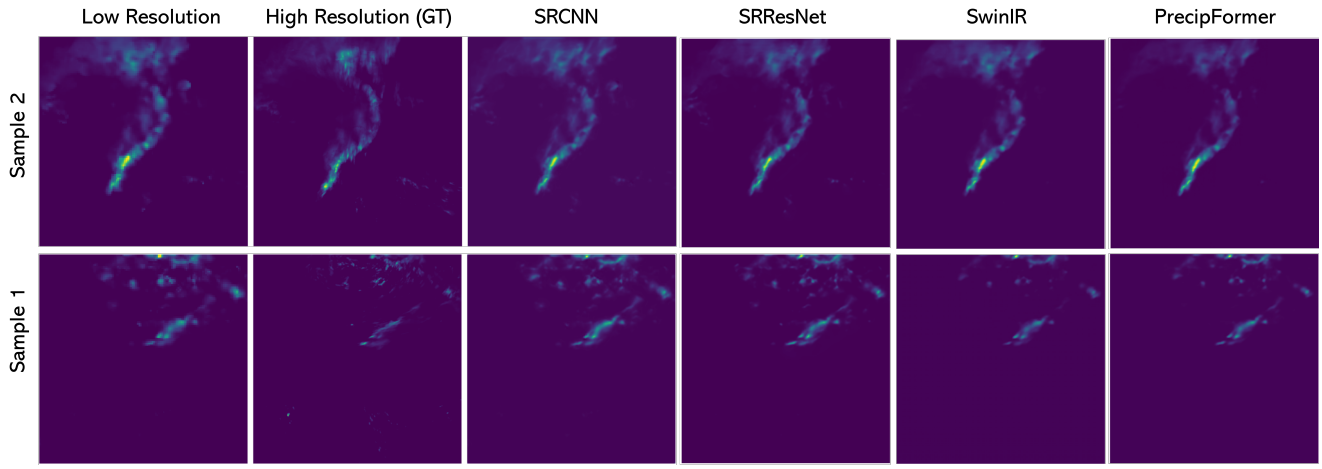

Figure 3. Qualitative comparison of precipitation downscaling results (3× upscaling). From left to right: Low-resolution input, high-resolution ground truth (GT), SRCNN, SRResNet, SwinIR, and our PrecipFormer, respectively. Two representative samples demonstrate our method's ability to better preserve fine precipitation patterns while maintaining structural coherence. Zoom in for better visualization.

Table 2. Ablation study demonstrating the impact of architectural components. The best results are shown in **bold**.

Model	MAE ↓	SSIM ↑	CRPS ↓
SwinIR	0.611	0.961	0.783
SwinIR + Parallel Processing	0.614	0.961	0.788
SwinIR + Parallel Processing + Multiscale	0.607	0.962	0.780
PrecipFormer (Ours)	**0.595**	**0.964**	**0.772**

enables the model to effectively balance local details and global context.

Effect of CBAM Integration: A separate ablated version removes the CBAM module from the shallow feature extractor, replacing it with conventional convolutional layers. Interestingly, this modification shows competitive performance with an MAE of 0.607 and SSIM of 0.962, demonstrating only marginal degradation compared to the full model. This suggests that while CBAM provides benefits in feature refinement, the core strength of our architecture lies in its efficient multi-scale processing capability. The relatively small performance gap indicates that the model's fundamental architecture is robust even without sophisticated attention mechanisms in the shallow features.

Full Model Analysis: Our complete PrecipFormer, incorporating both multi-scale parallel attention and CBAM, achieves the best performance across all metrics. The final model shows a 2.6% improvement in MAE over the baseline SwinIR while maintaining better or comparable performance in other metrics. Importantly, these improvements are achieved while significantly reducing computational complexity, as demonstrated in our efficiency analysis.

4.6. Computational Efficiency Analysis

Beyond downscaling performance, we analyze the computational efficiency of our proposed model in terms of parameter count, inference speed, and floating point operations (FLOPs). Table 3 presents these metrics for different models. PrecipFormer achieves remarkable efficiency across all computational metrics. With only 237K parameters, it is significantly lighter than both SRResNet (972K) and SwinIR (680K), representing parameter reductions of 75.6% and 65.1%, respectively. This substantial reduction in model size is achieved while maintaining superior downscaling performance. In terms of computational complexity, PrecipFormer requires only 59.01G FLOPs for processing a batch of 16 samples, which is significantly lower than SRResNet (266.29G) and SwinIR (177.64G), representing reductions of 77.8% and 66.8% respectively. Notably, while adding parallel processing and multi-scale features to SwinIR substantially reduces FLOPs to 69.32G, our complete PrecipFormer achieves a further modest reduction to 59.01G through efficient integration of these components with CBAM.

The FLOPs reduction translates to improved inference speed, with PrecipFormer processing a batch of 16 samples in 132.08ms compared to SwinIR's 294.76ms. In-

Table 3. Computational efficiency analysis comparing model complexity and runtime performance. Inference speed is measured on an NVIDIA RTX A6000 GPU with batch size 16, averaged over 100 iterations. PrecipFormer achieves significant reductions in parameters and FLOPs compared to SRResNet while maintaining superior downscaling performance.

Model	# Parameters	Inference Speed (ms)	FLOPs (G)
SRResNet	972K	**97.85**	266.29
SwinIR	680K	294.76	177.64
SwinIR + Parallel Processing	282K	281.45	71.78
SwinIR + Parallel Processing + Multiscale	**236K**	134.70	59.05
PrecipFormer (ours)	237K	132.08	**59.01**

terestingly, while SRResNet shows competitive inference speed (97.85ms) despite its high FLOP count, this can be attributed to its simpler architecture that avoids computationally intensive operations like layer normalization and attention calculations. However, this comes at the cost of a much larger model size and inferior downscaling performance. The variants of SwinIR with parallel processing and multi-scale features show computational metrics close to our final model, with only marginal differences in parameters (277K vs 237K) and FLOPs (69.32G vs 59.01G). However, PrecipFormer achieves significantly better inference speed (132.08ms vs 288.45ms) through optimized integration of these components, while maintaining slightly better performance metrics.

These results demonstrate that while the individual architectural components contribute to efficiency gains, their careful integration in PrecipFormer achieves an optimal balance between model complexity and downscaling performance. This makes our model particularly suitable for real-world precipitation downscaling applications where computational resources may be limited.

5. Conclusion

Due to the complex nature of the precipitation data, traditional computer vision or deep learning-based image super-resolution techniques fail to enhance it. In response, we present PrecipFormer, a novel transformer-based architecture that leverages parallel multi-scale attention processing and enhanced feature refinement for efficient precipitation downscaling. Our model achieves competitive performance while significantly reducing computational demands compared to state-of-the-art image enhancement algorithms. Through comprehensive ablation studies, we validated that both multi-scale parallel attention and CBAM contribute positively to model performance. In the future, we aim to extend the architecture of PrecipFormer by incorporating temporal modeling for sequences of precipitation maps and exploring multi-variate downscaling to jointly process related climate variables by leveraging topographical information.

References

[1] Arman Abdollahipour, Hassan Ahmadi, and Babak Aminnejad. A review of downscaling methods of satellite-based precipitation estimates. *Earth Science Informatics*, 15(1):1–20, 2022. 1

[2] Pierre Charbonnier, Laure Blanc-Féraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *IEEE International Conference on Image Processing (ICIP)*, pages 168–172, 1994. 5

[3] Jie Chen and Xunchang John Zhang. Challenges and potential solutions in statistical downscaling of precipitation. *Climatic Change*, 165(3):63, 2021. 1, 2

[4] Xuanhong Chen, Kairui Feng, Naiyuan Liu, Bingbing Ni, Yifan Lu, Zhengyan Tong, and Ziang Liu. Rainnet: A large-scale imagery dataset and benchmark for spatial precipitation downscaling. *Advances in Neural Information Processing Systems*, 35:9797–9812, 2022. 2, 5

[5] Chia-Hao Chiang, Zheng-Han Huang, Liwen Liu, Hsin-Chien Liang, Yi-Chi Wang, Wan-Ling Tseng, Chao Wang, Che-Ta Chen, and Ko-Chih Wang. Climate downscaling: A deep-learning based super-resolution model of precipitation data with attention block and skip connections. *arXiv preprint arXiv:2403.17847*, 2024. 3

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 1, 3, 5

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dmitry Weissenborn, Florian Zoph, Thomas Unterthiner, and Georg Heigold. An image is worth 16x16 words: Transformers for image recognition at scale. *CVPR*, 2020. 1, 2, 3

[8] Paula Harder, Qidong Yang, Venkatesh Ramesh, Prasanna Sattigeri, Alex Hernandez-Garcia, Campbell Watson, Daniela Szwarcman, and David Rolnick. Generating physically-consistent high-resolution climate data with hard-constrained neural networks. *arXiv preprint arXiv:2208.05424*, 18:109–122, 2022. 2

[9] B.C. Hewitson, J. Daron, R.G. Crane, et al. Interrogating empirical-statistical downscaling. *Climatic Change*, 122:539–554, 2014. 1

[10] Jeongsoo Kim, Jongho Nang, and Junsuk Choe. Lmlt: Low-to-high multi-level vision transformer for image super-resolution. *arXiv preprint arXiv:2409.03516*, 2024. 2, 3, 5

[11] Bipin Kumar, Kaustubh Atey, Bhupendra Bahadur Singh, Rajib Chattopadhyay, Nachiketa Acharya, Manmeet Singh, Ravi S. Nanjundiah, and Suryachandra A. Rao. On the modern deep learning approaches for precipitation downscaling. *Earth Science Informatics*, 16(2):1459–1472, 2023. 3

[12] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CVPR*, pages 4681–4690, 2017. 3, 5

[13] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016. 4

[14] Xintao Li, Hanhuo Zhang, Zhe Yu, Honglak Li, Chia-Wen Cheng, Chao Dong, Xian Lu, and Ming-Hsuan Liu. Swinir: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1833–1844, 2021. 3, 4, 5

[15] Pengyuan Lv, Wenjun Wu, Yanfei Zhong, Fang Du, and Liangpei Zhang. Scvit: A spatial-channel feature preserving vision transformer for remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. 2, 3

[16] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 5

[17] Marek Smid and Ana Cristina Costa. Climate projections and downscaling techniques: a discussion for impact studies in urban systems. *International Journal of Urban Sciences*, 22(3):277–307, 2018. 1

[18] Prakhar Srivastava, Ruihan Yang, Gavin Kerrigan, Gideon Dresdner, Jeremy J McGibbon, Christopher S Bretherton, and Stephan Mandt. Precipitation downscaling with spatiotemporal video diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[19] Francisco J. Tapiador, Andrés Navarro, Raúl Moreno, José Luis Sánchez, and Eduardo García-Ortega. Regional climate models: 30 years of dynamical downscaling. *Atmospheric Research*, 235:104785, 2020. 1

[20] Thomas Vandal, Evan Kodra, Sangram Ganguly, Adam Michaelis, Ramakrishna Nemani, and Arun Ganguly. Deepsd: Generating high resolution climate change projections through single image super-resolution. In *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1663–1672. ACM, 2017. 3

[21] Fang Wang, Di Tian, Lisa Lowe, Latif Kalin, and John Lehrter. Deep learning for daily precipitation and temperature downscaling. *Water Resources Research*, 57(4):e2020WR029308, 2021. 3

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 1, 4

[23] Jing Yao, Bing Zhang, Chenyu Li, Danfeng Hong, and Jocelyn Chanussot. Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 2

[24] Renjie Zou, Chunfeng Song, and Zhaoxiang Zhang. The devil is in the details: Window-based attention for image compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17492–17501, 2022. 2