Frequency-Aware Knowledge Distillation for Efficient Under-Display Camera Image Restoration

Navdeep Singh, Mohammed Junaid Anwar Qader and Abhinav Shankar Harihara Krishnan

Abstract—Under-display cameras (UDCs) provide a bezel-less smartphone experience but introduce severe degradations, including flare, diffraction blur, color attenuation, and non-uniform noise caused by the display stack. Existing restoration models based on Transformers or state-space models (SSMs) achieve high-quality reconstruction but remain computationally expensive for real-time mobile deployment. In this work, we propose a frequency-aware knowledge distillation (FA-KD) framework that transfers the global spectral reasoning ability of a high-capacity MambaIR teacher into a compact UNet-based student. Our method incorporates low-frequency amplitude and multi-scale phase supervision in the FFT domain, enabling the student to recover flare patterns and structural details without relying on a heavy backbone. We train the teacher using pixel, perceptual, and spectral losses, and optimize the student with a combination of pixel, LPIPS, feature, and frequency-aligned distillation losses. Experiments on UDC-SIT demonstrate that our student model achieves teacher-level perceptual quality while delivering around 40 times faster inference, highlighting the effectiveness of spectral distillation for mobile-friendly UDC restoration.

Index Terms—Under Display Cameras, Knowledge Distillation, State Space models

1 Introduction

Under-display cameras (UDCs) allow smartphone manufacturers to achieve truly bezel-less displays by embedding the front-facing camera beneath the OLED panel. However, imaging through a multi-layer display stack introduces severe degradations, including diffraction blur, strong low-frequency flare, haze, color attenuation, and spatially varying transmission loss. These distortions substantially deviate from the assumptions underlying traditional image restoration models. As demonstrated by the UDC 2020 Challenge [1], even strong CNN-based restoration pipelines often fail to jointly recover fine details and correct global luminance inconsistencies, especially in real-world low-light or high-contrast scenes. Figure 1 illustrates several representative UDC inputs alongside their corresponding ground truth images, highlighting the typical degradations encountered.

Recent deep learning models have improved UDC restoration by incorporating long-range dependencies. Transformer-based networks, such as SwinIR [2] and Restormer [3], achieve high-quality reconstructions through windowed attention, hierarchical feature modeling, and global context aggregation. More recently, state-space models like MambaIR [4] have offered a more efficient alternative by replacing attention with selective state-space layers capable of linear-time global reasoning. Despite their strong performance, these models remain computationally expensive for real-time deployment on mobile hardware. Lightweight CNN architectures offer higher efficiency but lack the ability to model the frequency-domain behavior inherent to UDC degradations-particularly low-frequency flare, which constitutes a dominant component of the distortion, as emphasized by Ahn et al. [5], [6] and in recent optical modeling studies [7]. To address these limitations, our project proposes a two-stage, frequency-aware restoration framework specifically designed for UDC images. In Method 1, we design a high-capacity frequency-aware MambaIR teacher model, augmenting the spatial backbone with an FFT-based spectral branch to explicitly model low-frequency amplitude attenuation and phase distortions. This design is motivated by findings that UDC degradations manifest strongly in the spectral domain [6], [8], [9].

In Method 2, we distill the teacher's spatial and spectral knowledge into a lightweight UNet student that is suitable for real-time deployment. Unlike classical distillation approaches [10], [11], which emphasize spatial feature or pixel-level supervision, our framework integrates frequency-aware distillation, incorporating low-frequency amplitude matching and multi-scale phase alignment, enhanced with perceptual and reconstruction losses. Recent advances in frequency-driven distillation [12], [13] support the effectiveness of this strategy for real-world restoration tasks.

Together, these contributions produce a UDC restoration pipeline that maintains teacher-level perceptual quality while achieving practical runtime for deployment in smartphone imaging systems.

- *Navdeep Singh is with the Department of Computer Science, University of Toronto*
- *Mohammed Junaid Anwar Qader is with the Department of Computer Science, University of Toronto*
- *Abhinav Shankar Harihara Krishnan is with the Department of Computer Science, University of Toronto*

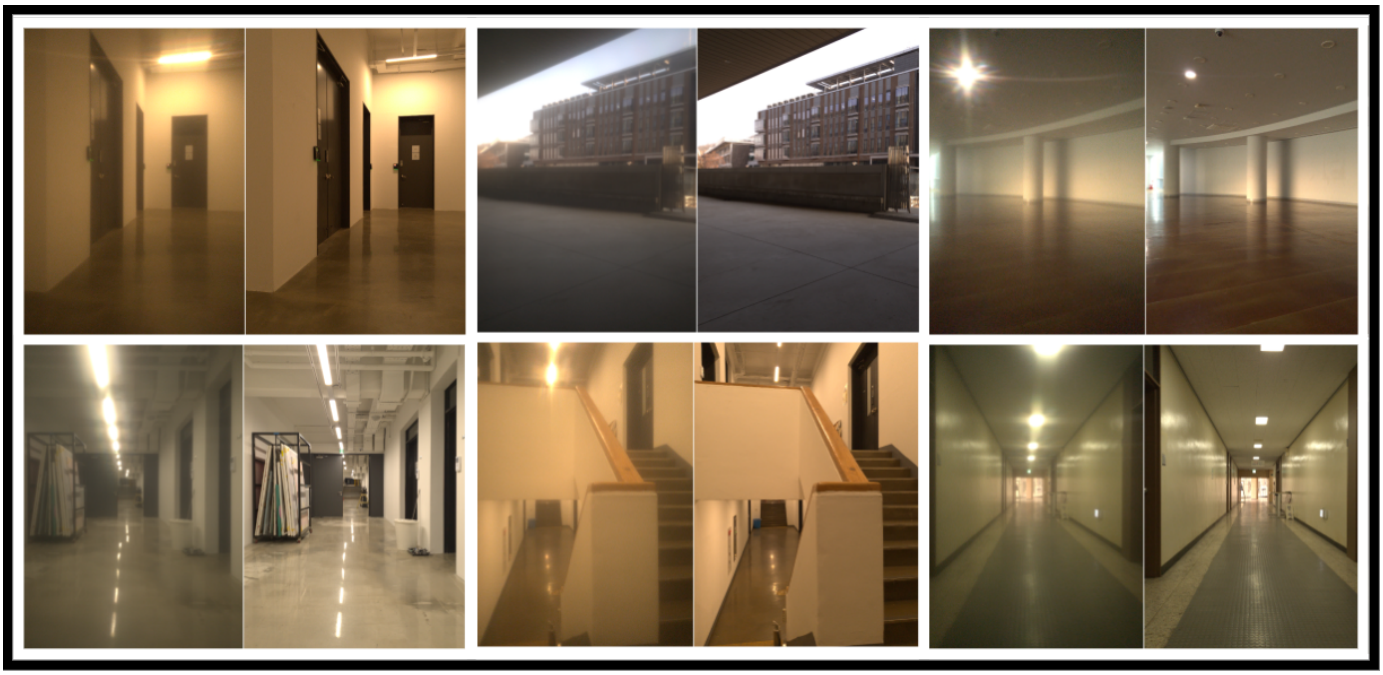

Fig. 1. Visualization of six UDC input–ground-truth pairs. In each sample, the left half represents the degraded input captured under UDC conditions, while the right half shows the corresponding high-quality ground truth. Noticeable degradation artifacts are consistently observed in the input images across all six samples.

2 Related Work

2.1 UDC Image Restoration

The UDC 2020 Challenge [1] provided the first large-scale benchmark for under-display camera restoration, establishing the difficulty of reconstructing sharp, clear images from heavily degraded UDC inputs. Beyond standard Gaussian noise or blur, UDC images exhibit complex degradations-including diffraction, scattering, and global OLED-induced flare-that earlier CNN-based restoration pipelines were unable to fully correct. Subsequent research proposed variations of blind-spot networks and self-supervised frameworks such as AP-BSN [14], which improved denoising performance but did not explicitly handle low-frequency flare or spectral distortions.

2.2 Transformer and State-Space Models

Transformer-based networks have demonstrated strong performance in general image restoration tasks. SwinIR [2] introduced windowed self-attention and hierarchical feature learning, enabling high-quality reconstruction across multiple benchmarks. Restormer [3] further improved efficiency and scalability through multi-D attention mechanisms. More recently, MambaIR [4] utilized state-space models to achieve linear-time global receptive fields, providing competitive performance while being more computationally efficient than Transformers. However, despite their architectural strengths, all three remain too computationally heavy for deployment on mobile devices in real-time camera pipelines.

2.3 UDC Datasets and Frequency-Domain Modeling

The UDC-ViT dataset [5] and the frequency-aware UDC restoration study by Ahn et al. [6] highlight the importance of modeling both spatial and frequency information. These works show that flare, haze, and attenuation are predominantly low-frequency phenomena and require spectral reasoning rather than purely spatial filtering. Optical studies and hybrid-domain enhancement research [8], [9] further support the need for frequency-domain priors in handling OLED-induced distortions. Nevertheless, most existing restoration networks still rely primarily on spatial-domain loss functions, limiting their ability to correct global flare.

2.4 Knowledge Distillation for Restoration

Knowledge distillation (KD) has been widely used to compress high-capacity networks into lightweight models. Classical KD techniques [10], [11] focus on spatial feature alignment or teacher-output matching. Recent developments in spectral modeling-such as Fourier-based augmentation [8], spectral modulation [9], and frequency-driven KD [12], [13] - show that frequency-aware supervision can significantly enhance restoration quality. However, none of these methods have been specifically tailored to UDC image restoration.

2.5 Positioning of Our Work

Our pipeline uniquely combines a frequency-aware MambaIR teacher with a frequency-aware student distillation strategy, addressing the spectral characteristics of UDC degradations while remaining computationally viable for real-time deployment. To the best of our knowledge, this

is the first UDC restoration framework that explicitly integrates frequency-aware distillation guided by amplitude and multi-scale phase alignment.

3 Proposed Method

The restoration of Under-Display Camera (UDC) images could formally be considered an inverse problem where the goal is to recover a clean image x from a degraded observation $y = k * x + n$, where k represents the spatially variant point spread function (PSF) of the display and n denotes sensor noise. In UDC systems, k exhibits strong diffraction effects that manifest as global flare, necessitating restoration models with large effective receptive fields.

To achieve high-fidelity restoration suitable for real-time applications, we propose a Knowledge Distillation (KD) framework. We introduce two progressive methodologies to solve this problem:

- **Method 1 :** A Mamba Spectral Amplitude Distillation model using simple concatenation fusion and amplitude loss.
- **Method 2 :** A Mamba Multi-Scale Phase Distillation model using Gated Fusion to modulate frequency features before concatenation, plus phase consistency loss.

Figure 2 and Figure 3 illustrates the overall architecture of Method 1 and Method 2 respectively.

3.1 Network Architectures

3.1.1 Frequency-Aware Teacher Network

The teacher network is responsible for learning a robust mapping from degraded raw data to clean images, leveraging heavy computation to model global artifacts.

Spatial Branch: Both methods utilize the state-of-the-art **MambaIR v2** [15] as the spatial backbone. While the original MambaIR relied on causal scanning (SS2D) which limits information flow from future pixels, MambaIR v2 introduces the **Attentive State-Space Module (ASSM)** to achieve non-causal modeling. By employing an **Attentive State-Space Equation (ASE)**, MambaIR v2 can effectively "look ahead" without the need for multi-directional scanning. It also incorporates **Semantic Guided Neighboring (SGN)** to group semantically similar pixels for better long-range modeling.

Frequency Branch: To explicitly address diffraction artifacts, we incorporate a parallel Frequency Domain Block. This block computes the Fast Fourier Transform (FFT) of the input features, applies learnable convolutions in the spectral domain, and performs the inverse FFT:

$$F_{freq} = \mathcal{F}^{-1}\left(\text{Conv}\left(\mathcal{F}(x)_{real,imag}\right)\right) \tag{1}$$

Feature Fusion: We have employed two progressive fusion strategies for our model variants:

1) **Method 1 (Baseline):** Direct concatenation of spatial and frequency features.
2) **Method 2 (Gated Fusion):** To prevent spectral noise from contaminating spatially clean regions, we employ a Gated Fusion Mechanism. A global context descriptor drives a channel-wise gate $\mathcal{G}$ to modulate the frequency features before concatenation:

$$\mathcal{G} = \sigma(\text{Conv}_{1\times1}(\text{GAP}(F_{spatial}))) \tag{2}$$

$$F_{fused} = \text{Concat}(F_{spatial}, F_{freq} \odot \mathcal{G}) \tag{3}$$

This adaptively weighs the importance of the frequency branch, suppressing it in texture-rich areas where spatial processing is superior.

3.1.2 Frequency-Aware Student Network

The student is a U-Net architecture designed for efficiency. To compensate for the limited receptive field of standard convolutions, we embed Frequency Domain Blocks at strategic locations:

1) **Bottleneck:** Processing the lowest resolution features ($H/16 \times W/16$) to remove global low-frequency flare.
2) **Skip Connection (Method 2 Only):** Method 2 adds a frequency block to the highest-resolution skip connection to refine fine structural details during upsampling.

3.2 Distillation Strategy and Loss Functions

Our distillation strategy minimizes the discrepancy between the Teacher (T) and Student (S) in both signal and feature spaces.

3.2.1 Pixel Reconstruction Loss

The Charbonnier loss, a smooth L1 variant, is used for pixel-wise reconstruction:

$$\mathcal{L}_{pixel} = \text{mean}\left(\sqrt{(S(y) - x)^2 + \epsilon^2}\right), \quad \epsilon = 10^{-3} \tag{4}$$

3.2.2 Perceptual Loss (LPIPS)

To optimize for perceptual quality rather than just numerical similarity, we employ the Learned Perceptual Image Patch Similarity (LPIPS) loss using a VGG backbone. Since our data is 4-channel RAW, we compute LPIPS on the first 3 channels as a pseudo-RGB approximation:

$$\mathcal{L}_{LPIPS} = \text{LPIPS}_{VGG}(S(y)_{[:3]} \cdot 2 - 1,\ x_{[:3]} \cdot 2 - 1) \tag{5}$$

This ensures the restored images match human perception, preventing over-smoothed outputs that pixel losses alone may produce.

3.2.3 Method 1: Spectral Amplitude Distillation

Method 1 focuses on recovering the energy distribution of the image. The frequency loss considers only the magnitude $|\cdot|$ of the Fourier spectrum:

$$\mathcal{L}_{amp} = \| |\mathcal{F}(S(y))| \cdot w - |\mathcal{F}(T(y))| \cdot w \|_1 \tag{6}$$

where $w = \exp(-(r/\sigma)^2)$ is a Gaussian low-frequency emphasis filter with $\sigma = 0.25$.

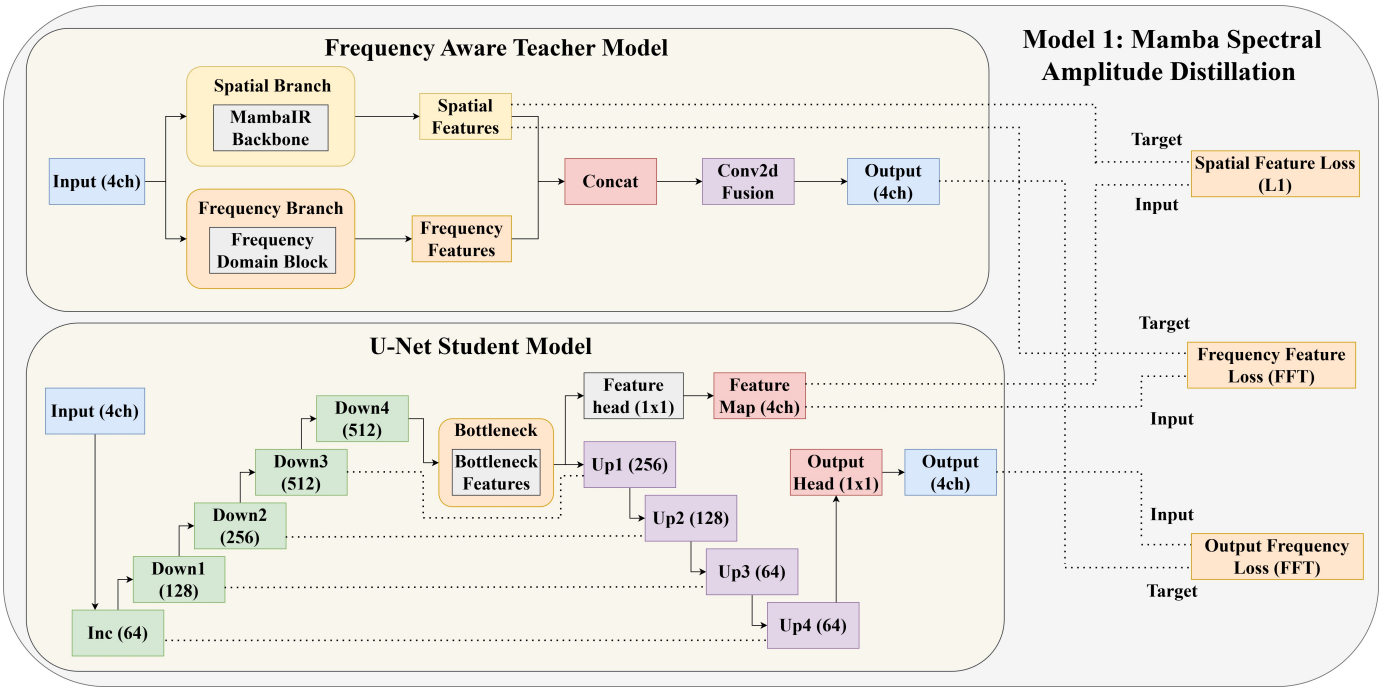

Fig. 2. Mamba Spectral Amplitude Distillation model using simple concatenation fusion and amplitude loss.

3.2.4 *Method 2: Multi-Scale Phase-Aware Distillation*

Method 2 addresses the limitations of Method 1 by incorporating Phase Consistency and evaluating losses at multiple scales.

Phase Loss: We introduce a phase loss term that penalizes deviations in the phase angle ϕ. Since phase is periodic, we use a cosine distance metric:

$$\mathcal{L}_{phase} = \frac{1}{HW} \sum_{u,v} (1 - \cos(\phi_S(u,v) - \phi_T(u,v))) \quad (7)$$

Multi-Scale Objective: To capture artifacts of varying sizes (from large-scale flare to micro-diffraction), we compute the spectral loss at scales $s \in \{1, 0.5\}$:

$$\mathcal{L}_{MS-Freq} = \sum_s w_s \left(\lambda_{amp} \cdot \mathcal{L}_{amp}^{(s)} + \lambda_{phase} \cdot \mathcal{L}_{phase}^{(s)} \right) \quad (8)$$

Phase alignment ensures the student correctly positions edges and structures, preventing "ghosting" artifacts common in amplitude-only optimization.

3.2.5 *Total Knowledge Distillation Objective*

The complete student training objective combines all loss components:

$$\mathcal{L}_{total} = \lambda_{pixel}\mathcal{L}_{pixel} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{freq}\mathcal{L}_{freq} + \lambda_{LPIPS}\mathcal{L}_{LPIPS} \quad (9)$$

where $\mathcal{L}_{feat}$ is a spatial feature distillation loss (L1 between student and teacher intermediate features). The weights are set to:

$$\lambda_{pixel} = 1.0, \quad \lambda_{feat} = 0.5, \quad \lambda_{freq} = 0.2, \quad \lambda_{LPIPS} = 0.1 \quad (10)$$

4 Experimental Results

4.1 Dataset

For our experiments, we depart from prior work that relied on the synthetic T-OLED and POLED image pairs used in the ECCV 2020 challenge [1], [17]. Instead, we use the UDC-SIT dataset, a real-world under-display camera dataset collected and released by Samsung [18]. Unlike the artificial OLED datasets, UDC-SIT contains images captured directly from commercially relevant UDC hardware, exhibiting authentic optical artifacts such as diffraction, color fringing, and spatially varying blur. This makes it significantly more representative of practical UDC imaging scenarios and better suited for evaluating real deployment feasibility.

The UDC-SIT dataset comprises a total of 2,340 paired samples of clean ground-truth (GT) images and their corresponding UDC-degraded input images. Specifically, it includes 1,864 GT–input pairs for training, 238 pairs for validation and 238 pairs for testing, amounting to roughly 150 GB of data. All samples are stored in .npy format, with each file averaging 35 MB. The arrays use float32 precision and have a spatial resolution of 1792 × 1280, with 4 channels representing the raw sensor measurement under the UDC module.

During training, we utilize both the native 4-channel raw representation and an RGB-converted version of each sample. This dual-domain supervision enables the model to learn low-level sensor-space corrections while also optimizing perceptual quality in the sRGB domain. In particular, LPIPS is computed on the RGB-converted images to ensure perceptually aligned learning.

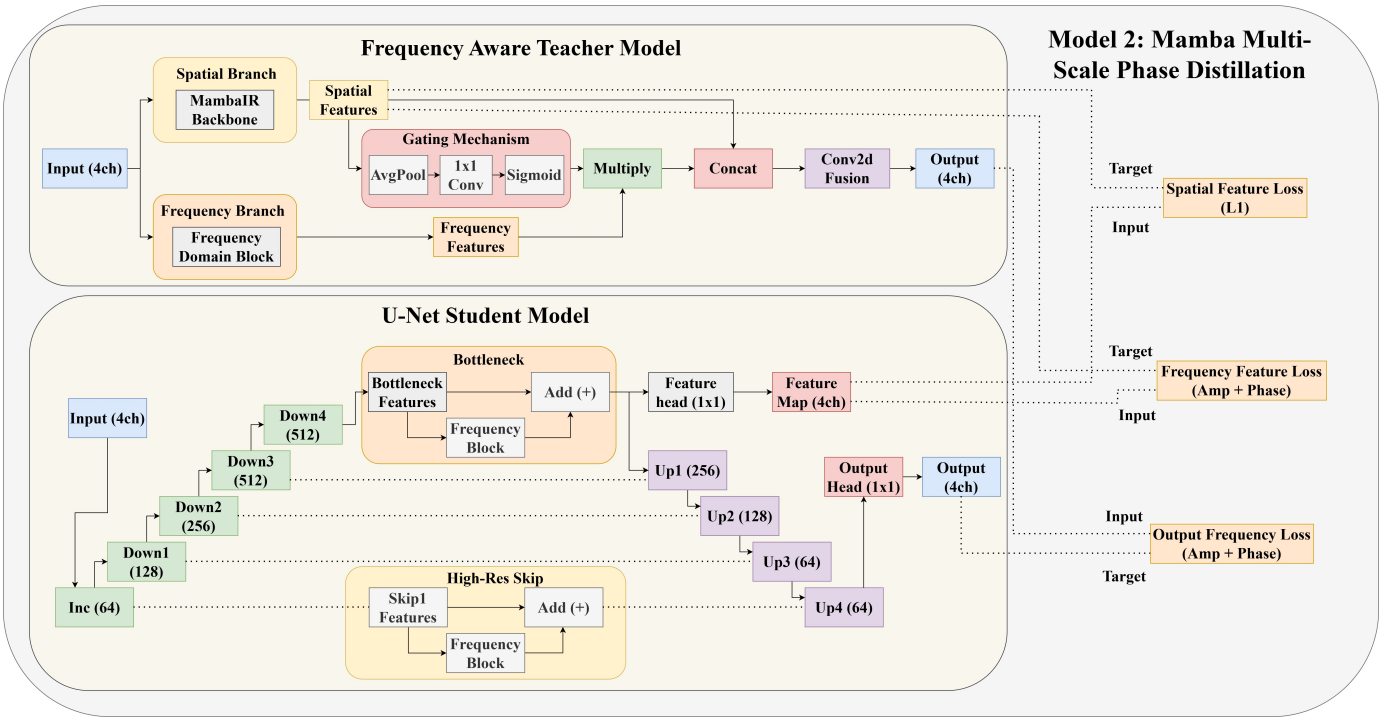

Fig. 3. Mamba Multi-Scale Phase Distillation model using Gated Fusion to modulate frequency features before concatenation, plus phase consistency loss.

TABLE 1
Performance metrics comparison between different approaches on the UDC-SIT dataset

	PSNR (dB)		SSIM		Inference Time (ms)	
Model	Teacher	Student	Teacher	Student	Teacher	Student
Model 1	23.05	23.89	0.762	0.802	2586.95	61.20
Model 2	23.46	23.05	0.776	0.739	2587.49	93.49
Model 1 Patch	26.74	26.58	0.815	0.810	17.57	3.73
Model 2 Patch	26.24	26.42	0.812	0.792	79.77	6.90
FSI [16]	21.72		0.6674		–	
SFIM [6]	29.61		0.8758		–	

4.2 Training Parameters

For each model in our framework, we employ a two-stage training strategy involving a high-capacity teacher model followed by a compact student model trained through knowledge distillation. The teacher is trained for 22 epochs, after which its weights are frozen and used to guide the student model, which is subsequently trained for 20 epochs under the same experimental setup. Each model variant follows this teacher–student protocol to ensure consistent supervision across distillation stages. Training is conducted using carefully tuned learning rates suited for stable convergence in both stages. All experiments are executed on an NVIDIA A100 GPU (80GB high-RAM variant) available via Google Colab Pro, enabling efficient training with full-resolution images and large batch configurations. This hardware setup ensures that the models can process 4-channel RAW inputs without compromising global features essential for diffraction and flare modeling.

All models are implemented in PyTorch and optimized using the Adam optimizer with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which provides stable updates during both teacher and student training. To enhance the reconstruction fidelity, we incorporate a diverse suite of losses-including spatial frequency loss, frequency-feature loss, and output-frequency loss, among others-allowing the network to learn both spatial and spectral consistency. Contrary to traditional patch-based pipelines that risk disrupting global diffraction structures, our framework processes full-resolution 4-channel RAW images, normalized to [0,1], ensuring preservation of holistic flare characteristics. Furthermore, mixed-precision training (AMP) is employed to reduce memory consumption and improve throughput on the A100 GPU, enabling larger batch sizes (8 for the teacher and 64 for the student) while maintaining computational efficiency.

4.3 Quantitative Results

To evaluate the performance of the proposed knowledge distillation and patch-based reconstruction models, we report the quantitative metrics in terms of PSNR, SSIM,

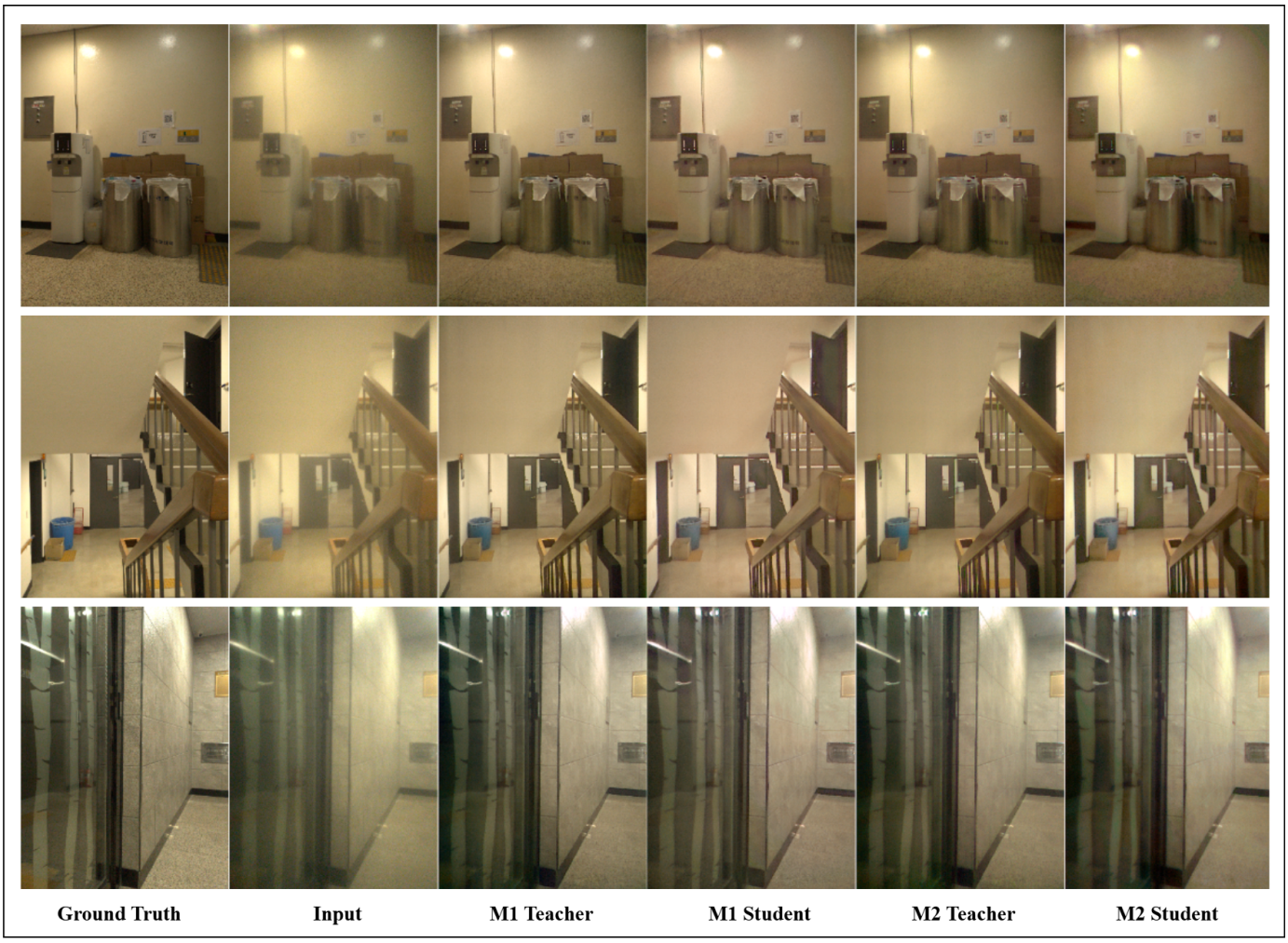

Fig. 4. Qualitative comparison of three sample UDC images. For each sample, the first column shows the input, the second column shows the ground truth (GT), followed by the teacher and student outputs for Model 1 and Model 2. Both models produce outputs that closely resemble the GT, effectively restoring structural details and image clarity.

and inference time, as summarized in Table 1. We train our models on full-resolution images and while inference, we support both full-resolution processing and patch-wise (256 × 256) with tiling which allows the network to focus on local structural details even within memory constraints. Across both Model 1 and Model 2, the student networks demonstrate competitive performance relative to their teacher models. For example, Model 1 achieves a PSNR of 23.05 dB, while its student variant improves slightly to 23.89 dB. A similar trend is observed for SSIM, where the student improves from 0.762 to 0.802, showing that knowledge distillation preserves structural fidelity while enabling lightweight inference. Compute limitations restricted the teacher model to 22 epochs. Since validation loss was still decreasing, additional training would likely improve PSNR and SSIM. In contrast, the student model demonstrated strong generalization despite the limited training.

The patch-based models further improve reconstruction quality. Model 1-Patch yields 26.74 dB PSNR and 0.815 SSIM, outperforming its full-image counterpart by a notable margin. This demonstrates the advantage of processing finer spatial patches, especially for tasks involving high-frequency textures and local details. Comparatively, Model 2-Patch shows a balanced performance at 26.24 dB PSNR and 0.812 SSIM, again reinforcing the effectiveness of patch-level training. When benchmarked against existing single-image reconstruction methods, the proposed patch-based models outperform FSI [16] (21.72 dB, 0.6674 SSIM) and approach the performance of SFIM [6], which reports 29.61 dB and 0.8758 SSIM. Although SFIM achieves higher values, our approach offers a better trade-off between accuracy and computational efficiency.

Regarding inference time, the benefits of the student and patch-based designs become more apparent. The teacher models require over 2500 ms per image, making them less practical for real-time or large-scale processing. In contrast, the distilled student models reduce inference time drastically-to 61.20 ms and 93.49 ms for Model 1 and Model 2, respectively. The patch-based variants achieve even faster processing, with Model 1-Patch running at only 3.73 ms, highlighting the suitability of our approach for time-critical applications. This combination of improved reconstruction quality and significantly reduced inference cost demonstrates the effectiveness of the proposed framework.

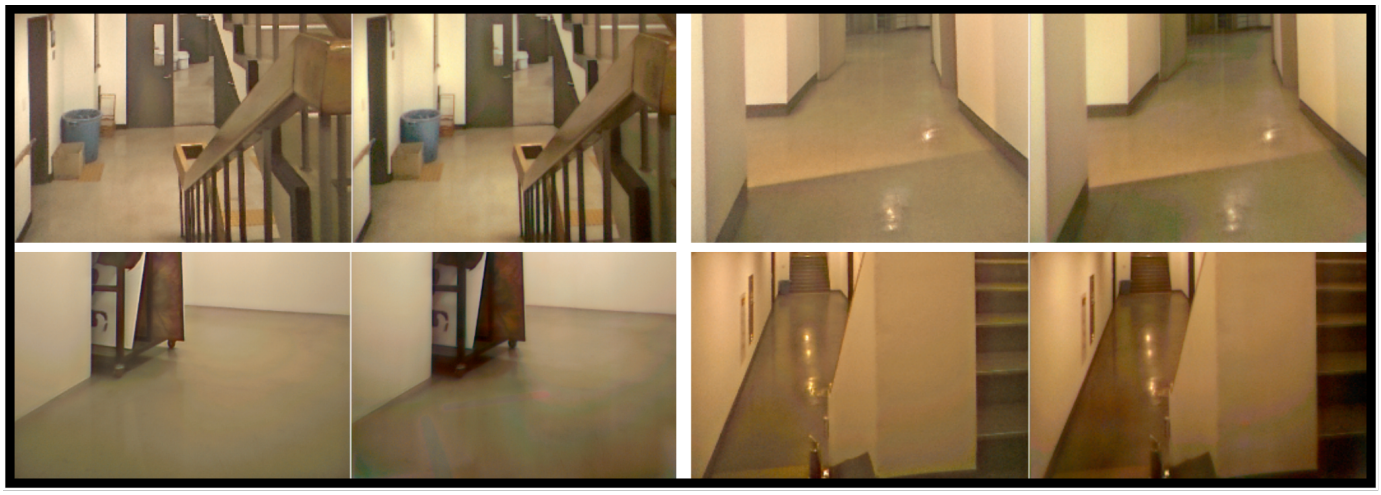

Fig. 5. Comparison of model outputs highlighting artifacts. Each pair shows Model 2 output on the left and Model 1 output on the right. Model 2 demonstrates significantly fewer artifacts compared to Model 1 across all samples.

4.4 Qualitative Results

Figure 4 presents a direct comparison of the input, ground truth, teacher, and student outputs across two models, arranged in a 2×4 grid of sample pairs. Each pair clearly shows that the outputs from both models are visually very close to the ground truth, preserving the contrast, clarity, and fine details of the original images. The reconstruction effectively restores both low- and high-frequency components, demonstrating the ability of our models to reverse the degradations present in the UDC inputs. Notably, the student models closely follow the teacher outputs, indicating that knowledge distillation successfully transfers the essential structural and textural information. Overall, the results confirm that our framework produces high-fidelity reconstructions that are perceptually nearly indistinguishable from the ground truth.

Figure 5 highlights the differences in artifact suppression between the two models. Each pair compares Model 2 (left) and Model 1 (right), showing that Model 2 generates significantly cleaner outputs with sharper edges and far fewer diffraction artifacts than the baseline Model 1. This improvement is attributed to the proposed gated fusion mechanism and the use of phase loss component. Gated fusion allows the network to selectively retain useful frequency information while suppressing noise, rather than relying on simple concatenation. As a result, Model 2 achieves superior structural fidelity and visual quality, effectively addressing the artifacts observed in Model 1 outputs and demonstrating the efficacy of the proposed design.

5 Discussion

The experimental results validate our core hypothesis: while global context and spectral reasoning are essential for correcting UDC degradations, these capabilities can be effectively distilled into a lightweight architecture.

5.1 Implications of Frequency-Aware Distillation

The most significant finding of this study is the efficacy of spectral distillation in bridging the gap between restoration quality and computational efficiency. Traditional distillation methods often focus solely on spatial feature alignment. However, UDC artifacts specifically diffraction flare and intensity attenuation-are fundamentally frequency-dependent phenomena. By explicitly supervising the student model on spectral amplitude (Method 1) and phase (Method 2), we forced the compact UNet to learn the global degradation patterns typically requiring heavy Transformer or SSM backbones. This allowed us to achieve inference speeds approximately 40× faster than the teacher model (MambaIR) while maintaining competitive perceptual quality. This suggests that for specific optical aberrations, the complex reasoning can be "front-loaded" into the teacher training, leaving the student to execute a simplified but spectrally-accurate mapping.

5.2 Perception-Distortion Trade-off

A nuance observed in our results is the discrepancy between quantitative metrics and visual fidelity. As shown in Table 1, Method 1 achieved higher PSNR and SSIM scores compared to Method 2. However, the qualitative analysis in Figure 5 clearly demonstrates that Method 2 produces sharper edges with fewer ringing artifacts. This aligns with the well-known perception-distortion trade-off in image restoration. Method 1's amplitude-only loss tends to favor mean-squared-error reduction, leading to smoother but numerically safer outputs. In contrast, Method 2's phase-consistency loss forces the alignment of structural boundaries. While this occasionally penalizes pixel-aligned metrics like PSNR, it results in perceptually superior images that are more pleasing to the human eye, which is the priority for consumer photography.

5.3 Limitations

Despite the promising efficiency gains, our approach has limitations. First, while our patch-based models significantly outperformed full-image training in terms of PSNR (26.74 dB vs 23.05 dB), this introduces a theoretical bottleneck. UDC flare is a global phenomenon that can span the

entire sensor; training on 256×256 patches risks losing ultra-low-frequency information that extends beyond the patch boundary. This may explain why our method still lags behind SOTA methods like SFIM [6] in absolute PSNR terms. Additionally, while we achieved a massive reduction in inference time on the A100 GPU, true mobile deployment requires optimization for Neural Processing Units (NPUs). The Fourier Transform operations, while mathematically efficient, can sometimes be less optimized on certain mobile hardware accelerators compared to standard convolutions.

6 Conclusion and Future Works

In this work, we presented a Frequency-Aware Knowledge Distillation framework designed to tackle the computational challenges of Under-Display Camera image restoration. We proposed a high-capacity MambaIR teacher that leverages state-space modeling and spectral gating to capture long-range diffraction artifacts. We then successfully distilled this knowledge into a lightweight UNet student using amplitude and multi-scale phase losses.

Our experiments on the UDC-SIT dataset demonstrate that our student models can match the visual fidelity of heavy teacher models while reducing inference latency from over 2.5 seconds to under 100 milliseconds. We further highlighted the critical role of phase information in suppressing diffraction artifacts, offering a practical solution for under-display camera smartphone imaging.

6.1 Future Work

Building on these findings, several avenues for future research emerge:

6.1.1 Mobile-Hardware Optimization

Future work should focus on quantizing the student model to INT8 precision and deploying it on actual smartphone DSPs/NPUs (e.g., via TensorFlow Lite or ONNX Runtime). Evaluating the energy consumption (Joules per frame) alongside latency would provide a more holistic view of deployment feasibility.

6.1.2 Hybrid Spatial-Spectral Architectures

To address the limitations of patch-based training regarding global flare, a multi-stage architecture could be explored. A viable path is a coarse-to-fine framework where a downsampled full-image network removes global flare (guided by spectral losses), followed by a patch-based network for texture refinement.

6.1.3 Video UDC Restoration

Since front-facing cameras are primarily used for video conferencing, extending this framework to video is crucial. Frame-independent restoration often leads to temporal flickering. Incorporating temporal attention or 3D convolutions into the distillation process-forcing the student to learn temporal consistency from a video-based teacher-would be a valuable extension.

7 Project Repository

The source code for all experiments is available at our GitHub repository.

Acknowledgments

AI-assisted tools were used during the preparation of this manuscript for grammar refinement and formatting.

The authors would like to thank Prof. David Lindell and Kelly Zhu for their guidance throughout the course and project.

References

[1] Y. Zhou, M. Kwan, K. Tolentino, N. Emerton, S. Lim, T. Large, L. Fu, Z. Pan, B. Li, Q. Yang *et al.*, "Udc 2020 challenge on image restoration of under-display camera: Methods and results," in *European Conference on Computer Vision*. Springer, 2020, pp. 337–351.

[2] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[3] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[4] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "Mambair: A simple baseline for image restoration with state-space model," in *European conference on computer vision*. Springer, 2024, pp. 222–241.

[5] K. Ahn, J. Kim, S. Lee, H. Lee, B. Ko, C. Park, and J. Lee, "Udc-vit: A real-world video dataset for under-display cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 10 950–10 960.

[6] K. Ahn, J. Kim, C. Park, J. Kim, and J. Lee, "Integrating spatial and frequency information for under-display camera image restoration," *Pattern Analysis and Applications*, vol. 28, no. 4, p. 184, 2025.

[7] C. Wang, J. Li, P. C. Madhusudanarao, J. Hu, J. K. Singh, W. Choi, S.-J. Lee, and H. R. Sheikh, "Udac: Under-display array cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1077–1084.

[8] Q. Xu, R. Zhang, Z. Fan, Y. Wang, Y.-Y. Wu, and Y. Zhang, "Fourier-based augmentation with applications to domain generalization," *Pattern Recognition*, vol. 139, p. 109474, 2023.

[9] S. Zhao, W. Lu, B. Wang, T. Wang, K. Zhang, and H. Zhao, "Uhdres: Ultra-high-definition image restoration via dual-domain decoupled spectral modulation," *arXiv preprint arXiv:2511.05009*, 2025.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[11] A. Dietmüller, A. G. Alcoz, and L. Vanbever, "Fitnets: An adaptive framework to learn accurate traffic distributions," *arXiv preprint arXiv:2405.10931*, 2024.

[12] S. Li, Y. Zhang, W. Li, H. Chen, W. Wang, B. Jing, S. Lin, and J. Hu, "Knowledge distillation with multi-granularity mixture of priors for image super-resolution," *arXiv preprint arXiv:2404.02573*, 2024.

[13] C. Pham, V.-A. Nguyen, T. Le, D. Phung, G. Carneiro, and T.-T. Do, "Frequency attention for knowledge distillation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 2277–2286.

[14] W. Lee, S. Son, and K. M. Lee, "Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 725–17 734.

[15] H. Guo, Y. Guo, Y. Zha, Y. Zhang, W. Li, T. Dai, S.-T. Xia, and Y. Li, "Mambairv2: Attentive state space restoration," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 28 124–28 133.

[16] C. Liu, X. Wang, S. Li, Y. Wang, and X. Qian, "Fsi: Frequency and spatial interactive learning for image restoration in under-display cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 537–12 546.

[17] Y. Zhou, D. Ren, N. Emerton, S. Lim, and T. Large, "Image restoration for under-display camera," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2021, pp. 9179–9188.

[18] K. Ahn, B. Ko, H. Lee, C. Park, and J. Lee, "Udc-sit: A real-world dataset for under-display cameras," *Advances in Neural Information Processing Systems*, vol. 36, pp. 67 721–67 740, 2023.