

WS-NeRF: A Mamba-Driven World-State-Aware Adaptive Deblurring Neural Radiance Field

Hang Jiang^{1,2}, Jinghao Wang^{1,2}, Yiming Zhang^{1,2},
Xinhong Wang^{1,2}, Luwei Ran^{1,2}, and Yinfeng Yu^{1,2,3,4,†}

Abstract—Neural Radiance Fields (NeRF) have attracted extensive attention in recent years due to their strong capability for high-quality 3D reconstruction and novel view synthesis from multi-view images. Existing methods usually rely on high-quality sharp inputs, while real-world image acquisition is highly susceptible to blur degradation, which severely affects the reconstruction quality of NeRF. In this paper, we propose a novel *Mamba-driven world-state-aware adaptive deblurring neural radiance field*, termed WS-NeRF, to address image degradation and 3D inconsistency. We formulate the alternating optimization of radiance fields as a dynamic evolution process with temporal memory, and jointly exploit comprehensive multi-dimensional world states and a mixture-of-experts mechanism to dynamically adjust the confidence of deblurring priors. Experimental results show that WS-NeRF significantly improves blurry radiance field reconstruction quality, achieving better performance on PSNR, SSIM, and LPIPS, while exhibiting more stable iterative recovery behavior.

I. INTRODUCTION

Since the seminal work of Mildenhall *et al.* [1], Neural Radiance Fields (NeRF) have achieved photorealistic 3D representation and novel view synthesis by encoding scene geometry and appearance into multilayer perceptrons. Owing to their strong self-supervised learning capability and impressive rendering quality, NeRF-based methods have been widely applied in robot navigation, urban modeling, and virtual reality. However, the high-fidelity rendering ability of NeRF strongly depends on high-quality sharp inputs. In real scenarios, blur frequently occurs due to long exposure or camera motion, which seriously violates the physical assumptions of NeRF and leads to severe degradation of 3D reconstruction quality. Therefore, improving the 3D reconstruction capability of NeRF under blurred observations has become an important research topic. The blur-aware reconstruction task is illustrated in Fig. 1.

Although many methods attempt to recover blurred radiance fields by alternating between 2D deblurring and 3D reconstruction, effectively coupling external image priors with 3D geometric constraints remains challenging. The intrinsic mismatch between single-view 2D restoration priors

and global 3D spatial consistency introduces three major difficulties. First, there is a dimensional mismatch between 2D restoration priors and 3D geometric constraints. Second, prior fusion lacks a confidence-aware closed loop. Third, temporal modeling over optimization states is absent.

To address these challenges, we model radiance field deblurring as a dynamic evolution system with temporal memory. During training, we jointly leverage comprehensive multi-dimensional world states and a mixture-of-experts (MoE) mechanism to adaptively adjust the confidence of deblurring priors. Specifically, we design a world-state estimator that deeply fuses low-level spatial cues, high-frequency edge features, and deep semantic discrepancies extracted by a pretrained large vision model, DINOv2 [2], to diagnose spatial inconsistency in real time. The resulting state features are then fed into a Mamba-based temporal inference network [3], endowing the system with long-range memory over optimization trajectories. Driven by both multi-dimensional state awareness and temporal memory, the adaptive controller collaborates with a MoE module to intelligently and dynamically allocate confidence to 2D restoration signals at each iteration. By balancing 2D prior restoration ability and 3D geometric constraints, the proposed method achieves both high-quality deblurring and stable scene convergence.

The main contributions of this paper are summarized as follows:

- 1) We propose **WS-NeRF**, a novel adaptive state-guided deblurring framework that fundamentally resolves the inconsistency between introducing external 2D priors and maintaining 3D geometric constraints through dynamic perception and control.
- 2) We construct a **multi-granularity world-state perception mechanism** that innovatively fuses spatial cues, restoration-aware signals, and deep semantic features from DINOv2 to sensitively monitor subtle geometric degradation during reconstruction.
- 3) We design a **Mamba-based temporal module** for radiance field optimization trajectory modeling, enabling the system to exploit historical evolution information and effectively alleviate over-correction and oscillation commonly observed in alternating optimization.

II. RELATED WORK

A. Neural Radiance Fields and Novel View Synthesis

NeRF implicitly models scene density and color with multilayer perceptrons, achieving remarkable success in

¹School of Intelligent Science and Technology, Xinjiang University, Urumqi, China. Email: jianghanglaa@icloud.com.

²School of Computer Science and Technology, Xinjiang University, Urumqi, China. Email: jianghanglaa@icloud.com.

³Joint Research Laboratory for Embodied Intelligence, Xinjiang University, Urumqi, China. Email: yuyinfeng@xju.edu.cn.

⁴Joint International Research Laboratory of Silk Road Multilingual Cognitive, Xinjiang University, Urumqi, China. Email: yuyinfeng@xju.edu.cn.

[†]Corresponding author: Yinfeng Yu (yuyinfeng@xju.edu.cn).

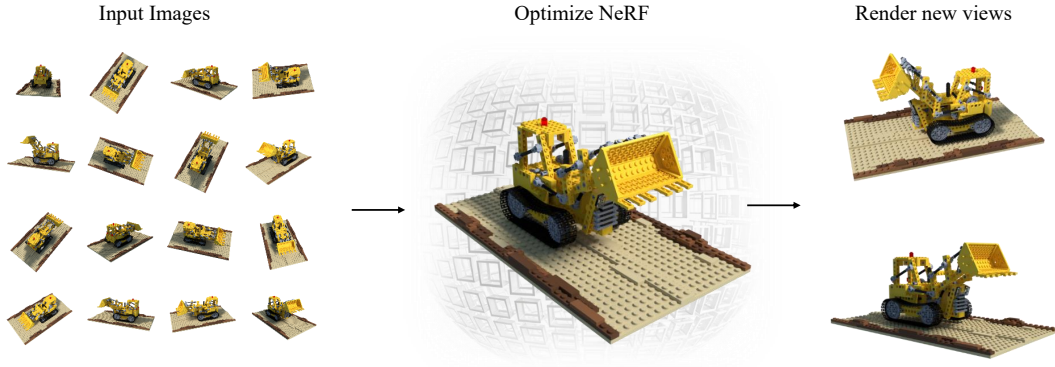


Fig. 1. Overview of the blur-aware radiance-field reconstruction task. Blurry input images are used to optimize the radiance field and synthesize restored novel views.

novel view synthesis [1]. Subsequent studies have further improved radiance-field representations in terms of rendering efficiency, representation capability, and view generalization.

B. Radiance Field Restoration Under Blur

Existing research on 3D reconstruction from blurry inputs can generally be divided into two categories. The first category is based on explicit degradation modeling. Methods such as Deblur-NeRF [4], DP-NeRF [5], and BAD-NeRF [6] simulate the physical degradation process by jointly optimizing blur kernels or continuous camera trajectories. Although these methods are physically interpretable, they rely heavily on parameterized assumptions, making it difficult to accurately fit the complex and non-uniform blur encountered in real-world scenarios. In addition, their training process is extremely time-consuming.

The second category is based on alternating optimization and prior guidance. To overcome the limitations of explicit modeling, methods such as PDRF [7], ExBluRF [8], Deep-DeblurRF [9], and AFDU-NeRF [10] introduce 2D image deblurring priors. Related Gaussian-splatting-based deblurring methods, including Deblurring-3DGS [11] and BAGS [12], further explore explicit scene representations under blur. By alternately optimizing the 2D deblurring network and the 3D radiance field, this strategy effectively combines the strong single-image restoration capability of 2D networks with the multi-view geometric consistency inherent in 3D representations.

C. Multimodal Perception and Dynamic Fusion

Related advances in embodied multimodal perception demonstrate the benefits of complementary sensory cues, collaborative observations, and external knowledge for robust navigation [13]–[20]. Dynamic gating and representation fusion have also been investigated for audio-visual source separation and incomplete multimodal understanding [21], [22], while multi-domain detail enhancement and attention-based feature fusion improve robustness in remote-sensing and speech applications [23]–[26]. Although these studies address tasks different from radiance-field restoration, their common principle of reliability-aware fusion motivates our

adaptive regulation of heterogeneous 2D restoration cues and 3D geometric evidence.

D. Error Accumulation in Iterative Restoration and Temporal State Modeling

Most existing alternating deblurring frameworks statically treat the reliability of rendered guidance. Blindly trusting unconverged 3D priors at early stages can easily cause serious error accumulation. On the other hand, although world models and state-space models have shown strong capability in representing latent system states and mitigating catastrophic forgetting in long-sequence tasks, their use in complex 2D–3D radiance field restoration remains largely unexplored. To address this gap, we are the first to formulate iterative radiance-field deblurring as a dynamic evolution system. We exploit a world model to implicitly perceive residual restoration states and introduce Mamba to propagate long-range states across iterations, thereby adaptively regulating the strength of radiance field priors and fundamentally blocking cumulative errors caused by unreliable guidance.

III. METHOD

A. Overall Framework

This paper proposes WS-NeRF, an adaptive deblurring Neural Radiance Field framework driven by Mamba-based world state awareness. By leveraging a multi-dimensional world state estimator for error perception and Mamba for long-range temporal memory, the framework employs an adaptive controller to dynamically modulate the fusion ratio between 2D deblurring priors and 3D radiance fields, thereby achieving stable and synergistic optimization. The overall architecture is shown in Fig. 2.

B. Multi-Dimensional World-State Observer

Traditional mean squared error is sensitive to high-frequency noise and cannot faithfully reflect the true quality of 3D reconstruction. Therefore, we build a multi-dimensional world-state observer that fuses high-level semantic cues and low-level physical features for state estimation. Acting as the sensing unit in the closed-loop control system, the observer innovatively combines deep

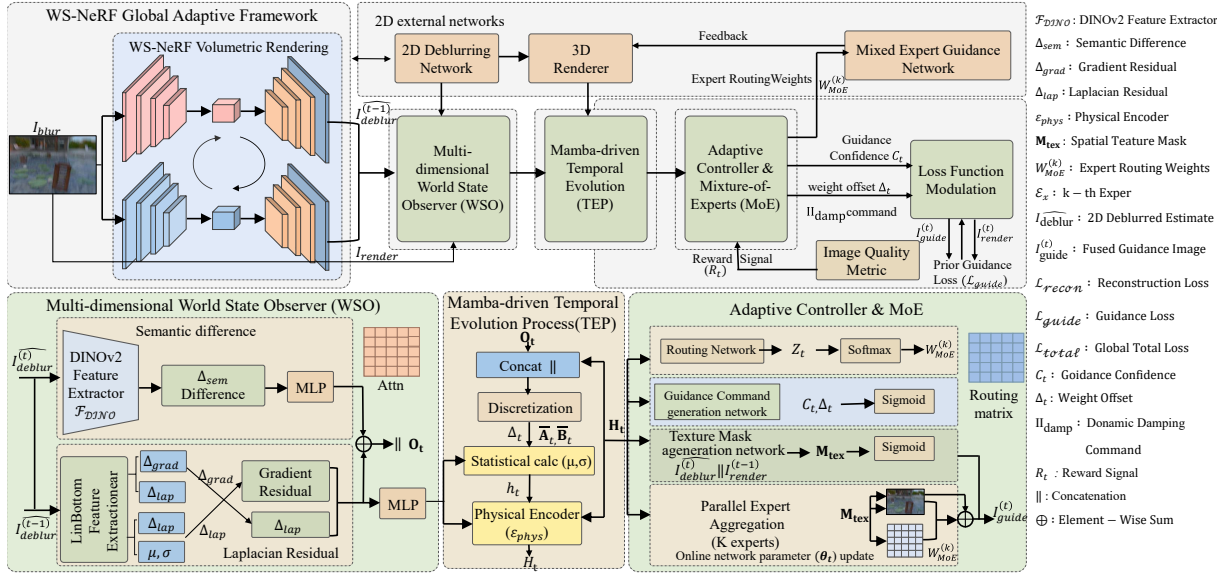


Fig. 2. Overall framework of WS-NeRF. The world-state observer, Mamba-driven temporal evolution process, and adaptive controller with mixture-of-experts jointly regulate prior guidance during radiance-field optimization.

semantic discrepancies extracted by a pretrained large vision model with low-level physical features derived from image gradients and Laplacian edges.

At the t -th iteration, the system receives the original blurry input I_{blur} , the current 2D deblurred estimate $\hat{I}_{deblur}^{(t)}$, and the rendered image from the 3D radiance field I_{render} . Through a dual-stream architecture, the observer extracts the instantaneous state O_t from both semantic and physical perspectives.

1) *Macroscopic Semantic and Microscopic Physical Dual-Stream Perception for State Estimation*: Macroscopic Semantic and Microscopic Physical Dual-Stream Perception To extract both global and local features, the estimator adopts a dual-stream architecture. Macroscopically, it leverages a pretrained DINOv2 model to extract high-dimensional features from the deblurred estimate and the rendered image, computing their absolute error norm Δ_{sem} . Being highly sensitive to global topological degradation, Δ_{sem} serves as an early warning to suppress 3D floaters and structural fragmentation caused by misleading 2D priors. Microscopically, Sobel and Laplacian operators are applied to extract first-order gradients and second-order high-frequency features, yielding low-level physical residuals Δ_{grad} and Δ_{lap} . These residuals precisely localize texture distortions, perfectly compensating for the semantic features' lack of fine-grained, high-frequency awareness.

2) *State Aggregation and Observation Generation*: After extracting the above features, the system computes global spatial statistics (mean and standard deviation) of the residuals. The physical residuals are encoded by dedicated encoders and concatenated with the statistics to form a comprehensive low-level physical state:

$$F_{phys} = [E_{grad}(\Delta_{grad}) \parallel \mu_{grad} \parallel \sigma_{grad} \parallel E_{lap}(\Delta_{lap}) \parallel \mu_{lap} \parallel \sigma_{lap}]. \quad (1)$$

Then, a multilayer perceptron fuses the physical state and the semantic discrepancy to generate the final instantaneous observation:

$$O_t = \text{MLP}([\Delta_{sem} \parallel F_{phys}]) \in \mathbb{R}^d. \quad (2)$$

This multi-dimensional observation escapes the limitations of single-direction pixel-level metrics.

C. Mamba-Based Temporal Evolution and Dynamic Adjustment

After obtaining the instantaneous observation O_t , directly making decisions based only on this signal would easily fall into the non-Markovian dilemma common in traditional alternating optimization. Therefore, WS-NeRF introduces a Mamba-based temporal propagation network as the state memory of the control system, enabling the framework to reference historical evolution trends and achieve temporally smooth, dynamically routed deblurring guidance.

1) *Selective State-Space Modeling and Gated Temporal Propagation*: The system treats the instantaneous observation O_t as sequence input and updates the basic temporal hidden state h_t via selective scanning based on a state-space model. To dynamically balance trust in current fluctuations and dependence on long-range memory, we further design an adaptive gated fusion mechanism on top of the Mamba output. Let H_{t-1} denote the final hidden state from the previous iteration. The dynamic gate g_t is computed as

$$g_t = \sigma(W_g[h_t, H_{t-1}] + b_g), \quad (3)$$

and the current coherent state is updated by

$$H_t = g_t \odot \tanh(W_o h_t) + (1 - g_t) \odot \tanh(W_h H_{t-1}). \quad (4)$$

When the current observation contains burst noise caused by misleading 2D features, the gate automatically suppresses

unreliable inputs and forces the system to rely more on the historical smooth state H_{t-1} .

2) *Dynamic Expert Weighting for Deblurring Prior Confidence*: The coherent state H_t implicitly encodes the dynamic evolution of error convergence. It is further fed into a dynamic adjustment module to produce preference logits for different deblurring experts. A temperature-controlled softmax is then used to obtain the routing weights $W_{\text{MoE}}^{(k)}$, which determine how aggressively the system should combine 3D priors and conservative 2D outputs at the current iteration.

D. Progress-Aware Mixture-of-Experts Decision

Based on H_t and the routing weights, the adaptive controller outputs the final control commands.

1) *Pseudo-Supervised Online Update from Environment Feedback*: The system predicts the guidance confidence C_t and a weight offset Δ_t from the coherent state. For online adaptation, the relative quality gain between consecutive rendered images is computed as an environment reward:

$$R_t = \text{Metric}(I_{\text{render}}^{(t)}) - \text{Metric}(I_{\text{render}}^{(t-1)}), \quad (5)$$

where $\text{Metric}(\cdot)$ denotes a no-reference image quality assessment function. When geometric corruption is detected ($R_t < 0$), the controller receives a penalty signal and rapidly decays the confidence to block the injection of erroneous priors.

2) *Mixture-of-Experts Guidance and Local Texture Fusion*: The controller maintains K deblurring strategy experts $\mathcal{E}_k(\cdot)$ in parallel. To achieve spatially aware fine-grained control, a convolutional layer predicts a spatial texture mask $M_{\text{tex}} \in [0, 1]^{H \times W}$:

$$M_{\text{tex}} = \text{Sigmoid}(\text{Conv}(\hat{I}_{\text{deblur}}^{(t)} \parallel I_{\text{render}}^{(t)})). \quad (6)$$

The final fused guidance image is then constructed as

$$I_{\text{guide}}^{(t)} = \sum_{k=1}^K W_{\text{MoE}}^{(k)} \left[M_{\text{tex}} \odot \mathcal{E}_k(I) + (1 - M_{\text{tex}}) \odot \hat{I}_{\text{deblur}}^{(t)} \right]. \quad (7)$$

This mechanism allows the system to aggressively inject details in high-confidence regions while enforcing conservative fusion in risky edge areas.

3) *Progress-Aware Dynamic Damping Trigger*: To prevent over-correction during late-stage optimization, the controller includes a progress-aware safety guard. The reconstruction quality gain is smoothed with an exponential moving average over a temporal window, and a damping indicator $I_{\text{damp}} \in \{0, 1\}$ is defined as

$$I_{\text{damp}} = \begin{cases} 1, & \text{if } \bar{R}_t < \epsilon \text{ or } \text{Var}(R_{t-W:t}) > \delta, \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where ϵ is a near-zero reward threshold and δ is an oscillation tolerance. When damping is activated, the confidence output is overridden by an exponential decay and the routing weights are forced toward a conservative expert distribution.

E. Joint Optimization Objective

To achieve stable joint optimization of the radiance field and camera trajectories, WS-NeRF formulates a compound loss function modulated by a closed-loop mechanism:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \alpha \cdot [(1 - \mathbb{I}_{\text{damp}}) \cdot C_t] \cdot \mathcal{L}_{\text{guide}} + \gamma \mathcal{L}_{\text{reg}} \quad (9)$$

Physical reconstruction loss ($\mathcal{L}_{\text{recon}}$) constrains the mean squared error (MSE) between the synthesized blurry rendering and the real blurry input, serving as the fundamental photometric anchor.

Prior guidance loss ($\mathcal{L}_{\text{guide}}$) leverages the fused image $I_{\text{guide}}^{(t)}$ to impose structural supervision, combining L_1 error and SSIM [27] on the intermediate sharp renderings to distill 2D high-frequency details into 3D space.

Closed-loop dynamic modulation adjusts the prior weights in real time via the confidence score C_t and the damping indicator $\mathbb{I}_{\text{damp}}$.

IV. EXPERIMENTS

A. Experimental Setup

1) *Datasets*: We adopt the benchmark dataset introduced by DeepDeblurRF [9]. Specifically, (1) BlurRF-Synth contains 65 training scenes and 10 test scenes, covering both 6-DOF motion blur and defocus blur. In addition, physical noise is explicitly injected in the RAW space to better approximate real-world degradation. (2) BlurRF-Real contains five real low-light scenes with severe motion blur and high-frequency noise. Each scene provides 20–40 multi-view images for training, while 3–5 clear images are strictly reserved for novel view synthesis evaluation. Representative samples are shown in Fig. 3. We evaluate synthetic scenes using PSNR, SSIM [27], and LPIPS [28], and employ NIQE [29] and ARNIQA [30] for blind quality assessment on real scenes. The full quantitative comparison and ablation summary are reported in Tables I and II.

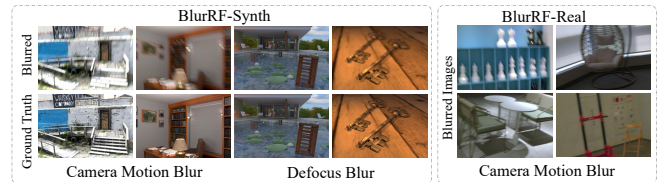


Fig. 3. Dataset examples used in the experiments, including synthetic motion blur, synthetic defocus blur, and real motion-blur scenes.

2) *Implementation Details*: WS-NeRF is implemented in PyTorch. To better support local feature extraction for the 2D deblurring network, we adopt image-patch-based ray sampling with a batch size of 4096 rays per iteration. The number of experts is set to $K = 3$. The Mamba state dimension is set to $d = 128$ [3]. We use Adam for optimization, with the radiance field learning rate exponentially decayed from 5×10^{-4} to 5×10^{-5} , and the adaptive controller learning rate fixed at 6×10^{-4} . Each scene is trained for 200k iterations on a single NVIDIA GPU. DINOv2 (ViT-Small) with frozen weights is used for high-level semantic

perception [2], NAFNet is adopted as the 2D deblurring executor, and the adaptive controller uses a dual-branch perception architecture with a shared 64-channel hidden layer and an additional Mamba state propagation module.

B. Result Analysis

Quantitatively, comparisons with existing mainstream baselines [4]–[9], [11], [12] demonstrate that WS-NeRF achieves strong overall performance in handling complex blur degradation, benefiting from its state-aware perception and adaptive decision-making mechanism. As shown in Table I, in motion blur scenarios, the proposed model attains an average PSNR of 30.04 and SSIM of 0.8699, consistently surpassing DeepDeblurRF. In defocus blur settings, WS-NeRF achieves superior LPIPS performance relative to the baseline methods. On the BlurRF-SB dataset, which involves more comprehensive degradation factors, our method exhibits a markedly better trade-off between perceptual quality and distortion than the competing approaches. Furthermore, under the real-world no-reference setting of BlurRF-Real, WS-NeRF demonstrates strong blind deblurring capability, achieving favorable performance in terms of NIQE.

Qualitatively, under severe nonlinear blur degradation, WS-NeRF achieves finer detail restoration while better preserving three-dimensional geometric structures. Visualization results on the BlurRF synthetic and BlurRF real datasets are shown in Fig. 4.

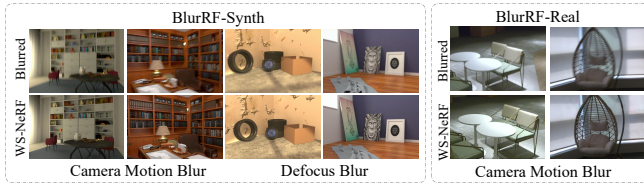


Fig. 4. Qualitative comparison on synthetic and real blurry radiance-field benchmarks. WS-NeRF restores sharper structures while preserving cross-view geometry.

C. Ablation Study



Fig. 5. Early-stage failure case in alternating radiance-field restoration. Unreliable rendered guidance can introduce artifacts and motivate state-aware adaptive control.

We further validate the contributions of world-state perception (WS), the Mamba temporal propagation module, and

the multi-strategy fusion mechanism. As shown in Table II, removing any component degrades performance on at least part of the benchmark suite. Among them, the multi-strategy fusion module contributes the largest single-module improvement, mainly because it dynamically balances the guidance ratio between 2D priors and 3D physical constraints. The Mamba module also yields a clear gain by establishing long-range temporal memory and reducing convergence oscillation. Fig. 5 further illustrates the early-stage artifacts caused by unreliable guidance, which motivates the adaptive damping design. Combined together, these modules form the complete WS-NeRF framework and deliver the most balanced overall performance.

V. DISCUSSION AND CONCLUSION

In this paper, we proposed WS-NeRF, an adaptive deblurring neural radiance field framework. By combining a multi-dimensional world-state evaluator with a Mamba-based temporal propagation network, the proposed method establishes a closed-loop dynamic routing mechanism for intelligent modulation of 2D deblurring priors and 3D physical constraints. The dynamic damping mechanism further suppresses systematic artifacts in alternating optimization. Extensive benchmark results indicate that WS-NeRF consistently improves reconstruction quality in extreme blur scenarios and significantly broadens the applicability of radiance fields in realistic degraded environments.

Limitations include sensitivity to initial hyperparameters and poor 2D prior generalization in textureless or underexposed scenes. Future work will investigate prior-free generative physical inversion and joint pose-radiance calibration under severe degradation for enhanced stability.

ACKNOWLEDGMENT

This research was financially supported by Xinjiang University Training Program of Innovation and Entrepreneurship for Undergraduates (Grant No. 202510755141).

REFERENCES

- [1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing scenes as neural radiance fields for view synthesis,” in *Proceedings of the European Conference on Computer Vision*, pp. 405–421, 2020.
- [2] M. Oquab *et al.*, “DINOv2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023.
- [3] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023.
- [4] L. Ma, X. Li, J. Liao, Q. Zhang, X. Wang, J. Wang, and P. V. Sander, “Deblur-NeRF: Neural radiance fields from blurry images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12861–12870, 2022.
- [5] D. Lee, M. Lee, C. Shin, and S. Lee, “DP-NeRF: Deblurred neural radiance field with physical scene priors,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12386–12396, 2023.
- [6] P. Wang, L. Zhao, R. Ma, and P. Liu, “BAD-NeRF: Bundle adjusted deblur neural radiance fields,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4170–4179, 2023.
- [7] C. Peng and R. Chellappa, “PDRF: Progressively deblurring radiance field for fast scene reconstruction from blurry images,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 2029–2037, 2023.

TABLE I

QUANTITATIVE RESULTS OF WS-NeRF AND COMPETING METHODS ACROSS ALL BENCHMARKS. THE BEST METRIC IN EACH COLUMN IS SHOWN IN BOLD.

Model	Camera Motion Blur			Defocus Blur			BlurRF-Real		BlurRF-SB		
	PSNR ($\uparrow$)	SSIM ($\uparrow$)	LPIPS ($\downarrow$)	PSNR ($\uparrow$)	SSIM ($\uparrow$)	LPIPS ($\downarrow$)	NIQE ($\downarrow$)	ARNIQA ($\uparrow$)	PSNR ($\uparrow$)	SSIM ($\uparrow$)	LPIPS ($\downarrow$)
Deblur-NeRF	27.67	0.8340	0.1450	30.03	0.8727	0.1137	6.341	0.279	25.43	0.7264	0.2319
BAD-NeRF	21.74	0.5298	0.3969	—	—	—	9.908	0.237	21.97	0.5231	0.4731
DP-NeRF	28.03	0.8412	0.1267	30.15	0.8763	0.0991	6.498	0.266	27.07	0.7940	0.1795
ExBluRF	26.56	0.7823	0.1955	—	—	—	7.479	0.279	24.16	0.6649	0.2659
PDRF-10	28.33	0.8435	0.1495	30.03	0.8750	0.1225	6.243	0.266	26.66	0.7730	0.2124
Deblurring-3DGS	26.30	0.7729	0.1728	29.37	0.8545	0.1470	6.086	0.262	23.87	0.6459	0.2475
BAGS	27.41	0.8108	0.1382	29.90	0.8638	0.1152	5.837	0.290	24.49	0.6748	0.2391
DeepDeblurRF	29.93	0.8663	0.1163	32.49	0.9054	0.0976	6.086	0.262	29.06	0.8489	0.1516
WS-NeRF (Ours)	30.04	0.8699	0.1152	32.50	0.9046	0.0966	6.049	0.347	29.24	0.8519	0.1496

TABLE II

ABLATION STUDY OF WS-NeRF. THE BEST METRIC IN EACH COLUMN IS SHOWN IN BOLD.

Variant	Camera Motion Blur			Defocus Blur			BlurRF-SB		
	PSNR ($\uparrow$)	SSIM ($\uparrow$)	LPIPS ($\downarrow$)	PSNR ($\uparrow$)	SSIM ($\uparrow$)	LPIPS ($\downarrow$)	PSNR ($\uparrow$)	SSIM ($\uparrow$)	LPIPS ($\downarrow$)
w/o World State	29.84	0.8687	0.1164	32.55	0.9048	0.0967	29.26	0.8507	0.1498
w/o Mamba	29.91	0.8672	0.1159	32.44	0.9051	0.0969	29.05	0.8502	0.1507
w/o Multi-strategy Fusion	29.93	0.8663	0.1163	32.49	0.9054	0.0976	29.06	0.8489	0.1516
Full Model	30.04	0.8699	0.1152	32.50	0.9046	0.0966	29.24	0.8519	0.1496

- [8] D. Lee, J. Oh, J. Rim, S. Cho, and K. M. Lee, “ExBluRF: Efficient radiance fields for extreme motion blurred images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17639–17648, 2023.
- [9] H. Choi, H. Yang, J. Han, and S. Cho, “Exploiting deblurring networks for radiance fields,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6012–6021, 2025.
- [10] Q. Cao, X. Ma, C. Wang, H. Kuang, and X. Liu, “AFDU-NeRF: Deblurring NeRF with adaptive-aware fusion and iterative data updating,” in *Proceedings of the IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology*, pp. 493–496, 2025.
- [11] B. Lee, H. Lee, X. Sun, U. Ali, and E. Park, “Deblurring 3D gaussian splatting,” in *Computer Vision – ECCV 2024*, pp. 127–143, 2025.
- [12] C. Peng, Y. Tang, Y. Zhou, N. Wang, X. Liu, D. Li, and R. Chellappa, “BAGS: Blur agnostic gaussian splatting through multi-scale kernel modeling,” in *Computer Vision – ECCV 2024*, pp. 293–310, 2025.
- [13] Y. Yu, W. Huang, F. Sun, C. Chen, Y. Wang, and X. Liu, “Sound adversarial audio-visual navigation,” in *The Tenth International Conference on Learning Representations*, 2022.
- [14] Y. Yu, C. Chen, L. Cao, F. Yang, and F. Sun, “Measuring acoustics with collaborative multiple agents,” in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 335–343, 2023.
- [15] D. Yang, Y. Yu, and L. Wang, “Beyond textual knowledge: Leveraging multimodal knowledge bases for enhancing vision-and-language navigation,” *Information Processing & Management*, vol. 63, no. 6, p. 104766, 2026.
- [16] Y. Yu and D. Yang, “DOPE: Dual object perception-enhancement network for vision-and-language navigation,” in *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 1739–1748, 2025.
- [17] J. Li, Y. Yu, L. Wang, F. Sun, and W. Zheng, “Audio-guided dynamic modality fusion with stereo-aware attention for audio-visual navigation,” in *International Conference on Neural Information Processing*, pp. 346–359, Springer, 2025.
- [18] H. Zhang, Y. Yu, L. Wang, F. Sun, and W. Zheng, “Advancing audio-visual navigation through multi-agent collaboration in 3d environments,” in *International Conference on Neural Information Processing*, pp. 502–516, Springer, 2025.
- [19] H. Zhang, Y. Yu, L. Wang, F. Sun, and W. Zheng, “Iterative residual cross-attention mechanism: An integrated approach for audio-visual navigation tasks,” *arXiv preprint arXiv:2509.25652*, 2025.
- [20] Y. Yu, H. Zhang, and M. Zhu, “Dynamic multi-target fusion for efficient audio-visual navigation,” *arXiv preprint arXiv:2509.21377*, 2025.
- [21] Y. Yu and S. Sun, “DGFNet: End-to-end audio-visual source separation based on dynamic gating fusion,” in *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 1730–1738, 2025.
- [22] X. Wang, L. Wang, Y. Yu, and X. Jiao, “Modality-invariant bidirectional temporal representation distillation network for missing multimodal sentiment analysis,” in *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, IEEE, 2025.
- [23] J. Fu, Y. Yu, and L. Wang, “FSDENet: A frequency and spatial domains based detail enhancement network for remote sensing semantic segmentation,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.
- [24] A. Mattarsun, L. Wang, and Y. Yu, “BSS-CFFMA: Cross-domain feature fusion and multi-attention speech enhancement network based on self-supervised embedding,” in *2024 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3589–3594, IEEE, 2024.
- [25] J. Zhang, L. Wang, Y. Yu, and M. Xu, “Nonlinear regularization decoding method for speech recognition,” *Sensors*, vol. 24, no. 12, p. 3846, 2024.
- [26] Y. Cao, Y. Li, L. Wang, and Y. Yu, “VNet: A GAN-based multi-tier discriminator network for speech synthesis vocoders,” in *2024 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 4384–4389, IEEE, 2024.
- [27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.
- [29] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a completely blind image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
- [30] L. Agnolucci, L. Galteri, M. Bertini, and A. Del Bimbo, “ARNIQA: Learning distortion manifold for image quality assessment,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 189–198, Jan. 2024.