ClarityNetSR: End-to-End Joint Deblurring and Super-Resolution for Real-World Degraded Frame Enhancement

Faisal Ayub Khan, Soo Young Shin*

Department of IT Convergence Engineering, Kumoh national Institute of Technology, Gumi, South Korea

faisal@kumoh.ac.kr , *wdragon@kumoh.ac.kr

Abstract

Frames captured in real-world scenarios such as CCTV/surveillance, UAV/drone footage, and low-bitrate streaming are often degraded by a mixture of motion/defocus blur and limited spatial resolution, which reduces the interpretability of critical details. In this paper, we propose ClarityNetSR, an end-to-end joint Deblur-and-Super-Resolution (Deblur-SR) framework that directly reconstructs a sharp high-resolution frame from degraded observations. When short temporal neighborhoods are available, ClarityNetSR leverages alignment-aware multi-frame fusion to exploit temporal redundancy and recover finer details while reducing misalignment-induced ghosting. We train the model under realistic mixed degradations and evaluate it against SR-only, deblur-only, and two-stage baselines using standard fidelity and perceptual metrics. Results demonstrate improved sharpness and more coherent textures with fewer restoration artifacts, supporting the practical use of ClarityNetSR for real-world frame enhancement.

1. Introduction

Real-world imaging systems frequently produce frames degraded by blur (motion/defocus), low spatial resolution, and compression artifacts [1]. This is common in settings such as CCTV/surveillance, dashcams/bodycams, drone imagery, and low-bandwidth streaming [2], where improving the clarity of a single frame can significantly increase interpretability [3]. Existing approaches typically treat deblurring and super-resolution (SR) as separate problems; however, SR applied to blurred inputs can amplify artifacts, deblurring alone cannot recover missing high-frequency details, and sequential pipelines (deblur → SR) often accumulate errors [4]. To address these challenges, we propose ClarityNetSR, Fig: [1], an end-to-end joint Deblur-and-Super-Resolution (Deblur-SR) framework that directly reconstructs a sharp high-resolution frame from degraded observations. When adjacent frames are available, ClarityNetSR leverages temporal redundancy through alignment-aware fusion to better recover fine details while reducing ghosting artifacts. Experiments against SR-only, deblur-only, and two-stage baselines demonstrate improved sharpness and detail reconstruction under realistic degradations involving blur, downsampling, and compression.

2. System Model

Fig. 1 illustrates the overall pipeline of ClarityNetSR, an end-to-end framework that performs joint deblurring and super-resolution (Deblur-SR) for real-world degraded frames. We consider a target degraded frame and (optionally) a small temporal neighborhood from the same capture source (e.g., CCTV, dashcam, UAV footage, low-bitrate streaming).

2.1. Degraded Observation Model

Let $x_t \in R^{H\times W\times 3}$ denote the unknown sharp high-resolution (HR) latent frame at time t. The observed low-quality low-resolution (LR) frame $y_t \in R^{h\times w\times 3}$ is modeled as a composition of blur, downsampling, and codec/noise corruption:

$$y_t = C((x_t * k_t) \downarrow_s + n_t) \qquad (1)$$

where, k_t is an unknown blur kernel (motion/defocus), * denotes convolution, $\downarrow_s$ is a downsampling operator with scale factor s, n_t models sensor noise and residual perturbations, and $C(.)$ represents compression distortion (e.g., blocking/ringing artifacts). This model reflects practical conditions where blur and resolution loss co-occur and are further affected by compression [4].

2.2. Multi-Frame Input (Optional Temporal Support)

To improve restoration of ambiguous details (e.g., faces, text edges), we exploit temporal redundancy when available [5]. Define a temporal window of $2K$+1 frames centered at t:

$$y_t = \{y_{t-k}, \dots, y_{t-1}, y_t, y_{t+1}, \dots, y_{t+k}\} \qquad (2)$$

When only one single frame is available, the ClarityNetSR model reduces to the special case which is $K = 0$.

2.3. Multi-Frame Input (Optional Temporal Support)

Neighboring frames are not spatially aligned due to motion, parallax, and occlusions. We therefore introduce an alignment operator to map each neighbor y_{t+i} to the reference time t. Let $W(.;\theta y_{t+i})$ denote an alignment function (e.g., deformable alignment or flow-guided warping) parameterized by θ_{t+i}:

$$\bar{y}_{t+i} = W(y_{t+i}\,;\theta_{t+i}), \quad i \in -K, K \qquad (3)$$

Each aligned frame is encoded into feature space using a shared encoder $E(\cdot)$:

$$f_{t+i} = E(\bar{y}_{t+i}) \qquad (4)$$

The set of aligned features is fused into a single representation via a fusion module $\Phi(.)$ (e.g., attention-based aggregation):

$$F_t = \Phi(\{f_{t-K}, \dots, f_{t,\dots}, f_{t+K}\}) \qquad (5)$$

This alignment-aware fusion helps recover fine details that may be missing or blurred in the reference frame y_t.

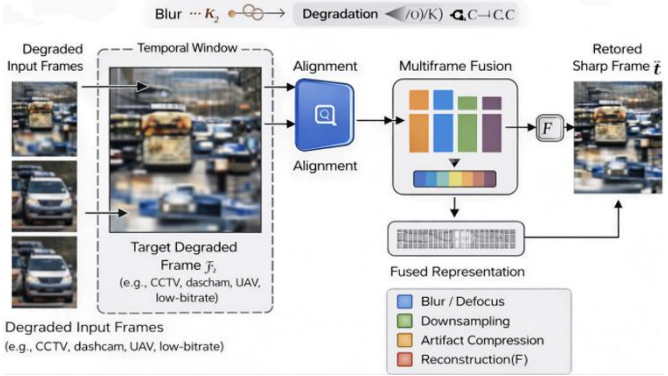

Figure 1: The Proposed System model for ClarityNetSR.

2.4. End-to-End Joint Deblur-SR Reconstruction

Given the fused representation F_t, ClarityNetSR reconstructs the HR output directly in a single network pass:

$$\hat{x}_t = G(F_t) \quad (6)$$

where $G(\cdot)$is the decoder/reconstruction head that jointly accounts for blur removal and resolution enhancement. So unlike sequential pipelines (deblur→SR), the proposed formulation learns a unified restoration mapping optimized for the final HR sharp output.

3. Dataset and Training Setup

We train ClarityNetSR end-to-end for joint deblurring and super-resolution using paired clean HR frames x_t and synthetically degraded observations y_t. Dataset contained approximately $N \approx 10{,}000$ high-quality RGB frames sampled from diverse scenes (day/night, indoor/outdoor, slow/fast motion). Following the degradation model in (1), each HR frame is corrupted by a randomized combination of: (i) motion blur and defocus blur via a sampled kernel k_t, (ii) downsampling by scale factor $s \in \{2,4\}$, (iii) additive noise n_t, and (iv) compression artifacts simulated through re-encoding at multiple quality levels. For the multi-frame setting, we sample a temporal window $\mathcal{Y}_t = \{y_{t-K}, \dots, y_{t+K}\}$ from the same sequence (typically $K = 2$, i.e., 5 frames). This enables the network to exploit temporal redundancy while learning alignment-aware fusion under realistic motion. The network is optimized using a weighted objective:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec} + \lambda_p \mathcal{L}_{perc} + \lambda_e \mathcal{L}_{edge} (+\lambda_t \mathcal{L}_{temp}) \quad (7)$$

where $\mathcal{L}_{\text{rec}}$ is an L1/Charbonnier reconstruction loss, $\mathcal{L}_{\text{perc}}$ is a perceptual loss computed in a fixed feature space $\psi(\cdot)$ and $\mathcal{L}_{\text{edge}}$ enforces sharper gradients; $\mathcal{L}_{\text{temp}}$is optionally used to reduce ghosting and improve temporal stability. We employ AdamW optimization with an initial learning rate of 1×10^{-4}(cosine decay), standard augmentations (random crops, flips, temporal reversal), and train for a fixed schedule (e.g., 200–300 epochs) with batch size determined by GPU memory.

4. Dataset and Training Setup

We evaluate ClarityNetSR on held-out data under both synthetic and realistic degradations that reflect operational imagery. Performance is reported using full-reference quality metrics: PSNR and SSIM, together with the perceptual metric LPIPS to capture visual fidelity beyond pixel-wise similarity. For multi-frame inputs, we additionally assess temporal robustness using a consistency score computed between reconstructed outputs and motion-aligned neighboring reconstructions, reflecting flicker and ghosting behavior. We compare ClarityNetSR against three categories of baselines: (i) SR-only methods applied directly to degraded inputs (ii) Deblur-only methods followed by bicubic upsampling, and (iii) Two-stage pipelines (deblur →SR). Across evaluations, ClarityNetSR achieves improved sharpness and detail recovery while reducing common artifacts such as ringing, residual blur, and misalignment-induced ghosting. Qualitative results show clearer edges and more coherent textures in challenging regions (e.g., faces, text, object boundaries), supporting benefit of end-to-end joint restoration with alignment-aware multi-frame fusion.

5. Conclusion

This paper presented ClarityNetSR, an end-to-end framework for joint deblurring and super-resolution targeting real-world degraded frames commonly encountered in surveillance, vehicular cameras, drones, and bandwidth-limited video systems. Unlike methods that treat deblurring and SR separately or rely on sequential pipelines, ClarityNetSR performs unified Deblur-SR reconstruction in a single network, reducing error propagation and suppressing artifact amplification. By incorporating alignment-aware multi-frame fusion, the proposed approach exploits temporal redundancy to recover details that are ambiguous or missing in a single degraded frame, while reducing ghosting effects caused by misalignment. Quantitative and qualitative evaluations against strong baselines show improved sharpness and detail reconstruction under mixed degradations involving blur, downsampling, and compression. Future work will explore broader real-world degradations (e.g., low-light noise and rolling-shutter effects), improved uncertainty estimation for reliability, and deployment-oriented optimization techniques.

6. Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00553810, 50%) This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program(IITP-2025-RS-2022-00156394, 50%) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation)

References

[1] S. Sinaei, D. Iwai and K. Sato, "Artificial Blur Effect for Optical See-Through Near-Eye Displays," in *IEEE Access*, vol. 13, pp. 140382-140391, 2025, doi: 10.1109/ACCESS.2025.3595907.

[2] H. H. Mehdi and S. Y. Shin, “Smart DRX Wakeup Signal Control with Latency-Aware DCP Signaling in B5G/6G Multi-Service Scenarios,” in Proc. 2025 Korea Institute of Communications and Information Sciences (KICS) Summer Conference (2025), Jeju Shinhwa World, Seogwipo-si, Jeju, Republic of Korea, Jun. 2025, pp. 659–660.

[3] A. Nahli *et al*., "Camera Sensor Raw Data-Driven Video Blur Effect Prevention: Dataset and Study," in *IEEE Access*, vol. 13, pp. 184762-184774, 2025, doi: 10.1109/ACCESS.2025.3622993.

[4] X. Xu, H. Liu, Y. Li and Y. Zhou, "Image deblurring with blur kernel estimation in RGB channels," *2016 IEEE International Conference on Digital Signal Processing (DSP)*, Beijing, China, 2016, pp. 681-684, doi: 10.1109/ICDSP.2016.7868645.

[5] F. A. Khan and S. Y. Shin, "Deep Learning Based Active Noise Cancellation for Reducing UAV Propeller Sound," 2024 15th International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, Republic of, 2024, pp. 2072-2077, doi: 10.1109/ICTC62082.2024.10827319