Clarity Without Reconstruction: Trust-Guided Temporal Fusion for Perceptual Enhancement

Andrew Chen

Abstract—In distant-capture video scenarios such as conferencing, faces often occupy only a small region of the frame, making them difficult to interpret when displayed on high-resolution screens. Conventional super-resolution and sharpening methods can increase apparent detail, but may also introduce hallucinated structure or temporal artifacts. We investigate whether temporal redundancy can instead be used to improve perceived clarity without reconstructing new spatial detail. We propose a lightweight temporal enhancement pipeline that aligns neighboring frames, estimates a per-pixel trust map from photometric agreement, motion magnitude, edge evidence, and temporal reliability, and uses this trust signal to guide temporal fusion and optional sharpening. In this work, we suggest that reliability-guided temporal fusion provides a practical operating point for improving the perceptual quality of low-resolution video.

Index Terms—Computational Photography, Video Processing, Temporal Fusion, Perceptual Enhancement

1 Introduction

In many video imaging scenarios, subjects of interest occupy only a small portion of the frame. For example, in video conferencing or surveillance settings, a person's face may appear at only a few hundred pixels even when the captured video is high resolution. In such cases, improving perceived visual clarity is challenging. Traditional approaches attempt to recover spatial detail through super-resolution or sharpening techniques [1], [2]. While these methods can produce visually sharp images, they often rely on hallucinated detail or aggressive contrast enhancement, which can introduce temporal artifacts when applied independently to each frame of a video.

Human perception of video quality is influenced not only by spatial resolution but also by temporal consistency. Frame-to-frame fluctuations such as flicker or unstable edges can reduce perceived clarity and visual comfort even when individual frames appear sharp [3], [4]. Consequently, improving perceived clarity in video requires leveraging temporal information while ensuring that enhancements remain temporally coherent. Furthermore, human vision research shows that perceived sharpness depends strongly on contrast structure and temporal stability rather than spatial resolution alone. Studies on contrast sensitivity and perceptual image quality demonstrate that temporally stable, high-contrast structures are more likely to be perceived as sharp and interpretable even when spatial detail is limited [5].

In this work, we investigate whether temporal redundancy can be used to improve perceived clarity without introducing instability. We propose a lightweight temporal processing pipeline that selectively integrates information across frames to reinforce stable structures while suppressing unreliable temporal evidence. Rather than synthesizing new spatial detail, the method focuses on enhancing temporally consistent features that contribute to perceptual clarity.

We evaluate the proposed approach on video sequences where faces occupy a small region of the frame. Using downscaled inputs derived from high-resolution ground truth, we compare our method against several baselines including bicubic interpolation, per-frame sharpening, naive temporal averaging, and a learning-based super-resolution model. Quantitative evaluation includes reconstruction metrics, region of interest fidelity measures, and temporal stability metrics such as flicker magnitude and edge variance.

Our results show that the proposed approach maintains reconstruction quality comparable to spatial baselines while reducing temporal instability relative to per-frame enhancement methods. Compared to naive temporal fusion, the method preserves higher structural fidelity while still improving temporal consistency. These findings suggest that selectively leveraging temporal information can improve perceptual clarity in video without relying on hallucinated detail or aggressive sharpening.

• *Andrew Chen is a student enrolling in the SCPD program of Stanford University, Stanford, CA, 94305.*
E-mail: ajychen@stanford.edu

2 Related Work

Improving image quality using information from multiple frames has been widely studied in both computational photography and video processing. Our work situates closely to previous research in multi-frame fusion, temporal reconstruction, and perceptual video quality assessment.

2.1 Multi-frame fusion and burst photography

Multi-frame fusion techniques combine multiple aligned frames to improve image quality by aggregating signal across observations. Burst photography pipelines have demonstrated that combining short image sequences can substantially reduce noise and improve effective image quality in handheld cameras. For example, Hasinoff et al. [6] introduced a practical burst photography pipeline for mobile imaging that aligns and merges multiple frames to produce higher quality HDR and low-light images. Similarly, Wronski et al. [7] showed that multi-frame alignment and confidence-weighted fusion can recover additional spatial detail in handheld multi-frame super-resolution. These approaches highlight the benefits of aggregating temporally

redundant observations while weighting contributions according to alignment reliability. However, burst pipelines typically operate on short frame sequences to produce a single reconstructed image and do not address the stability challenges of continuous video processing.

2.2 Temporal reconstruction and motion-compensated filtering

In continuous video streams, temporal information is commonly reused to improve signal quality through motion-compensated filtering. Classical approaches such as VBM4D [8] perform spatiotemporal denoising by grouping similar patches across frames and jointly filtering them in a transform domain. More recent reconstruction pipelines in graphics and video restoration similarly accumulate information from previous frames while applying reliability checks to avoid artifacts such as ghosting and temporal instability. For example, spatiotemporal variance-guided filtering (SVGF) accumulates temporally reprojected samples while using variance estimates to control the contribution of historical information [9]. These methods demonstrate that temporal accumulation can significantly improve signal quality, but require mechanisms to determine when temporal evidence should be trusted.

2.3 Temporal artifacts and perceptual video quality

While temporal aggregation can improve signal fidelity, it may also introduce artifacts such as ghosting, flicker, or unstable edges when motion or alignment errors occur. Prior work in image and video quality assessment has emphasized that perceived visual quality depends strongly on structural fidelity and temporal stability. Structural similarity metrics (SSIM) capture perceptually relevant differences in image structure [3], while studies of human contrast sensitivity show that the perception of sharpness is strongly influenced by contrast structure rather than spatial resolution alone [5]. These findings suggest that improving perceived clarity requires not only enhancing spatial structure but also maintaining temporal consistency.

2.4 Our work: Reliability-guided temporal aggregation

Across these areas, a common principle is that temporal information should be aggregated only when it is sufficiently reliable. Hence, we build on this idea by explicitly modeling per-pixel temporal reliability through a trust function derived from photometric consistency, motion magnitude, edge evidence, and accumulation stability. This is then used to guide both temporal accumulation and perceptual enhancement, which enables the possibility to reinforce temporally stable structures while suppressing unstable regions that may lead to visual artifacts.

3 METHOD

Our goal is to improve perceived visual clarity in video sequences where the subject occupies a small region of the frame. Rather than attempting to synthesize new spatial detail, we aim to enhance temporally consistent structures across frames while suppressing unreliable temporal evidence.

Given an input video sequence, we first spatially downsample frames to simulate low-resolution observations. The overall pipeline is illustrated in Fig. 1: Temporal information from neighboring frames is then aligned and selectively aggregated according to a per-pixel trust estimate. The resulting temporally fused frame is optionally enhanced through trust-guided sharpening to reinforce stable visual structures.

3.1 Temporal Alignment

To leverage temporal redundancy, frames must first be aligned to a common reference frame. For each frame I_t, the accumulated history buffer (encoding past frames) is warped into the current frame's coordinate system using a 2D affine motion estimate from ECC. This alignment allows corresponding image structures across frames to be aggregated while minimizing spatial misalignment artifacts.

Because motion estimation may be imperfect, especially in regions with large motion or occlusion, temporal contributions must be weighted according to their reliability. We therefore compute a trust map that determines how strongly information from each frame should contribute to the reconstruction.

3.2 Trust Map Estimation

We define a per-pixel trust value that estimates the reliability of temporal evidence. The trust value incorporates several factors including photometric consistency between frames, motion magnitude, local edge strength, temporal clamping, and edge-direction stability. We describe each component below, then present the combined formulation.

3.2.1 Frame alignment

Before computing any trust signal, the accumulated history buffer is warped into the current frame's coordinate system. We estimate a 2D affine transformation between consecutive luminance frames using the Enhanced Correlation Coefficient (ECC) algorithm [10], applied to Gaussian-smoothed, spatially downscaled copies of the luminance channel for efficiency. The resulting 2×3 warp matrix is used to warp the full-resolution history buffer, and a binary valid mask V is produced to flag pixels that fall outside the warped field of view.

3.2.2 Photometric difference

Let Y_t denote the luminance of the current frame and $\hat{Y}_t$ the luminance of the aligned, temporally accumulated history. Before measuring the difference, the warped history luminance is optionally clamped to the local statistics of the current frame. Specifically, a box filter of size $K \times K$ yields the local mean μ and standard deviation σ of Y_t, and the warped history is clamped element-wise to the interval $[\mu - k_\sigma \sigma,\ \mu + k_\sigma \sigma]$, producing $\hat{Y}_t^{\mathrm{c}}$. The photometric difference map is then

$$\mathrm{diff}(p) = \left| Y_t(p) - \hat{Y}_t^{\mathrm{c}}(p) \right| \tag{1}$$

Fig. 1. The general processing pipeline of our work. The sharpening block is optional.

3.2.3 Clamp amount

The clamping step also records how far the history luminance was shifted:

$$\text{clamp}(p) = \left|\hat{Y}_t(p) - \hat{Y}_t^{\text{c}}(p)\right| \tag{2}$$

A large clamp value indicates the temporal buffer had drifted significantly from the local statistics of the current frame, suggesting unreliable temporal evidence.

3.2.4 Motion magnitude

A global motion scalar is derived from the affine warp matrix $\mathbf{W} \in \mathbb{R}^{2\times 3}$ by measuring its deviation from the identity:

$$\begin{aligned} \boldsymbol{\Delta} &= \mathbf{W} - \mathbf{I}_{2\times 3} \\ \text{motion} &= \min\Big(\frac{\|\boldsymbol{\Delta}_{\text{RS}}\|_F}{0.1} + \frac{\|\boldsymbol{\Delta}_{\text{T}}\|_2}{10},\ 1\Big) \end{aligned} \tag{3}$$

where $\boldsymbol{\Delta}_{\text{RS}}$ is the upper-left 2×2 rotation/scale residual and $\boldsymbol{\Delta}_{\text{T}}$ is the two-element translation residual. This scalar is broadcast to a spatially constant map. If alignment fails, motion is set to 1 (maximum), causing the trust to drop to near zero so that the frame passes through without temporal fusion.

3.2.5 Edge magnitude

The normalised Sobel edge magnitude of the current luminance frame is

$$\text{edge}_{\text{mag}}(p) = \frac{\sqrt{g_x(p)^2 + g_y(p)^2}}{\max_{p'} \sqrt{g_x(p')^2 + g_y(p')^2}} \tag{4}$$

where g_x and g_y are horizontal and vertical Sobel responses. This term acts as a multiplicative boost: regions with strong edges receive slightly higher trust because they carry structural detail that benefits from temporal averaging.

3.2.6 Reliability

The reliability map tracks the temporal consistency of edge directions via an exponential moving average (EMA). At each frame the raw gradient components g_x, g_y and their magnitude $|g| = \sqrt{g_x^2 + g_y^2}$ are incorporated as

$$\begin{aligned} \overline{g}_x &\leftarrow \beta\, \overline{g}_x + (1-\beta)\, g_x \\ \overline{g}_y &\leftarrow \beta\, \overline{g}_y + (1-\beta)\, g_y \\ \overline{|g|} &\leftarrow \beta\, \overline{|g|} + (1-\beta)\, |g| \end{aligned} \tag{5}$$

where β controls the memory of the EMA. The reliability is the ratio of the magnitude of the vector average to the scalar average:

$$\text{reliability}(p) = \frac{\sqrt{\overline{g}_x(p)^2 + \overline{g}_y(p)^2}}{\overline{|g|}(p) + \epsilon} \tag{6}$$

When edges are temporally stable (consistent direction across frames), this ratio approaches 1. Flickering edges cause the vector average to cancel while the scalar average persists, driving the ratio toward 0. In flat regions where the edge magnitude is below a small threshold, reliability is set to 1 to avoid penalizing textureless areas.

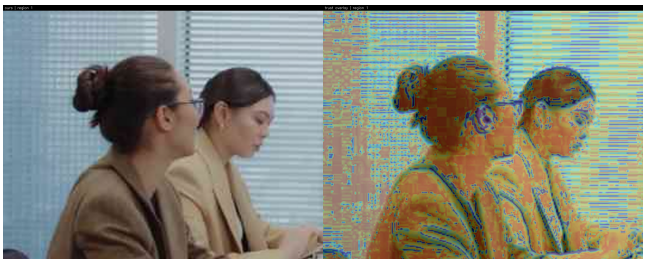

Fig. 2. Left: Region Processed with our temporal pipeline (no sharpening); Right: Trust map overlaid on original downscaled input

3.2.7 Combined trust

The final per-pixel trust map is

$$\begin{aligned} \text{trust}(p) = \exp\!\big(-k_{\text{diff}}\ \text{diff}(p)\big)\ \exp\!\big(-k_{\text{motion}}\ \text{motion}(p)\big) \\ \big(1 + w_{\text{edge}}\ \text{edge}_{\text{mag}}(p)\big) \\ \times\ \exp\!\big(-k_{\text{clamp}}\ \text{clamp}(p)\big)\ \text{reliability}(p)^{\gamma}\ V(p) \end{aligned} \tag{7}$$

where $V(p)$ is the binary valid mask from the affine warp. Each exponential factor acts as a soft gate: large photometric differences, high motion, or significant clamping independently suppress trust toward zero. The edge magnitude term provides a mild additive boost, and the reliability power-law term down-weights temporally unstable edges. The result is clipped to $[0, 1]$.

High trust drives strong temporal averaging (noise reduction and detail recovery via the history buffer), while low trust causes the pipeline to fall back to the current frame, preventing ghosting artefacts.

3.3 Temporal Fusion

Using the estimated trust map, aligned frames are aggregated through a weighted temporal fusion process. Let H_t be the fused (accumulated) frame at time t, $\widehat{H}_{t-1}$ the previous fused frame warped into the current frame, I_t the current frame, and $\text{trust}(p)$ the per-pixel trust at time t. At each time step the output is a single trust-weighted blend between warped history and current frame:

$$H_t(p) = \big(\alpha \cdot \text{trust}(p)\big)\, \widehat{H}_{t-1}(p) + \big(1 - \alpha \cdot \text{trust}(p)\big)\, I_t(p) \tag{8}$$

where $\alpha \in [0, 1]$ is a fixed history strength (e.g. 0.6–0.9). High trust increases the weight on the warped history (temporal smoothing); low trust reverts to the current frame, reducing ghosting. Compared to naive temporal averaging, this approach suppresses contributions that deviate from the reference observation.

3.4 Trust-Guided Enhancement

After temporal fusion, an optional sharpening stage can be applied to improve local contrast and enhance visible structures. Rather than applying sharpening uniformly, the enhancement is modulated by the trust map to avoid amplifying unstable regions.

In regions with high temporal trust, sharpening can safely reinforce structural features such as edges. Conversely, in regions with low trust, sharpening is suppressed to prevent amplification of noise or temporal artifacts. This selective enhancement allows the system to increase perceived clarity while maintaining temporal stability.

The final output frame is therefore produced by combining temporally fused information with trust-guided enhancement, yielding improved structural visibility without introducing significant flicker or instability.

4 Evaluation Setup

4.1 Quantitative Study

4.1.1 Evaluation Protocol

Since no paired ground-truth exists for real low-quality video, we adopt a pseudo-GT protocol. A single 4K clip (3840×2160, 25 fps, 15 frames) serves as the ground truth. Low-resolution (LR) inputs are synthesized by downscaling each frame to $0.33\times$ the original resolution with `INTER_AREA` interpolation, followed by JPEG compression at quality 30 to introduce realistic compression artifacts. Each method processes the LR sequence and upscales the result back to the original 4K resolution via bicubic interpolation, enabling direct comparison against the ground-truth frames.

4.1.2 Regions of Interest

To evaluate fine-detail reconstruction, we define three fixed spatial ROIs containing high-frequency content (edges, texture, and fine structure). All ROI metrics are computed within these crops at the native 4K resolution.

4.1.3 Baselines

We compare five methods:

1) **Bicubic**: standard bicubic upsampling of each LR frame (spatial-only baseline).
2) **Per-frame Sharpen**: unsharp-mask sharpening applied independently to each LR frame before bicubic upsampling.
3) **Naive Temporal**: temporal averaging with uniform trust (trust $= 1$ everywhere), then bicubic upsampling. This ablates the trust-gating mechanism.
4) **Real-ESRGAN**: a pre-trained deep super-resolution model (RealESRGAN_x4plus) applied per frame.
5) **Ours**: the full pipeline with trust-gated temporal fusion and adaptive sharpening. We also report an ablation without the sharpening stage (**Ours, no sharpen**).

4.1.4 Metrics

We report four metrics, averaged over random 10 sequential frames post frame 5 of the video to discard the initial ramp-up of the temporal buffer:

- **PSNR** (dB, ↑): peak signal-to-noise ratio measuring per-pixel fidelity.
- **SSIM** (↑): structural similarity index computed on grayscale frames with an 11×11 Gaussian window ($\sigma = 1.5$).
- **Flicker** ℓ_1 (↓): mean absolute difference between each frame and the motion-compensated previous frame, $\frac{1}{HW} \sum | I_t - \mathcal{W}(I_{t-1}) |$, where $\mathcal{W}$ denotes ECC-based affine warping. Lower values indicate greater temporal stability.
- **Edge Variance** (↓): per-pixel variance of Sobel edge magnitudes over time, averaged spatially. Captures temporal jitter in high-frequency detail.

All metrics are reported both full-frame and within each ROI.

4.2 Qualitative Study

We conduct a pairwise preference study to evaluate perceived visual quality beyond what full-reference metrics capture.

4.2.1 Methods

We compare three conditions:

1) **Bicubic** — bicubic upsampling of each LR frame (spatial-only baseline).
2) **Ours (temporal)** — trust-gated temporal fusion without sharpening.
3) **Ours (temporal + sharpen)** — the full pipeline with trust-gated fusion and adaptive sharpening.

4.2.2 Stimuli

From our pseudo-GT evaluation clip (Sec. 4), we extract three spatial regions of interest containing fine detail and edges. For each region we generate all three pairwise comparisons, yielding $\binom{3}{2} \times 3 = 9$ trials. Each trial is presented as a side-by-side video with the two methods randomly assigned to the left ("A") and right ("B") positions. Videos loop for approximately four seconds at the native frame rate.

4.2.3 Protocol

Participants are shown the 9 trials in a randomized order via an online form. For each trial they answer three forced-choice questions:

1) *Which side appears perceptually clearer?* (A / B / No difference)
2) *Which side appears perceptually sharper?* (A / B / No difference)
3) *Which side looks more temporally stable?* (A / B / No difference)

No time limit is imposed; participants are instructed to watch each clip at least twice before responding.

4.2.4 Analysis

For each method pair, we report the *win rate* of the more advanced method (fraction of responses preferring it) on all three questions. Statistical significance is assessed with a two-sided exact binomial test against the null hypothesis of equal preference ($p_0 = 0.5$), with "no difference" responses excluded from the test.

5 Experimental Results

5.1 Quantitative Evaluation

We first evaluate the methods using full-frame and region of interest (ROI) fidelity metrics with respect to the high-resolution reference, together with temporal stability metrics. Tables 1 and 2 summarize the results. Sample outputs can be seen in Fig. 3.

At the full-frame level, bicubic interpolation provides the highest PSNR and SSIM, which is expected under a reconstruction-based evaluation because it introduces minimal deviation from the degraded input while remaining stable. However, our temporal-only method remains close to bicubic in both PSNR and SSIM (33.09 vs. 33.47 PSNR; 0.9210 vs. 0.9227 SSIM) while reducing flicker from 0.0237 to 0.0203. This indicates that trust-guided temporal fusion can improve temporal stability with only a small loss in reconstruction fidelity.

Naive temporal averaging achieves the lowest flicker, but this comes at a substantial cost in reconstruction quality (29.59 PSNR, 0.8771 SSIM), indicating that ungated accumulation suppresses variation primarily by over-smoothing or introducing history leakage. In contrast, the trust-guided formulation provides a more balanced tradeoff between temporal stability and fidelity.

Applying sharpening after temporal fusion increases local contrast but also shifts the result further away from the reference frame. This is reflected in the drop from 33.09 to 31.93 PSNR and from 0.9210 to 0.9063 SSIM when sharpening is enabled. The Real-ESRGAN baseline performs poorly under the same evaluation protocol, likely because the model synthesizes high-frequency detail that differs substantially from the pseudo-ground-truth reference.

The ROI results show the same overall trend. Averaged across all regions, our temporal-only method remains close to bicubic in ROI PSNR and ROI SSIM (31.34 / 0.8795 vs. 31.60 / 0.8797) while reducing ROI flicker from 0.0332 to 0.0286. This suggests that trust-guided temporal fusion is particularly effective in the face regions relevant to our application, preserving structural fidelity while improving temporal consistency.

By comparison, per-frame sharpening consistently decreases both ROI fidelity and temporal stability, while naive temporal averaging again achieves very low flicker at the cost of substantial fidelity loss. The sharpened version of our method improves local appearance subjectively (as shown in the user study below), but the quantitative reconstruction metrics indicate that this gain comes from perceptual enhancement rather than closer agreement with the reference.

5.2 Qualitative Preference Study

Because the target application is perceptual clarity rather than strict reconstruction accuracy, we additionally conducted a small user study comparing bicubic interpolation, our temporal-only method, and our temporal+sharpen variant.

Participants viewed short video clips and answered pairwise preference questions along three dimensions: *clarity*, *sharpness*, and *temporal stability*. Responses were exported from the form and matched to the underlying method pairs. "No difference" responses were excluded from preference counts. For each pair and question, we computed the win rate of the first-listed method and used a two-sided exact binomial test against a null hypothesis of equal preference ($p_0 = 0.5$).

We collected responses from 17 participants over 9 trials each, resulting in 126 total response rows. Results are shown in Table 3.

The qualitative results reveal a different tradeoff from the quantitative reconstruction metrics. The full pipeline (temporal fusion + sharpening) is significantly preferred over bicubic interpolation for *sharpness* (76.7%, $p = 0.001$), and is also preferred over the temporal-only variant for sharpness (75.6%, $p = 0.001$). This indicates that the sharpening stage contributes substantially to perceived detail.

At the same time, the sharpened pipeline is significantly worse than bicubic in *temporal stability*, with only 18.5% of responses preferring it over bicubic ($p = 0.002$). A similar, though weaker, trend appears when comparing the sharpened pipeline to the temporal-only variant. Together, these results support a clear *sharpness–stability tradeoff*: sharpening increases perceived sharpness but also increases visible temporal variation.

For *clarity*, no comparison reached significance. In particular, temporal fusion with sharpening was roughly comparable to bicubic interpolation (57.1% preference, $p = 0.441$), suggesting that the increase in sharpness does not come at a large perceptual cost in overall interpretability, even though it reduces stability.

6 Discussion and Conclusion

In this work, we demonstrated that perceived clarity in low-resolution video can be improved without hallucinating spatial detail. By estimating a per-pixel trust map derived from photometric consistency, motion, and edge reliability, our lightweight pipeline guides temporal fusion and selective sharpening to reinforce temporally stable structures.

These findings are particularly relevant for low-resolution whole-room video viewed across heterogeneous displays. In many practical settings, a single wide-angle camera captures an entire room, so each face occupies only a small portion of the frame. When such video is shown on high-resolution laptops, desktop monitors, or televisions, standard bicubic upsampling often appears visually soft even though it is temporally stable.

Our results suggest that trust-guided temporal processing offers a more flexible operating point than bicubic interpolation alone. The temporal-only variant improves temporal consistency while preserving reconstruction fidelity, and the sharpened variant significantly improves perceived sharpness without a statistically significant loss in perceived clarity. This makes the sharpened pipeline particularly attractive for viewing conditions where apparent detail

Fig. 3. This shows the three regions we use for evaluation in this work. From left to right: 4k Ground Truth, Low Resolution Input, Bicubic Upscaling, Frame Sharpening, Naive Temporal Accumulation, Ours (No Sharpening). Ours (Sharpening), Real-ESRGAN

TABLE 1
Full-frame reconstruction and temporal metrics.

Method	PSNR ↑	SSIM ↑	Flicker L1 ↓	Edge Var ↓
Bicubic	33.47	0.9227	0.0237	0.001311
Per-frame Sharpen	33.15	0.9173	0.0263	0.001291
Naive Temporal	29.59	0.8771	0.0098	0.001277
Ours (no sharpen)	33.09	0.9210	0.0203	0.001168
Real-ESRGAN	17.58	0.6948	0.0657	0.019607
Ours (sharpen)	31.93	0.9063	0.0231	0.001325

TABLE 2
ROI metrics by region.

Method	ROI #	PSNR ↑	SSIM ↑	Flicker ↓
Bicubic	1	**31.25**	0.8619	0.0341
Per-frame Sharpen	1	30.97	0.8533	0.0374
Naive Temporal	1	27.29	0.8084	**0.0153**
Ours (no sharpen)	1	31.07	**0.8620**	0.0297
Real-ESRGAN	1	19.83	0.6620	0.0641
Ours (sharpen)	1	30.20	0.8409	0.0334
Bicubic	2	**31.08**	**0.8797**	0.0391
Per-frame Sharpen	2	30.79	0.8702	0.0432
Naive Temporal	2	26.85	0.8091	**0.0162**
Ours (no sharpen)	2	30.72	0.8794	0.0342
Real-ESRGAN	2	17.59	0.6418	0.0799
Ours (sharpen)	2	29.41	0.8523	0.0391
Bicubic	3	**32.48**	**0.8976**	0.0264
Per-frame Sharpen	3	32.16	0.8896	0.0292
Naive Temporal	3	29.00	0.8562	**0.0109**
Ours (no sharpen)	3	32.23	0.8970	0.0219
Real-ESRGAN	3	18.22	0.6383	0.0649
Ours (sharpen)	3	30.91	0.8759	0.0256
Bicubic	AVG	**31.60**	**0.8797**	0.0332
Per-frame Sharpen	AVG	31.31	0.8710	0.0366
Naive Temporal	AVG	27.71	0.8246	**0.0141**
Ours (no sharpen)	AVG	31.34	0.8795	0.0286
Real-ESRGAN	AVG	18.55	0.6474	0.0696
Ours (sharpen)	AVG	30.17	0.8564	0.0327

TABLE 3
Qualitative study: overall preference win rates (17 participants, 9 trials each). Win rate is the proportion of non-"no difference" responses that preferred the first-listed method. p: two-sided binomial test against $p_0 = 0.5$. [T+S = temporal + sharpen; T = temporal only. Bold p indicates significance at $\alpha = 0.05$.]

Pair	Question	Win%	W/N	p
Ours (T+S) vs Bicubic	Clarity	57.1	24/42	0.441
	Sharpness	**76.7**	33/43	**0.001**
	Stability	**18.5**	5/27	**0.002**
Ours (T+S) vs Ours (T)	Clarity	56.1	23/41	0.533
	Sharpness	**75.6**	31/41	**0.001**
	Stability	29.6	8/27	0.052
Ours (T) vs Bicubic	Clarity	42.4	14/33	0.487
	Sharpness	41.2	14/34	0.392
	Stability	42.9	6/14	0.791

matters, such as large or high-resolution displays, while the temporal-only mode may be preferable in applications where temporal smoothness is the dominant requirement.

Overall, the proposed method does not define a single universally optimal output. Instead, it exposes a useful tradeoff between sharpness and stability, allowing the operating point to be selected according to the display conditions and application requirements.

7 Limitations and Future Work

Due to time constraints, our evaluation is limited to a small pseudo-ground-truth setup based on a single short 4K clip with synthetic degradations, so the results may not fully generalize to real-world video. In addition, the current affine ECC alignment model may be less reliable under large motion, occlusion, or nonrigid scenes. Future work of this project includes testing the algorithm and pipeline on more diverse real video other than the single one used in this study, designing or tuning stronger motion estimation as well as having a video-based super-resolution baseline.

ACKNOWLEDGMENTS

The author would like to thank the course staff of EE367, namely Prof. Gordon Wetzstein and Sonia Kim, for their advising and support on this project.

REFERENCES

[1] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part V*. Berlin, Heidelberg: Springer-Verlag, 2018, p. 63–79. [Online]. Available: https://doi.org/10.1007/978-3-030-11021-5_5

[2] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[3] 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.

[4] Y. Zhao, L. Yu, Z. Chen, and C. Zhu, "Video quality assessment based on measuring perceptual noise from spatial and temporal perspectives," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 12, pp. 1890–1902, 2011.

[5] F. W. Campbell and J. G. Robson, "Application of fourier analysis to the visibility of gratings," *The Journal of Physiology*, vol. 197, no. 3, pp. 551–566, 1968.

[6] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Trans. Graph.*, vol. 35, no. 6, Dec. 2016.

[7] B. Wronski, I. Garcia-Dorado, M. Ernst, D. Kelly, M. Krainin, C.-K. Liang, M. Levoy, and P. Milanfar, "Handheld multi-frame super-resolution," *ACM Trans. Graph.*, vol. 38, no. 4, Jul. 2019. [Online]. Available: https://doi.org/10.1145/3306346.3323024

[8] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian, "Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 3952–3966, 2012.

[9] C. Schied, A. Kaplanyan, C. Wyman, A. Patney, C. R. A. Chaitanya, J. Burgess, S. Liu, C. Dachsbacher, A. Lefohn, and M. Salvi, "Spatiotemporal variance-guided filtering: real-time reconstruction for path-traced global illumination," in *Proceedings of High Performance Graphics*, ser. HPG '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3105762.3105770

[10] G. D. Evangelidis and E. Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1858–1865, 2008.