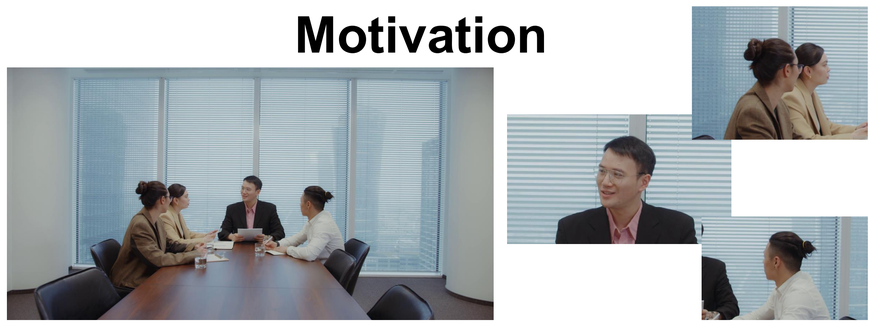
Trust-Guided Temporal Fusion in Video for Perceptual Enhancement

Andrew Chen
SCPD, Stanford University

Motivation



- In distant-capture video, such as whole-room conferencing, faces often occupy only a small fraction of the frame.
- Perceived clarity depends not only on spatial resolution, but also on **temporal stability** and **consistent edge structure** across frames.
- Goal: To improve the perceived sharpness of small faces by leveraging temporally reliable information, without reconstructing or hallucinating new spatial detail.

Related Work

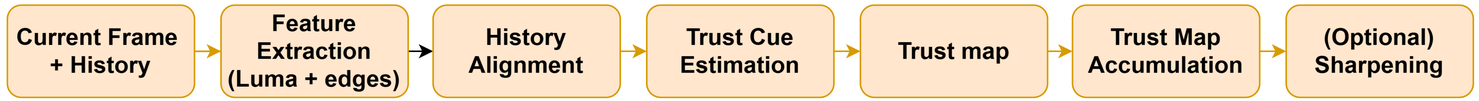
- Burst pipelines improve image quality by aligning and aggregating multiple frames, but typically operate on short fixed sequences and focus on reconstruction [1].
- Video pipelines reuse temporal information to reduce noise and stabilize structure, but require reliability checks to avoid ghosting and flicker [2].
- Human perception of sharpness depends strongly on contrast structure and temporal stability, suggesting that reliable temporal cues can improve perceived clarity even without recovering new spatial detail [3].

References

[1] Hasinoff, Sharlet, Geiss, Adams, Barron, Kainz, Chen, and Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," ACM Trans. Graph. (2016)
[2] Maggioni, Boracci, Foi, and Egiazarian, -- "Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms," IEEE Trans. Image Processing (2012)
[3] Campbell and Robson, "Application of fourier analysis to the visibility of gratings," The Journal of Physiology (1968)

New Technique

- We propose a lightweight temporal enhancement pipeline that aligns temporal evidence, estimates a per-pixel **trust map**, and uses that trust to control both temporal fusion and optional sharpening.
- Unlike naive temporal averaging, which can introduce ghosting, and per-frame sharpening, which can introduce flicker, our method selectively reinforces only **temporally stable** structures.

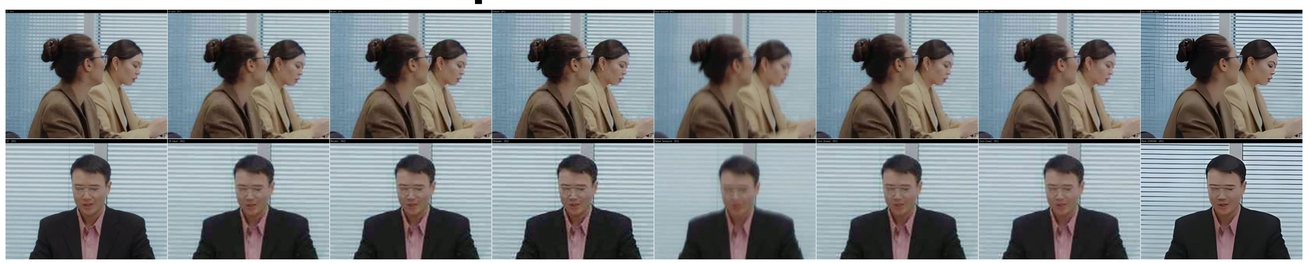


$$\mathrm{trust}(p) = \exp\big(-k_{\mathrm{diff}}\,\mathrm{diff}(p)\big)\ \exp\big(-k_{\mathrm{motion}}\,\mathrm{motion}(p)\big)\ \big(1 + w_{\mathrm{edge}}\,\mathrm{edge}_{\mathrm{mag}}(p)\big) \times\ \exp\big(-k_{\mathrm{clamp}}\,\mathrm{clamp}(p)\big)\,\mathrm{reliability}(p)^{\gamma}\,V(p)$$

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

L-R: Low Resolution Input, Trust Map, Ours (No Sharpening), Ours (Sharpening)

Experimental Results



L-R: 4k Ground Truth, Low Resolution Input, Bicubic Upscaling, Whole Frame Sharpening, Naive Temporal Accumulation, Ours (No Sharpening), Ours (Sharpening), Real-ESRGAN

Method	PSNR ↑	SSIM ↑	Flicker ↓
Bicubic	31.60	0.8797	0.0332
Sharpen	31.31	0.8710	0.0366
Naive Temporal	27.71	0.8246	0.0141
Ours (w/o sharpen)	31.34	0.8795	0.0286
Real-ESRGAN	18.55	0.6474	0.0696
Ours (sharpen)	30.17	0.8564	0.0327

Pair	Question	Win%	*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