FH University of Applied Sciences
TECHNIKUM WIEN

PROPOSAL

In Deep Learning

Temporal Retinex and Low-Rank Enhancement of B-mode Ultrasound Videos with Synthetic Gain/TGC Errors

By: Philip Donner & Martin Stasek

Vienna, November 4, 2025

Contents

1 Problem Description

Ultrasound is widely used in education and point-of-care settings because it is inexpensive, mobile, and safe compared to ionising imaging methods. In practice, however, many B-mode (Brightness-mode) scans, particularly those acquired by inexperienced users or with lower-end machines, are compromised by suboptimal device settings [1]. These settings include overall gain, time-gain compensation (TGC), dynamic range, and post-processing curves. They are adjusted under time pressure and with limited feedback. This produces inconsistent images across operators and probes [1].

The goal is to learn an automatic correction of exposure problems in ultrasound videos. The focus is on overall gain and TGC. These controls set the global brightness and the depth-dependent amplification. Wrong settings cause underexposed or overexposed regions and hide anatomy [1]. Automated correction could standardize brightness across depth and across frames. This would support diagnostics and training for novice users. The target is recovering visibility of structures while keeping the physical appearance of ultrasound.

Ultrasound images also contain speckle. Speckle is not simple noise. It carries information about tissue and the imaging system. Many filters reduce speckle but also remove important edges. Any correction must avoid hallucinations as well [2, 3].

Prior work shows that low-rank representation GANs can improve ultrasound quality by separating global structure from local texture. This helps preserve anatomy while changing contrast [4]. Low-rank modeling is attractive in cardiac videos because many frames share the same global layout with limited motion.

Retinex theory models an image as illumination times reflectance. In ultrasound, the illumination field is a good proxy for TGC. If we estimate this field, we can remove the exposure error and keep the tissue reflectance. This matches how users adjust gain and TGC on real devices [5].

We use a lightweight temporal transformer to combine information from neighboring frames. This reduces flicker and makes the correction stable over the cardiac cycle [6].

The main issues are incorrect brightness and depth-wise amplification. Wrong gain darkens or washes out the entire image. Wrong TGC creates bands that are too dark near the probe or too bright at depth. These errors vary across operators, probes, and vendors, so brightness is not consistent between patients or even between frames of the same clip. Over-correction can also clip highlights or boost noise in dark regions. Our goal is to standardize brightness across depth and time while keeping speckle realistic and avoiding hallucinated structures [1, 3, 6].

2 Dataset Description

We use the publicly available EchoNet-Dynamic Dataset [7] consisting of 10,030 deidentified echocardiogram videos. They are short ($\approx$ 3 seconds) grayscale videos in .avi format (112×112 pixels). These videos offer real B-mode cardiac ultrasound with consistent anatomy and motion, which is suitable for exposure correction and temporal modeling.

We preprocess the dataset by reading each video, sub-sampling by a fixed stride, and slicing fixed-length windows. We center-crop the heart region and resize to 128×128. We normalize frames to [0,1]. We then synthesize "bad$\rightarrow$good" pairs by simulating exposure errors: a global gain change (too dark or washed-out), a depth-wise TGC ramp that creates bright/dark bands, mild dynamic-range clipping that crushes shadows or highlights, and light sensor noise. We sample these effects per clip (with small frame-to-frame jitter) so the result looks like typical novice settings while keeping anatomy and speckle intact. This creates paired supervision while preserving speckle statistics and anatomy. Examples are shown in Figure 1.

Data split plan: Video-wise split 70/15/15 into training, validation, and test.

Figure 1: Paired examples — left: synthetic degradation; right: acquisition with expert settings.

3 Methodology

- **Architecture.** We use a Generative Adversarial Network (GAN) where the generator follows a Retinex–Low-Rank–Temporal design and the discriminator is a PatchGAN. The model takes a short grayscale clip and predicts a single corrected center frame. Figure 2 shows a simplified view of the proposed architecture.

  The generator consists of the following parts:

  Per-frame encoder. Each frame is processed by a shallow CNN at native resolution to extract feature maps without aggressive pooling. Operating at constant resolution helps preserve speckle statistics and fine boundaries that would be blurred by downsampling.

  Temporal fusion. We average-pool each frame's feature map over spatial dimensions and apply a small transformer encoder. The resulting temporal features are broadcast back and added to the spatial features for each time step. This lets the center-frame prediction leverage context from neighbors and reduces flicker compared to frame-wise processing [6].

  Three heads. From the fused features of the center frame, three small decoders predict:

  - **Reflectance R:** a detail map from a short CNN. It keeps edges and speckle and allows gentle local contrast changes [5].
  - **Illumination I:** a smooth exposure/TGC field. It should slowly vary across the image, in line with TGC [5].
  - **Low-rank LR:** a global trend map from a light low-rank factorization to capture broad intensity trends (like depth slopes) without changing edges [4].

  Composition layer. A composition layer fuses R, I, and LR with the center input to produce the final image.

  Discriminator (PatchGAN). A small PatchGAN [8] receives the corrected frame and outputs a dense map of real/fake logits on local patches. Using a patch-wise discriminator encourages realistic speckle and local contrast without forcing global hallucinations.

- **Training strategy.** Supervised training on synthetic pairs. An adversarial term with a small PatchGAN is used to enforce realistic speckle. Regularizers include total variation on illumination and an identity loss so good inputs stay unchanged.

- **Transfer learning.** Potentially initialize the CNN encoder from image models. Train temporal and heads from scratch.

- **Frameworks.** PyTorch and standard python libraries.

- **Evaluation metrics.** We report PSNR, SSIM, and Mutual Information [4], plus the Speckle Similarity Index and Edge Preservation Index for ultrasound texture and boundary sharpness [9].

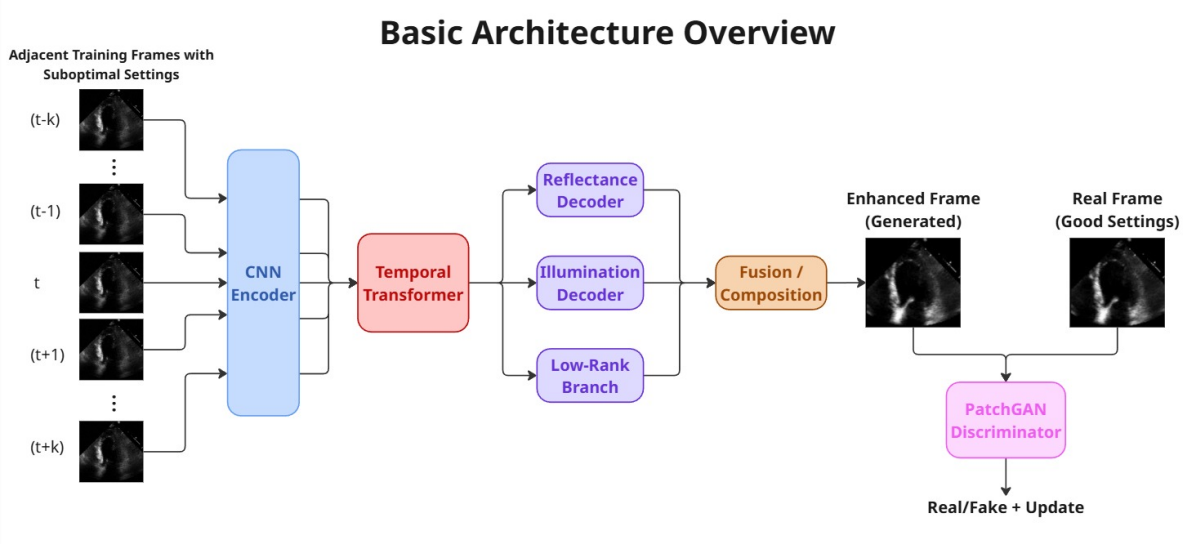

Figure 2: Simplified view of the proposed architecture for B-mode ultrasound quality correction: per-frame CNN encoders produce feature maps that feed three specialized heads (Reflectance, Illumination/TGC, Low-Rank). A temporal transformer fuses bottleneck features across the input window and the decoder outputs are merged to produce the enhanced center frame, which is compared to a reference "good-settings" frame of the same scene by a discriminator of PatchGAN-style.

These metrics balance fidelity and perceptual realism. PSNR and SSIM capture pixel- and structure-level similarity to the reference, which reflects basic restoration accuracy but can favor overly smooth results. Mutual Information measures shared information independent of a fixed linear mapping, so it is robust to small gain/contrast shifts and aligns with our exposure-correction goal. The Speckle Similarity Index focuses on ultrasound-specific texture statistics, penalizing outputs where speckle is lost, while the Edge Preservation Index quantifies whether anatomical boundaries remain sharp after correction.

- **Comparisons.** Single-frame UNet and/or a pix2pix [8] as a paired baseline.

4 Expected Outcomes

We expect more uniform brightness across depth and time, improved contrast, and preserved boundaries compared to non-GAN baselines. The PatchGAN should raise perceptual realism by restoring plausible speckle, while the Retinex and low-rank factors keep corrections physically plausible and reduce over-contrast [4, 5]. Temporal fusion is expected to lower flicker by pooling stable cues from neighboring frames [6].

Main risks are hallucinated structures, over- or under-correction in very dark/bright regions, and residual banding from TGC errors. We mitigate these with bounded illumination maps, low-rank regularization on global trends, identity loss on already-good inputs, conservative adversarial weight, and early stopping [3, 5].

We test three hypotheses:

- (H1) Adversarial training improves speckle realism, yielding higher speckle and edge scores, while causing only small changes in PSNR/SSIM compared with non-GAN training [4].
- (H2) The predicted illumination map is smooth along depth and correlates with TGC, supporting the Retinex interpretation [5].
- (H3) The method improves brightness uniformity across depth and restores usable dynamic range without clipping, as seen by flatter depth profiles and stable histogram percentiles [1, 5].

Bibliography

[1] David Zander et al. “Ultrasound Image Optimization (“Knobology”): B-Mode”. In: *Ultrasound International Open* 6.1 (June 2020), E14–E24. ISSN: 2509-596X. DOI: 10.1055/a-1223-1134. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7458857/ (visited on 10/15/2025).

[2] V. Damerjian et al. “Speckle characterization methods in ultrasound images – A review”. In: *IRBM* 35.4 (Sept. 2014), pp. 202–213. ISSN: 1959-0318. DOI: 10.1016/j.irbm.2014.05.003. URL: https://www.sciencedirect.com/science/article/pii/S1959031814000797 (visited on 10/15/2025).

[3] Rüdiger Göbl, Christoph Hennersperger, and Nassir Navab. *Speckle2Speckle: Unsupervised Learning of Ultrasound Speckle Filtering Without Clean Data*. arXiv:2208.00402 [eess]. July 2022. DOI: 10.48550/arXiv.2208.00402. URL: http://arxiv.org/abs/2208.00402 (visited on 10/10/2025).

[4] Zixia Zhou, Yi Guo, and Yuanyuan Wang. “Handheld Ultrasound Video High-Quality Reconstruction Using a Low-Rank Representation Multipathway Generative Adversarial Network”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.2 (Feb. 2021), pp. 575–588. ISSN: 2162-2388. DOI: 10.1109/TNNLS.2020.3025380. URL: https://ieeexplore.ieee.org/document/9210866 (visited on 10/10/2025).

[5] En Mou et al. “Retinex theory-based nonlinear luminance enhancement and denoising for low-light endoscopic images”. In: *BMC Medical Imaging* 24.1 (Aug. 2024), p. 207. ISSN: 1471-2342. DOI: 10.1186/s12880-024-01386-2. URL: https://doi.org/10.1186/s12880-024-01386-2 (visited on 10/10/2025).

[6] Jingyun Liang et al. *VRT: A Video Restoration Transformer*. arXiv:2201.12288 [cs]. June 2022. DOI: 10.48550/arXiv.2201.12288. URL: http://arxiv.org/abs/2201.12288 (visited on 10/10/2025).

[7] David Ouyang et al. “Video-based AI for beat-to-beat assessment of cardiac function”. en. In: *Nature* 580.7802 (Apr. 2020). Publisher: Nature Publishing Group, pp. 252–256. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2145-8. URL: https://www.nature.com/articles/s41586-020-2145-8 (visited on 11/02/2025).

[8] Phillip Isola et al. *Image-to-Image Translation with Conditional Adversarial Networks*. arXiv:1611.07004 [cs]. Nov. 2018. DOI: 10.48550/arXiv.1611.07004. URL: http://arxiv.org/abs/1611.07004 (visited on 10/10/2025).

[9] Midhila Madhusoodanan et al. *BEAM-Net: A Deep Learning Framework with Bone Enhancement Attention Mechanism for High Resolution High Frame Rate Ultrasound Beamforming*. arXiv:2507.15306 [eess]. July 2025. DOI: 10.48550/arXiv.2507.15306. URL: http://arxiv.org/abs/2507.15306 (visited on 10/15/2025).