Supplementary Material

Figure A. Selected frames from video stylization tests.

A. Style2Log Model Details

In this section, we provide detailed information about the Style2Log model used to generate synthetic log-space images from style references, as mentioned in the main paper.

A.1. Overview

The Style2Log model is a specialized neural network designed to transform standard images into their log-space representations by learning from style references. This model enables us to leverage unpaired data by generating synthetic training pairs that capture complex color grading characteristics found in professional photography and cinematography.

A.2. Architecture Components

To achieve higher-quality outputs, we integrate a NAFNet-based [3] refinement network into Style2Log. Our implementation employs a NAFNet with a width of 32, four middle blocks, and block configurations of [1,1,1,2] for the encoder and [1,1,1,1] for the decoder.

The refinement network receives the initial LUT-transformed image and generates the final log-space representation, enhancing both local details and global consistency.

We trained the Style2Log model using a curated dataset of log images combined with various LUTs. The training objective is formulated as a weighted combination of multiple loss functions:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{Perc}}, \tag{S1}$$

where $\mathcal{L}_1$ denotes the $\mathcal{L}_1$ loss and $\mathcal{L}_{\text{Perc}}$ represents the perceptual loss.

B. Limitations

Despite SA-LUT's strong performance, we identify several limitations that suggest directions for future research: (1) Challenging lighting conditions: As shown in Figure B, our method struggles with severely under/overexposed content images, where the Context Generator cannot produce meaningful spatial maps. Cross-attention fails to establish valid content-style feature correspondences when visual information is insufficient, resulting in poor stylization. (2) Semantic mismatch: Our cross-attention mechanism becomes less effective when content and style images differ dramatically in semantics, resulting in more global, less spatially-adaptive transformations. (3) Temporal consistency: While enabling real-time 4K processing, our method exhibits subtle frame-to-frame variations during rapid scene changes;

C. Visualization Results for Video Stylization

Figure A shows selected frames from our video stylization tests. For complete demonstrations of these results, please refer to the attached video.

D. Visualization Results for PST50 (Paired)

Figure C and Figure D show additional stylization results on the paired branch of our PST50 benchmark using SA-LUT.

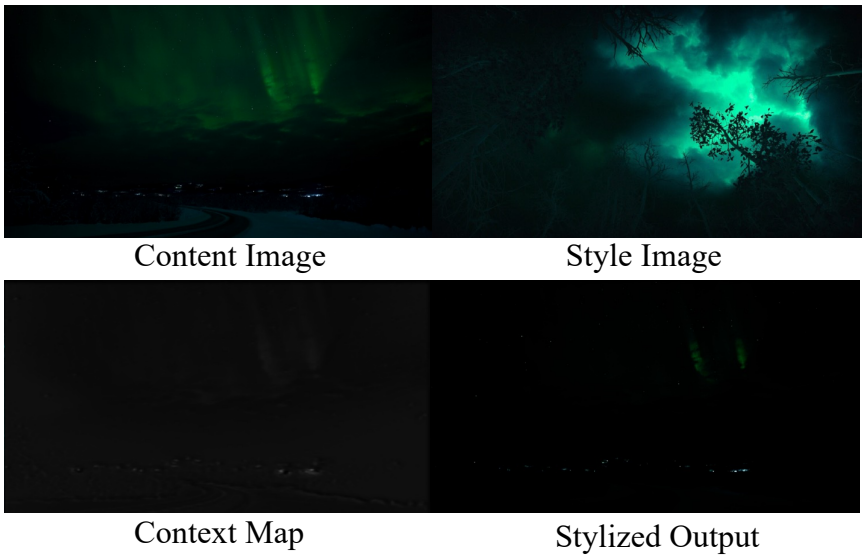

Figure B. Failure case under extreme lighting conditions.

E. Visualization Results for PST50 (Unpaired)

Figure E and Figure F present additional stylization results on the unpaired branch of our PST50 benchmark using SA-LUT.

Figure C. Stylization results on PST50 paired test set (page 1).

Figure D. Stylization results on PST50 paired test set (page 2).

Figure E. Stylization results on PST50 unpaired test set (page 1).

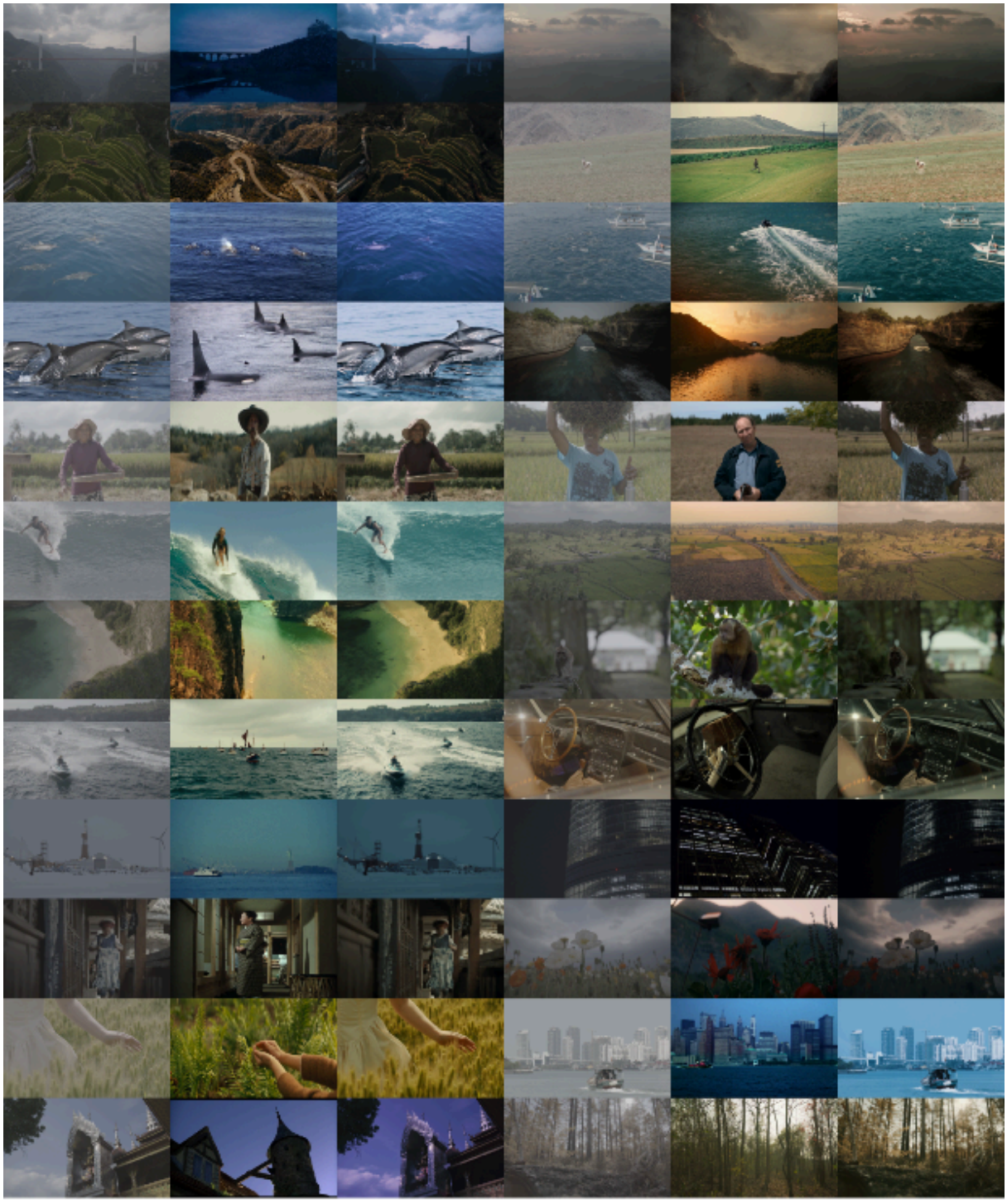

Figure F. Stylization results on PST50 unpaired test set (page 2).