Bringing Old Films Back to Life

CVPR 2022

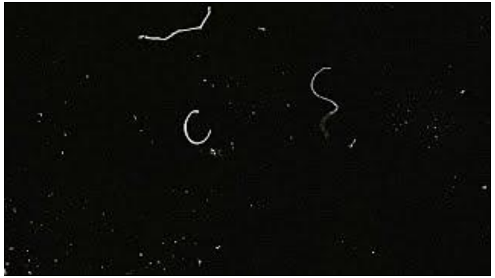
Presenter: Munkhsoyol Ganbat

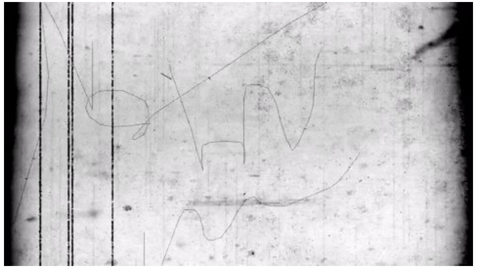
Vision Study Seminar
2022/08/29

DAVIAN

Tasks

- ## Video Restoration
  - Quality enhancement of the videos with blurriness, noises, scratches, cracks, dirt or dust and artifacts.





Tasks

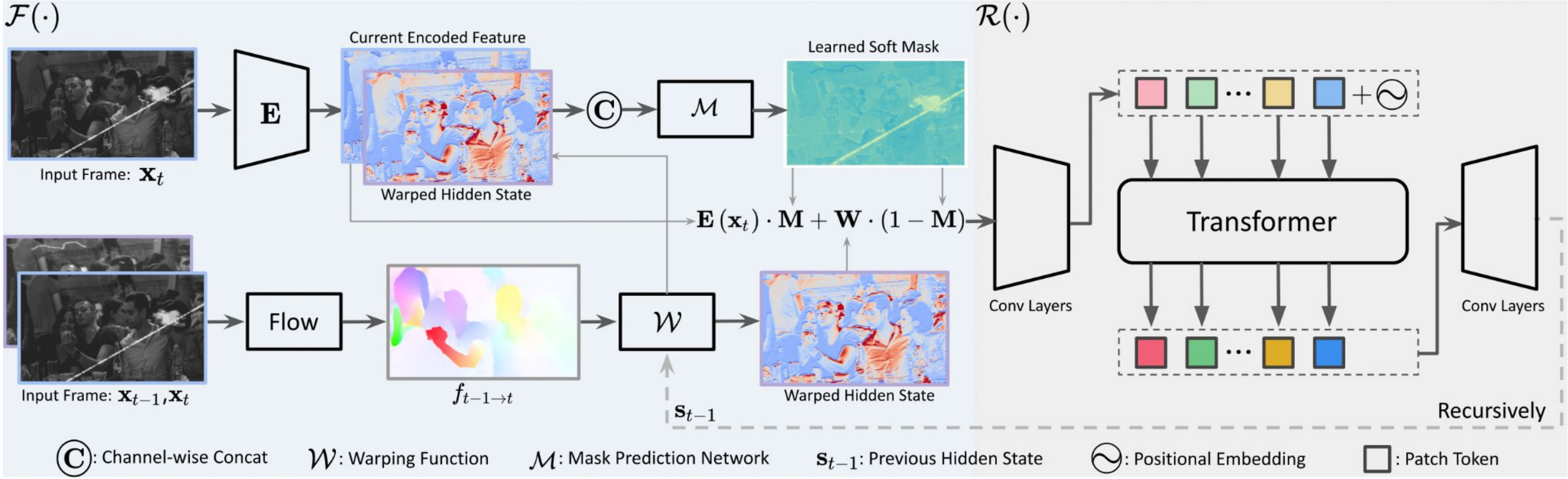
- ## Video Colorization
  - Automatic colorization, user-guide colorization and reference-based colorization.

Contributions

- Recurrent Transformer Network (RTN)
  - Robust to real-world old film degradation of large scratches and cracks
  - Same architecture for both restoration and colorization
  - Bidirectional RNN: effectively reduces old film flickering

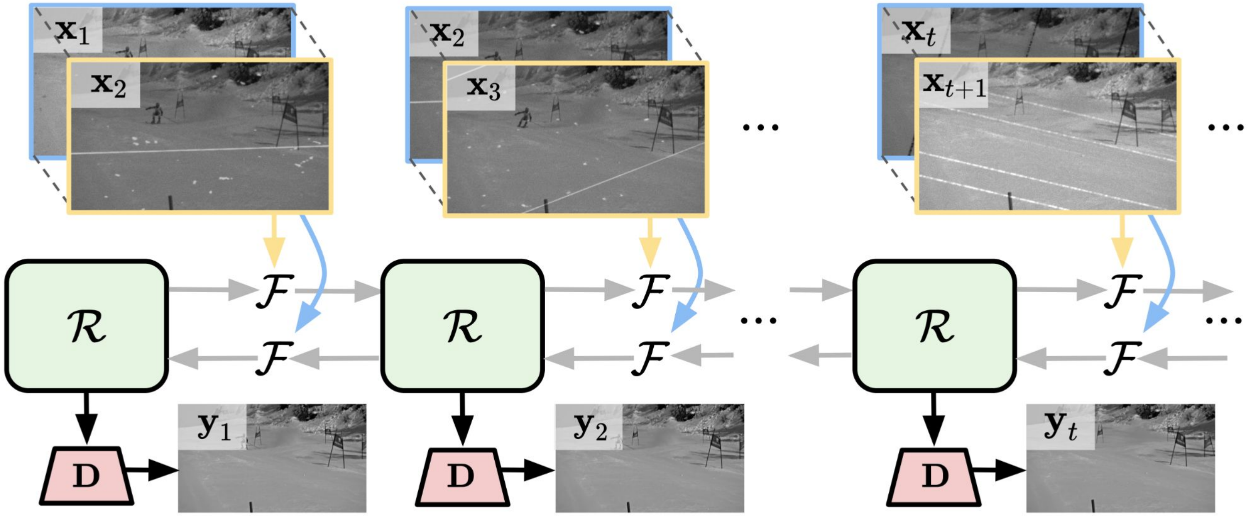
Model

- **Recurrent Transformer Network (RTN)**
  - F: Temporal Aggregation Module
  - R: Spatial Restoration Transformer



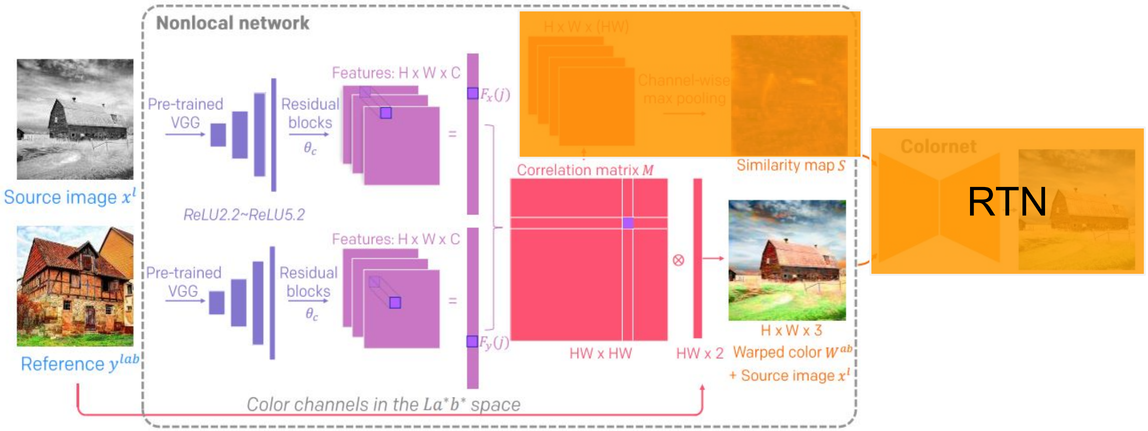
Model

- Model Pipeline
  - F: Temporal Aggregation Module
  - R: Spatial Restoration Transformer
  - D: Pixel Reconstruction Decoder



Model

- **Video Colorization**
  - Convert input color space RGB to LAB
  - Compare semantic similarity to predict AB channel
  - AB channel will be concatenated with gray input frame.



Training

- Loss
    - L1 Loss:
      Pixel-wise reconstruction loss between restored & GT frames

$$\mathcal{L}_1 = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{y}_t - \hat{\mathbf{y}}_t\|_1 .$$

    - Perceptual Loss:
      Perceptual loss between activation maps of VGG19
      p: selected layer (relu2_2 to relu5_2)
      w: importance of different layers

$$\mathcal{L}_{perc} = \frac{1}{T}\sum_{t=1}^{T}\sum_{p\in P} \omega_p \left\|\Phi_p^{\mathbf{y}_t} - \Phi_p^{\hat{\mathbf{y}}_t}\right\| ,$$

    - Spatial-Temporal Adversarial Loss:
      Temporal-PatchGAN
      Discriminator D (3D Conv): distinguish each spatial temporal feature as real or fake by hinge loss

$$\mathcal{L}_{\mathrm{D}} = \mathbb{E}_{\mathbf{y}\sim Y}[\mathtt{ReLU}(1 - D(\mathbf{y}))] + \mathbb{E}_{\hat{\mathbf{y}}\sim \hat{Y}}[\mathtt{ReLU}(1 + D(\hat{\mathbf{y}}))] ,$$

$$\mathcal{L}_{\mathrm{G}} = -\mathbb{E}_{\mathbf{y}\sim Y}[D(\mathbf{y})].$$

    - Full Objective: $\mathcal{L}_{total} = \lambda_1\mathcal{L}_1 + \lambda_p\mathcal{L}_{perc} + \lambda_G\mathcal{L}_G .$

Training

- **Video Degradation Model**
  - Contaminant Blending: 1k+ texture templates from internet and augmentation (rotation, crop, contrast change)
  - Video Quality Degradation:
    - Gaussian noise and Speckle noise,
    - Isotropic and anisotropic Gaussian blur kernels
    - Random JPEG compression level, downsampling and upsampling, brightness & contrast.
  - Temporal Frames Rendering: apply degradation for random consecutive temporal frames

Experiments

- Baselines
  - **Old Photo Restoration + TS**: Restore photo and blind temporal smoothing.
  - **BasicVSR**: Video super resolution method.
  - **Video Swin**: Attention mechanisms in both spatial and temporal dimensions.
  - **DeepRemaster**: State-of-the-art old film restoration method using 3D convolutions.
  - **DeOldify**: An open-source tool for restoring old films.

- Implementation
  - 20 epochs using the ADAM optimizer, learning rate 2e-4 for the first 20 epochs, batch size 4
  - Flow estimation:
    - RAFT (Recurrent All-Pairs Field Transforms for Optical Flow),
    - Fix parameters for first 5 epochs
  - Dataset: REDS video deblurring and super-resolution dataset (randomly crop 256 patches)
  - Training time: ~2 days on 4 RTX 2080Ti

Result

- ## Quantitative Result

Method	PSNR↑	SSIM↑	LPIPS↓	E_{warp} ↓
Input	19.982	0.699	0.456	0.0167
Old Photo+TS [23, 40]	21.962	0.768	0.315	0.0041
BasicVSR [4]	23.363	0.808	0.328	0.0053
Video Swin [29]	22.758	0.774	0.319	0.0061
DeepRemaster [17]	20.634	0.728	0.427	0.0066
DeOldify [1]	20.051	0.708	0.436	0.0149
Ours	**24.465**	**0.840**	**0.192**	**0.0019**
Ours w/o bi-direction	24.251	0.831	0.207	0.0036
Ours w/o soft mask	24.297	0.827	0.243	0.0025
Ours w/o transformer	24.342	0.830	0.229	0.0023

Table 1. Quantitative restoration comparisons on synthetic dataset. Our method achieves better performance on all metrics.

Method	PSNR↑	SSIM↑	LPIPS↓	FID ↓
Input	27.100	0.945	0.189	110.559
DeOldify* [1]	26.271	0.937	0.149	59.686
DeepExemplar [51]	30.064	0.952	0.091	37.971
DeepRemaster [17]	29.253	0.950	0.127	40.385
Ours	**32.838**	**0.977**	**0.065**	**31.992**

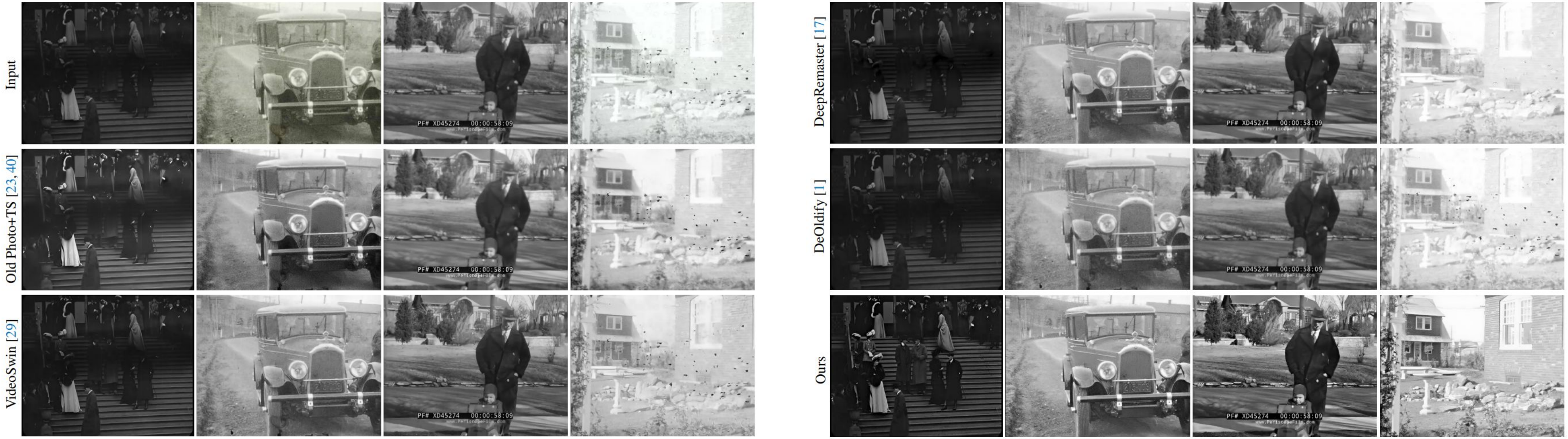
Table 2. Quantitative colorization comparisons on REDS [33] dataset. DeOldify*: Non-reference based video colorization.

Method	NIQE↓	BRISQUE↓
Input	18.9907	53.6776
Old Photo+TS [23, 40]	17.5110	48.1470
BasicVSR [4]	17.6842	62.7381
Video Swin [29]	18.9462	52.4758
DeepRemaster [17]	17.9697	49.9638
DeOldify [1]	17.9062	51.2813
Ours	**15.4254**	**42.1422**

Table 3. Quantitative restoration comparisons on real old films.

Result

- Qualitative Result: Video Restoration



Result

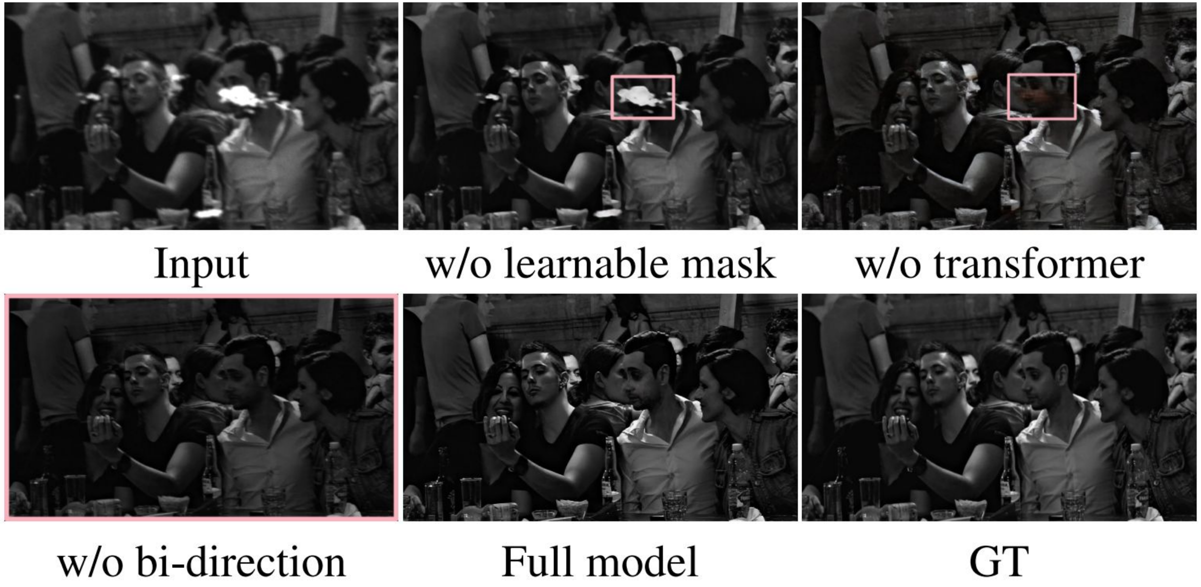
- Qualitative Result: Video Colorization

Input [1]* [17] [51] Ours

Input Mask Output

Result

- Ablation Study



Result

- Limitation
  - Content and noise ambiguity
  - Synthesize inadequate high-frequency details and artifacts
  - Still challenging to restore severely degraded frames

Conclusion

- Video Restoration
  - Temporal Bi-directional RNN: reduces flicker artifact of old films.
  - Learnable Guided Mask: accurate and effective to restore real-world noises.
  - Recurrent Spatial Transformer: Improved restoration ability for mixed degradations.
  - More stabilized training than CNN networks.
- Video Colorization
  - Not specifically designed for video colorization.
  - But more temporal consistent and reduced color bleeding.

Bringing Old Films Back to Life

Ziyu Wan[1], Bo Zhang[2], Dongdong Chen[3], Jing Liao[1]
[1]City University of Hong Kong, [2]Microsoft Research, [3]Microsoft Cloud+AI

CVPR 2022

Paper | arXiv | Code (Soon) | Data (Soon)

Thank you