Deep Exemplar-based Video Colorization

Bo Zhang[1,3], Mingming He[2,5], Jing Liao[2,3], Pedro V. Sander[1], Lu Yuan[3,4], Amine Bermak[1], Dong Chen[3]

[1]Hong Kong University of Science and Technology, [2]City University of Hong Kong, [3]Microsoft Research Asia, [4]Microsoft AI Perception and Mixed Reality, [5]USC Institute for Creative Technologies

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

香港城市大學 CityU City University of Hong Kong

Microsoft

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

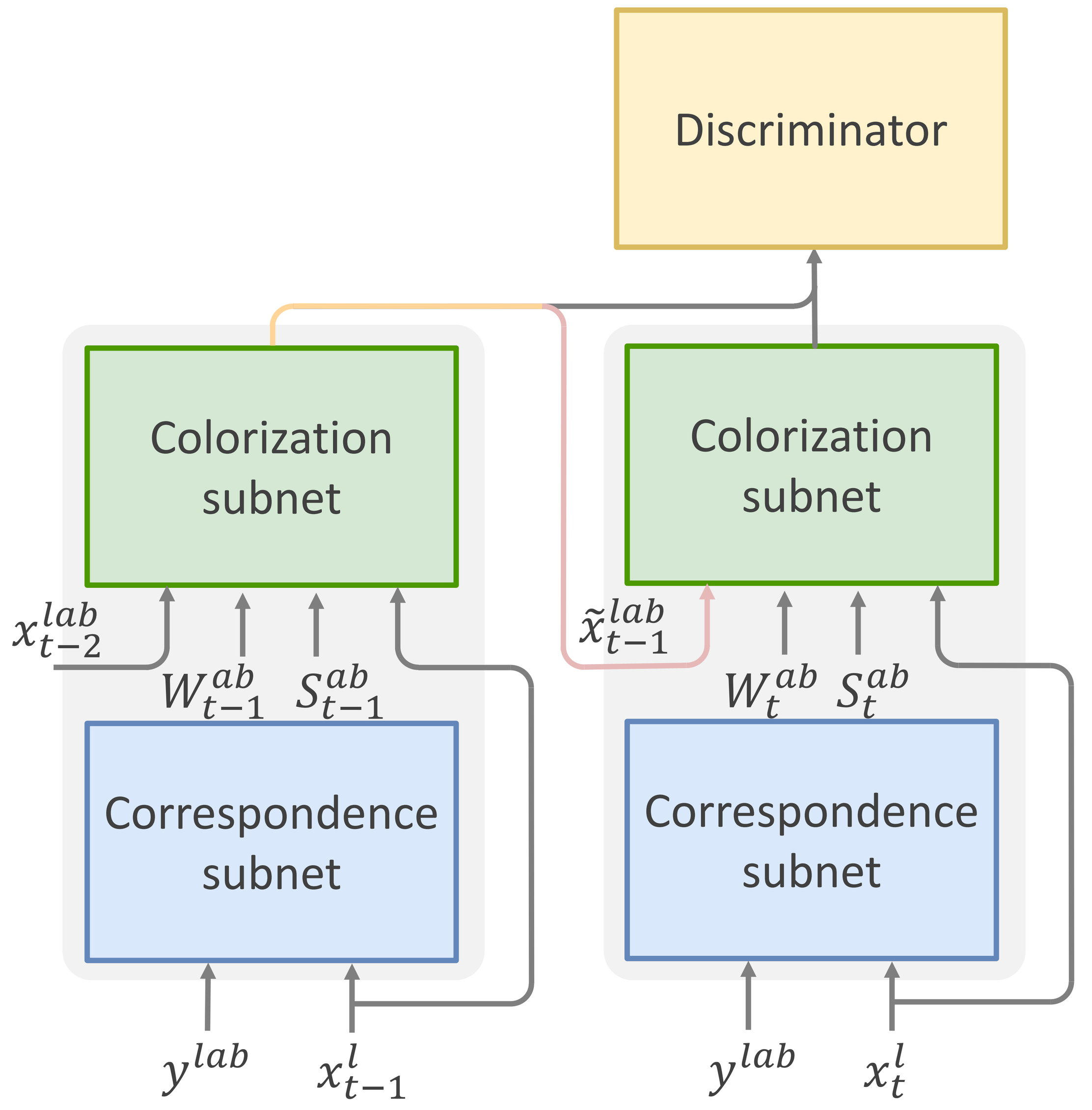
We propose the first end-to-end network for exemplar-based video colorization. The contributions are:

- Semantic dense correspondence in an unsupervised manner
- Correspondence and colorization are jointly trained, yielding faster inference speed and better quality
- State-of-the-art multi-modal video colorization
- Two modes: automatic colorization & color propagation

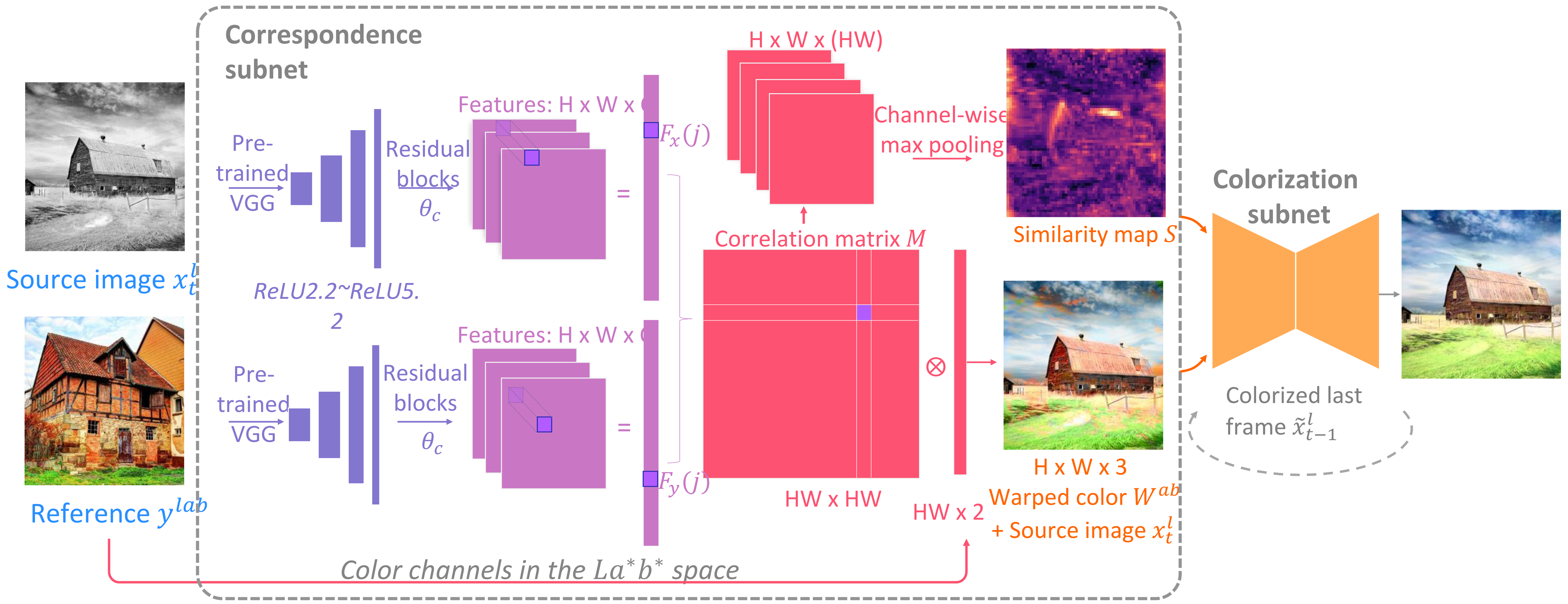
Overall Framework

The colorization for the current frame $x_t^{a,b}$ depends on the semantically corresponding region of the example image $y_{\{lab\}}$, and the historic colorization $\tilde{x}_t^{a,b}$

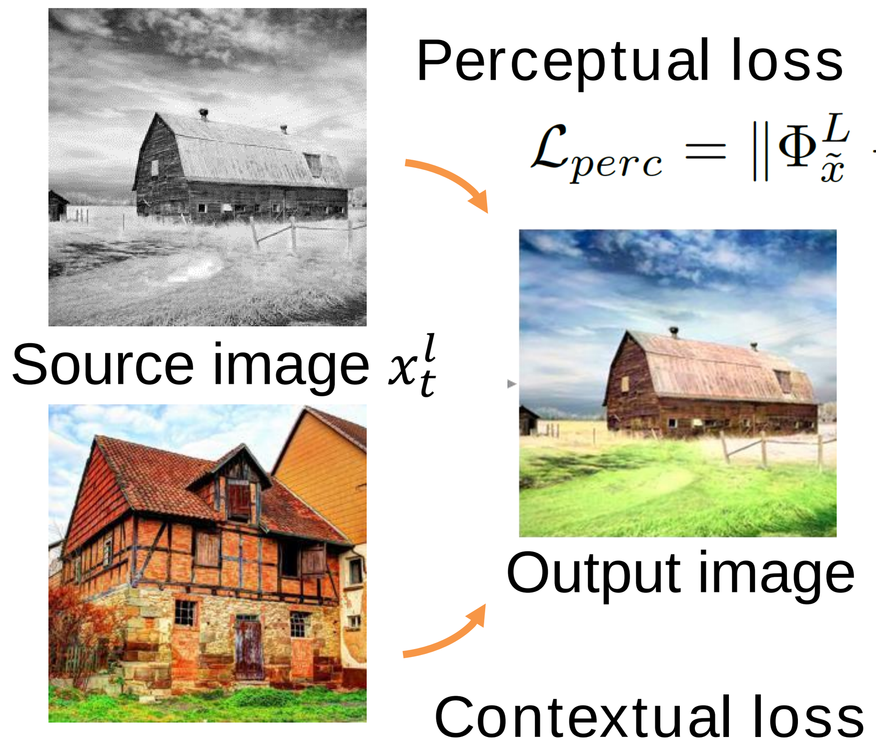
$$\tilde{x}_t^{ab} = \mathcal{C}(x_t^l, \mathcal{N}(x_t^l, y^{lab}) | \tilde{x}_{t-1}^{lab})$$



Joint Learning of Correspondence and Colorization



Loss Function



Source image x_t^l

Reference y^{lab}

Output image

Perceptual loss

$$\mathcal{L}_{perc} = \|\Phi_{\tilde{x}}^L - \Phi_x^L\|_2^2$$

Contextual loss

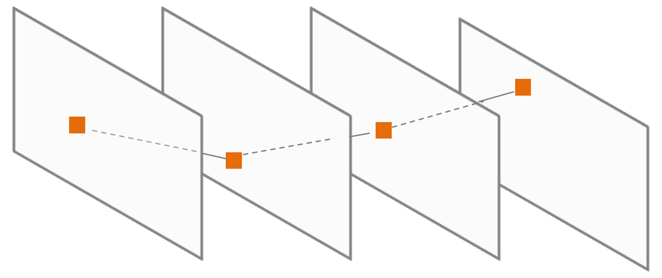
$$A^L(i,j) = \mathrm{softmax}_j(1 - \tilde{d}^L(i,j)/h)$$

$$\mathcal{L}_{context} = \sum_l w_L \left[-\log\left(\frac{1}{N_L}\sum_i \max_j A^L(i,j)\right)\right]$$

Smoothness loss: coherent region should have similar color

$$\mathcal{L}_{smooth} = \frac{1}{N}\sum_{c\in\{a,b\}}\sum_i \left(\tilde{x}_t^c(i) - \sum_{j\in\mathbb{N}(i)} w_{i,j}\tilde{x}_t^c(j)\right)$$

Temporal loss



$$\mathcal{L}_{temporal} = \|m_{t-1}\odot W_{t-1,t}(\tilde{x}_{t-1}^{ab}) - m_{t-1}\odot\tilde{x}_t^{ab}\|$$

Video adversarial loss

$$\begin{aligned}\mathcal{L}_{adv}^G = {} & \mathbb{E}_{(\tilde{x}_{t-1},\tilde{x}_t)\sim\mathcal{P}_{\tilde{x}}}[(D(\tilde{x}_{t-1},\tilde{x}_t) \\ & - \mathbb{E}_{(z_{t-1}z_t)\sim\mathcal{P}_z} D(z_{t-1},z_t) - 1)^2] \\ & + \mathbb{E}_{(z_{t-1}z_t)\sim\mathcal{P}_z}[(D(z_{t-1},z_t) \\ & - \mathbb{E}_{(\tilde{x}_{t-1},\tilde{x}_t)\sim\mathcal{P}_{\tilde{x}}} D(\tilde{x}_{t-1},\tilde{x}_t) + 1)^2]\end{aligned}$$

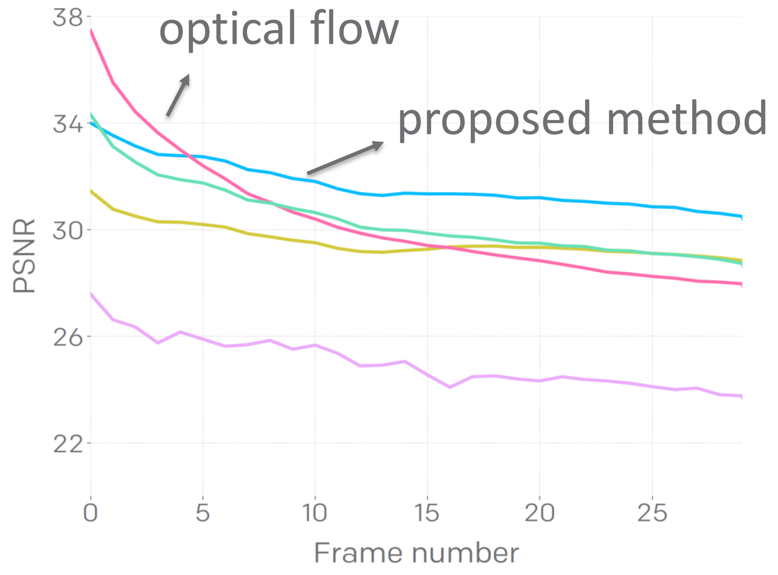
Introduction

Image colorization

input | reference | ours | He. 2018 | Iizuka 2016 | Larsson 2016 | Zhang 2016

Quantitative comparison

	Top-5 Acc(%)	Top-1 Acc(%)	FID	Colorful	Flicker
GT	90.27	71.19	0.00	19.1	5.22
[15]	85.03	62.94	7.04	11.17	7.19/5.69+
[16]	84.76	62.53	7.26	10.47	6.76/5.42+
[17]	83.88	60.34	8.38	**20.16**	7.93/5.89+
[30]	85.08	64.05	4.78	15.63	NA
Ours	**85.82**	**64.64**	**4.02**	17.90	5.84



Legacy film colorization

input | reference | ours