TWO-STREAM CONVOLUTIONAL NETWORKS FOR DYNAMIC TEXTURE SYNTHESIS

Matthew Tesfaldet, Marcus A. Brubaker - York University; Konstantinos G. Derpanis - Ryerson University
{mtesfald, mab}@eecs.yorku.ca, kosta@scs.ryerson.ca

Abstract

We introduce a two-stream model for dynamic texture synthesis based on pre-trained convolutional networks (ConvNets) that target two independent tasks: object recognition and optical flow prediction. Given an input dynamic texture, the object recognition ConvNet models the per-frame appearance of the input texture, while the optical flow ConvNet models its dynamics. To generate a novel texture, a noise sequence is optimized to match the feature statistics from each stream of the input texture.

Contributions

1. Motivated by the recent successes in texture synthesis using ConvNets [1, 2], we present a novel, two-stream model of **dynamic texture synthesis** to capture both appearance and dynamics.

2. A novel network architecture (which is motivated by the spacetime oriented energy model of [3]) designed to compute optical flow in an **appearance-invariant manner**, serving as the dynamics stream of our dynamic texture synthesis model.

3. A two-stream model that enables **dynamics style transfer**, where the appearance and dynamics from different sources can be combined to generate a novel texture.

Key Idea

Iteratively coerce an initial Gaussian noise sequence such that its spatiotemporal statistics from each stream match those of an inputs dynamic texture. This is done by optimizing (3) w.r.t. the spacetime volume (initially noise).

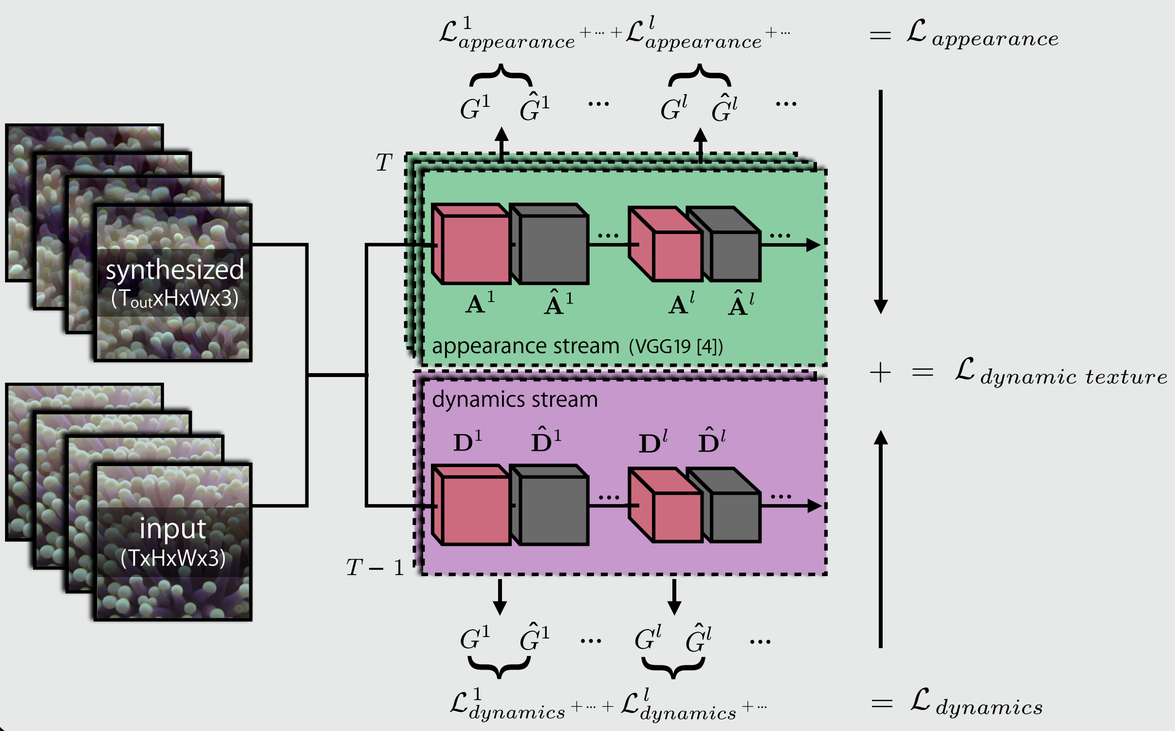
(1) $\mathcal{L}_{\text{appearance}} = \frac{1}{L_{\text{app}} T_{\text{out}}} \sum_{t=1}^{T_{\text{out}}} \sum_{l} \|\mathbf{G}^l - \hat{\mathbf{G}}^{lt}\|_F^2$ where L_{app} is the number of ConvNet layers being used in the dynamics stream, T_{out} is the number of generated frames, $\|\cdot\|_F$ is the Frobenius norm, $\hat{\mathbf{G}}^{lt}$ is the Gram matrix that models the synthesized texture appearance, and $\mathbf{G}^l$ models the target texture appearance averaged across time.

(2) $\mathcal{L}_{\text{dynamics}} = \frac{1}{L_{\text{dyn}}(T_{\text{out}}-1)} \sum_{t=1}^{T_{\text{out}}-1} \sum_{l} \|\mathbf{G}^l - \hat{\mathbf{G}}^{lt}\|_F^2$ where L_{dyn} is the number of ConvNet layers being used in the appearance stream, $\hat{\mathbf{G}}^{lt}$ models the synthesized texture dynamics, and $\mathbf{G}^l$ models the target texture dynamics averaged across time.

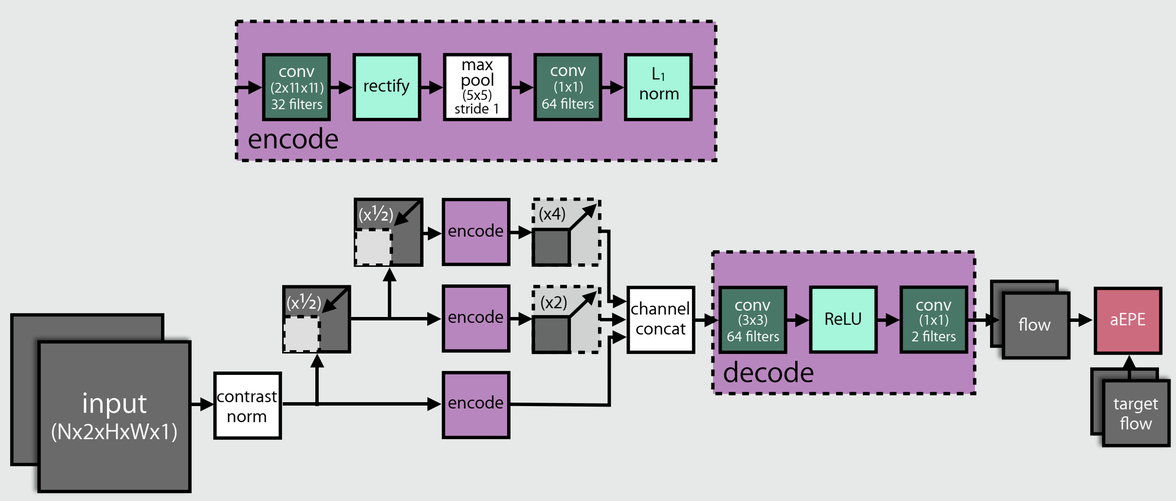
(3) $\mathcal{L}_{\text{dynamic texture}} = \alpha \mathcal{L}_{\text{appearance}} + \beta \mathcal{L}_{\text{dynamics}}$

Network Design

Overall Architecture



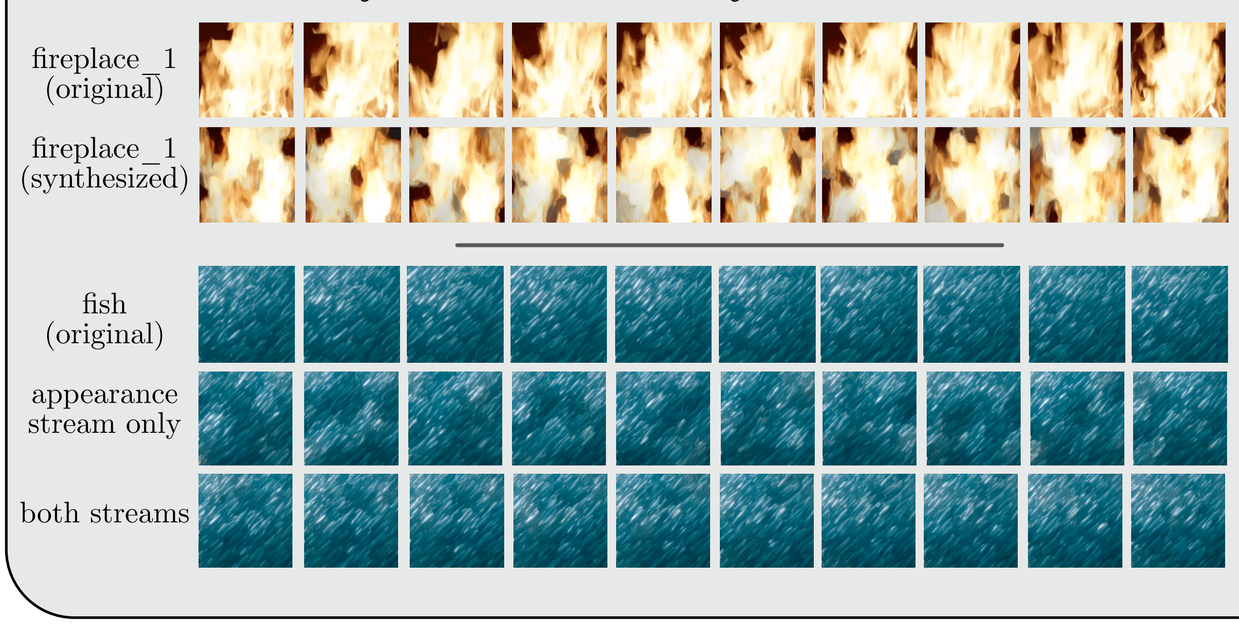
Dynamics Stream



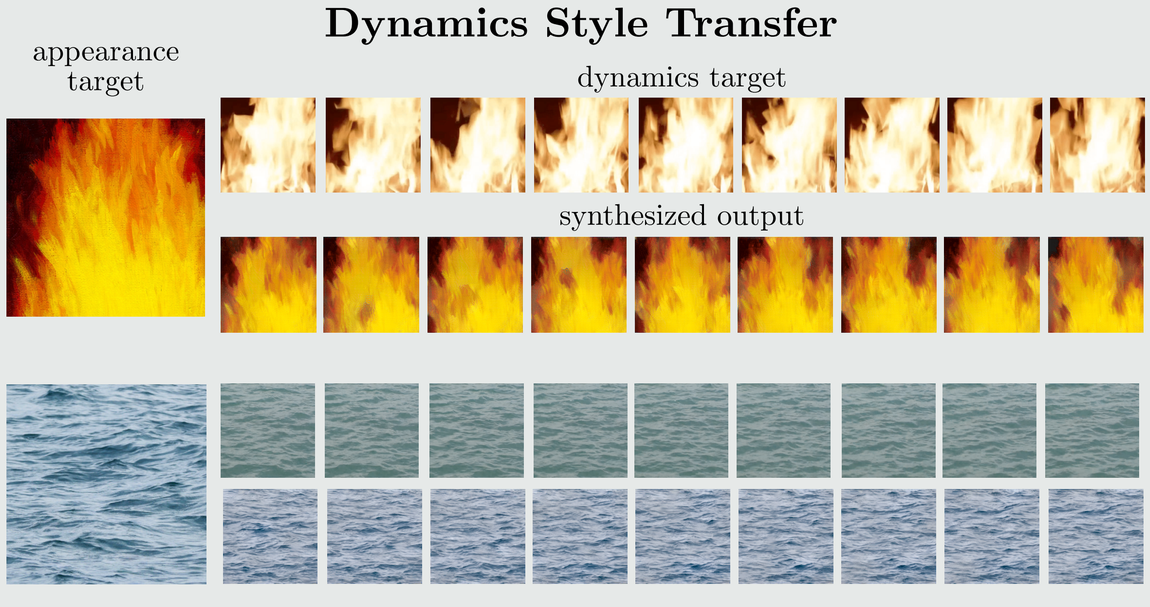
Dynamics Stream network trained for optical flow prediction in an appearance-invariant manner. Models the dynamics of the input dynamic texture.

Results

Dynamic Texture Synthesis



Dynamics Style Transfer



References

[1] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *NIPS* 2015.
[2] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. *ICML* 2016.
[3] Konstantinos G. Derpanis and Richard P. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *PAMI* 2012.
[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.