DeepFovea: Neural Reconstruction for Foveated Rendering and Video Compression using Learned Natural Video Statistics

Anton Kaplanyan
Facebook Reality Labs. Redmond, WA

Anton Sochenov
Facebook Reality Labs. Redmond, WA

Thomas Leimkühler
Facebook Reality Labs. Redmond, WA

Mikhail Okunev
Facebook Reality Labs. Redmond, WA

Todd Goodall
Facebook Reality Labs. Redmond, WA

Gizem Rufo
Facebook Reality Labs. Redmond, WA

Figure 1: Foveated reconstruction with DeepFovea. Left to right: (1) sparse foveated video frame (gaze in the upper right) with 10% of pixels; (2) frame reconstructed from (1) using our reconstruction method; and (3) full resolution reference. Our method in-hallucinates missing details based on the spatial and temporal context provided by the stream of sparse pixels. Zoom-ins show the foveal and a far periphery regions with different pixel densities.

ACM Reference Format:
Anton Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. 2019. DeepFovea: Neural Reconstruction for Foveated Rendering and Video Compression using Learned Natural Video Statistics. In *Proceedings of SIGGRAPH '19 Talks.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3306307.3328186

1 INTRODUCTION

Recent advances in head-mounted displays (HMDs) provide new levels of immersion by delivering imagery straight to human eyes. The high spatial and temporal resolution requirements of these displays pose a tremendous challenge for real-time rendering and video compression. Since the eyes rapidly decrease in spatial acuity with increasing eccentricity, providing high resolution to peripheral vision is unnecessary. Upcoming VR displays provide real-time estimation of gaze, enabling gaze-contingent rendering and compression methods that take advantage of this acuity falloff. In this setting, special care must be given to avoid visible artifacts such as a loss of contrast or addition of flicker.

In this talk, we present DeepFovea, a universal method that can be used efficiently for foveated rendering and foveated video compression. A perceptual function is used to spatially remove pixels in the input video stream, creating a sparse input signal which is then applied to DeepFovea. Our method *reconstructs* the original input by considering the best match given a learned manifold of natural videos. DeepFovea is a video reconstruction network that leverages recent advances in adversarial generative models, allowing it to inpaint and in-hallucinate peripheral details while maintaining fidelity at fixation. We observe that this method allows for a significant reduction in required input information while maintaining high quality across the visual field.

SIGGRAPH '19 Talks, July 28 - August 01, 2019, Los Angeles, CA, USA

ACM ISBN 978-1-4503-6317-4/19/07.
https://doi.org/10.1145/3306307.3328186

2 SETTING

We reduce the information required to encode a video signal by randomly sampling each frame using a perceptually weighted blue noise mask. This weighting exploits the cell density layout of the human retina, allowing perceptual bit allocation (Figure 1, left) with even spatial coverage. This sparse video stream is applied to DeepFovea, a neural network which has been trained to reconstruct the missing pixels (Figure 1, center).

Since this network is targeted for HMDs, we impose two main constraints on our reconstruction methodology. First, it must be able to operate in online mode, i.e., it can only rely on the history of previous frames. Second, the network must be able to operate at typical HMD refresh rates. For many existing devices, this requires an execution time of 11 ms per reconstructed frame.

3 MODEL AND TRAINING

DeepFovea is designed based on U-Net [Ronneberger et al. 2015], using an encoder-decoder architecture capable of encoding information over multiple spatial scales. With the addition of skip connections, this representation facilitates better gradient flow during training. The basic architecture is outlined in Figure 2.

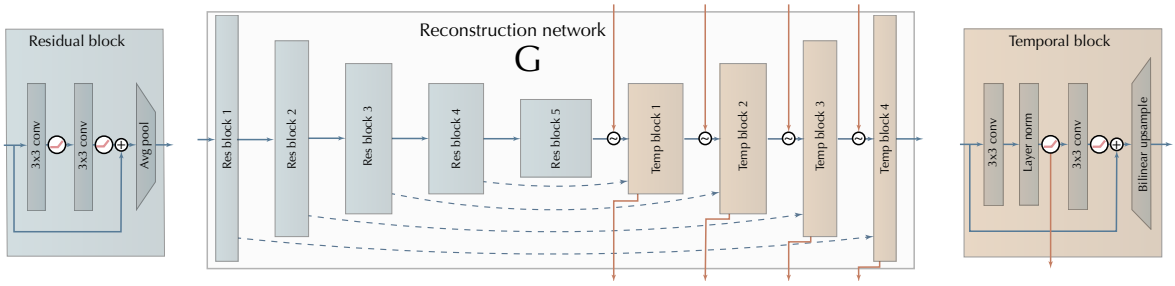

Figure 2: DeepFovea is a recurrent video encoder-decoder network architecture with skip connections (based on U-Net) targeting efficient video reconstruction from sparse pixels. The decoder is modified to be stateful and hierarchically retains temporal context using recurrent connections (orange).

Notably, the decoder employs recurrent connections. Each block at every scale retains its output activations from the previous frame. This allows the network to super-resolve features through time, capturing the spatio-temporal redundancy of natural videos while also achieving higher temporal stability.

The model has 3.2M parameters and requires 336 GFlops for an inference pass on a 1920x1080 frame.

We optimize the reconstruction network using a weighted sum of three losses $L = w_{\text{adv}} \cdot L_{\text{adv}} + w_{\text{VGG}} \cdot L_{\text{VGG}} + w_{\text{flow}} \cdot L_{\text{flow}}$. The *adversarial loss* L_{adv} is modeled by a discriminator network. The discriminator allows the reconstruction network to learn the spatio-temporal manifold of natural videos by classifying videos into fake and real. We use a Wasserstein GAN design [Arjovsky et al. 2017] with Spectral Normalization [Miyato et al. 2018], which implements a stabilizing distance measure. Since natural videos exhibit characteristic Fourier spectra, we additionally employ a spatio-temporal Fourier-domain discriminator.

To promote similarity of the reconstructed frames to the source frames, we use a *VGG loss* L_{VGG}, by transforming the frames into internal representations of the VGG network [Simonyan and Zisserman 2014]. This roughly perceptual representation stabilizes training, while providing enough freedom to the reconstruction.

Finally, we use the *optical flow loss* L_{flow} to stimulate temporal consistency across frames. Optical flow is only computed during training by requiring the network to match a reconstructed frame with previous reconstructed frames warped by the optical flow.

Training follows a standard adversarial approach of interleaving updates from reconstruction network and discriminator. During training, we reduce the valid pixels present in the input video stream.

4 RESULTS

User Study. We conduct a user study to validate DeepFovea and compare it to two other foveation algorithms: Multiresolution [Guenter et al. 2012] and Concentric H.265 compression. Participants were asked to detect artifacts and to rate the visual quality of videos generated using these methods. As depicted in Fig. 3, we find that DeepFovea achieves a 50% detectability threshold at a 47x compression rate, which is significantly better than the other tested methods at the same compression rate. Furthermore, the subjective ratings for DeepFovea indicate better or comparable quality of experience for almost all video content for 50x compression compared to the other methods.

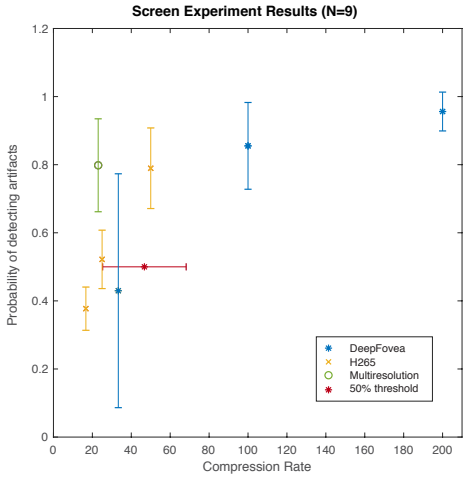

Figure 3: A summary of detectability results from screen experiment.

Runtime Performance. DeepFovea runs at 11ms on four Nvidia Tesla V100 GPUs per frame and achieves 90Hz in an Oculus Rift HMD.

5 CONCLUSION

We show that the spatio-temporal statistics of natural videos can be leveraged to achieve efficient video reconstruction for foveated rendering and reconstruction. Our method demonstrates temporally stable reconstruction from a noisy input and sets a new bar of 16x compression rate in savings achievable for foveated rendering with no significant degradation in perceived quality.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, 214–223.

Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D Graphics. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 31, 6, Article 164 (2012), 164:1–164:10 pages.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral Normalization for Generative Adversarial Networks. *CoRR* abs/1802.05957 (2018).

O. Ronneberger, P. Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI) (LNCS)*, Vol. 9351. 234–241.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).