DeepFovea++: Reconstruction and Super-Resolution for Natural Foveated Rendered Videos

TECHNISCHE UNIVERSITÄT DARMSTADT

Christoph Reich Marius Memmel Jonas Henry Grebe

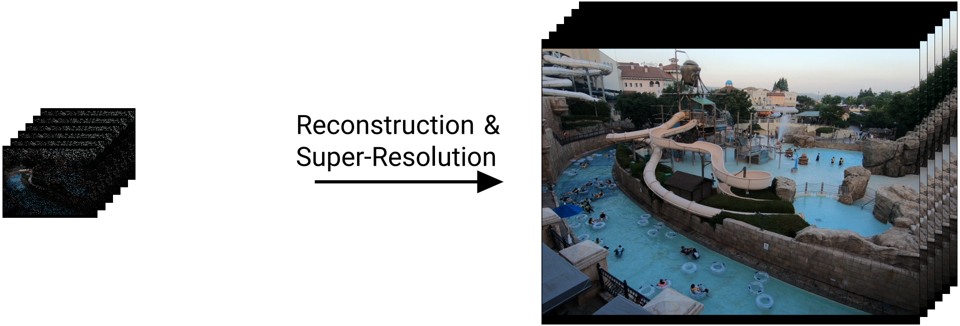
etit

Fovea Rendered Video Reconstruction and Super-Resolution

Problem Setting REDS dataset [1]

Fovea sampled input sequence

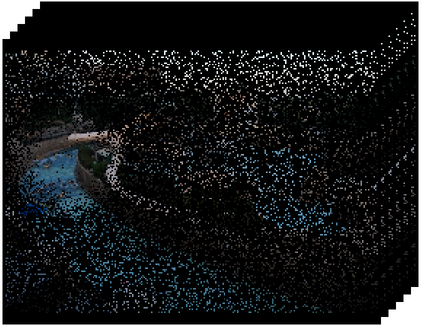
Reconstructed & upsampled sequence



Fovea Rendered Video Reconstruction and Super-Resolution

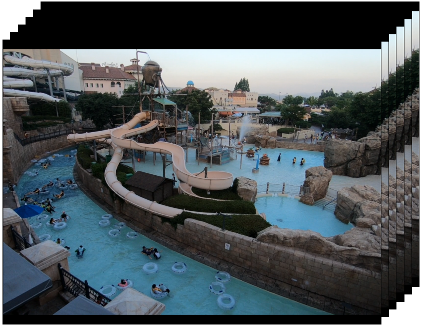
Problem Setting REDS dataset [1]

Fovea sampled input sequence



Reconstruction & Super-Resolution

Reconstructed & upsampled sequence



Fovea sample reconstruction
Related Work

DeepFovea - Kaplanyan et al. [2] (Facebook AI)

- Reconstructions of most plausible peripheral video from small portion of pixels in each frame
- Able to reconstruct video sequences (128×128)
- Recurrent U-Net reconstruction network
- Loss combination (adversarial, perceptual, optical flow)

TECHNISCHE UNIVERSITÄT DARMSTADT

Detail-revealing deep video super-resolution - Tao et al. [3] (2017)

- Sub-pixel motion compensation layers (improved frame alignment)

Video restoration with enhanced deformable convolutional networks - Wang et al. [4] (2019)

- Enhanced deformable convolution (frame alignment, temporal and spatial attention)

etit

Video Super-Resolution
Related Work

Detail-revealing deep video super-resolution - Tao et al. [3] (2017)

- Sub-pixel motion compensation layers (improved frame alignment)

Video restoration with enhanced deformable convolutional networks - Wang et al. [4] (2019)

- Enhanced deformable convolution (frame alignment, temporal and spatial attention)

Additional Work

- Video super-resolution with convolutional neural networks - Kappeler et al. [5] (2016)
- Video super-resolution via deep draft-ensemble learning - Liao et al. [6] (2015)
- Dictionary-based multiple frame video super-resolution - Dai et al. [7] (2015)

Reconstruction Model Architecture
Method

TECHNISCHE UNIVERSITÄT DARMSTADT

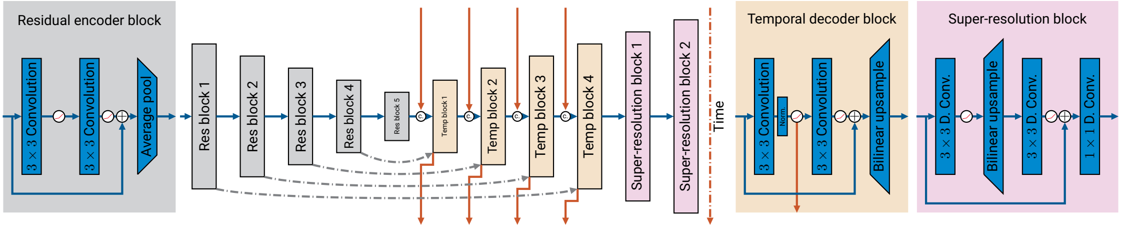


Figure: Architecture of the reconstruction network.

etit

Training process
Method

Loss function

$$\mathcal{L} = w_{\text{sv}}\,\mathcal{L}_{\text{sv}} + w_{\text{adv}}\,\mathcal{L}_{\text{adv}} + w_{\text{adv fft}}\,\mathcal{L}_{\text{adv fft}} + w_{\text{flow}}\,\mathcal{L}_{\text{flow}} + w_{\text{LPIPS}}\,\mathcal{L}_{\text{LPIPS}} \tag{1}$$

Loss function

$$\mathcal{L} = w_{\text{sv}}\,\mathcal{L}_{\text{sv}} + w_{\text{adv}}\,\mathcal{L}_{\text{adv}} + w_{\text{adv fft}}\,\mathcal{L}_{\text{adv fft}} + w_{\text{flow}}\,\mathcal{L}_{\text{flow}} + w_{\text{LPIPS}}\,\mathcal{L}_{\text{LPIPS}} \tag{1}$$

- Supervised loss (general and adaptive robust loss function - Barron [8])

  $$\mathcal{L}_{\text{sv}} = p(x, \alpha, c) + \log Z(\alpha), \;\; p(x = \sum_{i \in (c\,h\,w)} \hat{\mathbf{I}} - \mathbf{I}, \alpha, c) = \frac{|\alpha - 2|}{\alpha}\left(\left(\frac{(x/c)^2}{|\alpha-2|} + 1\right)^{(\alpha/2)} - 1\right)$$

- Adversarial loss [9] $\mathcal{L}_{\text{adv, adv fft}} = -\mathbb{E}\left[\log\left(\mathrm{D}\left(\hat{\mathbf{I}}\right)\right)\right]$
- Optical flow loss [10, 2] $\mathcal{L}_{\text{flow}} = \frac{1}{5}\sum_{i=1}^{5}\frac{1}{c_{\text{rgb}}\,h_i\,w_i}\left\|\hat{\mathbf{I}}_i - \mathrm{Warp}\left[\hat{\mathbf{I}}_{i+1}\right]\right\|_1$
- Perceptual loss [11] $\mathcal{L}_{\text{LPIPS}} = \frac{1}{5}\sum_{i=1}^{5}\frac{1}{b\,c_i\,h_i\,w_i}\left\|\mathrm{VGG}_{i,2}\left(\hat{\mathbf{I}}\right) - \mathrm{VGG}_{i,2}\left(\mathbf{I}\right)\right\|_1$

Discriminator Architecture
Method

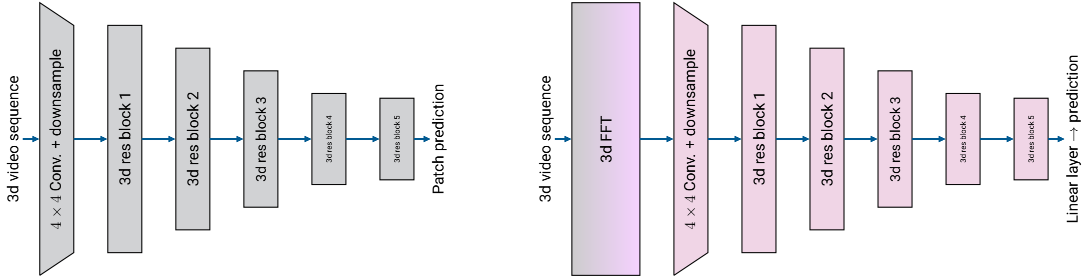


Figure: Discriminator network on the left and FFT discriminator network on the right.

Dataset and Fovea Sampling

Method

TECHNISCHE UNIVERSITÄT DARMSTADT

REDS dataset - Nah et al. [1]

- 300 sequences of 100 high-quality natural RGB video frames each
- Resolution 720×1280

Fovea Sampling

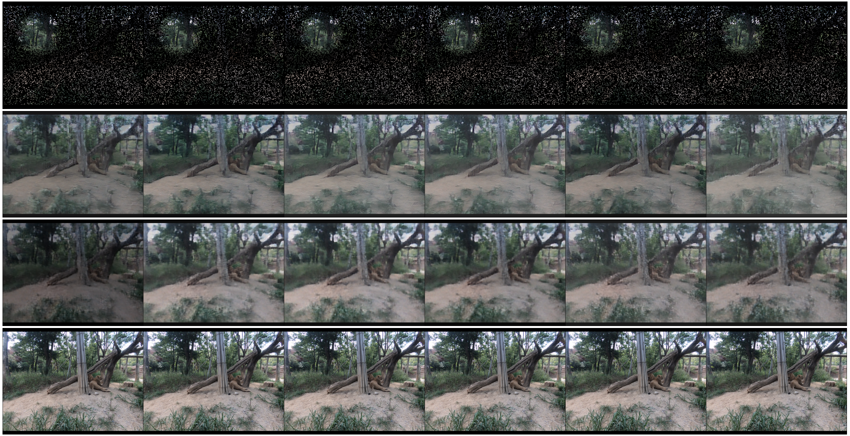
- Low chance of pixels not being masked out in or closer to focus point
- Higher chance of being masked out moving away from focus point
- Mask generated on downsampled image (relates to approximately 19% of the information in the low-res image and to **1.1%** of pixels compared to the high-res image)

Results

Qualitative results - with and without reset

Results

Qualitative results - with and without reset



Results

Qualitative results - with and without reset

Results
Quantitative results

TECHNISCHE UNIVERSITÄT DARMSTADT

Reset	L1 $\downarrow$	L2 $\downarrow$	Peak signal-to-noise ratio PSNR $\uparrow$	Structural similarity SSIM $\uparrow$
✓	0.0701	0.0117	22.6681	0.9116
✗	**0.061**	**0.009**	**23.8755**	**0.9290**

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\max\left\{\hat{\mathbf{I}}\right\}^2}{\mathrm{L2}\left(\hat{\mathbf{I}}, \mathbf{I}\right)}\right)$$

$$\mathrm{SSIM} = \frac{4\mathbb{E}\left[\hat{\mathbf{I}}\right]\mathbb{E}\left[\mathbf{I}\right]\mathrm{Cov}\left[\hat{\mathbf{I}}, \mathbf{I}\right]}{\left(\mathbb{E}\left[\hat{\mathbf{I}}\right]^2 + \mathbb{E}\left[\mathbf{I}\right]^2\right)\left(\mathrm{Var}\left[\hat{\mathbf{I}}\right] + \mathrm{Var}\left[\mathbf{I}\right]\right)}.$$

Conclusion and Possible Future Research

Conclusion

- Solid deep learning baseline to a novel problem
- Good result on the REDS dataset [1] which can be used as a benchmark
- Fast inference (reconstruction network 2.3M parameters)

Conclusion and Possible Future Research

Conclusion

- Solid deep learning baseline to a novel problem
- Good result on the REDS dataset [1] which can be used as a benchmark
- Fast inference (reconstruction network 2.3M parameters)

Possible Future Research

- Handle high memory consumption at training time
- Optimize the architecture of the reconstruction network

References I

[1] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. M. Lee, "Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[2] A. S. Kaplanyan, A. Sochenov, T. Leimkühler, M. Okunev, T. Goodall, and G. Rufo, "Deepfovea: neural reconstruction for foveated rendering and video compression using learned statistics of natural videos," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–13, 2019.

[3] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4472–4480.

[4] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

References II

TECHNISCHE UNIVERSITÄT DARMSTADT

[5] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, 2016.

[6] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 531–539.

[7] Q. Dai, S. Yoo, A. Kappeler, and A. K. Katsaggelos, "Dictionary-based multiple frame video super-resolution," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 83–87.

[8] J. T. Barron, "A general and adaptive robust loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4331–4339.

etit

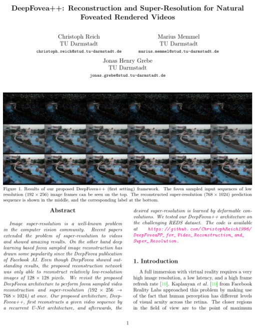
References III

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, “Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.

[11] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

Code Availability & Additional Resources



DeepFovea++: Reconstruction and Super-Resolution for Natural Foveated Rendered Videos

Christoph Reich
TU Darmstadt
christoph.reich@stud.tu-darmstadt.de

Marius Memmel
TU Darmstadt
marius.memmel@stud.tu-darmstadt.de

Jonas Henry Grebe
TU Darmstadt
jonas.grebe@stud.tu-darmstadt.de

Figure 1. Results of our proposed DeepFovea++ (first setting) framework. The fovea sampled input sequences of low resolution (192×256) image frames can be seen on the top. The reconstructed super-resolution (768×1024) prediction sequence is shown in the middle, and the corresponding label at the bottom.

Abstract

Image super-resolution is a well-known problem in the computer vision community. Recent papers extended the problem of super-resolution to videos and showed amazing results. On the other hand deep learning based fovea sampled image reconstruction has drawn some popularity since the DeepFovea publication of Facebook AI. Even though DeepFovea showed outstanding results, the proposed reconstruction network was only able to reconstruct relatively low-resolution images of 128×128 pixels. We revisit the proposed DeepFovea architecture to perform fovea sampled video reconstruction and super-resolution ($192 \times 256 \rightarrow 768 \times 1024$) at once. Our proposed architecture, DeepFovea++, first reconstructs a given video sequence by a recurrent U-Net architecture, and afterwards, the desired super-resolution is learned by deformable convolutions. We tested our DeepFovea++ architecture on the challenging REDS dataset. The code is available at https://github.com/ChristophReich1996/DeepFoveaPP_for_Video_Reconstruction_and_Super_Resolution.

1. Introduction

A full immersion with virtual reality requires a very high image resolution, a low latency, and a high frame refresh rate [10]. Kaplanyan *et al.* [10] from Facebook Reality Labs approached this problem by making use of the fact that human perception has different levels of visual acuity across the retina. The closer regions in the field of view are to the point of maximum

1

https://github.com/ChristophReich1996/DeepFoveaPP_for_
Video_Reconstruction_and_Super_Resolution