Towards Large-Scale Super Resolution Datasets via Learned Downsampling of Ray-Traced Renderings

Vaibhav Vavilala
Pixar Animation Studios
vibe@pixar.com

Mark Meyer
Pixar Animation Studios
mmeyer@pixar.com

Low-res input *Rendered (ours)* *Synthetic (our best)* *Blind (sharp)* *Blind (soft)*

Figure 1: ***Learning the downsampling operator to produce synthetic data for super resolution results in quality matching state of the art blind techniques but with fewer artifacts. The input (first) is compared with the upscaled result when trained on 3500 rendered pairs only (second), 25k synthetic pairs (third, our best), a blind network (fourth, shows some edge artifacts) and a softer blind network (fifth) that reduces artifacts but sacrifices some sharpness. In the first row, the blind networks show edge artifacts on text. In the second row, the sharper blind network shows edge artifacts on the character's eye, horn, and hat. ©Disney/Pixar***

ABSTRACT

Delivering high resolution content is a challenge in the film and games industries due to the cost of photorealistic ray-traced rendering. Image upscaling techniques are commonly used to obtain a high resolution result from a low resolution render. Recently, deep learned upscaling has started to make an impact in production settings, synthesizing sharper and more detailed imagery than previous methods. The quality of a super resolution model depends on the size of its dataset, which can be expensive to generate at scale due to the large number of ray-traced pairs of renders required. In this report, we discuss our experiments training an additional neural network to learn the degradation operator, which can be used to rapidly generate low resolution images from existing high resolution renders. Our testing on production scenes shows that super resolution networks trained with a large synthetic dataset produce fewer artifacts and better reconstruction quality than networks trained on a smaller rendered dataset alone, and compare favorably to recent state of the art blind synthetic data techniques.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**.

SIGGRAPH '21 Talks, August 09-13, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8373-8/21/08.
https://doi.org/10.1145/3450623.3464631

ACM Reference Format:
Vaibhav Vavilala and Mark Meyer. 2021. Towards Large-Scale Super Resolution Datasets via Learned Downsampling of Ray-Traced Renderings. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks (SIGGRAPH '21 Talks), August 09-13, 2021.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3450623.3464631

1 BACKGROUND & RELATED WORK

To deliver high resolution content, upscaling techniques are frequently used. Total Variational Inpainting is a standard fixture in production pipelines (as a Nuke node, TVIScale), generally avoiding artifacts and requiring a few seconds per image to run. TVI however cannot match the ground truth level of detail, particularly at higher upscaling factors, which motivates neural approaches that Pixar has been developing in recent years. Further, we are also exploring upscaling daily renders to support a growing number of shows simultaneously in production as well as increasing scene complexity - both of which burden the render farm.

Training a super resolution network requires corresponding pairs of low res and high res images. The most common ways of generating these pairs is to either render them directly, which is expensive, or to downsample existing high res imagery to create the low res. Much of the super resolution literature assumes the downsampling kernel is known and instead focuses on architectural innovations [Wang et al. 2018]. However, to deploy super resolution in the real world, the degradation operator must be known to obtain compelling results. The family of blind super resolution techniques has shown to be highly effective in our testing, whereby many random degradations are applied on the fly during the training process to obtain low resolution images from existing high

resolution images [Cornillere et al. 2019]. There have also been attempts to learn the degradation operator by modeling a camera's point spread function as well as using neural networks to predict the degradation operator. However, there is limited work on doing so with rendered images in production settings - which we explore here, showing that it produces high-quality results.

2 DOWNSAMPLING NETWORK

Our downsampling network is similar to recent super resolution architectures employing deep residual networks as the basic backbone, but we replace the upsampling unit at the end of the network with a downsampling unit. We follow a similar training procedure that we used to train our upscaling networks [Vavilala and Meyer 2020]. Our downsampling network is RGBA and single frame in and out. Our training set consists of 3500 rendered pairs spanning 5 recent Pixar features and includes solid elements, volumes, and final composites with film grain. We ensure the dataset is diverse in textures, characters, lighting conditions, and use of motion blur.

Our trained network takes 2 seconds for inference per frame on the GPU (from 2K to 1K), enabling us to efficiently generate a synthetic dataset of 25k pairs spanning over a dozen Pixar features and shorts by reusing high res renders already on our disk farm. We then trained a super resolution network on these synthetic pairs, and observed production-quality results with minimal hyperparameter tuning. When compared with recent state of the art blind techniques (we tested the cubic downsampling kernel as part of the degradation-aware upscaler in [Cornillere et al. 2019]), our network trained on learned-downsampled data matches the sharpness but avoids certain edge artifacts shown in Fig. 1. When testing various configurations of the blind network, we were able to remove many instances of artifacts at the cost of some sharpness.

While the synthetic set contains renders from 1K to 2K resolution, at inference time we had success creating 4K outputs from a 2K source and anticipate high quality results at a wide range of source dimensions. We note that rendering training data requires dozens of CPU hours per frame, and emphasize that the high res and low res must be rendered simultaneously to avoid image differences from production churn. There may still be unusable pairs due to uncorrected resolution-dependent parameters in the lighting and shading that require an additional data cleaning step. Learned downsampling and blind techniques can largely avoid these troubles, simplifying the data collection process for super resolution.

With additional tuning and experimentation, it may be possible to produce a blind network, or a collection of blind networks with sufficient coverage of production scenes, that eliminates the need to obtain a rendered pairs dataset and train a downsampling network. Nevertheless, this is the first successful experiment, to our knowledge, to show that the image degradation operator for ray-traced renders can be learned by a neural network, and production-quality super resolution results can be created using such data. Our work in efficiently scaling up training sets has shown to improve inference quality, which ultimately results in more faithful results and less production time spent in fixing artifacts.

3 DISCUSSION

We share additional details about our super resolution efforts here. In our previous attempts, we tried jittering the colors of the inputs and targets dynamically during the training process (color shift data augmentation). However, recent work on training GANs with limited data has shown that data augmentations, when applied to generator inputs instead of discriminator inputs, can "leak" into the generated distribution, essentially introducing color shifts, which we observed in our testing. Further, recent work shows that data augmentations help less and less as the size of the training set increases, and can even harm inference quality beyond a point. Thus, we removed the color jittering data augmentations. Only random flips and rotates remain as our augmentations.

We anticipate learning the residual - having the network predict the difference between the downsampled high res input and low res target instead of predicting the final colors directly - will accelerate training the downsampling network. Our super resolution networks learn the residual, and since they learn an easier task, they converge faster. In our testing, learning the residual has also removed the need for a color shift loss term that enforces a cycle consistency between the downsampled network output and network input. However, our tests showed learning the residual still required range-compressing the input via a log transform to avoid high dynamic range artifacts.

Our testing shows negligible advantage in using cross frames as inputs to the upscaler, and thus far not worth the added pipeline complexity. However, this is an active area of research since cross frames have shown material improvements in similar tasks like denoising. Running inference in single frame mode has not produced noticeable temporal coherence issues in our testing, though upscaling may exacerbate problems already present in the input.

4 CONCLUSION

The primary finding of this work is that degradation operators for rendered scenes can be learned by neural networks, which can enable scaling up super resolution training sets using a smaller render farm footprint. The upscaling networks currently deployed into Pixar's production pipeline were all trained using synthetic data generated from learned downsampling, and have shown higher quality results than training on a small rendered dataset alone and fewer artifacts than blind techniques. While we worked solely with RenderMan, we anticipate the degradation operator can be learned using other renderers like Hyperion and Arnold as long as a sufficiently large and diverse dataset is available. Our experiments motivate synthetic data approaches to augment training sets for other image-space tasks and merit further analysis.

ACKNOWLEDGMENTS

The authors thank Christopher Schroers, Aziz Djelouah, and Jeremy Newlin for helpful suggestions and support.

REFERENCES

Victor Cornillere, Abdelaziz Djelouah, Wang Yifan, Olga Sorkine-Hornung, and Christopher Schroers. 2019. Blind image super-resolution with spatially variant degradations. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–13.

Vaibhav Vavilala and Mark Meyer. 2020. Deep Learned Super Resolution for Feature Film Production. In *ACM SIGGRAPH 2020 Talks (SIGGRAPH '20)*. Association for Computing Machinery, New York, NY, USA, Article 41, 2 pages. https://doi.org/10.1145/3388767.3407334

Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. 2018. A fully progressive approach to single-image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 864–873.