Light Stage Super-Resolution: Continuous High-Frequency Relighting Supplementary Material

TIANCHENG SUN and ZEXIANG XU, University of California, San Diego
XIUMING ZHANG, Massachusetts Institute of Technology
SEAN FANELLO, CHRISTOPH RHEMANN, and PAUL DEBEVEC, Google
YUN-TA TSAI and JONATHAN T. BARRON, Google Research
RAVI RAMAMOORTHI, University of California, San Diego

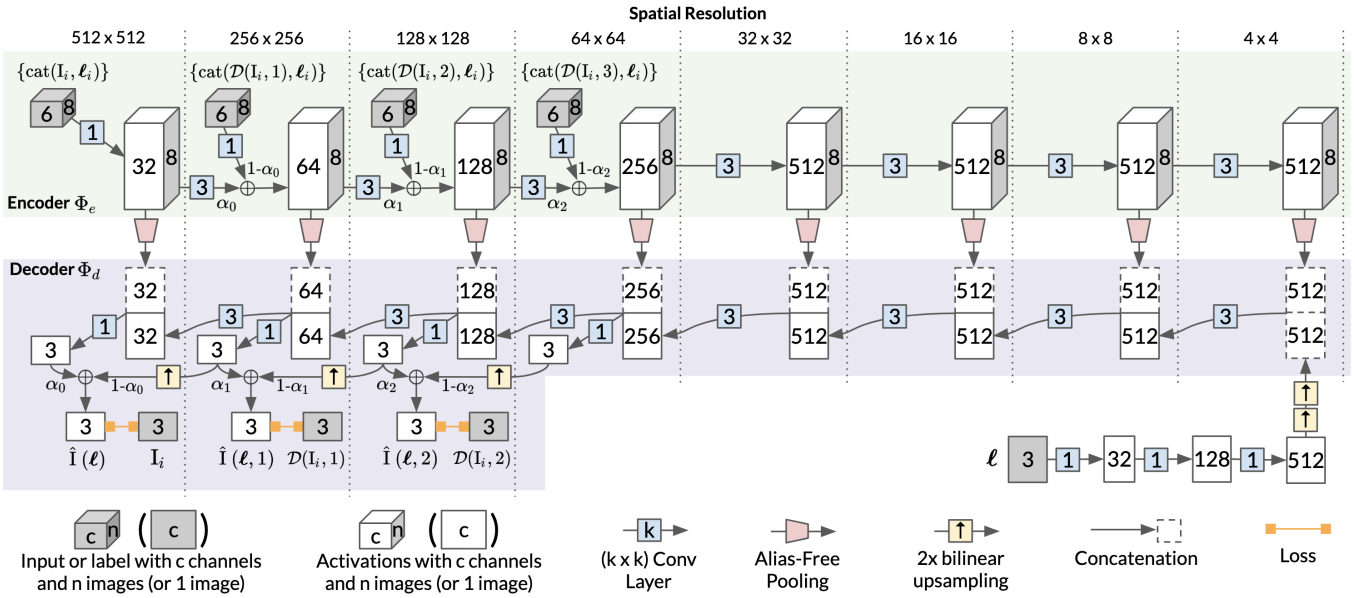

Fig. 1. A visualization of our model architecture and our progressive training scheme. The α_d parameters control the progressive training and growing of the network for each scale d of the network by modulating the resolution at which input images are used and output images are compared to the ground truth.

1 PROGRESSIVE TRAINING

To train our model, we use a progressive approach similar to that of [Karras et al. 2018]. Instead of simply training our model in one "stage" to minimize some loss between the full resolution output image $\mathbf{I}(\boldsymbol{\ell}_i)$ and the true image from the light stage $\mathbf{I}_i$, we train our model using multi-stage in a coarse-to-fine approach, wherein our model is progressively trained from low resolutions to high resolutions. To do this, we use an auxiliary set of 1×1 convolutional layers from the decoder branch of our network that produce a 3-channel image from the higher-dimensional neural activations at each level of the decoder (see Fig. 1).

Let $\mathbf{I}(\boldsymbol{\ell}_i, d)$ be the auxiliary predicted image for each level d, and let the full-resolution "auxiliary" image at the very end of the decoder be just the final predicted image itself: $\mathbf{I}(\boldsymbol{\ell}_i) = \mathbf{I}(\boldsymbol{\ell}_i, 0)$. Here d simultaneously indicates the depth of our encoder/decoder, the stage of our progressive training, and the degree of spatial downsampling. During the d'th stage of training, we use a convex combination of the auxiliary image at level d and an upsampled version of the auxiliary image at level $d+1$ as the current model prediction. Our loss at stage d is imposed between that combined image and the true image, downsampled to the native resolution of level d of our network. This approach ensures that the internal activation of our decoder at level d is sufficient to enable the reconstruction of an accurate RGB image (via the auxiliary branch), which means that the training of stage d results in network weights that are well-suited to initialize the as-yet-untrained model weights on level $d-1$ of the decoder in the next stage.

Formally, our loss at level d is:

$$\mathcal{L}_d = \left\| \left(\mathcal{D}\left(\mathbf{I}_i, d\right) - \left(\alpha_d \mathbf{I}\left(\boldsymbol{\ell}_i, d\right) + (1-\alpha_d) \mathcal{U}\left(\mathbf{I}\left(\boldsymbol{\ell}_i, d+1\right), 1\right) \right) \right) \right\|_1 \tag{1}$$

where $\mathcal{D}(\cdot, d)$ is bilinear downsampling by a factor of 2^d and $\mathcal{U}(\cdot, d)$ is bilinear upsampling by a factor of 2^d. When computing the loss over the image, we mask out pixels that are known to belong to the background of the subject. For each stage, the blending factor α_d is linearly interpolated from 0 to 1, which means that at the

Authors' addresses: Tiancheng Sun, tis037@cs.ucsd.edu; Zexiang Xu, zexiangxu@cs.ucsd.edu, University of California, San Diego; Xiuming Zhang, xiuming@csail.mit.edu, Massachusetts Institute of Technology; Sean Fanello, seanfa@google.com; Christoph Rhemann, crhemann@google.com; Paul Debevec, debevec@google.com, Google; Yun-Ta Tsai, yuntatsai@google.com; Jonathan T. Barron, barron@google.com, Google Research; Ravi Ramamoorthi, ravir@cs.ucsd.edu, University of California, San Diego.

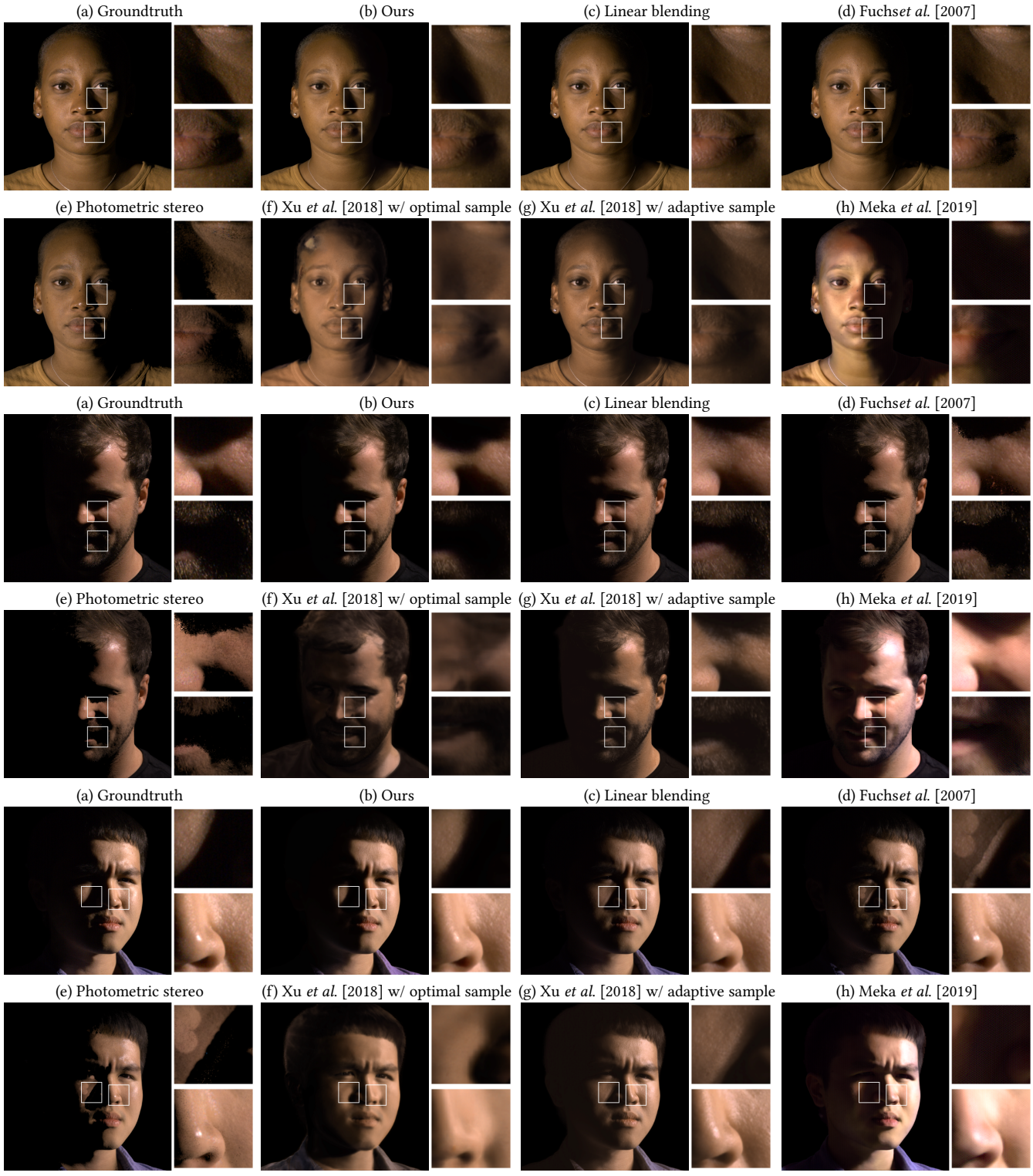

Fig. 2. Full resolution qualitative comparison between our method and other light interpolation algorithms.

beginning of that stage's training the loss is imposed entirely on an upsampled version of the last stage's predicted image, but at the end of that stage's training the loss is imposed entirely on the current stage's predicted image. These α_d factors also modulate the input to the encoder: as indicated in Fig. 1, the input to each level of the encoder is a weighted average of the output of the earlier level and a downsampled version of the input images. This means that the annealing of each α_d value has a similar effect on the progressive growing of the encoder as it does for the decoder—the deeper layers of the decoder are trained first using downsampled images, and then each finer layer of the decoder is added and blended in at each stage of training.

Our model is trained using a single optimizer instance with 4 stages, each of which corresponds to a spatial scale. For the first three stages, we train in two parts: first, 3×10^4 iterations at that stage's spatial resolution, then 2×10^4 iterations as α_d is linearly interpolated from that scale to the next. At our final stage, we train for 5×10^4 iterations. At each stage d, our model minimizes only $\mathcal{L}_d$. Note that this gradual annealing of each α_d during each scale means that the loss is always a continuous function of the optimization iteration, as $\mathcal{L}_d$ at the beginning of training for stage d is equal to $\mathcal{L}_{d+1}$ at the end of training for stage $d + 1$. In total, we train our network for 20,0000 iterations.

2 RELATED WORK COMPARISON

In Fig. 2, we present the full resolution comparison between our model and related works.

REFERENCES

Martin Fuchs, Hendrik PA Lensch, Volker Blanz, and Hans-Peter Seidel. 2007. Super-resolution reflectance fields: Synthesizing images for intermediate light directions. In *Computer Graphics Forum*, Vol. 26. Wiley Online Library, 447–456.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*.

Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. 2019. Deep Reflectance Fields: High-Quality Facial Reflectance Field Inference from Color Gradient Illumination.

Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. 2018. Deep image-based relighting from optimal sparse samples. In *SIGGRAPH*.