Seeing Through Obstructions with Diffractive Cloaking - Supplemental Document -

ZHENG SHI, YUVAL BAHAT, SEUNG-HWAN BAEK, Princeton University
QIANG FU, HADI AMATA, King Abdullah University of Science and Technology
XIAO LI, PRANEETH CHAKRAVARTHULA, Princeton University
WOLFGANG HEIDRICH, King Abdullah University of Science and Technology
FELIX HEIDE, Princeton University

ACM Reference Format:
Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Qiang Fu, Hadi Amata, Xiao Li, Praneeth Chakravarthula, Wolfgang Heidrich, and Felix Heide. 2022. Seeing Through Obstructions with Diffractive Cloaking - Supplemental Document - . *ACM Trans. Graph.* 41, 4, Article 37 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530185

In this Supplemental Document, we present additional results and details on the proposed method in the main manuscript. Specifically, we describe the

- fabrication of diffractive optical element,
- PSF calibration,
- model finetuning,
- reconstruction network architecture,
- additional simulations, and
- additional experimental validation.

1 FABRICATION OF DIFFRACTIVE OPTICAL ELEMENT

The DOE is fabricated on a 4 inch 0.5-mm-thick fused silica wafer. First, four master masks are fabricated on soda lime plates by laser direct writing on a mask maker (Heidelberg μPG 501). Second, in the photolithography step, the wafer is cleaned in Piranha solution at 115°C for 10 min to remove contaminants, and then dried with N_2 for 7 min. An auxiliary 200-nm-thick Chromium (Cr) layer is deposited by sputtering on the wafer. A 0.6-μm-thick photoresist (AZ1505) is then spin-coated on the Cr film, after HMDS (Hexamethyldisilazane) vapor priming for 20 min. We align the wafer with the master mask on a contact aligner (EVG6200 ∞) in the hard+vacuum mode, and then apply UV exposure with a dose of 9 mJ/cm^2. The patterns are then transferred from the master mask to the photoresist. We develop the photoresist in AZ726MIF for 17 sec, and clean it with De-Ionized water, and remove residual water with N_2 drying. We etch the Cr film under the photoresist with Cr etchant (mixtures of $HClO_4$ and $(NH_4)_2[Ce(NO_3)_6]$)

Authors' addresses: Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Princeton University, zhengshi@princeton.edu,yb6751@princeton.edu,sb38@princeton.edu; Qiang Fu, Hadi Amata, King Abdullah University of Science and Technology, qiang.fu@kaust.edu.sa,hadi.amata@kaust.edu.sa; Xiao Li, Praneeth Chakravarthula, Princeton University, xl2710@princeton.edu,praneethc@princeton.edu; Wolfgang Heidrich, King Abdullah University of Science and Technology, wolfgang.heidrich@kaust.edu.sa; Felix Heide, Princeton University, fheide@princeton.edu.

0730-0301/2022/7-ART37
https://doi.org/10.1145/3528223.3530185

for 1 min, and remove the remaining photoresist with Acetone. The patterns are then transferred from the photoresist to the Cr layer. Later, in the reactive-ion etching step, we dry etch the materials in the wafer with plasma of 15 sccm CHF_3 and 5 sccm O_2 at 10 °C. Only the open areas in the wafer without Cr covering are selectively etched. Once the etching is finished, we remove all the auxiliary layers. The above steps are repeated for 4 iterations to form the 16-level structures. For the design wavelength of 550 nm, the etching depths are 75 nm, 150 nm, 300 nm, and 600 nm, respectively. Finally, an additional Cr layer is deposited and etched with the aperture size to match the lens pupil. See Figs. 1 for a 3D visualization of the measured height of a section of the fabricated DOE from Zygo interfemometer.

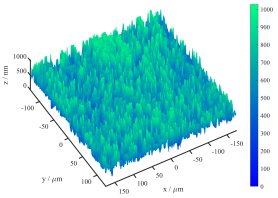

Fig. 1. Zygo interfemometer height measurement of a portion of the fabricated DOE.

2 PSF CALIBRATION

After building the experimental prototype with the fabricated DOE, we calibrated the depth-varying PSFs of our camera. To this end, we place a point light at a known distance from the camera and acquire a raw HDR image. For the point light source, we use a LED light source (Thorlabs QTH10) equipped with a diffuser and a precision pinhole (Thorlabs P500K). Cropping the high-intensity region results in a PSF at that distance and we repeat this procedure for both the background depth (5 m) and the foreground depth (15 cm).

3 MODEL FINETUNING

To improve the network performance on real captures, aside from feeding the network the PSF measurements, we also capture a real-world occlusion-free dataset to finetune our network. Specifically, we use a monitor to display different high-resolution images from DIV2K dataset [Agustsson and Timofte 2017] as the target scene, and capture 500 pairs of captures with and without using the proposed DOE. See Fig. 2 for the scene setup as well as an example image pair. During finetuning, we omit the obstruction loss, since both images are occlusion-free. Additionally, we downsample the image by 4 times before computing the per pixel ℓ_1 loss, to account for alignment inaccuracies.

Fig. 2. Left: Scene setup for the finetune data capture, Middle and Right: images captured with and without using the proposed DOE.

4 RECONSTRUCTION NETWORK ARCHITECTURE

Our reconstruction network follows a hierarchical structure. We start by performing Wiener deconvolution of the captured image with the PSFs corresponding to near (occluder) depths, which is concatenated to the captured

Table 1. **Network architecture description.** Rows enclosed between dashed lines signify non-learned operations performed on the output of preceding layers. In the table, "conv-k(a)-s(b)-IN-LRelu" represents a convolution layer with an $a \times a$ kernel window, using stride b, followed by instance normalization and a Leaky Relu ($\alpha = 0.02$) activation function. We use convT to denote transposed convolution, and Wdeconv to denote Wiener-deconvolution using the given far PSF, performed in the frequency domain.

Name	Layer Type	Output Channels
Inputs: concatenated captured image, deconvolved image		$3 \times 2 = 6$
down0_0	conv-k7-s1-IN-LRelu	12
down0_1	conv-k3-s1-IN-LRelu	12
deconv0	Wdeconv	12
down1_0	conv-k3-s2-IN-LRelu	24
down1_1	conv-k3-s1-IN-LRelu	24
down1_2	conv-k3-s1-IN-LRelu	24
deconv1	Wdeconv	24
down2_0	conv-k3-s2-IN-LRelu	48
down2_1	conv-k3-s1-IN-LRelu	48
down2_2	conv-k3-s1-IN-LRelu	48
deconv2	Wdeconv	48
down3_0	conv-k3-s2-IN-LRelu	96
down3_1	conv-k3-s1-IN-LRelu	96
down3_2	conv-k3-s1-IN-LRelu	96
deconv3	Wdeconv	96
down4_0	conv-k3-s2-IN-LRelu	144
down4_1	conv-k3-s1-IN-LRelu	144
down4_2	conv-k3-s1-IN-LRelu	144
deconv4	Wdeconv	144
Concatenating down4_2, deconv4		288
bottleneck_0	conv-k3-s1-IN-LRelu	288
bottleneck_1	conv-k3-s1-IN-LRelu	144
up4_0	convT-k2-s2-IN-LRelu	96
Concatenating up4_0, down3_2, deconv3		288
up4_1	conv-k3-s1-IN-LRelu	96
up3_0	convT-k2-s2-IN-LRelu	48
Concatenating up3_0, down2_2, deconv2		144
up3_1	conv-k3-s1-IN-LRelu	48
up2_0	convT-k2-s2-IN-LRelu	24
Concatenating up2_0, down1_2, deconv1		72
up2_1	conv-k3-s1-IN-LRelu	24
up1_0	convT-k2-s2-IN-LRelu	12
Concatenating up1_0, down0_2, deconv0		36
up1_1	conv-k3-s1-IN-LRelu	12
up0_0	conv-k3-s1-IN-LRelu	3
Concatenating up0_0, captured image		6
up0_1	conv-k5-s1-IN-LRelu	3
Summing up0_1 and captured image		3
Concatenating previous output and the captured image		6
res0	ResNet-k3-s1-d1-LRelu	6
res1	ResNet-k3-s1-d1-LRelu	6
res2	ResNet-k3-s1-d1-LRelu	6
out	conv-k7-s1-d1-Relu	3

image itself, amounting to two concatenated versions of the captured image. The spatial dimension of the network layers is gradually reduced, while additionally applying Wiener deconvolution in feature space before each

downscaling, using the far (background) PSF. The number of channels increases with each spatial dimension decrease, until reaching a bottleneck block. Subsequent layers then gradually grow in spatial dimensions, while using skip connections to exploit information from corresponding-dimension layers in the downscaling stream. We further refine the output using 3 residual blocks. See Tab. 1 for a full network specification.

5 ADDITIONAL SIMULATION

In addition to the results presented in the main manuscript, we present additional qualitative simulation results, demonstrating the capability of the proposed method to handle different types of obstructions. Figs. 3 & 4 present additional results on handling dirt splash obstructions, demonstrating how merely using our DOE (optimized to operate w/o or w/ a subsequent neural network, top rows, columns 2 & 3, respectively) already reveals some of the occluded background, while using the complete approach (including the complementary neural network, bottom rows, middle) then greatly improves reconstruction performance, bringing the background image quality close to that of the obstruction-free version (bottom right). In contrast, existing inpainting methods either result in non-plausible artifacts (CTSDG [Guo et al. 2021]) or at best hallucinate wrong details and fail to reconstruct significant content like pedestrians, cars and trees (LaMa [Suvorov et al. 2021]).

Similarly for mitigating raindrop obstructions, we show in Figs. 5 & 6 how our method can reconstruct the degraded background content (e.g. people, license plate), significantly outperforming existing inpainting methods [Guo et al. 2021; Suvorov et al. 2021] or even the AttentiveGAN method [Qian et al. 2018], designated for raindrop removal. Finally, Figs. 7 & 8 show more examples of coping with fence obstructions in a point-and-shoot photography scenario, demonstrating the advantage of our method in reconstructing occluded background content (e.g. porch roof, soccer player's body, facial details) over existing inpainting methods [Guo et al. 2021; Suvorov et al. 2021], as well as over the designated visual fence removal DefenceNet method [Matsui and Ikehara 2020].

We further present ablation results in Fig. 9 & 10 to clarifying the role of different components in the proposed framework. In particular, we show the role of the complementary neural network by comparing to results obtained by applying classic Wiener filtering or Richardson Lucy deconvolution on our DOE captures. We also compare our results with those of an image-to-image translation [Ronneberger et al. 2015] and image deblurring [Kupyn et al. 2019] methods applied *on top of our* DOE captures, which seem to struggle to remove the color aberrations induced by the foreground obstructions. We additionally qualitatively demonstrate the benefit of jointly, rather than sequentially, optimizing the DOE and subsequent reconstruction network.

6 ADDITIONAL EXPERIMENTAL VALIDATION

We present additional experimental results obtained using our prototype in Fig. 11. The results on six captured scenes correspond to the dirt splash, raindrop and fence obstruction scenarios, demonstrating our method's capability to produce high quality reconstructions of the occluded background scenes. This is in contrast to capturing with a conventional camera (4^{th} column), even when post processing using the LaMa method [Suvorov et al. 2021] (middle column). Since Lama is an image inpainting method, its performance depends on the additional inpainting mask input we provide. As we show in Fig 12, using a conservative (thinner) mask may allow it to exploit information from partially occluded regions, but typically does not yield much improvement compared to the degraded input, while using a more aggressive (thicker) mask may exhibit a more significant change, at the cost of being unfaithful to the existing latent scene. Therefore, for the comparison in Fig. 11 we experimented with various mask thicknesses, and hand-picked the one yielding the most appealing result for each scene.

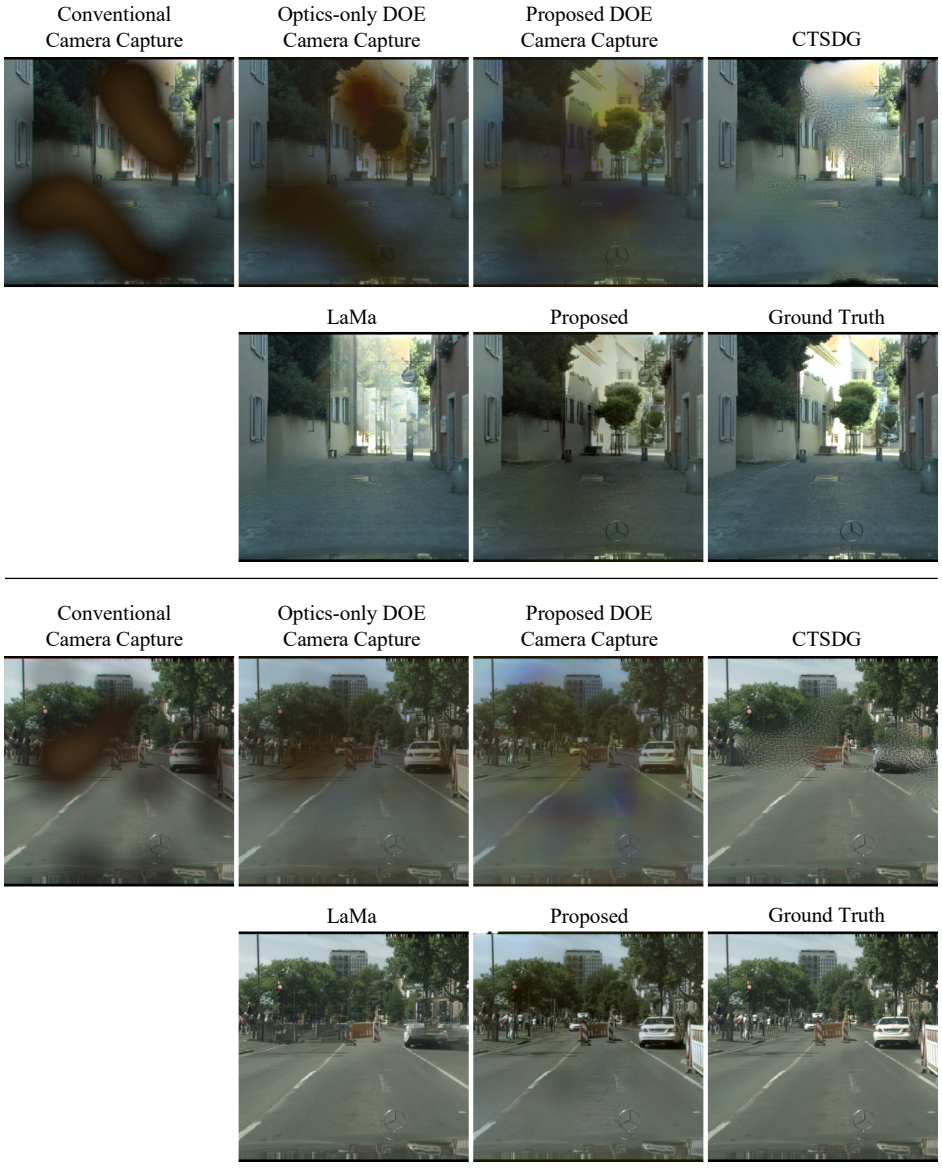

Fig. 3. **Additional dirt-splash obstructions results.** Our DOE (optimized to operate with or without a subsequent neural network) allows us to see through occlusions. Feeding the captured image into our neural network results in an almost obstruction-free image. Existing inpainting methods instead hallucinate or deblur the occluded background regions. Results of existing methods (CTSDG & LaMa) that try to hallucinate the occluded regions often lack some significant content (e.g. missing trees in the top example, and pedestrians in the bottom example) or even exhibit significant artifacts.

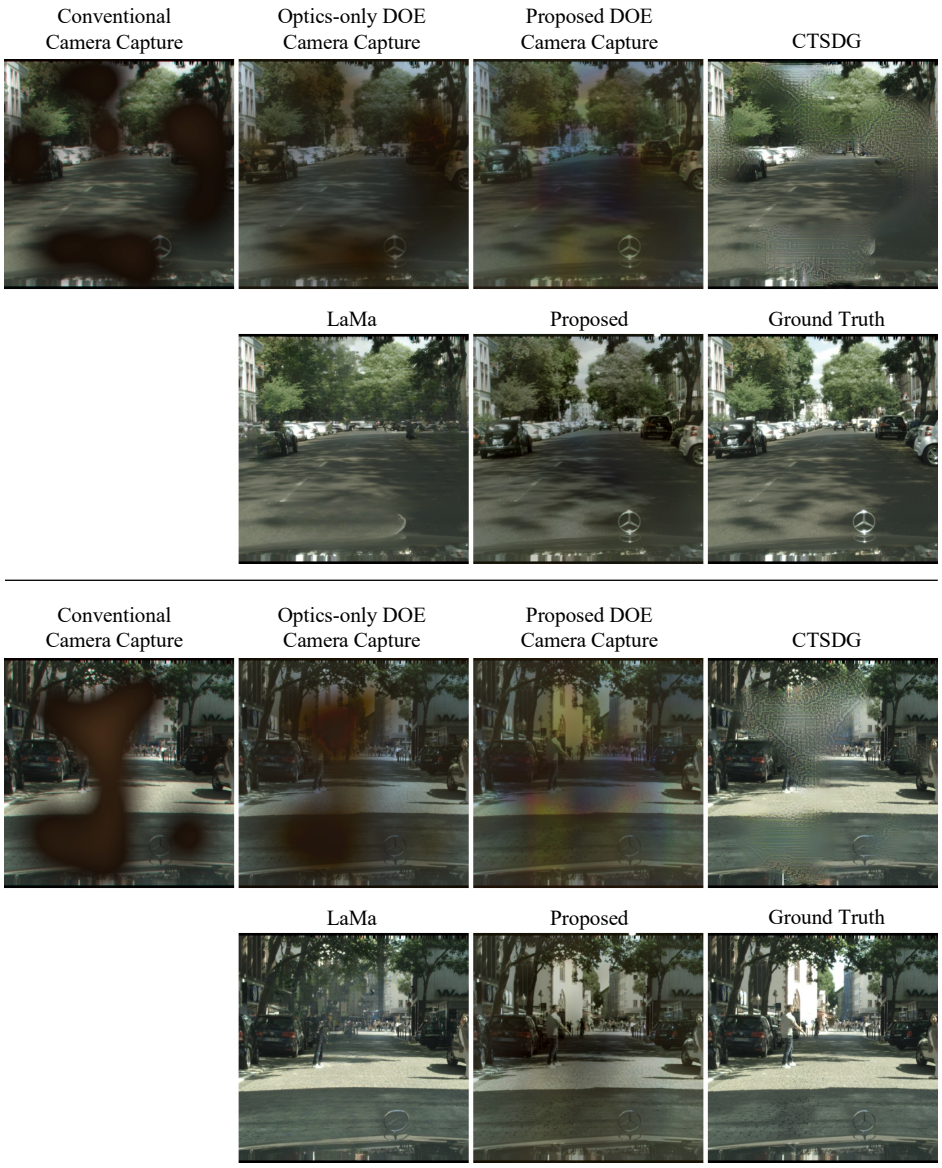

Fig. 4. **Additional dirt-splash obstructions results.** Our DOE (optimized to operate with or without a subsequent neural network) allows us to see through occlusions. Feeding the captured image into our neural network results in an almost obstruction-free image. Results of existing methods (CTSDG & LaMa) that try to hallucinate the occluded regions often lack some significant content (e.g. missing cars in the top example, and pedestrians in the bottom example) or even exhibit significant artifacts.

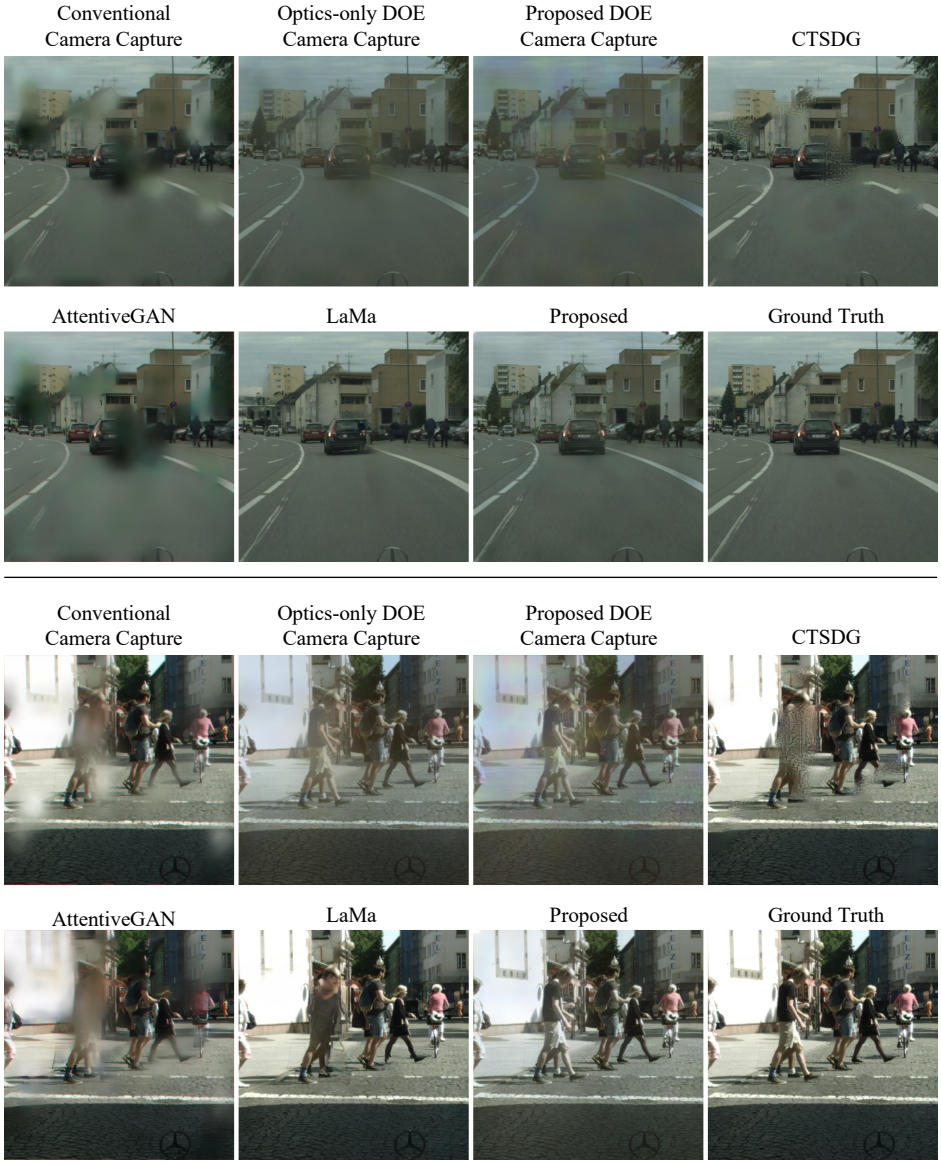

Fig. 5. **Additional rain-drop obstructions results.** Our DOE (optimized to operate with or without a subsequent neural network) allows us to see through occlusions. Feeding the captured image into our neural network results in an almost obstruction-free image. Existing inpainting methods (CTSDG and LaMa), as well as a designated raindrop removal AttentiveGAN method [Qian et al. 2018], often fail to reconstruct important visual details, e.g. license plates in the top example and people in the bottom example.

Fig. 6. **Additional rain-drop obstructions results.** Our DOE (optimized to operate with or without a subsequent neural network) allows us to see through occlusions. Feeding the captured image into our neural network results in an almost obstruction-free image. Existing inpainting methods (CTSDG and LaMa), as well as a designated raindrop removal AttentiveGAN method [Qian et al. 2018], often fail to reconstruct important visual details, e.g. license plates in both examples.

Fig. 7. **Additional fence obstructions results.** Our DOE (optimized to operate with or without a subsequent neural network) allows us to see through occlusions. Feeding the captured image into our neural network results in an almost obstruction-free image. Existing inpainting methods (CTSDG and LaMa), as well as a designated visual fence removal DefenceNet method [Matsui and Ikehara 2020], merely attempt to hallucinate the occluded regions, which leads to loss of significant content, e.g. the soccer-player's body in the top example and the porch's roof in the bottom example.

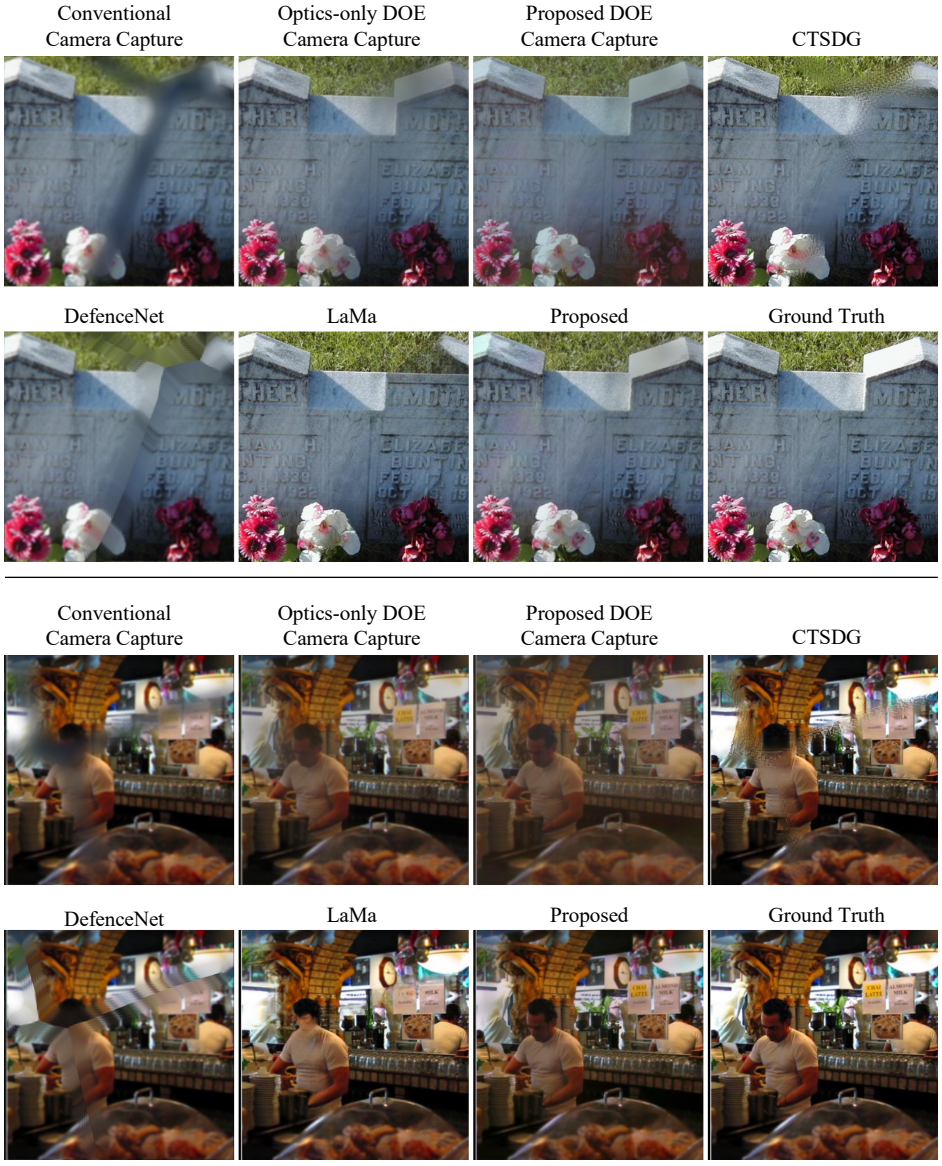

Fig. 8. **Additional fence obstructions results.** Our DOE (optimized to operate with or without a subsequent neural network) allows us to see through occlusions. Feeding the captured image into our neural network results in an almost obstruction-free image. Existing inpainting methods (CTSDG and LaMa), as well as a designated visual fence removal DefenceNet method [Matsui and Ikehara 2020], merely attempt to hallucinate the occluded regions, which leads to loss of significant content, e.g. the right-side gravestone's top in the top example, and the person's facial details in the bottom example.

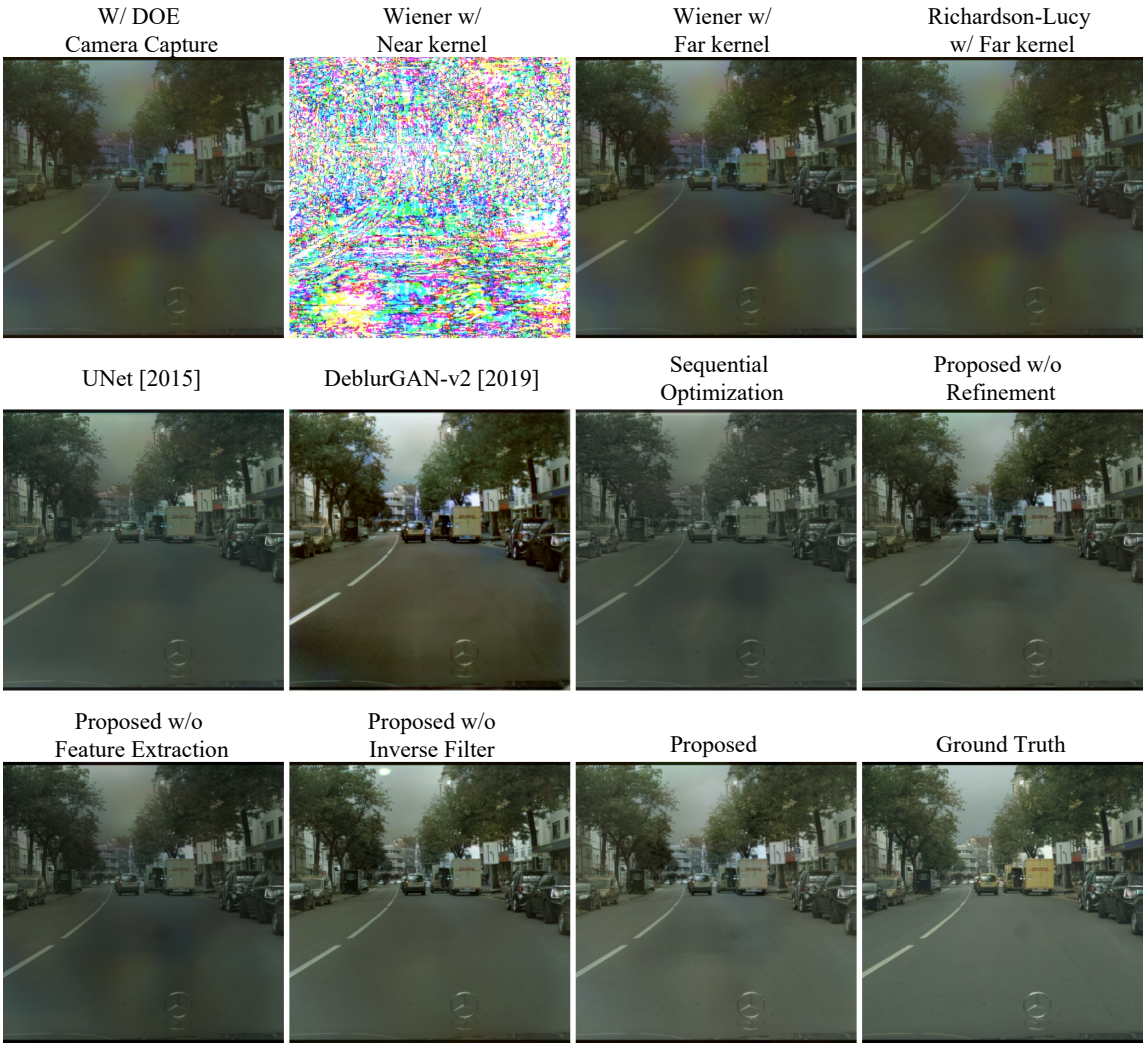

Fig. 9. **Additional ablation examples.** Applying Wiener filtering or Richardson-Lucy deconvolution by itself using the DOE PSF corresponding to far distances fails to remove the chromatic aberrations from the near scene obstructions. Using the near-scene PSF results in uninformative reconstructions to human observers. Conventional image-to-image mapping (UNet [Ronneberger et al. 2015]) and SOTA image deblurring (DeblurGAN-v2 [Kupyn et al. 2019]) struggle to recover the true color of the background due to the aberrated foreground obstructions. Sequentially optimizing the DOE and reconstruction network produces inferior results compared to the proposed end-to-end approach. We attribute this to the difficulty of designing optimal intermediate losses (on the PSFs) for this multi-objective problem.

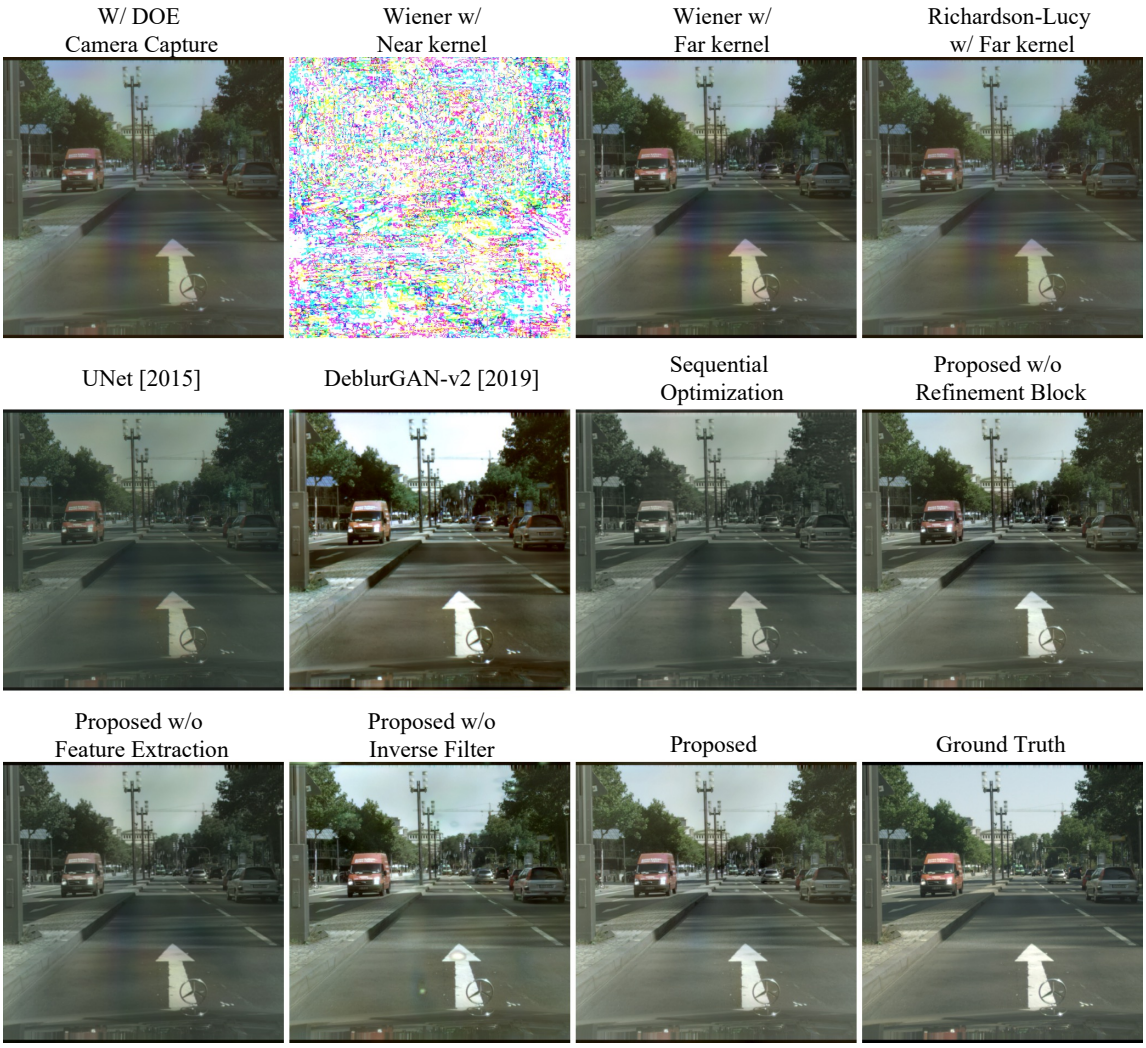

Fig. 10. **Additional ablation examples.** Applying Wiener filtering or Richardson-Lucy deconvolution by itself using the DOE PSF corresponding to far distances fails to remove the chromatic aberrations from the near scene obstructions. Using the near-scene PSF results in uninformative reconstructions to human observers. Conventional image-to-image mapping (UNet [Ronneberger et al. 2015]) and SOTA image deblurring (DeblurGAN-v2 [Kupyn et al. 2019]) struggle to recover the true color of the background due to the aberrated foreground obstructions. Sequentially optimizing the DOE and reconstruction network produces inferior results compared to the proposed end-to-end approach. We attribute this to the difficulty of designing optimal intermediate losses (on the PSFs) for this multi-objective problem.

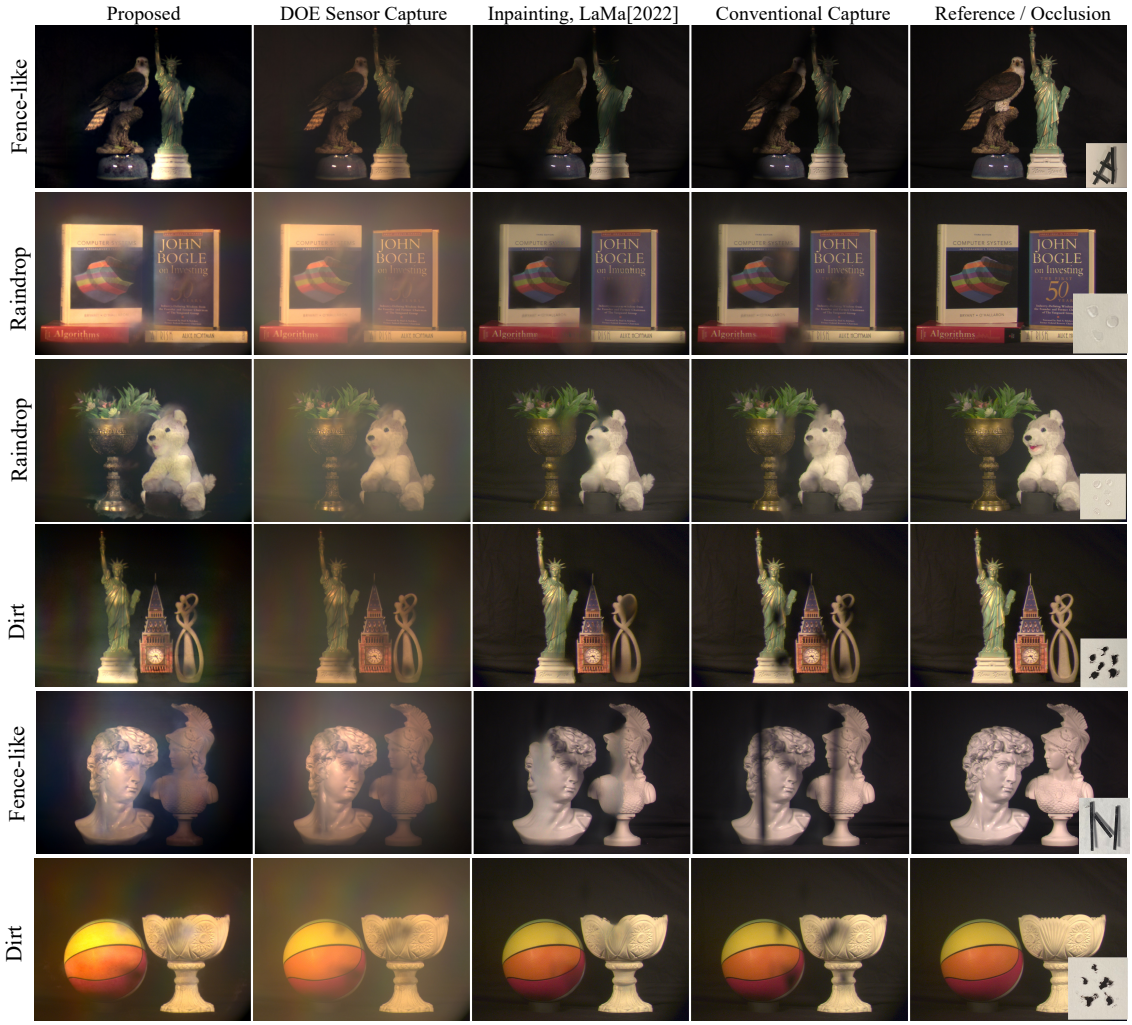

Fig. 11. **Additional experimental assessment results.** Our method (left column) is able to restore image regions that would otherwise be occluded by the obstruction, see the conventional camera column. Merely using the DOE without subsequent processing (second column) already provides more visual information compared to using a conventional camera. With help of the reconstruction network, the proposed computational camera significantly outperforms recent image inpainting methods LaMa [2021]. Post-capture inpainting fails to recover scene details (e.g., text) and sometime the entire regions due to lack of information.

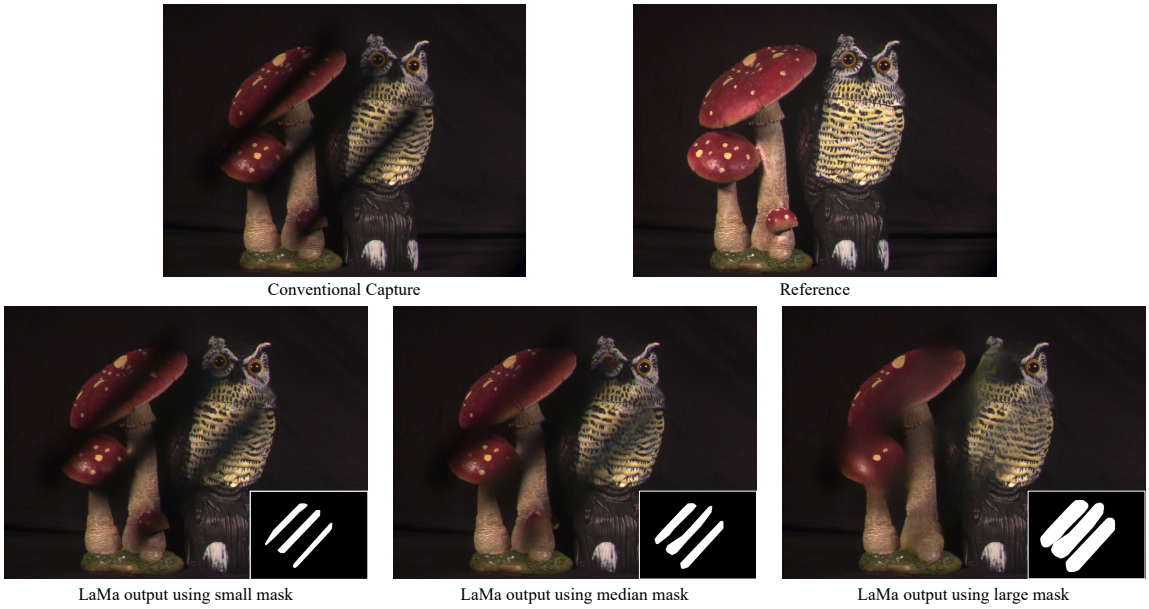

Fig. 12. **Mask-dependent results using the LaMa [2021] inpainting method.** To compare our method to the Lama image inpainting method, we manually marked masks indicating the occluded regions. However, the performance of Lama depends heavily on the input mask. Here we show that a very conservative mask (bottom left) can result in an output visually no different to the occuded input, while an aggressive mask (bottom right) can trade potentially useful information for a visually smoother output. For a fair comparison, we created several masks for each scene and picked the one that results in the best visual quality.

REFERENCES

Eirikur Agustsson and Radu Timofte. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 126–135.

Xiefan Guo, Hongyu Yang, and Di Huang. 2021. Image Inpainting via Conditional Texture and Structure Dual Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14134–14143.

Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. 2019. DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. In *The IEEE International Conference on Computer Vision (ICCV)*.

T. Matsui and M. Ikehara. 2020. Single-Image Fence Removal Using Deep Convolutional Neural Network. *IEEE Access* 8 (2020), 38846–38854.

Rui Qian, Robby T. Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. 2018. Attentive Generative Adversarial Network for Raindrop Removal from a Single Image. arXiv:cs.CV/1711.10098

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions. *arXiv preprint arXiv:2109.07161* (2021).