Supplementary: Stochastic Light Field Holography

Florian Schiffers, Praneeth Chakravarthula, Nathan Matsuda, Grace Kuo,
Ethan Tseng, Douglas Lanman, Felix Heide, Oliver Cossairt

Abstract—This document provides further information on the implementation of our holographic display prototype. This document provides further information on the implementation of our holographic display prototype. It includes discussions on the optimization models, as well as additional experimental results. The supplementary is also accompanied by further discussions that are too detailed for the main manuscript. For a better visualization of the 3D prototype, please refer to the accompanying video, as this provides the best direct comparison between the methods.

Index Terms—Computational Display, Holography, Light Field, Wigner Distributions, Near-Eye Display, VR/AR

1 Executive Paper Summary

This research paper contributes to advancing the field of near-eye holography by addressing an important gap in the current literature: It is the first to identify accurate 3D-ocular parallax-and defocus as a gap in current near-eye holography research and hence introduces a solution for optimizing holograms with accurate 3D-ocular parallax and focus cues. Our core idea is to enforce radiometric consistency between refocused light field and holographic display. The paper's contributions include developing novel CGH algorithms based on this approach and addressing the problem of providing accurate 3D-ocular parallax, which is a necessary step toward building practical holographic displays.

2 Algorithm Recap

Our algorithm optimizes for a phase-only holograms to produce high-fidelity LF-reconstructions with a uniform energy distribution across the eyebox. The core idea is a novel intensity-based reconstruction loss where the projection operators for the light fields and the Wigner-function are matched. The Wigner projection for the coherent field is implemented as a memory-efficient convolutional image formation model based on the angular spectrum method [1]. The corresponding light field projections are synthesized on the fly from the underlying light field representation [2]. By stochastically sampling different pupil states in each iteration, a phase-only hologram is optimized that provides accurate depth cues within a large eye-box.

In addition to the main paper, we show the flow of our algorithm in Fig. 1 as well as an algorithm block in 1.

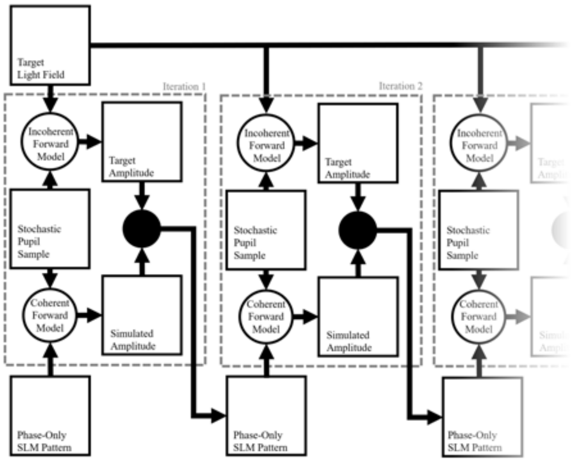

Fig. 1: *Algorithm Flow* A schematic that sketches the algorithm flow of stochastic light field holography. At each iteration in the algorithm, a random set of pupils are stochastically sampled. Both light field and Wigner projection operators are computed on-the-fly to compute the loss-function that depends on the sampled pupil state. Then, the loss is back-propagated until the SLM-pattern. Note that only the coherent forward model needs to be differentiable.

ALGORITHM 1: Stochastic-Light Field Holography (SLFH)

```
Input: Image Sampler $S$, regularization parameters $\lambda$,
       learning rate $\eta$, phase initialization variance $\sigma$
Output: Model parameters $\Theta = (\phi, s)$; Phase pattern $\phi$,
        intensity scale $s$
$\phi \sim \mathcal{N}(0, \sigma)$, $s \leftarrow 1$; repeat
    $(p, q, z, a) \leftarrow$ SampleRandomAperture();
    $I_{\text{target}} \leftarrow$ ComputeRefocusedImage(p,q,z,a);
    $I_{\text{model}} \leftarrow$ ComputeForward($\Theta$; p,q,z,a);
    $\Theta \leftarrow \Theta - \eta \left( \frac{\partial}{\partial \Theta} L(I_{\text{target}}, I_{\text{model}}) \right)$ ;
until stopping criterion is not met;
```

- *F. Schiffers is with the Computer Science Department at Northwestern University*
- *F. Schiffers, N. Matsuda, G. Kuo, D. Lanman and O. Cossairt are with Reality Labs Research, Meta*
- *P. Chakravarthula, E. Tseng and F. Heide are with Princeton University*

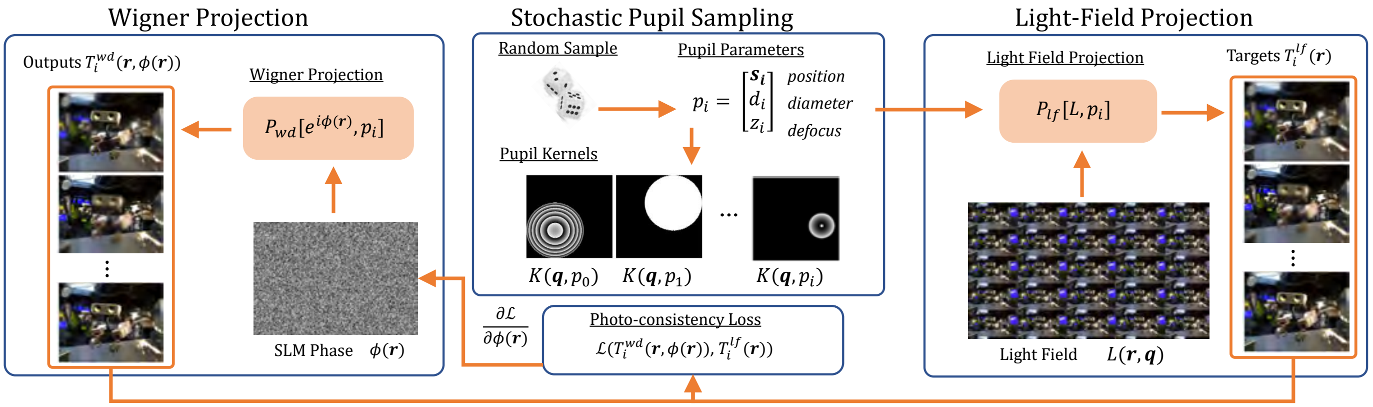
Wigner Projection
Stochastic Pupil Sampling
Light-Field Projection
Outputs $T_i^{wd}(\boldsymbol{r}, \phi(\boldsymbol{r}))$
Wigner Projection
$P_{wd}[e^{i\phi(r)}, p_i]$
SLM Phase $\phi(\boldsymbol{r})$
Random Sample
Pupil Parameters
$p_i = \begin{bmatrix} \boldsymbol{s}_i \\ d_i \\ z_i \end{bmatrix}$
position
diameter
defocus
Pupil Kernels
$K(\boldsymbol{q}, p_0)$
$K(\boldsymbol{q}, p_1)$
...
$K(\boldsymbol{q}, p_i)$
$\frac{\partial \mathcal{L}}{\partial \phi(\boldsymbol{r})}$
Photo-consistency Loss
$\mathcal{L}(T_i^{wd}(\boldsymbol{r}, \phi(\boldsymbol{r})), T_i^{lf}(\boldsymbol{r}))$
Light Field Projection
$P_{lf}[L, p_i]$
Targets $T_i^{lf}(\boldsymbol{r})$
Light Field $L(\boldsymbol{r}, \boldsymbol{q})$

Fig. 2: *A more detailed schematic that explains our loss-function.* At the core is the random sampling of pupils which gives rise to the projection geometry for both the coherent (Wigner projection) and incoherent (light field projection) light transport. At each iteration new projection are generated which then defined the photo-consistency loss (l_2). Auto-differentiation of this loss through the coherent forward model gives the update rule for the phase-only SLM pattern.

3 Implementation Details

The algorithm is implemented in PyTorch [3], and all gradients are computed via auto-differentiation using Wirtinger derivatives [4]. **All code will be made available open-source** (at date of publication) using easy-to-run Jupyter notebooks to enhance reproduction of our results.

3.1 Used Metric for Loss

As a general metric for our loss-function, we chose L_2 for simplicity and a fair comparison among different experiments. We acknowledge that more sophisticated metrics to measure image similarity, such as perceptual losses [5], are likely leading to slight performance improvements at the cost of larger computing. However, we argue that the majority of improvement comes in engineering more sophisticated losses that target the actual content and model.

3.2 Global scale optimization

As we are optimizing for random phase-only patterns, the integrated (total) energy contained in the holographic field is constant at all times as no energy is absorbed. Our goal is to optimize for an eye-box with roughly uniform energy distribution as there shouldn't be any preference for pupil positions. In other words, independent of where we sample the holographic field, we should get roughly consistent intensities in the eye-box if each pupil as has the same pupil diameter. However, it is also important to note that the image intensity scales *quadratically* with the pupil radius. Fortunately, if done correctly, light field refocusing correctly accounts for the pupil radius and the refocused/synthesized target images should scale in intensity accordingly.

However, as the SLM is phase-only, we still have to learn a scale variable [6] representing the product of exposure time and laser-power to correctly balance out the intensities of each color-channel. For this reason, we introduce two learning rates: One for the SLM ($\text{lr}_{SLM} = 0.1$) and one for the scale ($\text{lr}_{scale} = 1.0$).

3.3 Optimizer

As an optimizer, ADAM [7] is used for faster convergence compared to gradient descent. Convergence is typically achieved within a few hundred iterations. However, as with most gradient descent algorithms, the exact convergence rates depend heavily on the chosen learning rate and the specific image content to be learned.

4 Light field Refocusing

This section outlines more details on how the Light Field refocusing is performed.

4.1 Light Field Resolution

Our Light Fields are originally rendered with different numbers of views. E.g. the DeepSpaces dataset [8] consists of $9x9$ views and the robot scene is a $7x7$ scene. For a fair companion with STFT, we sub-sample our light fields to a $7x7$ light field as memory constraints for STFT limit the achievable angular resolution. E.g., a $9x9$ light field was already too much to fit into a 48GB GPU when the STFT-algorithm was used. Furthermore, **we want to stress** that this is not a limitation of our approach, as our algorithm is agnostic to the angular resolution of the light field.

The spatial resolution of each light field was matched to the resolution of the SLM $(1080 \cdot 1920)$. However, as our sampled pupils are much smaller than the full aperture, the expected resolution during play-back is actually smaller than the chosen light field resolution. This is because the pupil effectively band-limits our signal.

E.g. if we would like to encode $10 \cdot 10$ view (100 views), the maximal information that we could potentially encode into each view cannot be larger than $108 \cdot 192$ pixels.

4.2 Defocusing/Aliasing Artifacts

Our paper uses the shift-and-add algorithm [9] to synthesize novel views from the given light fields. This algorithm is a commonly used method for refocusing images from a light field dataset. This algorithm involves shifting the sub-aperture images in the light field along their corresponding epipolar lines and adding them together to form a final refocused image. However, this process can lead to aliasing artifacts in the final image.

Aliasing artifacts occur when the sub-aperture images are undersampled or when the sampling rate is too low. This can happen when the pixel size of the sub-aperture images is larger than the Nyquist frequency of the light field data.

When the pixel size of the sub-aperture images is too large, high-frequency information is lost, and the shift-and-add algorithm cannot recover it. This can lead to aliasing artifacts such as moiré patterns or jagged edges in the final refocused image. These artifacts occur because the sub-aperture images are not accurately sampled and reconstructed.

The easiest solution to avoid aliasing artifacts would be to use a Light Field with a larger angular resolution. However, this is not possible using the STFT approach due to memory limitations.

In addition to a fair comparison with STFT, we wanted to fix the angular resolution to something reasonable, as many views are not always attainable due to practical constraints.

Interestingly, even though some of the synthesized target images are aliased, the optimized hologram often doesn't exhibit the same aliasing artifacts when the image is reconstructed (both in simulation and experiment).

Future investigations on the importance of avoiding these aliasing artifacts are warranted but out of the scope of this submission. There are many approaches that tackle angular aliasing in Light Field refocusing which are potential directions to look into in future work. Among those are removing angular aliasing [10], directly synthesizing refocused images using neural-networks [8], super-resolve in angular resolution using prior information [11] or fast neural-radiance fields [12] storing a highly compressed light-field.

4.3 Geometry for Light Field projections

4.3.1 Wigner Projection (Coherent light)

In order to match the incoherent and coherent projection operator, we need to set the geometry correctly. The projection geometry is defined in physical world coordinates. For the coherent model, the aperture plane coordinate is defined by the eyebox size given by

$$\text{size}_{\text{eyebox}}(\lambda) = \frac{\lambda \cdot f}{dx}, \tag{1}$$

where λ corresponds to wavelength (660nm, 520nm, 440nm for our laser module), f to the focal length of the first lens, and dx the SLM-pixel spacing. Hence, the size of the eye-box depends on the wavelength. In the experimental setup, an arbitrary, physical aperture with a finite size d at a location $s = (x, y)$ is sampling the eye-box to generate different views. This has implications for how the frequency space (in normalized coordinates) has to be sampled for each wavelength. In Fig. 5, we show the effects of this wavelength dependence. First, as predicted in Eq. 1, the size of the employed aperture change with wavelength. Second, and this is slightly less intuitive, the shift of the physical apertures *also* corresponds to different spatial frequency coordinates. In other words, the sampled ASM-kernels do not have only different sizes for each wavelength but are also centered at different locations.

We confirm this behavior in an experiment by capturing images of the eye-box using a second arm in our experiment. For this, we compute a Fourier-CGH (in Far-Field Hologram) of two calibration targets and display them for each color channel, see Fig. 6. Note how the calibration target changes in size with changing wavelengths. The second row shows circles of different sizes that reflect the behavior seen in Fig. 5.

As a consequence, the SBP of the red and green channels cannot efficiently be used, as the maximal eye-box that can be employed for an RGB-images is defined by the smallest wavelength (*blue*). The ratio between the largest (red, 660nm) and the smallest (blue, 440nm) wavelength is 0.66, which actually means that only $44\% = 0.66^2$ of the red eye-box size can actually be used.

4.3.2 Light field projection geometry

In order to match the projection geometries for both Wigner and light field, we need to establish the correct coordinate systems. While the aperture coordinates for the Wigner-projection were wavelength dependent, the light field itself does not depend on the wavelength. This is because, in practice, each rendered/captured RGB-image in the LF-scene was captured from the same viewpoint.

However, this is at odds with the coherent model: If we would sample pupils with the same normalized-spatial frequency coordinates, as it was e.g. done in [13] or is done inherently in STFT approaches, we would create a mismatch with the light field projection.

The geometry for the light field needs to be defined for both the coordinate system in the (s,t)-plane and the aperture plane (u,v). Here, we list the parameters defined by those values:

- *Base-line spacing (u,v):* spacing between rendered view-points
- *Plane Spacing:* Distance between aperture and image plane
- *Pixel spacing (s,t):* The pixel spacing of the light-field resolution

For our setup, the imaging lens corresponds to a $200mm$ lens which will define the **plane spacing** between the two geometries. The **baseline spacing** is defined by the rendering aperture. In our implementation, we roughly match it to the eye-box size for the blue channel.

Let us consider the following example: For a 22mm eyebox, the baseline spacing is defined by

$$\text{dx}_{\text{baseline}} = \frac{\text{size}_{eyebox(blue)}}{N_{views} - 1} \tag{2}$$

where N_{views} defines the number of views horizontally. The baseline spacing in vertical direction $textdy$ is computed accordingly.

The pixel spacing is matched to the size of the FoV of the SLM. If the SLM has a height of $\text{size}_{SLM} = 8um * 1080(\text{pix}) \approx 8mm$, the pixel spacing is simply defined as

$$\text{dx}_{lightfield} = \frac{\text{size}_{SLM;x}}{N_{lightfield;x}} \tag{3}$$

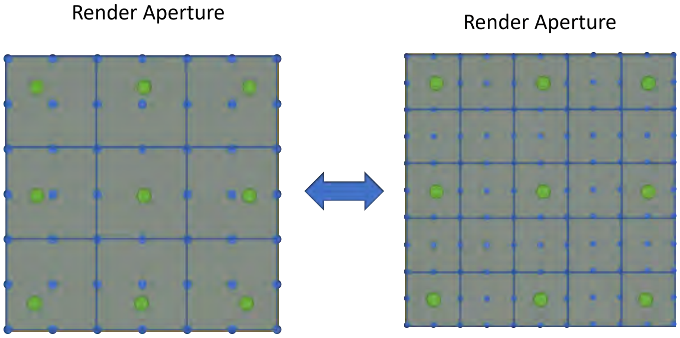

Fig. 3: A simple downsampling operation of the light field will result in wrong pupil centers. This is here visualized via the green centers which are mismatched with the blue pupil centers in the interpolated coordinate grid.

where $N_{\text{light field};x}$ corresponds to number of spatial pixels in the light field.

4.3.3 Resampling Light Fields to different angular resolutions

Typically a light field is given with a specific angular resolution, e.g. a $9x9$ light field. In normal vision applications, one typically interpolates the angular resolution to upsample or downsample the light field. Downsamplig/Upsampling of light fields, works without artifacts as long as the baseline is small enough to ensure that the light field doesn't alias. We already discussed aliasing artifacts earlier in Sec. 4.2.

However, in the case of holography, one cannot simply apply a cropping or downsampling operation to the light field. This is because the geometry is defined by the holographic field, and the center of the rendered images is not aligned with the edges of the physical aperture. In Fig. 3, we visualize the center positions that the Wigner projections assume for a $3x3$ and a $5x5$ light field. If the interpolation is not done correctly, conversions between different subsampled light fields lead to view inconsistencies.

This makes it particularly hard (or at least cumbersome) if one would like to train holograms using STFT optimization with different-sized angular light field resolutions. This becomes even more problematic when light fields with large baselines are used, where aliasing occurs quickly.

4.4 STFT and light fields

In the main paper, we discussed how the short-time-Fourier-transform [14], [15] is used to compute a light field from a complex field. The STFT, together with the window function, inherently defines the pupil that is applied in Fourier domain. Furthermore, by nature, the STFT always samples a grid of different pupil positions. We show examples of the effective pupil function for a $5x5$ and a $9x9$ 7. Note that the pupil function is effectively close to a Gaussian. In reality, a pupil/iris would be close to a binary aperture function. However, a binary aperture function cannot be used in an STFT approach as it contains high frequencies that would require a huge window function, drastically increasing the required memory space.

5 Comparison to Smooth Phase-Holograms

5.1 Comparison to Point-Integration Methods

One might argue that the point integration method, such as those proposed by Maimone et al. [16], is designed in a way that they are insensitive to pupil movement. However, this is not the case, as we will argue now. A single quadratic phase point (or plane wave) would indeed be pupil insensitive. However, this breaks down in an image with many interfering quadratic phase points. To understand this better, let us revisit how [16] works. Each point in the 3D object is projected onto the SLM-image plane using a quadratic phase with a depth-dependent slope. For a 3D object, we can control the phase of each object point. Let us first study the case where the object carries a constant phase. Here, the intensity in the eyebox resembles the Fourier transform of the image. I.e., the vast majority of the energy is contained in the very center of the eyebox for natural images, and high frequencies are contained in the edge of the eyebox, and are very low energy. If the eye moves around so that it does not consume the whole eyebox, various image frequencies will be lost, making the system effectively pupil insensitive. Mathematically, the point-wise integration method is the convolution of the target image I with the quadratic phase function QP, described as

$$\mathcal{F}\left(\mathcal{F}\{I\} * \mathcal{F}\{QP\}\right), \tag{4}$$

In the scenario of using a lens to direct rays to the eyebox as in Maimone's work, which approximates a Fourier transformer, the distribution in the eyebox can be approximated as:

$$\mathcal{F}\{I\} \cdot \mathcal{F}\{QP\} \tag{5}$$

If the phase of the image is constant, $\mathcal{F}\{\text{image}\}$ will be the Fourier transform of the real target image, which tends to be highly concentrated in the low-frequency regions for natural images. Consequently, the distribution

$$\mathcal{F}\{\ I\} \cdot \mathcal{F}\{QP\} \tag{6}$$

will also have a high concentration of energy in low frequencies, which is in the center of the eyebox when Fourier is transformed by the lens. While one could potentially try to assign a random phase to each object point, each wavelet will now start to interfere with neighboring points. Essentially a speckle field will be created, and the resulting complex wavefront could potentially be used to display a 3D wavefield. However, as of today, we do not have access to a complex modulator; hence it is unclear how to display such a wavefront. Hence, to the best of our knowledge, the pointwise integration method of by [16] can only work with a smooth-phase approximation, such that DPAC can be efficiently used to encode the complex-wavefront into a phase-only modulation.

5.2 DPAC-modulation

The DPAC method, e.g. used by [16], encodes the complex field using a phase-only field that is then propagated from the target plane onto the SLM plane. In DPAC, a high-frequency checkerboard pattern allows the complex encoding into the phase-only modulation. However, it also requires a physical aperture to low-pass filter the image such that neighboring pixels can interfere and form the desired complex number. The low-pass filter effectively filters out the higher-order signal copies that are created at the edges of the eye-box by the employed checkerboard pattern. Since we typically assume a zero phase on the target, the propagated field will be smooth. As a consequence, the field's Fourier spectrum will be concentrated around the zero domain. It is possible to introduce phase noise into the target plane to broaden the spectral domain. However, this will quickly cause the image quality of DPAC-created holograms to deteriorate. This is because the more noise is introduced into the field, the bandwidth of the holograms becomes larger. The peaks are created by the checkerboard pattern overlap and reduce image quality.

5.2.1 Comparisons to recent 3D holograms

In recent years, many 3D holographic algorithms similar to [17], [18] have been proposed. Most of these algorithms create holograms that are based on the assumption of a smooth object smooth phase. The goal of those papers is to show outstanding image quality; hence they don't have any consideration or discussion of whether their holograms are smooth-phase or random-phase. Recent studies, such as [19] investigated smooth-phase vs random-phase holograms and concluded, that smooth-phase holograms might not be able to drive accommodation for the human perceptual vision system. We hence do not compare algorithms that produce a smooth phase to our method, which assumes a random phase for the object phase.

To prove our claim that these methods produce random phase, we have implemented [17] for a layered focal-stack scene and report qualitative results obtained by a simulation in Fig. 4 We evaluate the hologram for front and back focus and then move the pupil slightly outside of the eyebox. For pupils that fully sample the DC-peak, image quality is great for both focus positions. However, once the pupil moves, the low-frequencies are no longer sampled. This will greatly reduce image quality to a point, where the image cannot be recognized anymore at all; second, the image intensity drops drastically as almost light is centered around the DC term.

In summary: Smooth phase holograms are able to produce best-of-class image quality for the center view, which is basically indistinguishable from the ground truth. However, they can only do that as long as the center view is sampled, and once the pupil slightly moves away, they completely lose all image contrast. For this reason, smooth-phase holograms are a somewhat orthogonal research direction, and we do not consider them in our evaluation.We acknowledge, though, that further perceptual studies are required to analyze the perceptual response between smooth-phase and random-phase holograms beyond existing work such as [19].

6 Eye-Box size and practical issues in LCoS-based SLMs

This section focuses on a practical issue of light distribution within the eye box that we discussed shortly in the main paper. Many recent studies on holography do not address this issue, and it is often assumed that the eye box is solely determined by the system Etendue, which defines the maximum diffraction angle given by the SLM-pixel pitch. However, for most near-field hologram algorithms, the light distribution in the eye-box is non-uniform and heavily concentrated around the DC-term. This concentration of spectral energy leads to a significant reduction in the "effective eye-box size".

In addition to the limitation of a small eyebox, there is another compelling argument against the use of smooth-phase holograms. This is due to the susceptibility of a tiny eyebox to floaters caused by scattering particles within the human eye, as well as other factors such as strong ringing artifacts from eyelashes. While previous work has demonstrated exceptional quality using smooth-phase holograms, their practicality is questionable. It is imperative that the effective eyebox be larger than a concentrated peak to avoid these issues, and randomized modulation is the only viable solution. There is a series of papers employing random-phase holograms, such as the works dealing with expansion setups [20], [21], [22]. However, it is important to notice that their image quality is also of significantly lower visual quality than papers such as those presented by [6], [18], [23], [24]. However, we emphasize again that the random phase hologram is required to unlock the true powers of holographic display. Hence, future research will have to tackle the problem of of how comparable image quality is achieved with random phase holograms.

6.1 Eye-Box using DPAC

6.2 Eye-Box using Gradient Descent

Similar characteristics are observed when computing CGHs based on gradient-descent style algorithms. The optimized hologram depends heavily on the chosen initial values. If we start with a constant phase or relatively small noise perturbation around zero, the optimized SLM patterns also tend to be smooth phases. A significant increase in the initial noise at the initial phase patterns leads to more random-looking SLM patterns, which have wider support in their Fourier Spectrum. If a wide eye box is a goal, one would hence optimize their holograms using a fully random initialization. This would inherently allow computing pupil-invariant 2D images. However, a noise-free image can only be computed if the full-aperture is sampled. Once smaller pupils sample the eye-box, speckle inherently starts to occur. To the best of our knowledge, temporal multiplexing [14] is currently the only way to effectively reduce speckle in this content. Partial coherence [23] could potentially be used to reduce speckle, but it likely comes at the cost of significant low-pass filtering of the displayed image.

6.3 Field-Fringing limits image quality

Here we discuss the important problem of field-fringing which comes with SLM based on LCOS-technology. Phase-SLMs based on LCOS-technology often exhibit strong field-fringing artifacts which drastically reduces image quality for holographic displays, especially when high-frequency content is displayed.

Field-fringing, also known as cross-talk, is a phenomenon that occurs naturally due to the crystalline material used in LCOS technology, which causes a phase shift

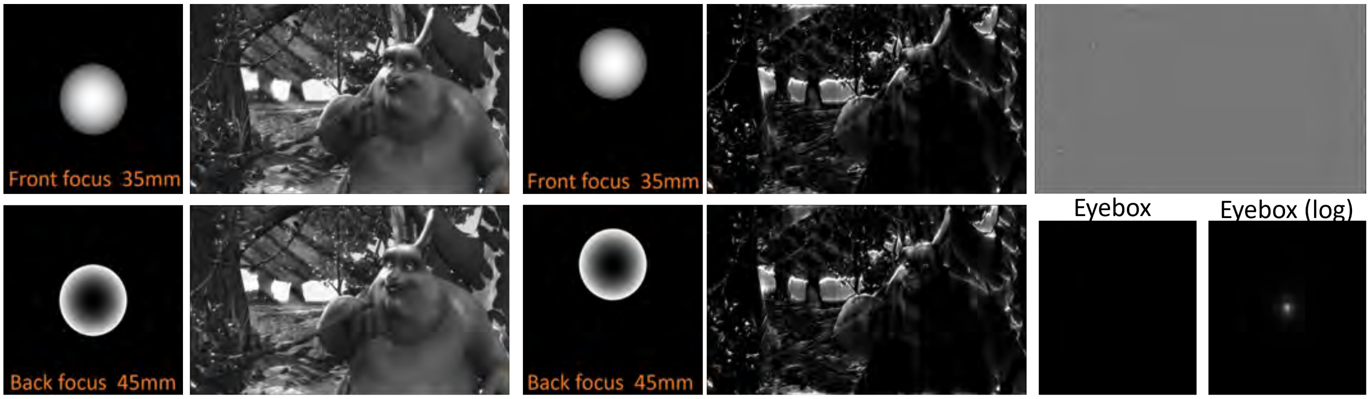

Fig. 4: **Simulation:** An example of a focal-stack hologram generated by [17]. When the pupil location is centered, one can focus on both the front and back of the hologram without artifacts. However, [17] achieves this by enforcing a smooth phase at the propagated hologram plane for each focal plane. This leads to smooth-phase holograms, which in turn leads to a highly localized eyebox. Note that the eyebox is heavily localized even in log-visualization, meaning almost all energy is centered at one peak. This again has catastrophic consequences when they pupil doesn't sample this DC-peak anymore. The image is directly lost. In fact, it's not just the image contrast that is gone, also because almost all of the hologram's light energy is centered at the DC, the image at off-center pupils is going to have barely any energy. The total energy is many magnitudes smaller than for the center pupils.

based on the applied voltage to the crystal. Several papers have investigated field-fringing in more detail and found that approximate models, such as convolutional models, can provide sufficient results [25], [26], [27].

One unique aspect of field-fringing is that it acts as a convolution in the phase domain rather than the complex domain. This results in non-linear behavior in the wavefront-domain that can sometimes introduce frequencies that are not present in the original phase pattern. This can be problematic for gradient-descent optimization for random-phase holograms, which is required for a wide eyebox. Field-fringing can destroy the information encoded in random-phase holograms, leading to artifacts such as incorrect contrast, unwanted point sources because of phase-wrapping, and in general strong speckles in the resulting images.

Recent papers based on neural-holography style system calibration [6], have mainly focused on other aspects than field-fringing. As a result, these papers typically use high-capacity neural networks to model and compensate for its effects. Yet, they allow smooth-phase holograms, which drastically reduce the effects of field-fringing. Note that we can label any hologram as *smooth phase* where image content is still visible on the optimized SLM-pattern. Likewise, we consider a phase hologram to be *random phase hologram*, if the holograms has uniform phase-distribution over 2pi and the optimized patterns basically looks like high-frequency noise.

We believe that investigating field fringing and its compensation models are exciting interesting work, as this is likely required to increase the perceived image quality performance on LCOS-based holographic displays with uniform eyebox. We believe that the image quality of all our implemented algorithms could be drastically increased if proper forward model calibration is performed. However, this is out-of-scope of this paper and left for future work. Alternatively, [14] uses fast, low-bit-depth modulators that induce the phase change using small mirrors which do not suffer from field-fringing. Unfortunately, we didn't have access to this technology at the time of writing but believe that this offers great opportunities to further increase the image fidelity of our algorithms.

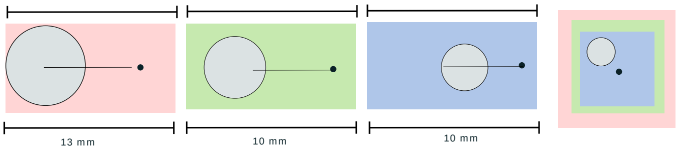

Fig. 5: **Eye-box** The eye-box size depends on the wavelength. When the eye-box is sampled by a pupil at a specific physical location, this pupil hence corresponds to a different spatial frequency for each wavelength. This means, that center of the pupil are at different spatial frequency locations and the size of the pupil also changes with wavelength.

7 Implementation

Here we provide further implementation details regarding the experimental setup which was used in our main paper.

7.1 Experimental Setup

We assess our SLFH-algorithm using a near-eye holographic prototype. For this, we followed the conventional design for near-eye holographic display as outlined in [6], [18]. A simplified rendering of our setup is shown in Fig. 8.

The FISBA Ready Beam Laser (450nm, 520nm, 660nm) with low temporal coherence is used as the light source. The large temporal bandwidth blurs the image and reduces the amount of speckle and interference (e.g. fringes due to reflections). To clean up the beam of the fisba-laser (coming out roughly at an 8deg angle), we use a $30mm$ lens to expand the beam and a $200mm$ for subsequent collimation. To ensure that the SLM functions in phase-only mode, we use

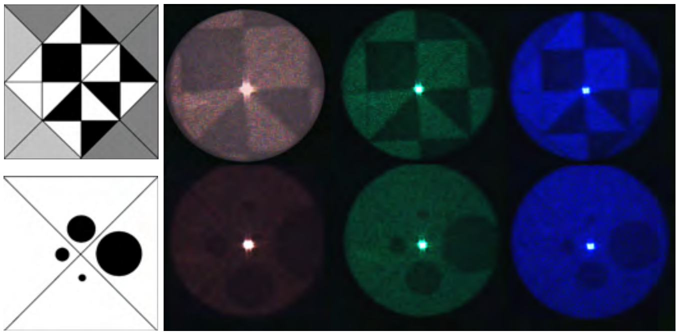

Fig. 6: **Eye-box in experiment:** We first compute Fourier-CGH holograms for two calibration targets. The images of the aperture were captured by adding a second path to the setup focused on the aperture. In this experiment, we opened our aperture to roughly 22mm. The eye-box size depends on the wavelength. For the red wavelength ,the eye-box is maximal, for the bule it is the smallest. The white peak in the center is the DC-term which stems from unmodulated light (SLM fill-factor ¡ 1) as well as light coming from the unmodulated boundaries of the SLM.

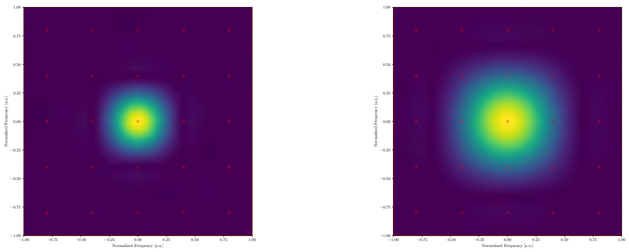

Fig. 7: When using STFT the window function (here, hamming window) implicitly defines the sampled pupil (left with 9x9 LF, right 5x5 LF). The STFT also defines the grid where the pupils are located. As the effective support of each pupil is slightly larger than the spacing between pupil locations, there will be pupil-wrapping at the outermost pupils leading to artifacts.

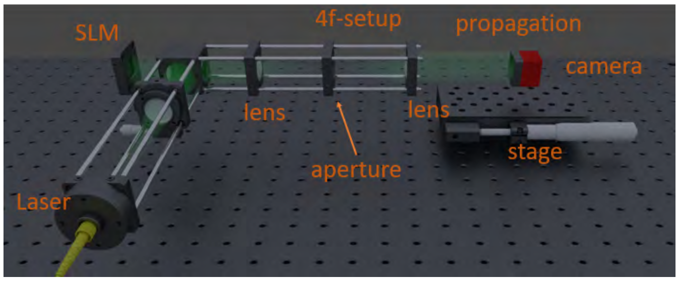

Fig. 8: **Rendering of the experimental setup.** We build a conventional near-eye holographic display. Defocus is measured by moving the camera on a 1D translation stage. Parallax cues are measured by moving a 2D-translation in the aperture domain. We additionally add a second beamsplitter after the second lens, to build another arm that captures an image of the aperture.

a polarizer after collimation to filter out wrongly polarized light. Note that the FISBA laser is already polarized, so we additionally adjusted the polarization direction of the FISBA laser to roughly align with our polarizer. The collimated beam then passes through a beamsplitter and impinges on an LCOS-based SLM (Holoeye Pluto-2.1-Vis-016 , 1080x1920 pixels with an 8um pitch).

The modulated beam then passes through a 4f system consisting of a 400mm (first) and 200mm (second) focal length. The change in focal length leads to an effective demagnification factor of 2 which allows to capture the full FoV with out camera. A motorized aperture (Standa, 8MID22-0-H) with a maximum diameter of 25mm is placed at the Fourier plane. The aperture can be moved laterally using two motorized-translation stages (Thorlabs, MTS25-Z8). The image sensor (MC089MG-SY by Ximea) is mounted on a high-speed brushless translation stage (Thorlabs, DDS050). The sensor is mounted on a rotation stage to correct for slight off-axis misalignment.

A $0mm$ standoff distance (measured in SLM space with 8um pitch) is chosen for the zero disparity plane, and the focal volume is defined over $20mm$ from $0mm$ to $20mm$. With the chosen magnification of 2, $20mm$ propagation corresponds to a $5mm$ travel on the translation stage. Additionally, we employ a second camera image, the pupil-sampled eyebox in order to calibrate the orientation of the 2D-translation stage and aperture.

7.2 Resolution Detail

All images are optimized with the full resolution on the SLM-plane (i.e. $1080x1920$) pixels. The pixel pitch for all our simulations and experiments is henceforth fixed to $8um$. However, as we're employing a pupil sampling approach where the pupils are default smaller than the maximal eye-box, the pixel resolution of our holograms is actually a misleading measure, as each hologram must be bandlimited. E.g., if we would evaluate our hologram with a pupil that is the $1/3$ of the eye-box size, this corresponds to a bandpass filter of $1/3$ of the fully available spectrum. The maximal image content that we could expect in such a case would hence be equivalent to a downsampled image of $360x640$. However, note that while image-detail is traded here, we can also move the pupil around to different views and could actually create $3x3 = 9$ independent views. As our stochastic light field algorithm randomly samples different radii and positions, the effective resolution of each sampled image actually varies. However, this is automatically accounted for in our loss-function as in each step a novel "image view" is synthesized using the LF-rendering algorithm.

7.3 Spatial Light Modulator Calibration

We assume the same specifications for the SLM (Holoeye Pluto-2.1-Vis-016) in both simulation and experiment. As the phase delay of the liquid crystals are dependent on the wavelength, we need to perform a basic calibration for the phase-range (Look-up-Table) for each color. To keep it simple, we assume that the phase always starts at 0 and is linear up. Hence, only one scalar value for each channel needs to tuned. As our optimization routines are not constrained in the phase values, we can wrap the phase-values after optimization to be matched the phase-range of the SLM. We tune the phase-range by hand for each color channel and found rough estimates for phase-LUT with $2.0\pi, 4.9\pi, 6.1\pi$ for red, green, and blue channels respectively.

We note that this calibration method is quite improper and likely contributes to a large model mismatch. However, as the same calibration settings were assumed in the simulation and also applied to each method during capture, we ensure fairness between comparisons. Still, we acknowledge that overall experimental image quality likely suffers from inaccurate calibration. For future work, Neural Holography style calibration [6] might be employed to further increase image fidelity. However, to the best of our knowledge, model calibration with full-random phase holograms on LCOS-displays is still an unresearched area and needs further investigation, which is beyond the scope of this paper.

8 Implementation details for STFT and Focal Stack

Watching the supplementary videos, one quickly notices that the STFT method struggles when dealing with extreme pupil shifts. On the other hand, our SLFH approach recovers good image quality even at the boundaries of the eyebox.

This is because we designed the propagation distances from the 4f-plane (where the SLM image is formed) to be small. This is e.g. beneficial to reduce artifacts that would arise due to partial coherent effects that increase with the propagation distance.

However, the STFT-algorithm [14] struggles with small propagation distances. This is because STFT doesn't have any inherent bandlimited during propagation and hence it is hard to form amplitude modulation with SLM's phase-only modulation for small propagation distances. To solve this problem, we adapt [14] and added a circular Fourier filter (roughly 95% eyebox radius), which allows pushing some energy into the *NO-CARE-AREAS* at the edges. Hence, STFT cannot optimize for holographic content when these *NO-CARE-AREAS* are sampled, but allows for content creation with large areas of the eyebox.

If we do not include this Fourier-filter in STFT, the performance is roughly 15dB worse as images are extremely low-contrast and almost look a high-pass filter.

This is a general shortcoming of STFT (as well as for the focal-stack loss, but less pronounced), which our SLFH approach does not suffer from. This is because the pupil-aware formulation automatically band limits the signal in each sampled pupil-position of during optimization with stochastical gradient descent.

In Fig. 9 we show simulation results for the comparison methods (STFT) and FS for small propagation distances without introducing no-care-areas. As your stochastic pupil sampling approach implicitly samples the pupils, each view is already implicitly bandlimited, hence our approach can inherently deal with small propagation distances.

In Fig. 10 we show the artifacts that occur when the *no-care areas are introduced*. STFT performs significantly worse at the edges of the eyebox. In Fig. 11 we include captures of the actual eyebox for the STFT-content which show high-energy peaks at the boundary of our fourier-filter which introduces the artifacts. Despite our best efforts, we were not able to reduce these peaks in the Fourier-domain.

9 Additional Discussions

9.1 Speckle due to Alignment

Random phase holograms (like ours) are actually fairly robust to misalignment as they act more like incoherent displays: A much larger-SLM area contributes to forming an image point compared to smooth-phase (where the scattering is small). In fact, all of our experimental results were done without any calibration tied to alignment. However, this doesn't mean that full-random phase holograms (including ours) aren't speckle free. In random-phase holograms, speckles occur naturally, just like in incoherent light, which is, by definition, speckly (just on a much smaller scale). Yet, we note that future developments in speckle reduction (let it be smaller SLM-pixels or novel system design like [28]) are fully compatible with our approach and hopefully will help achieve image quality close to smooth-phase holograms.

9.2 Pupil-Steering

Pupil-steering approaches might be an alternative solution to solve the 3D ocular parallax problem. However, there is no free lunch as it requires additional scanning hardware, which must be very accurate, robust, and run with very high FPS. It is potentially the only viable way to overcome etendue limits in short-term, but we consider it more a technological challenge than answering fundamental research questions in holography. Hence, we do not compare against pupil-steering technologies as those are complementary to our research.

9.3 Comparison to Pupil-Aware Holography

[13] lacks sophisticated discussion for how to achieve true 3D-content and their experiments show only 2-plane representations, which are not optimized with focal-stacks. Furthermore, in order to achieve good image quality, they use an additional DPAC-layer to encode complex-fields. This is because DPAC inherently leads to low-bandwidth holograms, which decrease artifacts that arise e.g., due to field-fringing. Unfortunately, low-bandwidth holograms come with a much-decreased eye-box size. As the goal of light-field holography is to utilize the maximal achievable eye-box, the algorithm proposed in [13] pupil-aware holography cannot be applied for our purposed.

Moreover, their proposed optimization uses a shift-invariant multi-layer representation. This is highly unphysical as parallax induced by FS-cues is ignored. Our results show that holograms with correct FS-cues automatically demand parallax (and vice-versa). Hence, pupil-aware holography is at odds with physical constraints and cannot produce image quality over the full-eyebox.

9.4 Comparison to other Light Field papers

There are several other papers that propose light field losses such as [22]. In fact, [22] show experimental results with

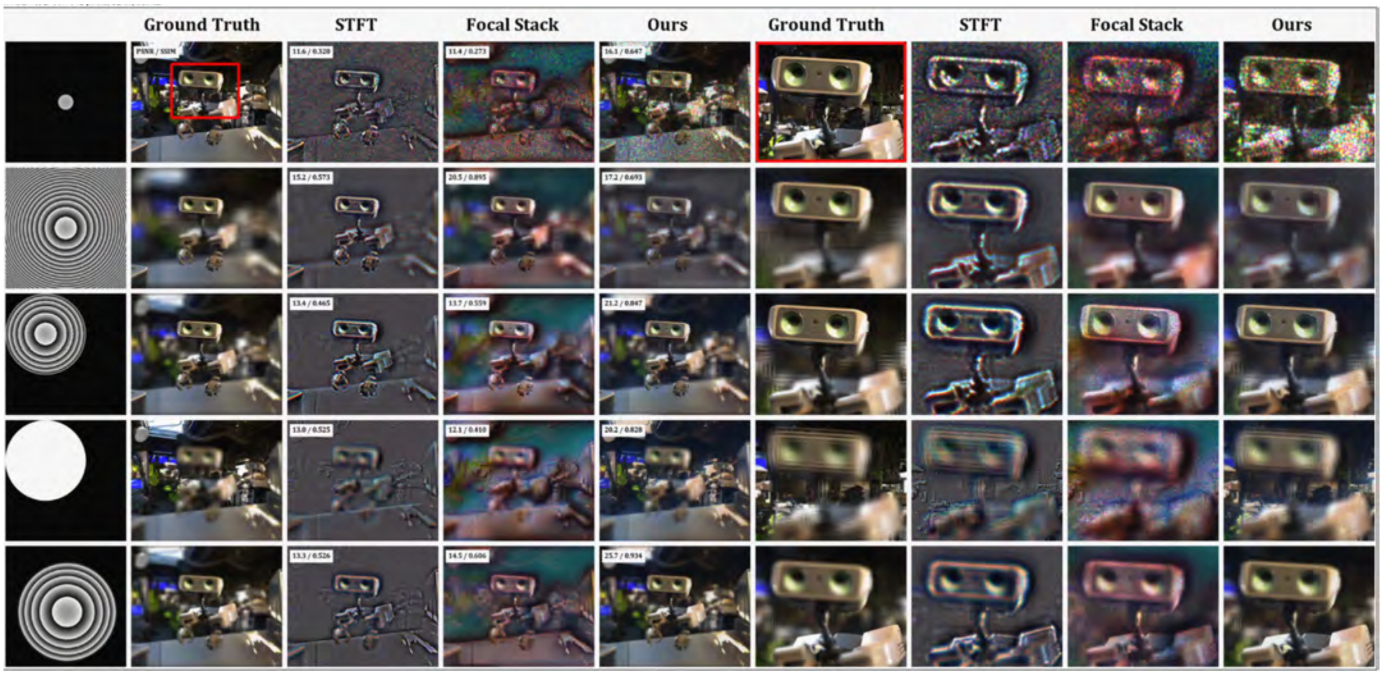

Fig. 9: **Artefacts without No-Care-Areas (Simulation):** This figure shows the poor performance of STFT and FS-approaches for small propagation distances without accounting for ”NO CARE AREAS” within the eyebox. Only by introducing a no-care area where the eye-box cannot be sampled, good images can be optimized (see other simulation and experimental results).

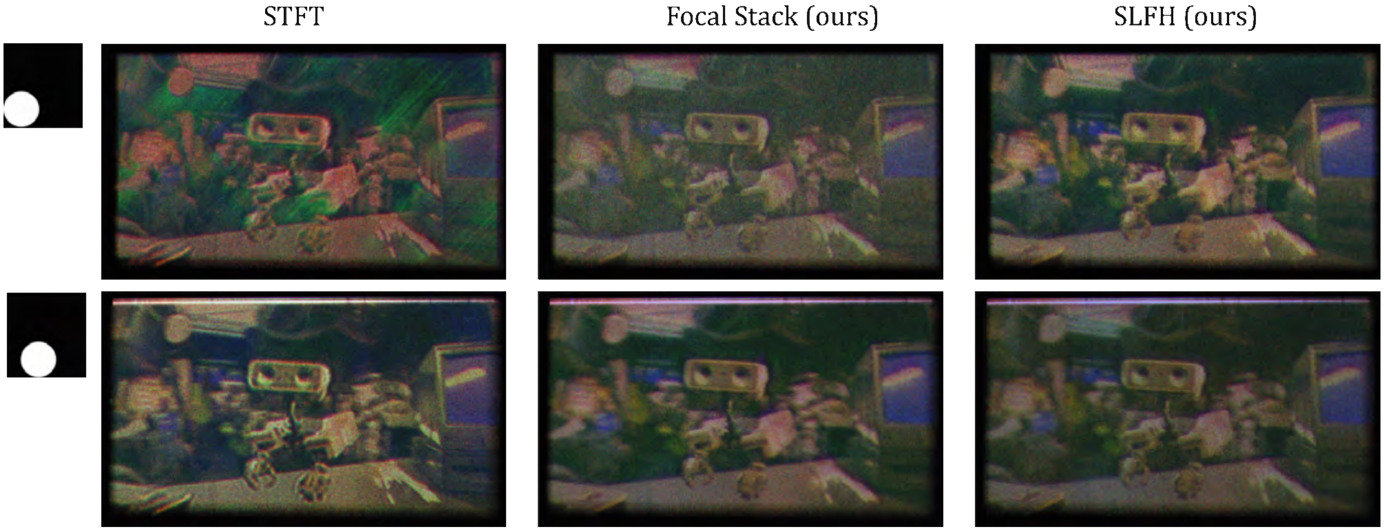

Fig. 10: **Artefacts with No-care-areas (Experiment):** Figure shows artefacts caused by introducing no-care areas, affecting STFT performance at the edges of the eyebox. To address this issue, a circular Fourier filter was added to push energy into the no-care areas, enabling content creation with larger areas of the eyebox.

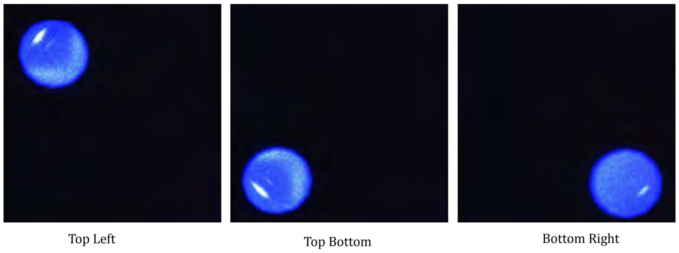

Fig. 11: **STFT-Artifacts in Fourier domain(Experiment).** Here, we show the displayed eye-box (aperture plane) for three different pupil locations with the STFT-reconstructions when including the NO-CARE-AREAS. In the ideal case, the Fourier spectrum would be uniform (noisy everywhere), but because of the additional Fourier filter the optimization algorithm tends to create high-energy regions inside of the eye box. Here we only show the blue channel. The location of the high-energy peaks changes accordingly with different wavelengths (red and green) as discussed above. Whenever a pupil samples these high-energy regions, strong streak artifacts occur. This drastically reduces the region where STFT can be evaluated.

both focal stacks and light fields. However, those experiments are done isolated from each other. The focal stack experiments and light fields scene are from totally different scenes and there's no cross-examination how well a light field scene works on a focal-stack evaluation and vice versa. For this reason, we consider light field papers [22] to be conceptually equivalent to STFT [14].

The proposed method using the Wigner-geometry, does much more as we indeed couple the effects of both focal stacks and light fields.

9.5 Clarifications on Far-Field / Fourier Holograms

In holographic near-eye displays there are two main classes of how the holographic image is created: Near-eye and Far-field holograms.

Most recent near-eye display holography papers work in the near-field configuration (sometimes also labeled Fresnel holograms). Here, the image is created at or close (propagation distances up to 200mm) to the relayed SLM-plane. This allows to compute holograms with closed-form solutions such as DPAC-encoding [16], real-time applications using Neural Networks [18] and iterative methods such as [4], [6]. As discussed above, Near-eye holograms can often achieve great image quality, especially when the chosen phase-distribution is rather smooth.

The other class is far-field holograms, where the display (SLM) modulates the wavefront in the spectral domain. Recent examples of near-eye displays that work in far-field configuration would be [20], [21], [22] that use iterative approaches. However, there are also many closed-form solutions out there, with [29] being a recent one that uses the Wigner-formulism directly to compute wavefronts from light fields.

9.6 Stochastic Light Field Holography with STFT-formulation

Let us shortly revisit the STFT equation as introduced in the main paper:

$$\begin{aligned} \mathrm{STFT}[u(\mathbf{r})]\left(\mathbf{r},\mathbf{q}\right) &= \int u\left(\mathbf{r}'\right) w\left(\mathbf{r}'-\mathbf{r}\right) e^{-j2\pi\mathbf{q}\cdot\mathbf{r}'} d\mathbf{r}' \\ &= \mathcal{F}^{-1}\left\{U\left(\mathbf{q}'\right)\cdot W\left(\mathbf{q}'-\mathbf{q}\right)\right\}, \end{aligned} \tag{7}$$

$$L\left(\mathbf{r},\mathbf{q}\right) = \left|\mathrm{STFT}\left[u(\mathbf{r})\right]\left(\mathbf{r},\mathbf{q}\right)\right|^2. \tag{8}$$

This equation is concise, simple and intuitive. In theory, by choosing different Window functions, different pupil functions can be realized. Note that the chosen window size limits what pupil functions could be achieved as the chosen window essentially also acts as a bandpass filter.

In any case, one could certainly vary windows-size, grid-density, and propagation- distance. However, these modifications would inherently demand a change in the LF-targets for each chosen window configuration. While the STFT-equation could be modified to be equivalent to our stochastic light field approach, it would require the same consideration on how the incoherent LF-part has to be geometrically coupled with the coherent part. At this point, even the STFT-formulation needs to revisit the Wigner-formulation argument to enforce a correct geometrical coupling. We conclude that the Wigner-approach provides only the theoretical framework - whether one uses STFT or pupil-aware representation - those are only semantics. However, we believe that explicit modeling of the pupil in the spectral domain is more intuitive than doing stochastic pupil sampling using the STFT-formulation.

10 Additional Results

10.1 Experimental Results

Changing aperture size: We show additional results for the robot scene where we've changed the aperture size and varied the back and front focus, see Fig. 15

Greenery Scene: In Fig. 14, we show a comparison between the front and back focus between the focal stack and our SLFH. There is higher color fidelity at leaves for the SLFH method as it probably has seen this particular focused region of the images more often than only one time when optimized with a focal stack only.

10.2 Simulation Results

Chaing aperture size: We show results where we compare the implemented algorithms in simulation for changing the aperture from small to large in Fig. 16. Differences are best noticeable when looking at noise level and fine details such as the occlusions at the wireframe.

Arbitrary pupil states In Fig. 12 to Fig. 21 we show additional simulation results evaluated under arbitrary pupil states for a large variety of different scenes.

PSNR statistic plot

For completeness, we show in Fig. 12 the corresponding statistical evaluation on the Deep Spaces [8] dataset with PSNR values instead of the SSIM which we have reported in the main paper. The findings are similar, but we believe that SSIM is a better image metric for visual performance evaluation than PSNR.

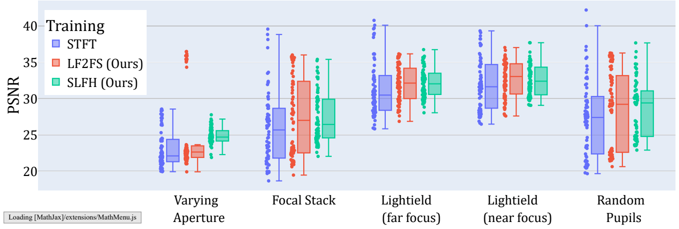

Fig. 12: *Statistical Evaluation:* Statistical performance evaluation of SSIM percpetual similarity metric over the Deep Spaces [8] light field dataset. Holograms are trained using STFT, Focal Stack supervision (LF2FS), and Stochastic Light Field Holography (SLFH). Statistics are plotted for various aperture types: Varying Aperture (focus and position are fixed), Focal Stack (aperture diameter and position are fixed), Light Field (aperture diameter and focus are fixed), and random pupils. Each dot represents the average PSNR of all evaluated pupil-positions for a specific aperture type for one specific light field within the Deep Spaces [8] dataset. Hence, each dot corresponds to the average performance for a specific scene. Our method always produces the best worse case performance, and produces a better mean and variance for random pupils, which means that our algorithm performs better for a wider variety of near-eye viewing conditions.

References

[1] J. W. Goodman, *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[2] R. Ng, “Fourier slice photography,” *Arxiv*, pp. 735–744, 2005.

[3] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

[4] P. Chakravarthula, Y. Peng, J. Kollin, H. Fuchs, and F. Heide, “Wirtinger holography for near-eye displays,” *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, p. 213, 2019.

[5] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European Conference on Computer Vision (ECCV)*, 2016.

[6] Y. Peng, S. Choi, N. Padmanaban, and G. Wetzstein, “Neural holography with camera-in-the-loop training,” *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, p. 185, 2020.

[7] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[8] L. Xiao, A. Kaplanyan, A. Fix, M. Chapman, and D. Lanman, “Deepfocus: Learned image synthesis for computational display,” in *ACM SIGGRAPH 2018 Talks*, 2018, pp. 1–2.

[9] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, “Light field photography with a hand-held plenoptic camera,” Ph.D. dissertation, Stanford University, 2005.

[10] Z. Xiao, Q. Wang, G. Zhou, and J. Yu, “Aliasing detection and reduction in plenoptic imaging,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3326–3333.

[11] P. Shedligeri, F. Schiffers, S. Ghosh, O. Cossairt, and K. Mitra, “Selfvi: Self-supervised light-field video reconstruction from stereo video,” in *ICCV*, 2021, pp. 2491–2501.

[12] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, “Plenoctrees for real-time rendering of neural radiance fields,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5752–5761.

[13] P. Chakravarthula, S.-H. Baek, E. Tseng, A. Maimone, G. Kuo, F. Schiffers, N. Matsuda, O. Cossairt, D. Lanman, and F. Heide, “Pupil-aware holography,” *arXiv preprint arXiv:2203.14939*, 2022.

[14] S. Choi, M. Gopakumar, Y. Peng, J. Kim, M. O’Toole, and G. Wetzstein, “Time-multiplexed neural holography: a flexible framework for holographic near-eye displays with fast heavily-quantized spatial light modulators,” in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–9.

[15] R. Ziegler, S. Bucheli, L. Ahrenberg, M. Magnor, and M. Gross, “A bidirectional light field-hologram transform,” in *Computer Graphics Forum*, vol. 26. Wiley Online Library, 2007, pp. 435–446.

[16] A. Maimone, A. Georgiou, and J. S. Kollin, “Holographic near-eye displays for virtual and augmented reality,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 85, 2017.

[17] S. Choi, M. Gopakumar, Y. Peng, J. Kim, and G. Wetzstein, “Neural 3d holography: Learning accurate wave propagation models for 3d holographic virtual and augmented reality displays,” *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–12, 2021.

[18] L. Shi, B. Li, C. Kim, P. Kellnhofer, and W. Matusik, “Towards real-time photorealistic 3d holography with deep neural networks,” *Nature*, vol. 591, no. 7849, pp. 234–239, 2021.

[19] D. Kim, S.-W. Nam, B. Lee, J.-M. Seo, and B. Lee, “Accommodative holography: improving accommodation response for perceptually realistic holographic displays,” *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[20] G. Kuo, L. Waller, R. Ng, and A. Maimone, “High resolution étendue expansion for holographic displays,” *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, p. 66, 2020.

[21] S. Monin, A. C. Sankaranarayanan, and A. Levin, “Analyzing phase masks for wide étendue holographic displays,” in *2022 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2022, pp. 1–12.

[22] ——, “Exponentially-wide étendue displays using a tilting cascade,” in *ICCP*, 2022.

[23] Y. Peng, S. Choi, J. Kim, and G. Wetzstein, “Speckle-free holography with partially coherent light sources and camera-in-the-loop calibration,” *Science advances*, vol. 7, no. 46, p. eabg5040, 2021.

[24] P. Chakravarthula, E. Tseng, H. Fuchs, and F. Heide, “Hogel-free holography,” *ACM Transactions on Graphics (TOG)*, 2022.

[25] B. Apter, U. Efron, and E. Bahat-Treidel, “On the fringing-field effect in liquid-crystal beam-steering devices,” *Applied optics*, vol. 43, no. 1, pp. 11–19, 2004.

[26] S. Moser, M. Ritsch-Marte, and G. Thalhammer, “Model-based compensation of pixel crosstalk in liquid crystal spatial light modulators,” *Optics express*, vol. 27, no. 18, pp. 25 046–25 063, 2019.

[27] M. Persson, D. Engström, and M. Goksör, “Reducing the effect of pixel crosstalk in phase only spatial light modulators,” *Optics express*, vol. 20, no. 20, pp. 22 334–22 343, 2012.

[28] Y. Jo, D. Yoo, D. Lee, M. Kim, and B. Lee, “Multi-illumination 3d holographic display using a binary mask,” *Optics Letters*, vol. 47, no. 10, pp. 2482–2485, 2022.

[29] K. Min, D. Min, and J.-H. Park, “Wigner inverse transform based computer generated hologram for large object at far field from its perspective light field,” *Optics Communications*, vol. 532, p. 129229, 2023.

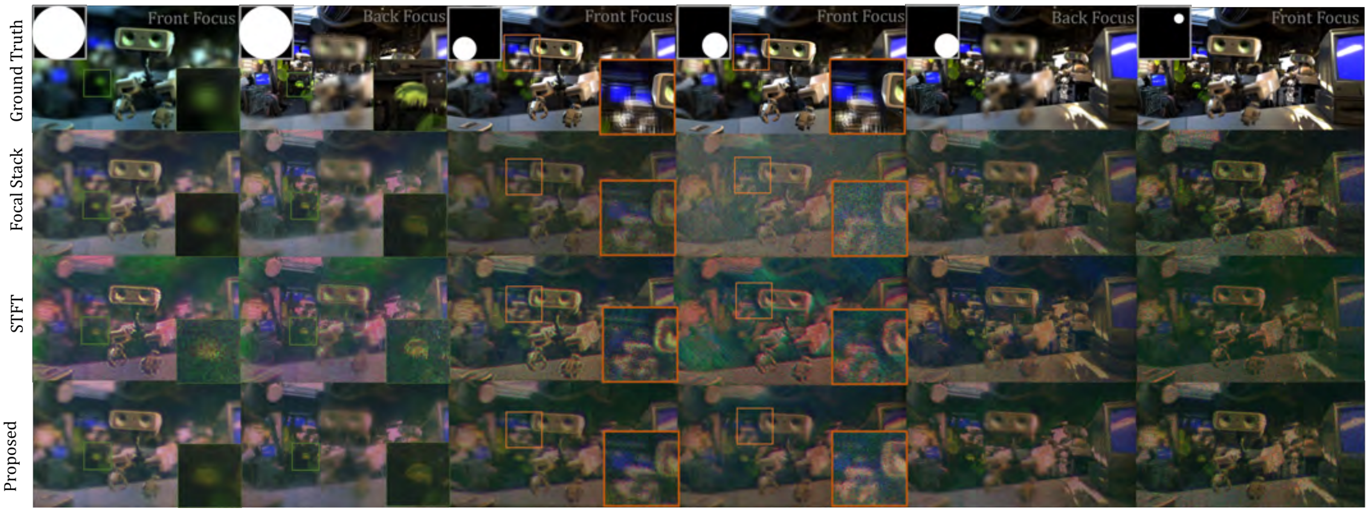

Fig. 13: **Experiment:** We show experimental results for a variety of different pupil states for STFT, focal stack, and light field supervision. Thy differences are slight, but one can see that our proposed method consistently shows accurate image quality.

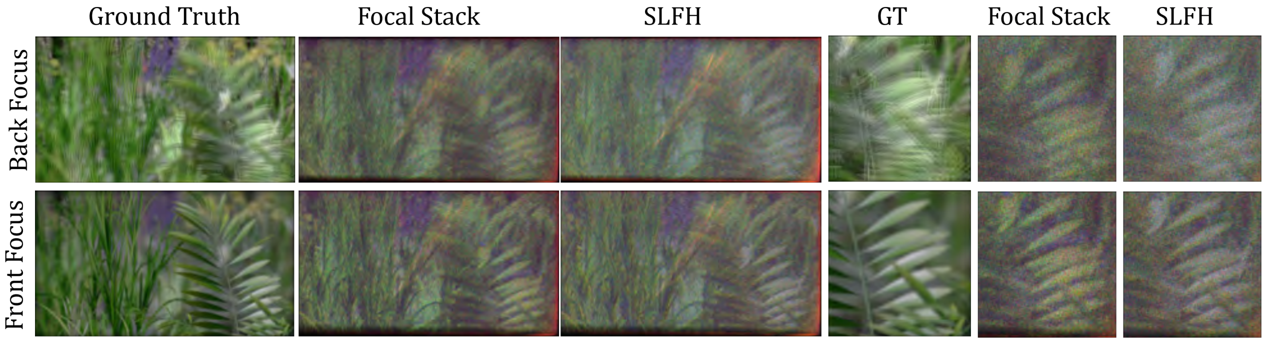

Fig. 14: **Experiment:** This example shows a scene with a lot of occlusions.The most striking difference is that Focal-Stack optimization fails to produce a correct color at the leaves (see inset on the right).

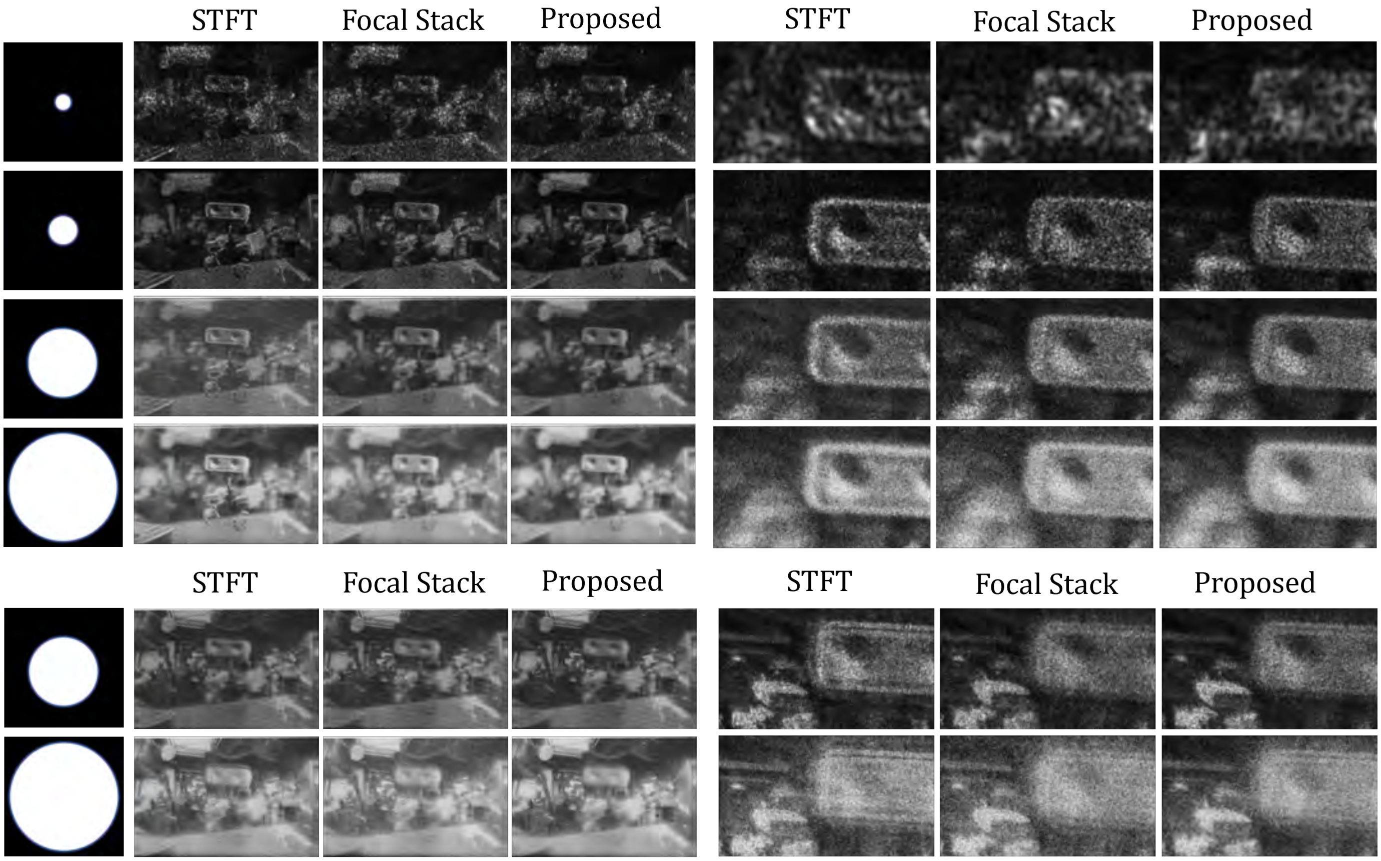

Fig. 15: **Experiment:** *Changing aperture size (Front/Back focus* In this experiment, we change the aperture size from very small to full aperture. Note how the second row looks almost the same for both back and front focus. One can also see that the Depth-of-field is large for small pupils, albeit the images are very speckly. The Focal Stack optimized images are slightly noisier than the proposed full-light image.

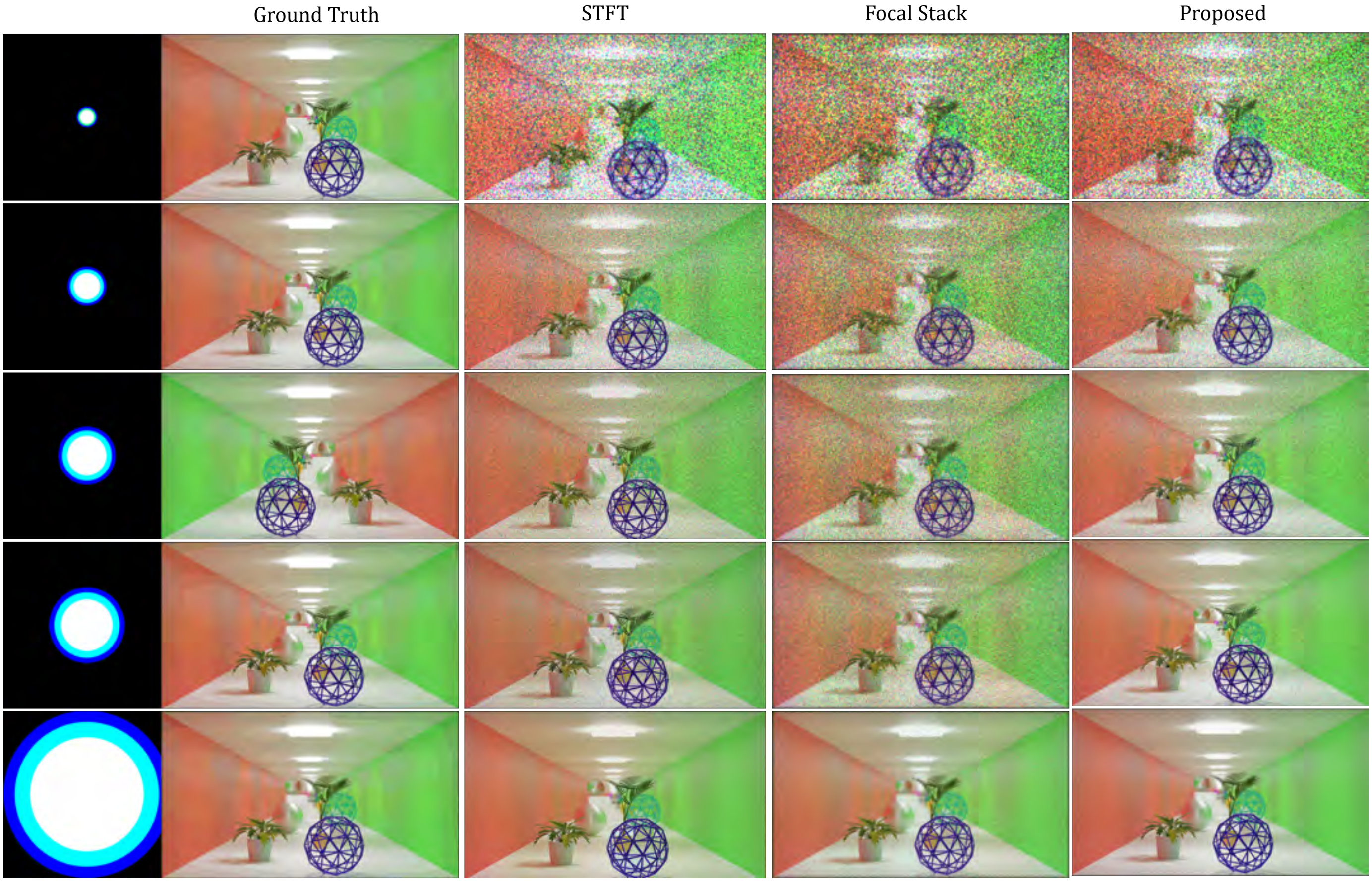

Fig. 16: **Simulation:** Additional results showing the reconstruction quality for different approaches when changing the pupil diameter. The proposed method SLHF produces imagery with signficantly less noise and less ringing.

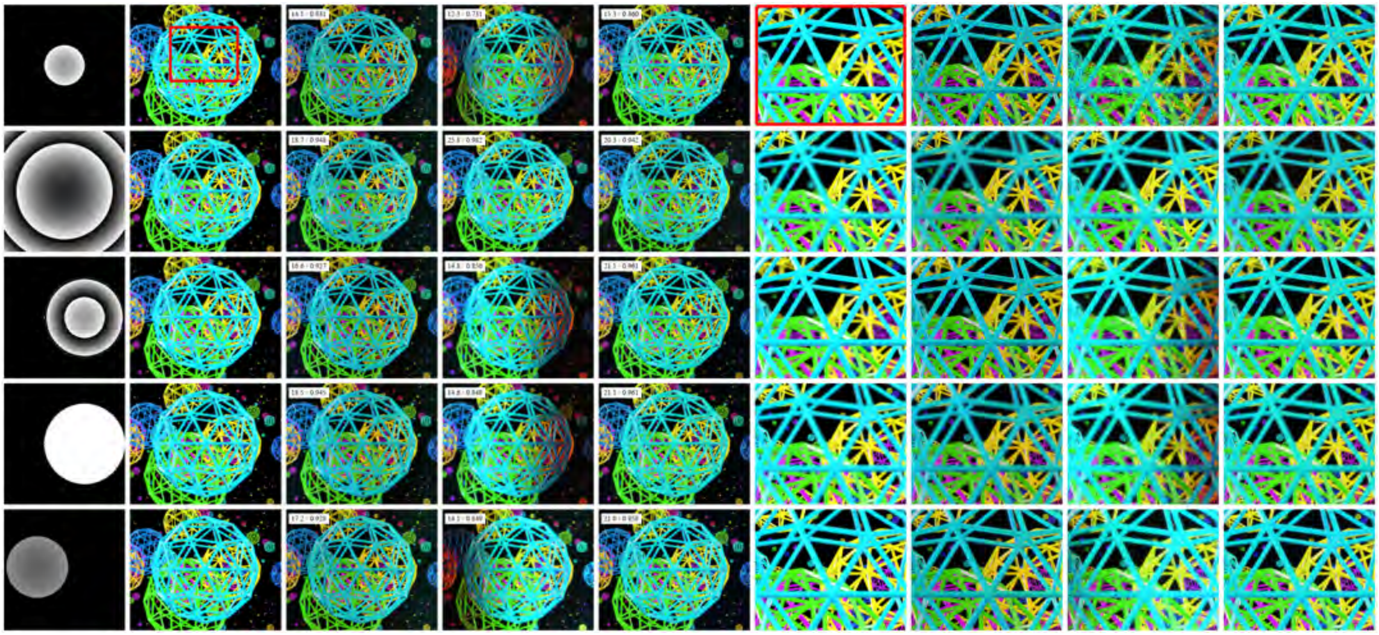

Fig. 17: **Simulation:** *Arbitrary pupils for a dense scene with occlusions.*

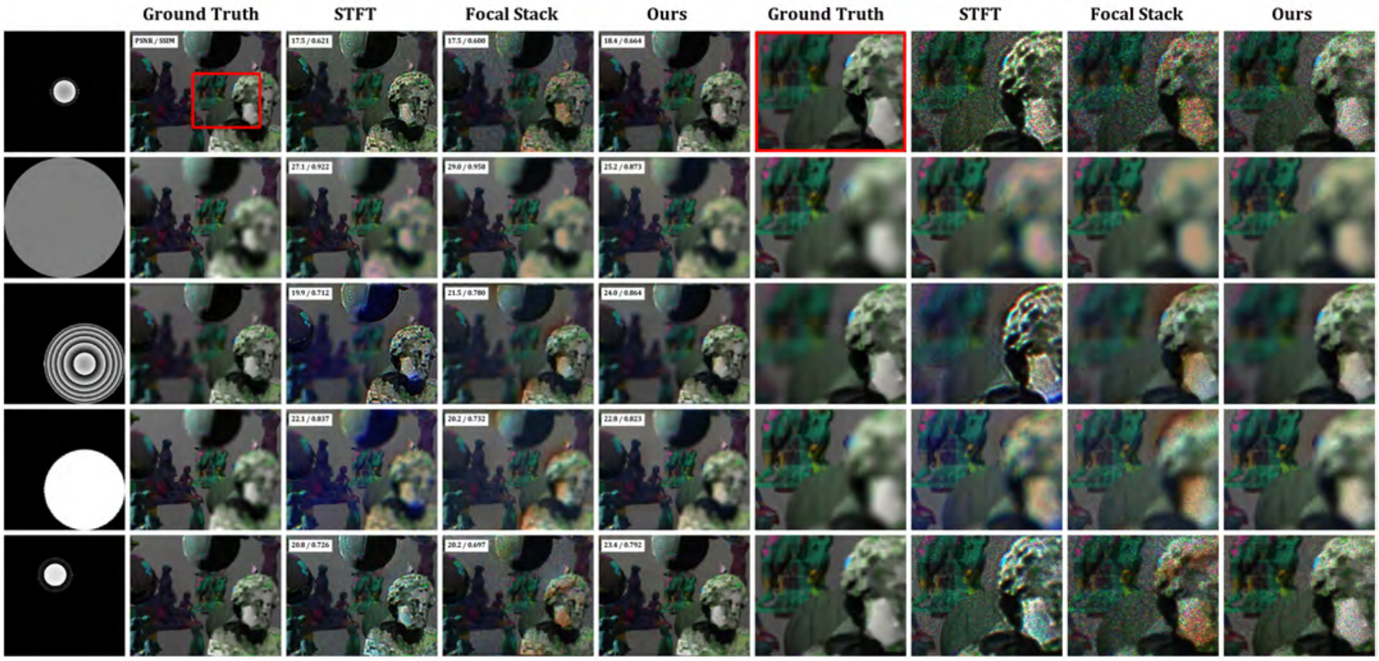

Fig. 18: **Simulation:** *Arbitrary pupils for a scene from the Deep Focus dataset.*

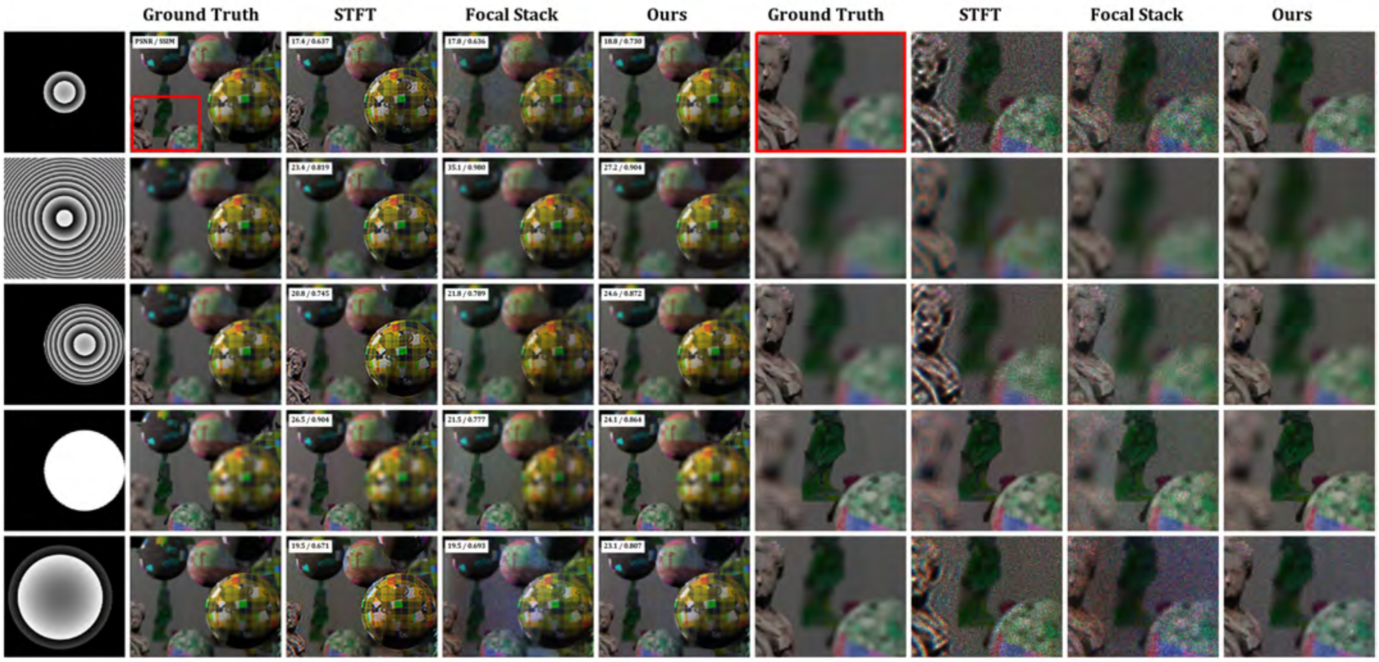

Fig. 19: **Simulation:** *Arbitrary pupils for a scene from the Deep Focus dataset.*

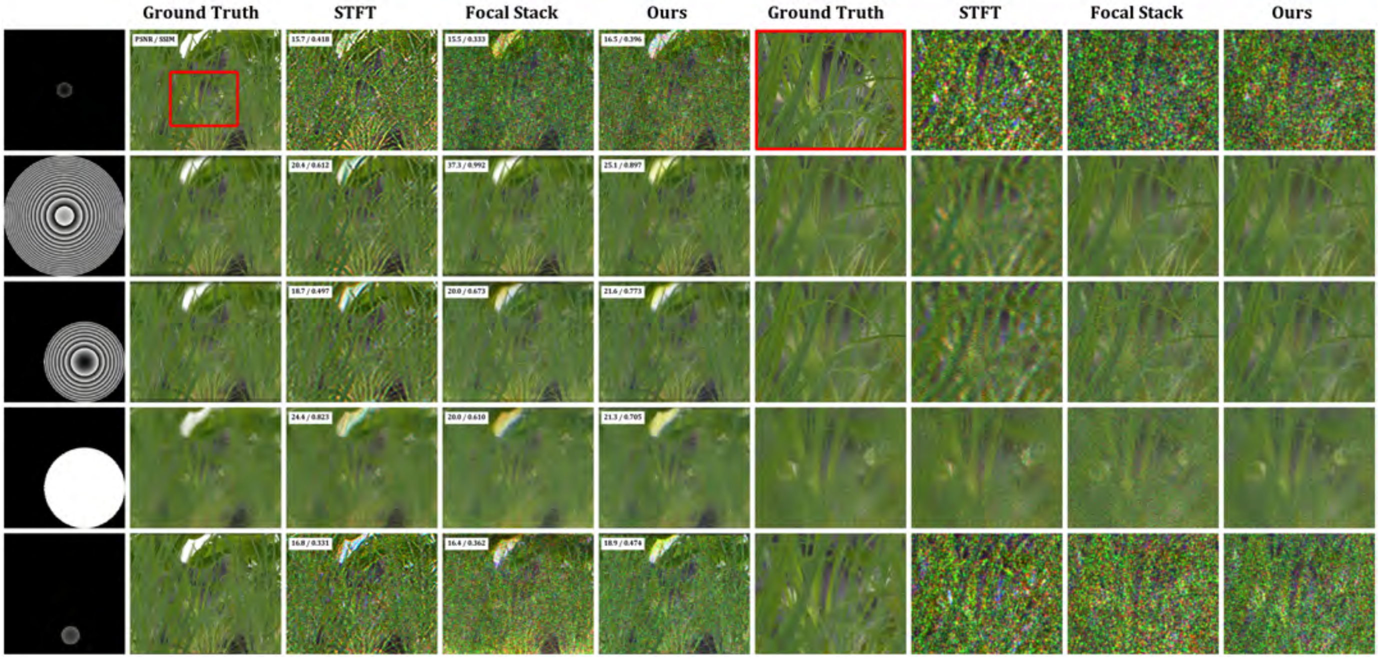

Fig. 20: **Simulation:** *Arbitrary pupils for a scene from same scene as in the paper to show more pupil states.*

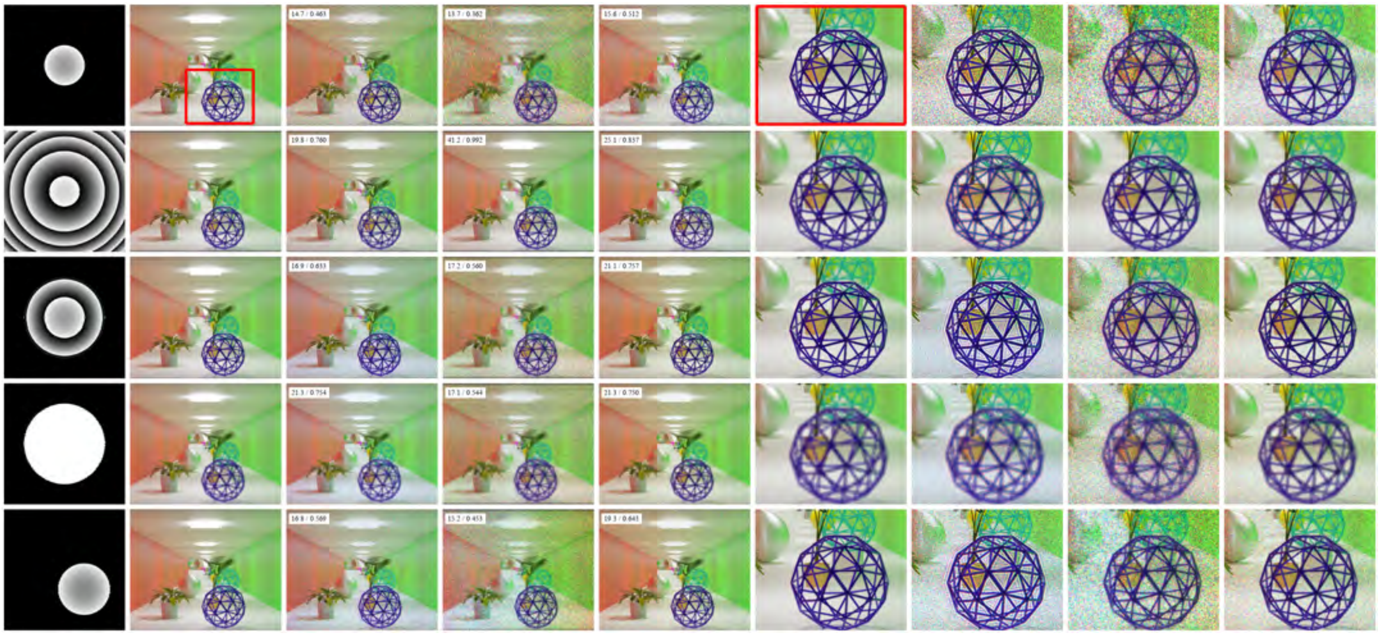

Fig. 21: **Simulation:** *Arbitrary pupils for a scene from same scene as in the paper to show more pupil states.*