# Supplementary Information – The Differentiable Lens: Compound Lens Search over Glass Surfaces and Materials for Object Detection

Geoffroi Côté[1,2] Fahim Mannan[3] Simon Thibault[1] Jean-François Lalonde[1] Felix Heide[2,3]
[1]Université Laval [2]Princeton University [3]Algolux

This supplemental document provides additional information in support of the findings from the main manuscript. Specifically, we describe the impact of the proposed ray-aiming approach in more detail, give further description of the wavelength selection for the proposed method, and discuss the assumptions made when deriving the optical image formation model. Furthermore, we provide a tolerancing analysis, visualizations of our ablation experiments, additional comparisons to Li et al. [4] and Tseng et al. [5], and additional details on the experiments from the main manuscript.

## 1. Code and Videos

We provide access to our code[1] for simulating and optimizing compound refractive lenses in an end-to-end manner using exact differentiable ray tracing.

Additionally, our project webpage[2] includes videos that illustrate the joint optimization of the Doublet, Cooke, and Tessar lenses for object detection on the BDD100K dataset under regular (1×) resolution. In particular, the videos illustrate how the selection of catalog glasses is handled through quantized continuous glass variables. Whereas the Doublet lens constantly varies throughout the optimization process, the Cooke and Tessar lenses exhibit a different behavior in which the state of the lens changes sporadically and abruptly and then quickly stabilizes. This behavior can be partly attributed to using the Adam optimizer [3] with high $\beta_1$ and $\beta_2$ values (0.9 and 0.999, respectively), where small perturbations can quickly add up due to slowly adapting learning rates. We empirically found this behavior helpful in maximizing object detection performance, possibly due to beneficial regularization on the object detector.

## 2. Ray Aiming

For any field angle within the full field of view, in the absence of optical vignetting, simulated rays incident upon a lens should precisely span the entire clear area of the aperture stop. To fulfill this condition, it is common in conventional ray tracing to initialize the rays at the entrance pupil of the system, whose size and position are found through a paraxial ray-tracing operation. Under strong pupil aberrations—that is, the aberrations between the entrance pupil and physical aperture stop—rays that are initialized naively at the entrance pupil may strongly deviate from their corresponding location on the aperture stop and, as such, skew the results of the ray-tracing operation. In contrast to previous works that tackle the joint design of compound optics through ray tracing [2, 4], we compensate for pupil aberrations with an accurate ray-aiming procedure, which consists in correcting the coordinates of the rays at the entrance pupil so that they adequately span the aperture stop. Similar to Côté et al. [1], we assume an elliptic shape for the corrected entrance pupil; thus, we find approximations for the displacements of the top $\Delta y^h_{\text{p,top}}$, bottom $\Delta y^h_{\text{p,bottom}}$, and side $\Delta x^h_{\text{p,side}}$ of the pupil for each off-axis field value $h$. Due to rotational symmetry, we have $\Delta x^h_{\text{p,right}} = -\Delta x^h_{\text{p,left}} = \Delta x^h_{\text{p,side}}$.

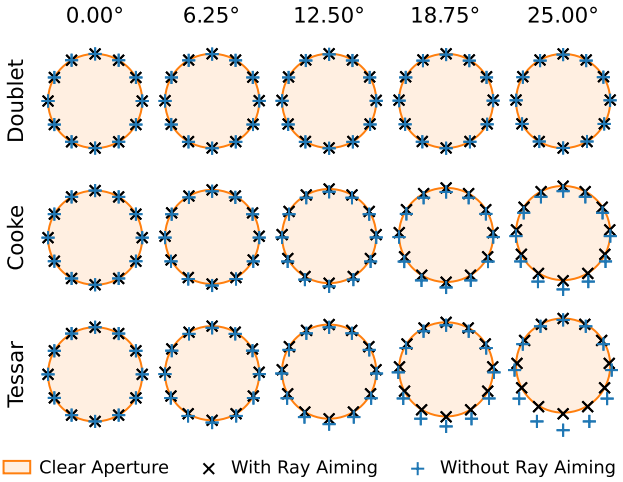


Figure 1. Ray-aiming error of rays aimed at the circular edge of the aperture stop with/without the ray-aiming correction step, for the f/2 Doublet, Cooke, and Tessar lenses used in our experiments and different field angles. While the ray-aiming correction step is not required for the Doublet, it prevents moderate errors with the Cooke lens and large errors with the Tessar lens.

[1] https://github.com/princeton-computational-imaging/joint-lens-design

[2] https://h nj gubnjnnae pg.jollibee ood.rest/joint-lens-design

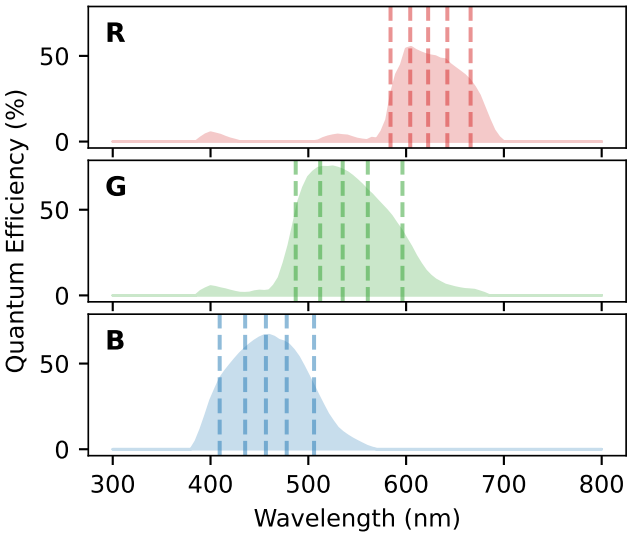


Figure 2. Selected wavelengths (indicated with vertical lines) for our experiments, which are based on the quantum efficiency spectrum of a typical sensor, here the Sony IMX172.

| R | G | B |
|---|---|---|
| 584.1 | 487.1 | 409.4 |
| 604.2 | 512.1 | 435.4 |
| 622.5 | 535.1 | 456.6 |
| 642.2 | 560.8 | 477.9 |
| 665.9 | 596.3 | 505.9 |

Table 1. Selected wavelengths (in nm) for each color channel R, G, and B.

As shown in Fig. 1, we find that the assumption of a linear relationship between the entrance pupil coordinates $x_\mathrm{p}$, $y_\mathrm{p}$ and the aperture stop coordinates $x_\mathrm{s}$, $y_\mathrm{s}$ provides sufficient accuracy for the lenses used in our experiments, that is

$$\Delta x_\mathrm{s} \approx \Delta x_\mathrm{p} \frac{\mathrm{d}x_\mathrm{s}}{\mathrm{d}x_\mathrm{p}}\,; \tag{1}$$

$$\Delta y_\mathrm{s} \approx \Delta y_\mathrm{p} \frac{\mathrm{d}y_\mathrm{s}}{\mathrm{d}y_\mathrm{p}}\,. \tag{2}$$

To compute the ray-aiming errors $\Delta x^h_\mathrm{s,side}$, $\Delta y^h_\mathrm{s,top}$, and $\Delta y^h_\mathrm{s,bottom}$, we trace a sagittal ray and two meridional (upper and lower) rays for each field $h$, respectively, then compare their coordinates at the aperture stop to the aperture stop diameter—computed by tracing an on-axis meridional ray. The derivative terms are obtained through automatic differentiation. Then, Eq. (1) and Eq. (2) are used to recover the field-wise entrance pupil displacements.

## 3. Wavelength Selection

In our experiments, we perform wavelength sampling that is representative of compound lenses. To this end, we rely on the quantum efficiency spectrum $Q(\lambda)$ of a typical sensor (here, the Sony IMX172), which is visualized in Fig. 2.

For each of the R, G, and B color channels, we select 5 wavelengths by computing all the odd-numbered 10-quantiles of $Q(\lambda)$. The selected wavelengths are given in Tab. 1. While these wavelengths adequately represent the spectrum for our task, we note that the proposed method supports denser wavelength sampling without any changes, though at the cost of additional compute overhead.

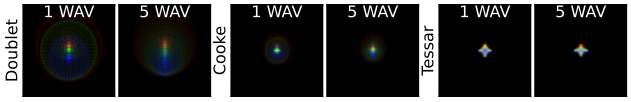


Figure 3. Multispectral sampling. The PSFs (shown at 12.5° field angle) are *more spread out* when chromatic aberrations are accurately captured with multispectral sampling (5 wavelengths).

## 4. Image Formation Model

Here we further discuss the assumptions made in the main paper and how they can impact our findings.

In our approach, we design all lenses for imaging at optical infinity. This assumption—common and often safe in lens design and computational imaging—is adequate beyond the hyperfocal distance $H = f^2/Nc$, where $f$ is the focal length and $N$ is the f-number. We can estimate the hyperfocal distance by setting an appropriate value for the tolerated circle of confusion. As we consider that a circle of confusion smaller than the spot size diameter (i.e., twice the spot size radius) of a lens will have limited impact on optical performance, we set $c$ as the mean spot size diameter of the 2-, 3-, and 4-element baseline lenses, with $f = 17.2\,\mathrm{mm}$ and $N = 2$ for all lenses, and estimate 0.9, 2.4, and 5.0 m for $H$, respectively. For the envisioned object detection applications, most small objects that have to be located are typically found beyond this range, thus justifying this assumption.

In the optical formation model used in this work, we implicitly consider the RGB values of an input image to be proportional to the luminance of the virtual scene. An underlying assumption is that the spectra for each of the R, G, and B channels are uniform and do not depend on the scene content, which is not the case in practice. Nonetheless, even though we contend with only three spectral bands (RGB), we accurately model chromatic aberrations with multispectral sampling. We assume the worst-case scenario for broadband spectrum and *overestimate* chromatic aberrations in the general case (see Fig. 3). As such, with real scenes rather than virtual ones, the actual chromatic aberrations would generally be smaller and object detection performance would presumably not be adversely impacted.

| Optics | Baseline w/ tolerancing | Proposed w/ tolerancing | Margin | Margin w/ tolerancing |
|---|---|---|---|---|
| Doublet (1× res.) | 30.3 (-0.0) | 32.0 (-0.0) | +1.7 | +1.7 |
| Cooke (1× res.) | 33.0 (-0.0) | 33.3 (-0.0) | +0.3 | +0.3 |
| Tessar (1× res.) | 33.4 (-0.0) | 33.6 (-0.0) | +0.2 | +0.2 |
| Doublet (2× res.) | 25.0 (-0.0) | 28.1 (-0.0) | +3.1 | +3.1 |
| Cooke (2× res.) | 31.5 (-0.0) | 31.7 (-0.0) | +0.2 | +0.2 |
| Tessar (2× res.) | 31.1 (-0.2) | 32.1 (-0.1) | +0.9 | +1.0 |

Table 2. Monte-Carlo tolerancing analysis (n=10 000). We report the change in average AP on the BDD100K dataset w.r.t. Tab. 2 of the main paper when including tolerancing. We note that the Tessar lens is more sensible to fabrication tolerances due to having more lens elements.

## 5. Tolerancing Analysis

We include a Monte-Carlo tolerancing analysis in Tab. 2, which supports that fabrication would marginally impact object detection performance while maintaining the margins that are reported in the main manuscript. We note that object detection performance with our joint optimization method, compared to the alternative of fine-tuning the object detector but fixing the lens, does not suffer more from fabrication tolerances.

Precisely, before evaluating each image (n=10 000), we apply random perturbations to each lens design by uniformly sampling across standard tolerances from Optimax and Schott (curvature: 0.2 %; glass thickness: 0.05 mm; refractive index: $5{\cdot}10^{-4}$; Abbe number: 0.5 %). We note that our designs are sensitive to changes in glass thickness due to their small total track length, so we employ only the second-most economical option (0.05 mm instead of 0.15 mm). In this analysis, we neglect airspace tolerances and use the paraxial image solve as we consider that the focus could be adjusted during fabrication.

Incidentally, we note that hypothetical discrepancies between the designed and manufactured lens could be accounted for by fine-tuning the downstream model using the measured optical performance metrics (PSF, distortion, etc.), as is common in joint design methodology [5].

## 6. Qualitative Ablation Experiments

In Fig. 4, we report the final 2D lens layouts that accompany each of the ablation experiments from the main document, for the joint design of the Tessar lens on 2× simulated resolution. For this experimental setting, *removing any component of the proposed method harms the stability* of the optimization process and results in a poorly behaved lens.

## 7. Comparison to Li et al. [4]

In this section, we present additional comparisons to the ray-tracing approach proposed by Li et al. [4]. In Tab. 3, we report Tessar lens experiments by making two changes to our ray-tracing algorithm: as in [4], we fill the entrance pupil with a square grid of rays (such that the corners of the square grid hit the circular edge of the aperture stop), and ignore accurate ray aiming. We train the Tessar lens and object detector jointly using this modified ray-tracing (MRT) algorithm, then evaluate the trained model using both the MRT algorithm and our complete ray-tracing (CRT) algorithm. Under both 1× and 2× simulated resolution, the MRT algorithm leads to an underestimated spot size (14.2 µm and 21.0 µm instead of 16.0 µm and 26.5 µm on 1× and 2× simulated resolution, respectively). Likewise, the average precision (AP) is overestimated when using the MRT instead of the CRT (33.4 and 28.3 instead of 33.2 and 27.5, respectively). This validates the proposed method as a more accurate ray-tracing algorithm to investigate task-driven optical design.

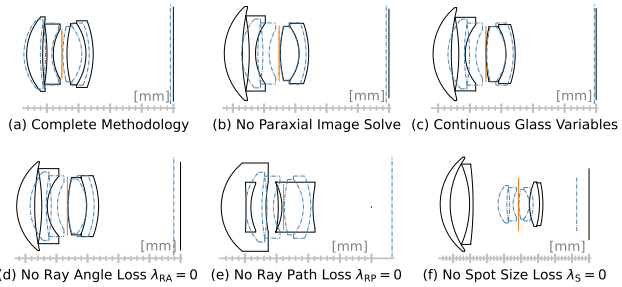


Figure 4. Qualitative ablation experiments. Illustrated are the lens layouts after the joint optimization of the Tessar lens on the BDD100K dataset under 2× simulated resolution. In this experimental setting, our complete methodology (a) favors a lens that resembles the baseline lens and starting point (shown in blue with the aperture stop as the reference plane). In (b), (c), and (d), in addition to vignetted rays (not shown), the lens deviates further from the starting point and ends up with strong pupil aberrations that cannot be handled by a single ray-aiming correction step. In (e), overlapping lens elements result from the removal of the ray path loss. In (f), without the spot size loss, only the noisy object detection loss drives the optimization process; as a result, the lens diverges significantly from the starting point and ends up with a mean spot size 9.6 times the one of the baseline lens.

| Setting | Optics | Eval. with MRT | | Eval. with CRT | |
|---|---|---|---|---|---|
| | | Spot (µm)$_\downarrow$ | AP$_\uparrow$ | Spot (µm)$_\downarrow$ | AP$_\uparrow$ |
| MRT [4] | Tessar (1× res.) | 14.2 | 33.4 | 16.0 | 33.2 |
| **Ours** | | | | 14.8 | **33.6** |
| MRT [4] | Tessar (2× res.) | 21.0 | 28.3 | 26.5 | 27.5 |
| **Ours** | | | | 24.7 | **32.2** |

Table 3. Comparison on the joint optimization of the Tessar lens with a modified ray-tracing (MRT) algorithm that follows the methodology in Li et al. [4]. We report the AP on BDD100K, where aberrations are modeled using either the MRT algorithm or our complete ray-tracing (CRT) algorithm. We also report the mean spot size evaluated using both ray-tracing algorithms.

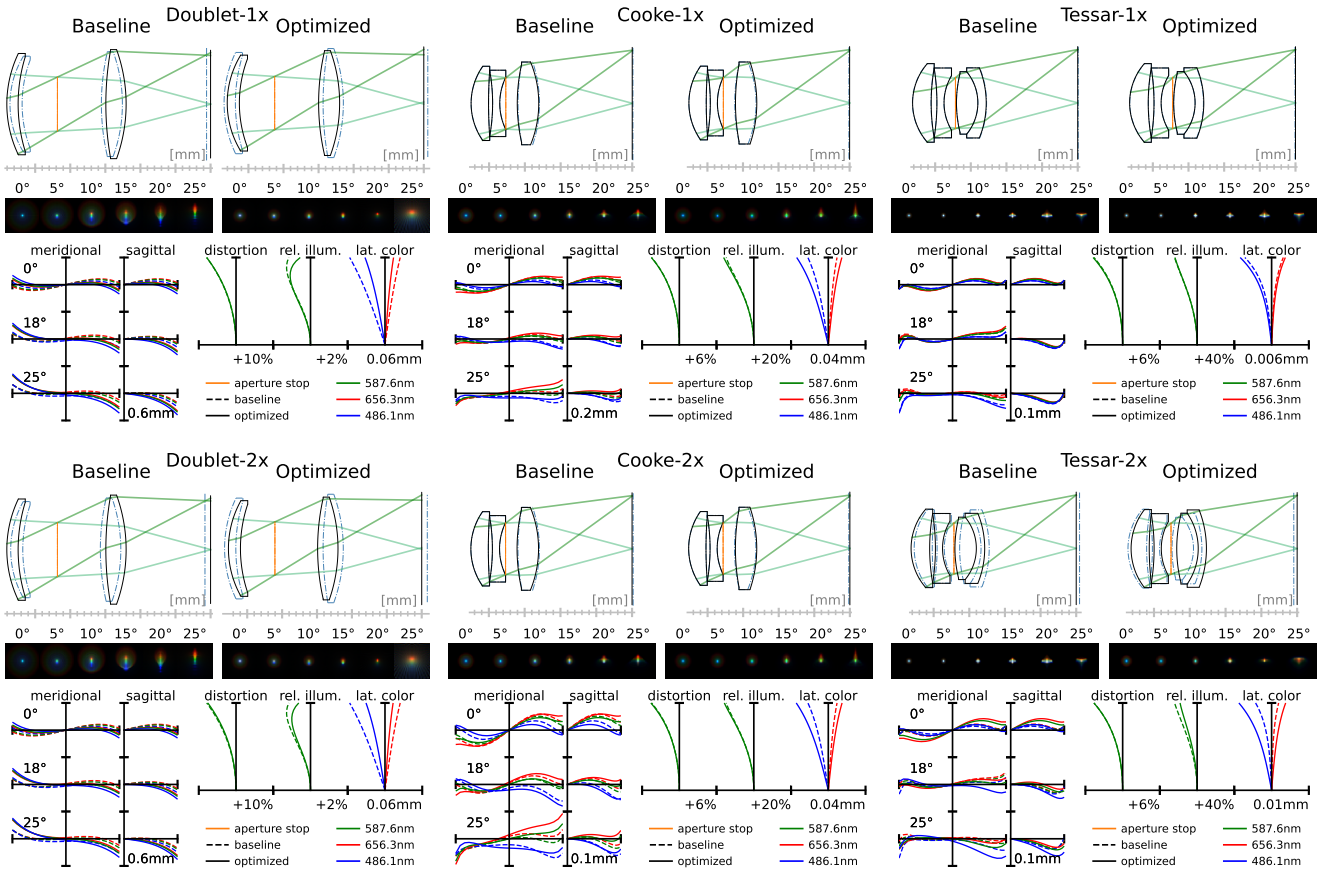


Figure 5. Baseline and optimized lenses. From top to bottom, we show 1) the lens designs (dashed lines represent the baseline/optimized counterpart); 2) PSFs for different fields; and 3) aberration charts (left: ray fan plots; right: field curves).

| | $c'$ | | $s'$ | | $g_1$ | | $g_2$ | |
|---|---|---|---|---|---|---|---|---|
| | min | max | min | max | min | max | min | max |
| 1 | 1.72 | 2.37 | 0.140 | 0.164 | -1.17 | -0.71 | -1.11 | -0.65 |
| 2 | 0.28 | 0.93 | 0.035 | 0.059 | | | | |
| 3 | -0.88 | -0.23 | 0.046 | 0.071 | -1.14 | -0.68 | 0.93 | 1.39 |
| 4 | 1.96 | 2.62 | 0.081 | 0.105 | | | | |
| 5* | | | -0.004 | 0.021 | | | | |
| 6 | 0.66 | 1.31 | 0.163 | 0.187 | -1.17 | -0.71 | -1.11 | -0.65 |
| 7 | -2.81 | -2.16 | 0.056 | 0.081 | -3.64 | -3.18 | -0.93 | -0.47 |
| 8 | | | -0.028 | -0.004 | | | | |

Table 4. Predefined boundaries for each of the 22 normalized Tessar lens variables, which we use to replicate the proxy model approach of Tseng et al. [5]. As in [5], the boundaries have two purposes: to sample the lens variables used to train the proxy model, and to limit the allowed range during joint optimization experiments. Note that there is no curvature for the flat aperture stop (denoted *) nor for the last optical surface, which is computed using a paraxial ray-tracing operation to enforce the desired focal length.

## 8. Proxy Model

In this section, we provide additional details on our comparison with the proxy model of Tseng et al. [5], closely adapted here for fair comparison.

We first generate 10k variations of the baseline Tessar lens by uniformly sampling each of the 22 lens variables between predefined boundaries as given in Tab. 4. The variable boundaries are centered on the lens parameters of the baseline Tessar lens. The allowed range for each lens variable is set to 0.4 times the standard deviation of each variable group: 6 normalized curvatures $c'$, 8 normalized spacings $s'$, and 8 normalized glass variables $g$. In joint optimization experiments, these variable boundaries are also used to clip each lens variable after each optimization step. In contrast to the use of a proxy model, we note that our ray-tracing approach does not require predefined boundaries: the ray-tracing algorithm works for all of the solution space (as long as ray aiming remains accurate), and manufacturing constraints are handled with carefully designed losses instead of limiting the solution space within a predefined region.

Quantized continuous glass variables do not synergize well with a proxy model, so we use standard continuous relaxations instead. However, we do use the paraxial image solve as in other experiments.

**Model Architecture and Training** As in Tseng et al. [5], our proxy model consists of a multilayer perceptron (MLP) followed by a convolutional decoder. The MLP takes as

| | $1/c$ mm | $s$ mm | Glass | $n_\mathrm{d}$ | $v_\mathrm{d}$ | $1/c$ mm | $s$ mm | Glass | $n_\mathrm{d}$ | $v_\mathrm{d}$ | $1/c$ mm | $s$ mm | Glass | $n_\mathrm{d}$ | $v_\mathrm{d}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Doublet (Baseline)** | | | | | **Doublet (Optimized, 1× res.)** | | | | | **Doublet (Optimized, 2× res.)** | | | | |
| 1 | 16.71 | 1.61 | S-LAL12 | 1.678 | 55.3 | 13.14 | 1.73 | S-TIM1 | 1.626 | 35.7 | 14.01 | 1.72 | S-TIM1 | 1.626 | 35.7 |
| 2 | 22.92 | 5.60 | | | | 17.78 | 5.03 | | | | 19.83 | 4.94 | | | |
| 3* | inf | 6.90 | | | | inf | 6.27 | | | | inf | 6.12 | | | |
| 4 | 44.34 | 2.89 | S-LAH92 | 1.892 | 37.1 | 32.88 | 2.99 | S-LAH96 | 1.764 | 48.5 | 33.02 | 2.92 | S-LAH96 | 1.764 | 48.5 |
| 5 | -22.87 | 12.03 | | | | -21.76 | 12.00 | | | | -22.17 | 12.00 | | | |
| | **Cooke (Baseline)** | | | | | **Cooke (Optimized, 1× res.)** | | | | | **Cooke (Optimized, 2× res.)** | | | | |
| 1 | 9.10 | 2.44 | S-LAH96 | 1.764 | 48.5 | 8.98 | 2.44 | S-LAH96 | 1.764 | 48.5 | 8.98 | 2.44 | S-LAH96 | 1.764 | 48.5 |
| 2 | 67.86 | 0.57 | | | | 58.92 | 0.57 | | | | 58.52 | 0.57 | | | |
| 3 | -26.08 | 1.00 | S-TIM1 | 1.626 | 35.7 | -27.40 | 1.00 | S-TIM1 | 1.626 | 35.7 | -27.33 | 1.00 | S-TIM1 | 1.626 | 35.7 |
| 4 | 8.35 | 0.84 | | | | 8.40 | 0.83 | | | | 8.40 | 0.83 | | | |
| 5* | inf | 1.60 | | | | inf | 1.71 | | | | inf | 1.73 | | | |
| 6 | 25.01 | 3.00 | S-LAH96 | 1.764 | 48.5 | 29.63 | 3.00 | S-LAH92 | 1.892 | 37.1 | 29.35 | 3.00 | S-LAH92 | 1.892 | 37.1 |
| 7 | -15.20 | 13.06 | | | | -17.97 | 13.06 | | | | -18.02 | 13.09 | | | |
| | **Tessar (Baseline)** | | | | | **Tessar (Optimized, 1× res.)** | | | | | **Tessar (Optimized, 2× res.)** | | | | |
| 1 | 8.39 | 2.61 | S-LAH96 | 1.764 | 48.5 | 8.40 | 2.61 | S-LAH96 | 1.764 | 48.5 | 8.89 | 2.53 | S-LAH96 | 1.764 | 48.5 |
| 2 | 28.27 | 0.81 | | | | 28.14 | 0.81 | | | | 39.54 | 0.68 | | | |
| 3 | -30.99 | 1.00 | S-TIM1 | 1.626 | 35.7 | -31.12 | 1.00 | S-TIM1 | 1.626 | 35.7 | -30.34 | 1.00 | S-TIL25 | 1.581 | 40.7 |
| 4 | 7.49 | 1.60 | | | | 7.47 | 1.62 | | | | 7.69 | 1.36 | | | |
| 5* | inf | 0.14 | | | | inf | 0.14 | | | | inf | 0.80 | | | |
| 6 | 17.43 | 3.00 | S-LAH96 | 1.764 | 48.5 | 17.33 | 3.00 | S-LAH96 | 1.764 | 48.5 | 19.53 | 3.00 | S-LAH96 | 1.764 | 48.5 |
| 7 | -6.90 | 1.17 | S-LAH88 | 1.916 | 31.6 | -6.95 | 1.22 | S-LAH88 | 1.916 | 31.6 | -8.72 | 1.15 | S-LAH88 | 1.916 | 31.6 |
| 8 | -13.00 | 12.84 | | | | -13.02 | 12.86 | | | | -14.85 | 12.68 | | | |

Table 5. Complete list of lens parameters for all experiments: radii $1/c$, spacings $s$, and glass materials along with the refractive index $n_\mathrm{d}$ and Abbe number $v_\mathrm{d}$. The aperture stop surface is denoted with *.

inputs the 22 lens variables as well as the field value, and is composed of two hidden layers with 128 units and an output layer with $32 \cdot 32 \cdot 3 + 2 = 3074$ units. The last 2 units are used for the relative illumination factor and distortion shift. The rest are reshaped ($32 \times 32 \times 3$), then fed into the decoder with the following architecture:

- two $3 \times 3$ convolutional layers with 64 output channels (output size is $32 \times 32 \times 64$);
- transposed convolutional layer for 2× upsampling;
- two $3 \times 3$ convolutional layers with 64 output channels (output size is $64 \times 64 \times 64$);
- transposed convolutional layer for 1× upsampling to closely follow Tseng despite a smaller PSF size (output size is $65 \times 65 \times 64$);
- $3 \times 3$ convolutional layer with 3 output channels (output size is $65 \times 65 \times 3$).

We obtain the PSFs by normalizing the outputs using a softmax operation so that their channel-wise area is 1, as is the case for the ground truth PSFs. The proxy model is trained on 10 epochs using the Adam optimizer [3] with a learning rate of 0.001 and a batch size of 10.

## 9. Additional Results

Fig. 5 provides detailed aberration charts and lens layouts for the Doublet, Cooke, and Tessar lenses reported in the main paper, optimized either for spot size (baseline) or object detection with 1× or 2× simulated resolution. Tab. 5 lists the corresponding lens parameters for each experimental setting.

## References


[1] G. Côté, J.-F. Lalonde, and S. Thibault. Deep learning-enabled framework for automatic lens design starting point generation. *Opt. Express*, 29(3):3841–3854, Feb 2021. doi: 10.1364/OE.401590. 1

[2] A. Halé, P. Trouvé-Peloux, and J.-B. Volatier. End-to-end sensor and neural network design using differential ray tracing. *Optics express*, 29(21):34748–34761, 2021. 1

[3] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *Proccedings of the 3rd International Conference on Learning Representations*, 2015. 1, 5

[4] Z. Li, Q. Hou, Z. Wang, F. Tan, J. Liu, and W. Zhang. End-to-end learned single lens design using fast differentiable ray tracing. *Optics Letters*, 46(21):5453–5456, 2021. 1, 3

[5] E. Tseng, A. Mosleh, F. Mannan, K. St-Arnaud, A. Sharma, Y. Peng, A. Braun, D. Nowrouzezahrai, J.-F. Lalonde, and F. Heide. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Trans. Graph.*, 40(2):1–19, jun 2021. ISSN 0730-0301. doi: 10.1145/3446791. 1, 3, 4