# Supplementary Information Differentiable Compound Optics and Processing Pipeline Optimization for End-to-end Camera Design

ETHAN TSENG*, Princeton University, United States
ALI MOSLEH* and FAHIM MANNAN*, Algolux, Canada
KARL ST-ARNAUD, AVINASH SHARMA, and YIFAN PENG, Algolux, Canada
ALEXANDER BRAUN, Hoschüle Dusseldorf, Germany
DEREK NOWROUZEZAHRAI, McGill University, Canada
JEAN-FRANÇOIS LALONDE, Université Laval, Canada
FELIX HEIDE, Princeton University, United States and Algolux, Canada

**ACM Reference Format:**
Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide. 2021. Supplementary Information Differentiable Compound Optics and Processing Pipeline Optimization for End-to-end Camera Design. *ACM Trans. Graph.* 38, 6, Article 1 (August 2021), 44 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

In this document we provide additional discussion and results in support of the primary text:

- EMVA 1288 (Section 1)
- Sensor Characterization (Section 2)
- Differentiable Poisson Layer (Section 3)
- Compound Optics Modeling (Section 4)
- Optics Meta-Network Validation (Section 5)
- Larger Field of View (Section 6)
- Cooke Triplet Optimization with Manufacturing Constraints (Section 7)
- Analysis and Synthetic Validation (Section 8)
- Experimental Validation (Section 9)
- Optical Properties of Optimized Lens Designs (Section 10)
- PSF Calibration (Section 11)

*indicates equal contribution.

Authors' addresses: Ethan Tseng, Princeton University, United States; Ali Mosleh; Fahim Mannan, Algolux, Canada; Karl St-Arnaud; Avinash Sharma; Yifan Peng, Algolux, Canada; Alexander Braun, Hoschüle Dusseldorf, Germany; Derek Nowrouzezahrai, McGill University, Canada; Jean-François Lalonde, Université Laval, Canada; Felix Heide, Princeton University, United States, Algolux, Canada.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0730-0301/2021/8-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 EMVA 1288

The sensor is simulated using the EMVA 1288 python library [1, 2]. Fig. 1 show the different steps of the EMVA 1288 that generate the image from a spectral image $S(x, y, \lambda)$. All variables in Fig. 1 that are framed in red were added to the original EMVA1288 model to obtain a better fit of the experimental raw captures. They will be described in more detail in the following sections.

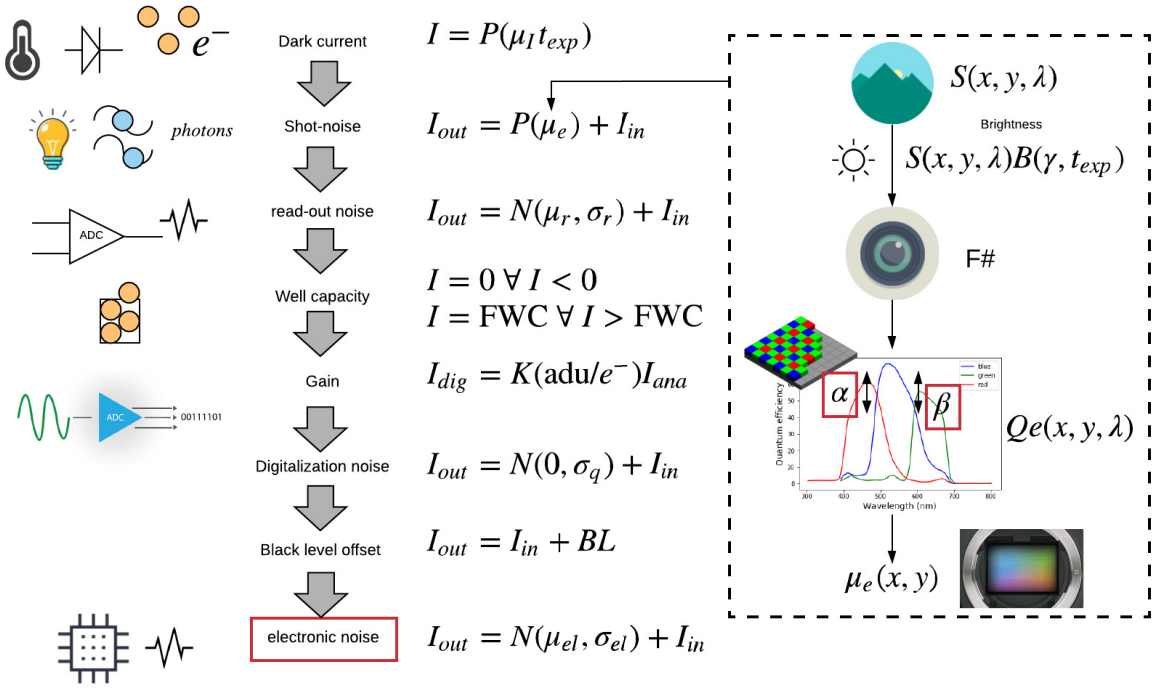



Fig. 1. EMVA1288 workflow to generate simulated raw from a spectral image $S(x, y, \lambda)$.

Following from Fig. 1, Tab. 1 describes all parameters that need to be evaluated in order to simulate the raw image. Those parameters were evaluated in four different ways:

(1) **Documentation**: A few parameters were directly taken from the camera specification sheet.
(2) **Observation**: Some parameters are directly determined from camera captures such as the black level, the size of the image, and the bit depth.
(3) **EMVA 1288**: The EMVA 1288 model provides procedures for estimating certain parameters such as the gain factor $K$, the read-out noise $\sigma_r$, and dark current $\mu_I$.
(4) **Optimization**: The remaining parameters were empirically estimated from the experimental raw captures.

Steps 1 to 3 allowed us to determine most of the camera model parameters. We performed optimization (4) on the remaining parameters to find the parameters set that best represents the real experimental captures.

Table 1. List of all the parameters of EMVA 1288 model.

| Parameters | Description | Comment |
|---|---|---|
| $\mu_I$ | Dark current [$e^-$ / pixel s] | Usually depends on the temperature but we assume the temperature to be constant. |
| $\gamma$ | Light intensity [$W/\text{cm}^2/sr$] | This value is manually adjusted to match the captured image. |
| $t_{\text{exp}}$ | Exposure [ms] | Real exposure as set on the camera. |
| $f\#$ | f-number | This is arbitrary set as it is only a scaling factor that can be included inside $\gamma$. |
| $Qe(x, y, \lambda)$ | Quantum efficiency | Quantum efficiency is given for every pixel of the image as a spectrum. |
| $\mu_r, \sigma_r^2$ | Read-out noise [$e^-$ / pixel] | - |
| $FWC$ | Full well capacity [$e^-$] | - |
| $K$ | Gain [$DN/e^-$] | - |
| $\sigma_q^2$ [$e^-$ / pixel] | Digitization noise | Given by EMVA 1288 documentation as (1/12) [1]. |
| $BL$ | Black level [$DN$] | Set by the camera. |
| $\mu_e, \sigma_e^2$ | Electronic noise [$e^-$ / pixel] | This is a noise that comes after the digitization so it is not affected by the gain parameters. |
| $\alpha, \beta$ | R/G, B/G | Ratio red/green and blue/green. |

Table 2. Parameters directly taken from documentation.

| Parameter | Value |
|---|---|
| Full well capacity | 33723 $e^-$ |
| Pixel area | $5.86^2$ |

## 2 SENSOR CHARACTERIZATION

### 2.1 Fixed Sensor Parameters

We used a BFLY-PGE-23S6C camera which has defined EMVA specifications [3]. However, we needed to adjust most of the parameters in order to obtain a simulated model that best represents experimental captures. In the end, only three parameters were directly taken from the documentation, the Full well capacity ($FWC$), the pixel area, and the quantum efficiency ($Qe$). These values are presented in Tab. 2. Furthermore, we used two additional parameters ($\alpha$ and $\beta$) to add flexibility to the quantum efficiency values.

The quantum efficiency is given for every pixel of the camera in the EMVA 1288 model. Fig. 2 shows the quantum efficiency for red, green, and blue channels. There are a few details about this quantum efficiency in the documentation [3], but the clipping of red at 700 nm shows that it includes the IR filter which is placed in front of the camera. With a maximum quantum efficiency of ~70%, we assume that this document value also includes the sensor quantum efficiency.

### 2.2 Calibrated Parameters from a Single Capture

From a single capture, certain parameters can be determined such as the image size, bit depth, and black level. These parameters are presented in Tab. 3.

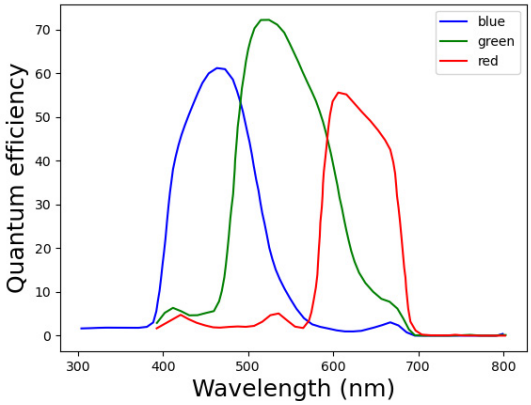



Fig. 2. Quantum efficiency for BFLY-PGE-23S6C camera.

Table 3. Parameters determined from observation.

| Parameter | Value |
|---|---|
| Width, Height | 1920, 1200 |
| Bit depth | 12 |
| Black level | 119 |

## 2.3 Parameter Estimates from EMVA 1288 Calibration

EMVA 1288 provides procedures for estimating some parameters such as gain factor ($K$) and the dark current.

*2.3.1 Gain.* The gain factor of the camera is the factor that exists between the number of electron generated in each pixel and the final Digital Number ($DN$). The procedure to evaluate this parameter is given in Sec. 6.6 of the EMVA 1288 documentation [1].

*Derivation.* The mean digital signal $\mu_y$ captured by the sensor is equal to the gain constant $K$ multiplied by the number of electrons generated without light $\mu_d$ and the number of electrons generated by the incident signal $\mu_e$:

$$\mu_y = K(\mu_d + \mu_e). \tag{1}$$

The noise $\sigma_y$ associated with this mean digital signal is given by

$$\sigma_y^2 = K^2\sigma_d^2 + \sigma_q^2 + K^2\mu_e \tag{2}$$

where $\sigma_d$ is the dark current and electronic noise of the sensor and $\sigma_q$ is the digitization noise. Applying Eq. (1) allows us to rewrite Eq. (2) as

$$\sigma_y^2 - (K^2\sigma_d^2 + \sigma_q^2) = K^2(\mu_y - \mu_d). \tag{3}$$

In that equation $K\mu_d$ corresponds to the mean signal generated by a black image and $(K^2\sigma_d^2 + \sigma_q^2)$ is the noise associated with that black image. We can further rewrite Eq. 3 as:

$$\sigma_y^2 - \sigma_{y,dark}^2 = K^2(\mu_y - \mu_{y,dark}) \tag{4}$$

where $\sigma^2_{y,dark}$ and $\mu_{y,dark}$ are respectively the dark image noise and mean value. Given at least two images of gray and black images we can isolate $K$ in Eq. (4) as the slope between difference of noise and difference of mean signal between gray and dark images. To be able to measure the slope the images need to be taken at different exposure times.

*Calibration Captures.* For both dark image and gray background the lens was removed from the camera to get the true sensor response without lens effects (transmission curve, vignetting). Dark captures were taken by covering the lens mount of the camera with an opaque lens cover as shown in Fig. 3.

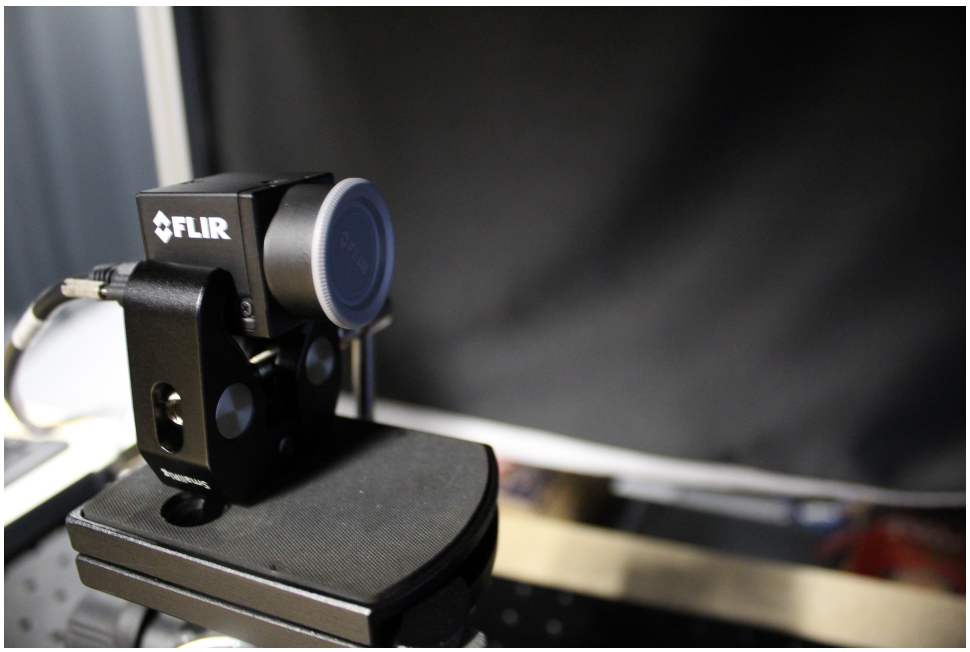



Fig. 3. Camera was covered with an opaque surface to take the dark frames.

Gray images were taken by placing the camera in front of an LCD screen displaying a gray patches. The size of the gray patches and the distance between the screen and camera were adjusted to get the most uniform illumination on the full sensor and minimize mechanical vignetting of the lens mount. To be able to later generate the spectral image from this capture, a spectral measurement of the screen was taken with a spectroradiometer (JETI Spectraval-1511). Fig. 4 shows the camera placed in front of the screen along with the JETI spectroradiometer.

Gray images and dark images were taken with 4 different exposures (2.5, 5, 7.5, 10 ms) and 7 different linear gain values (1.00, 1.77, 3.16, 5.62, 10.00, 17.78, 31.62). For every exposure and gain value, two images $A$ and $B$ were taken.

The mean signal of every image was calculated using the two captures $y_A$ and $y_B$ as

$$\mu = \frac{1}{2MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} (y^A[m][n] + y^B[m][n]). \tag{5}$$

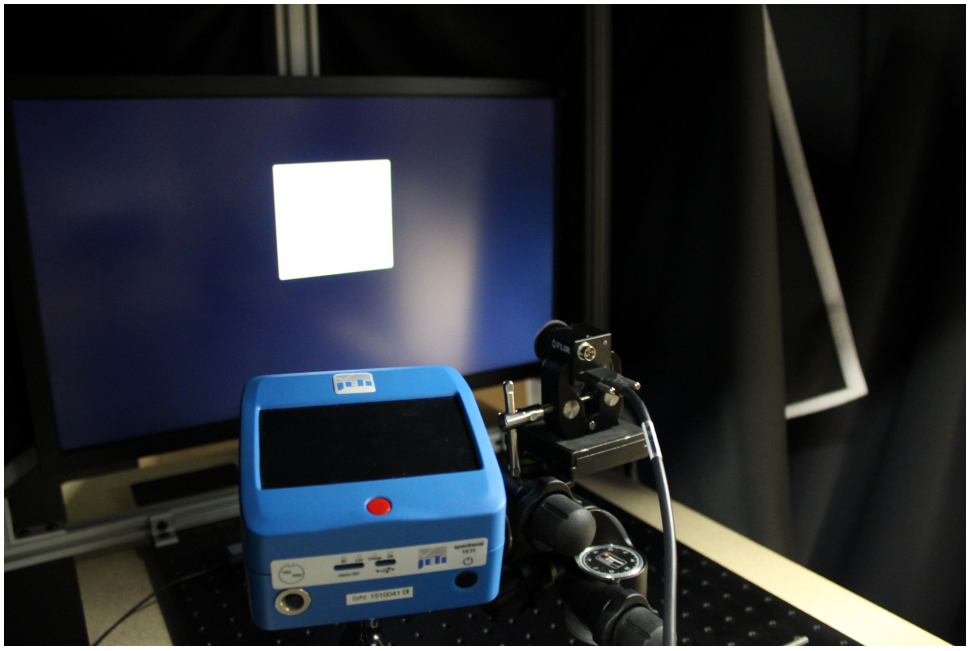


Fig. 4. Lab setup to take flat field image. A gray patches is displayed on the screen and camera is places in front of the LCD without lens. A JETI spectroradiometer is used to measure spectrum from the screen.

For stationary and homogeneous noise, the temporal variance of noise can be estimated from the difference of two images as

$$\sigma^2 = \frac{1}{2MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} (y^A[m][n] - y^B[m][n])^2 \tag{6}$$

*Calibration Result.* Fig. 5 presents the variation $(\sigma_y^2 - \sigma_{y_d ark}^2)$ as a function of $(\mu_y - \mu_{y,dark}^2)$ for different gains with the linear fit shown as a dotted line for every linear gain value.

From this graph it is possible to see that $K$ is gain dependent. However, after normalizing the $K$ constant by the gain value we obtain a value which is approximately constant for every gain ($K \approx 0.122$).

*2.3.2 Dark current.* The noise $\sigma_d$ is the combination of the read-out noise of the camera $\sigma_r$ and the dark current noise which depends on the exposure value $t_{\text{exp}}$:

$$\sigma_d^2 = \sigma_r^2 + \mu_I t_{\text{exp}}. \tag{7}$$

From Eq. (2) and Eq. (7), the dark image noise can be rewritten using $\mu_e$ is equal to 0 since there is no signal:

$$\sigma_{y,dark}^2 = (K^2\sigma_r^2 + \sigma_q^2) + K^2\mu_I t_{\text{exp}}. \tag{8}$$

By taking multiple dark images at multiple exposures times it is possible to estimate the dark current (slope) and read-out noise (offset).

*Calibration Capture.* Dark frame captures were taken as previously described for several exposures (2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20, 22.5, 25, 27.5, 30 ms) and for each of those exposures we use several gains (1.00, 1.77, 3.16, 5.62, 10.00, 17.78, 31.62).

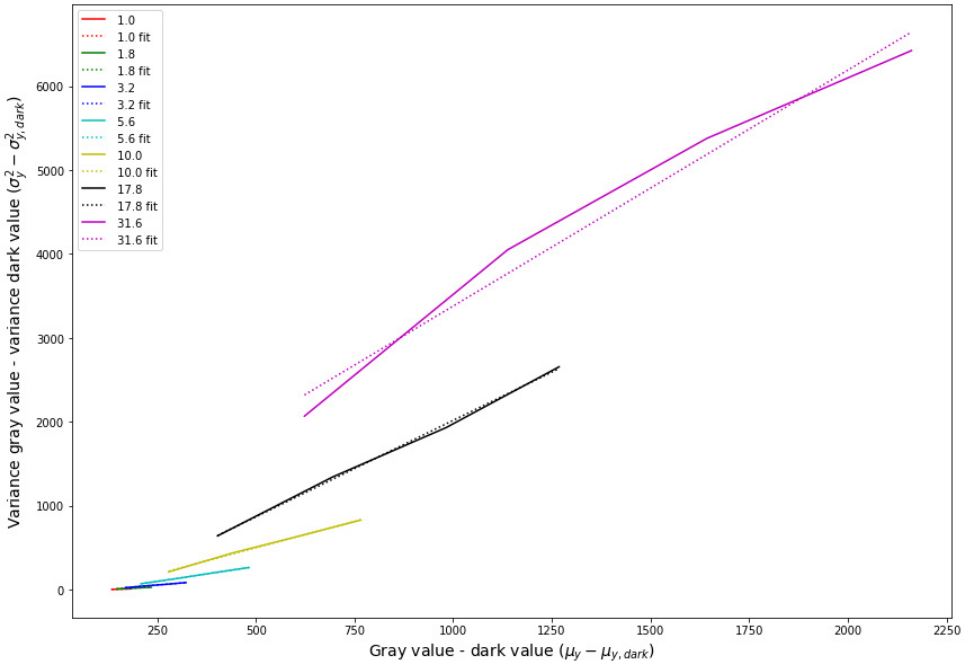



Fig. 5. Gain estimation from slope between mean signal and noise.

*Calibration Result.* For every gain a linear fit was performed between the value of $\sigma^2_{y,dark}$ and the exposure time $t_{\text{exp}}$. Fig. 6 shows an example for gain 1.8. At the end a final value of $\mu_I = 25e^-$ and $\sigma_r^2 = 16e^-$ was evaluated as the average of each value measured for every gain.

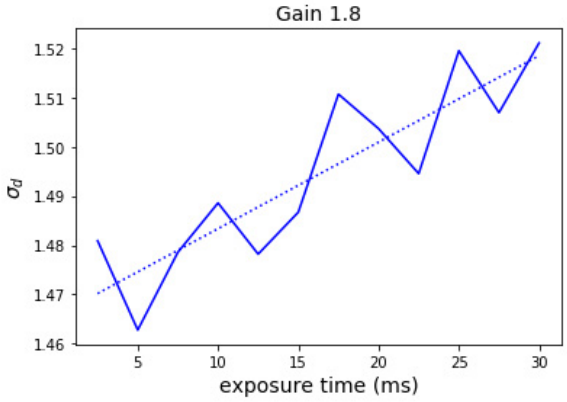



Fig. 6. Linear fit between the dark frame noise $\sigma^2_{y,dark}$ and the exposure time $t_{\text{exp}}$ for a gain of 1.8.

To get a better fit at every gain it was found that another source of noise independent of the gain factor $K$ should be added. This noise is generated after the digitization of the signal by the sensor. This corresponds to the electronic noise added in Fig. 1.

Table 4. Parameters evaluated using EMVA 1288 recommended procedures.

| Parameter | Value |
|---|---|
| Dark current ($\mu_I$) | 25 $e^-$ / pixel s |
| Read-out noise ($\sigma_r$) | 16 $e^-$ / pixel |
| Gain factor ($K$) | 0.122 |

## 2.4 Parameter Optimization and Validation

All of the above parameters were used to build a first model of the camera. They were then used to generate a flat field image and a dark frame which was compared against real captures of a flat field and a dark frame. All comparisons were conducted directly on the raw captures without any post-processing. The parameters of the model were then manually optimized to obtain a good match between capture and simulated raw.

*2.4.1 Flat Field Frames.* To generate the simulated flat field image, the spectrum measured previously with the JETI spectroradiometer was used. The spectrum was normalized and intensity was directly adjusted as one of the camera parameters to generate the image. It was adjusted empirically to match the real raw capture. The flat field images were principally useful for validating the quantum efficiency parameters and gain factor constant ($K$).

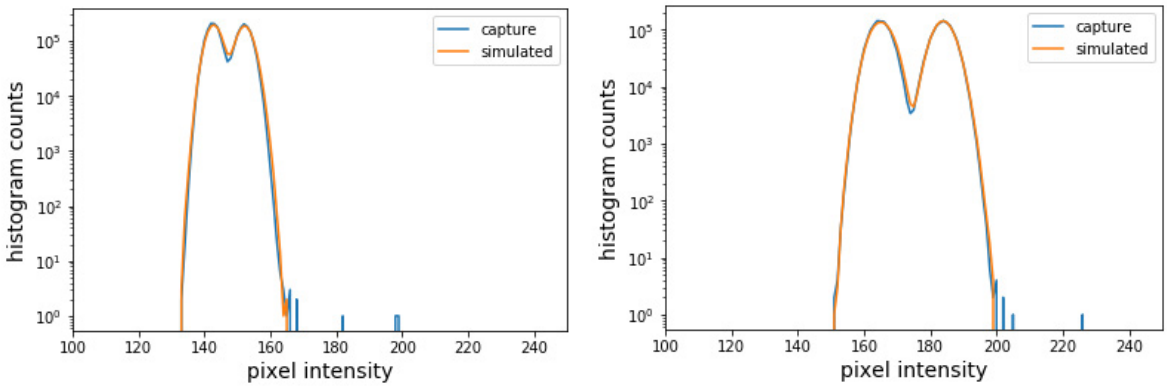
*Quantum Efficiency Adjustment.* The first parameters that needed to be adjusted from their theoretical values were the ratios in intensity between the red, blue, and green channels. Two parameters were added ($\alpha$ and $\beta$) to adjust the ratio between the blue and green and the ratio between red and green. Final values of 0.82 for R/G and 1.25 B/G have allowed for good color reproduction in the simulated raws. Those ratios are essential as the quantum efficiency curve provided by the vendor does not always accurately reflect the real quantum efficiency of the camera.

*Gain Adjustment.* The gain parameters also needed to be slightly adjusted from 0.125 to 0.14.

*Histogram Comparison.* Comparisons between the histograms of the simulated raw and the real captured raw are shown in Fig. 7. The simulated raw shows a strong similitude with the real captured raw.

*2.4.2 Dark Frames.* A similar procedure of comparison was performed on the dark frame. This comparison was principally useful to validate noise parameters and find some of the parameters that were not estimate up to now such as the electronic noise ($\sigma_e$ and $\mu_e$) and the read-out noise offset $\mu_r$. The spectral image for the dark image is simply null since there are no photons falling onto the sensor. The dark current and electronic noise were kept to the same value as estimate previously. By matching the simulated raw to the real capture, we were able to estimate the values of $\mu_r = 3.25$, $\sigma_e = 0.92$, and $\mu_e = 1.15$.

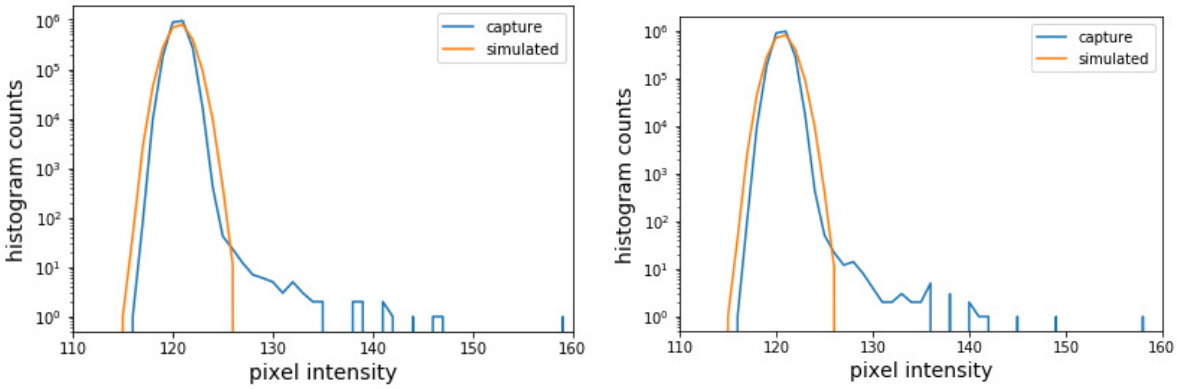
*Histogram Comparison.* Fig. 8 presents the histogram comparison between simulated raw and real raw capture for a dark frame. For the dark frame the simulated raw describes a perfect bell shape while the original raw capture presents the same bell shape with an extension towards large pixel values. These large pixel values come from the dark signal non uniformity (DSNU) caused by manufacturing imperfections of the camera. We did not characterize the DSNU of the camera.




(a) 5 ms capture

(b) 10 ms capture

Fig. 7. Comparison between histogram counts for simulated raw and captured raw of flat field at different exposure times for a linear gain of 1.




(a) 5 ms capture

(b) 10 ms capture

Fig. 8. Comparison between simulated and capture histogram of dark frame at different exposure time for a linear gain of 1.

## 2.5 Overall Sensor Model Calibration Results

Tab. 5 lists the final calibration results for the proposed sensor model.

Table 5. Final list of all estimated parameters of EMVA 1288 model.

| Parameter | Value |
|---|---|
| $\mu_I$ | 25 [$e^-$ / pixel s] |
| $\mu_r$ | 3.25 [$e^-$ / pixel] |
| $\sigma_r^2$ | 16 [$e^-$ / pixel] |
| $FWC$ | 33723 [$e^-$] |
| $K$ | 0.14 [$DN/e^-$] |
| $\mu_e$ | 1.15 [$e^-$ / pixel] |
| $\sigma_e$ | 0.92 [$e^-$ / pixel] |
| $\alpha$ | 0.82 |
| $\beta$ | 1.25 |

## 3 DIFFERENTIABLE POISSON LAYER

Our end-to-end optimization method requires that gradients can flow from the endpoint loss all the way back to the optics. Thus, for end-to-end differentiability we require that each compartment is itself differentiable. In order to make the sensor model described in Eq. (3) of the main document differentiable we implemented the Poissonian noise as a differentiable layer (differentiation with respect to the rate parameter). This was done with the following code segment:

```
import tensorflow_probability as tfp
p = tfp.distributions.Poisson(rate=photon_means)
number_of_photons = tfp.monte_carlo.expectation(f=lambda x: x, samples=p.sample(1),
                        log_prob=p.log_prob, use_reparametrization=False)
```

A conventional Gaussian approximation to the Poisson distribution works as well for our optimization scheme but should only be used if the datasets are well-exposed, on average, and contain a negligible amount of low intensity regions.

While traditional image processing and computer vision algorithms handle noise in the loss function (e.g. by using hand-engineered regularizers such as total variation), this approach is unsuitable for our framework. It is not immediately obvious how noise would affect a feature extraction loss such as LPIPS or a semantic loss such as IoU for object detection. Thus, we incorporate sensor noise as part of the image formation pipeline which allows our end-to-end optimizer to decide how to best meet the endpoint objective in the presence of sensor noise.

## 4 COMPOUND OPTICS MODELING

Our optics model takes as input the optical parameters and outputs PSFs for three different wavelengths with the appropriate vignetting factor applied to the PSFs. We obtained the ground truth PSFs from Zemax's optical simulation tool for different wavelengths and field parameters. In order to scale to different PSF resolution the parameters are first fed to a MLP network which then outputs spatial PSFs that can be scaled to different resolution. The overall network architecture is as follows,

The MLP consists of 2 layers of 128 hidden units, the output layer contains $32 \cdot 32 \cdot 3 + 3 = 3075$ units. The last 3 units are used for the RGB vignette factor and the rest are reshaped into a $32 \times 32 \times 3$ tensor which is fed into a decoder with 2 upsampling layers.

The decoder architecture is:

(1) $2 \times 3 \times 3$ conv layers with output channel 64 (feature map size is $32 \times 32 \times 64$)
(2) conv transpose upsampling by $2\times$
(3) $2 \times 3 \times 3$ conv layers with output channel 64 (feature map size is $64 \times 64 \times 64$)
(4) conv transpose upsampling by $2\times$
(5) $3 \times 3$ conv with 3 output channels (output map size is $128 \times 128 \times 3$)

The output is then channel-wise normalized by the channel-wise sum and multiplied by the vignetting factor. For the loss function, instead of taking the $\mathcal{L}_2$ loss between the estimated PSF and the Zemax output, we split the multichannel PSFs into energy preserving PSFs and per channel vignetting factors. This allows us to model the characteristics of the optical design from a sparse sampling of the field. Separating the vignetting factor from the PSFs also allows us to train the network to model the shape of the PSFs correctly. Without separating the vignetting factor, the network fails to accurately predict the shape of low energy PSFs at the periphery. Additionally, we use spatial gradient loss in order to give more weight to the high frequency components of the PSFs.

For PSF training we sampled different optical tolerances that produce a valid lens design and generated the PSFs by simulating the designs using Zemax. The input parameters are then normalized by the mean and variance and used for training. One could go further and incorporate robustness to manufacturing errors by adding noise to the input parameters, however we did not find this necessary.

Once trained, our optics meta-network computes spatial PSFs much faster than using full ray-tracing with Zemax. For computing a single spatial PSF, Zemax took 0.464 seconds running on an Intel Xeon Processor E5-1620 using 4 CPU cores, whereas the trained optics meta-network took 0.002 seconds running on a single Nvidia P100 GPU, providing a $200\times$ speedup. In our experiments we computed as many as 49 unique spatial PSFs per iteration, and the number of PSFs grows quadratically with sensor size and field of view. Thus, in addition to differentiability, the computational speed of our optics meta-network is an important attribute that enables efficient end-to-end optimization of our imaging pipelines.

As described in the main document, we chose to predict the spatial PSFs at $K = 13$ discrete locations across the field of view. This allows for sufficient coverage of the full scene while still remaining computationally efficient, as training for and simulating the unique PSF for each pixel location would require significantly more training time and training data from Zemax.

## 5 OPTICS META-NETWORK VALIDATION

Fig. 9 shows PSFs obtained using the optics meta-network for three randomly generated lens designs whose parameters are within the ranges provide in tab. 1 of the main document. For each lens design, the PSFs obtained using ray tracing in Zemax for various fields are also presented in this figure. These qualitative results demonstrate that our optics model can accurately reproduce spatially-varying PSFs.

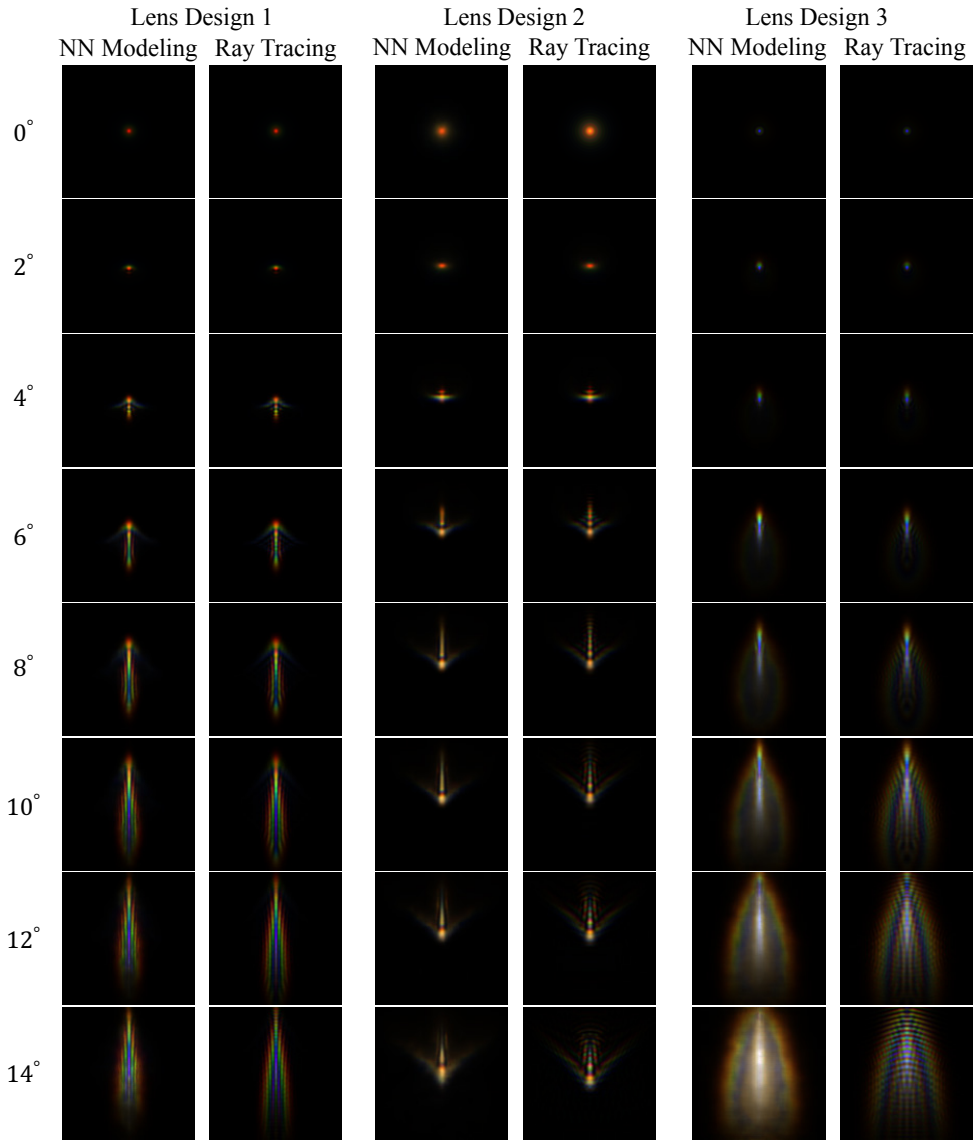



Fig. 9. PSFs across various vertical fields for three randomly generated lens designs. For each lens design, PSFs were obtained using the proposed learned optics meta-network (left) and ray tracing in Zemax (right). The center of the FOV corresponds to 0°. Each PSF is displayed in a $128 \times 128$ grid representing an area of $128 \times 128$ $\mu m^2$ on the imaging plane.

## 6 LARGER FIELD OF VIEW

We catered our lens designs towards the sensor size that we used for real-world experiments. Within our experimental setup our $1200 \times 1920$ sensor corresponds to a $25°$ FOV, typical in automotive mid-to-far range sensing. We demonstrate in this section that our optimization scheme also works for larger fields of view. Specifically, we optimize for a $2400 \times 3480$ sensor (same pixel pitch) corresponding to $50°$. For these experiments we jointly optimize the optics and post-processors in the same manner as described in the main document for the smaller sensor size, and we expert-optimize the nominal design for this larger field of view. Figs. 10 and 11 illustrate improved qualitative results for image quality with neural network and hardware ISP respectively, along with corresponding PSFs. Tab. 6 shows improved quantitative results on these same tasks.

The experiments with the larger field of view also demonstrated an interesting effect of the optimized PSFs. We observed that the optimized PSFs for image quality with neural network are the most compact not at the center of the FOV, but instead approximately halfway between the center and the periphery of the FOV. Due to the larger field of view, the off-center PSFs have a greater impact on image quality than the center PSFs. As such, the end-to-end optimization sharpens these off-center PSFs and the neural network compensates for the slight decrease in sharpness in the center of the FOV. However, the hardware ISP has lower capacity to encode such tailored deconvolution compared to the neural network and consequently the optimization produces a more conventional optic where the sharpest PSF is in the center of the FOV. These experiments again demonstrate that our end-to-end optimization methodology can produce specialized optics that are uniquely catered to different downstream processors and applications.

Table 6. Larger Field of View Evaluations. Quantitative evaluation of end-to-end design and nominal design on an unseen validation set for image quality using simulated measurements on a 2× expanded field of view compared to the prototype experiments in the main document.

| Methods | 1 - LPIPS | PSNR | SSIM |
|---|---|---|---|
| End-to-End with Neural Network | 0.902 | 30.4 | 0.884 |
| Nominal with Neural Network | 0.819 | 29.4 | 0.823 |
| End-to-End with Hardware ISP | 0.676 | 17.9 | 0.659 |
| Nominal with Hardware ISP | 0.575 | 17.9 | 0.623 |

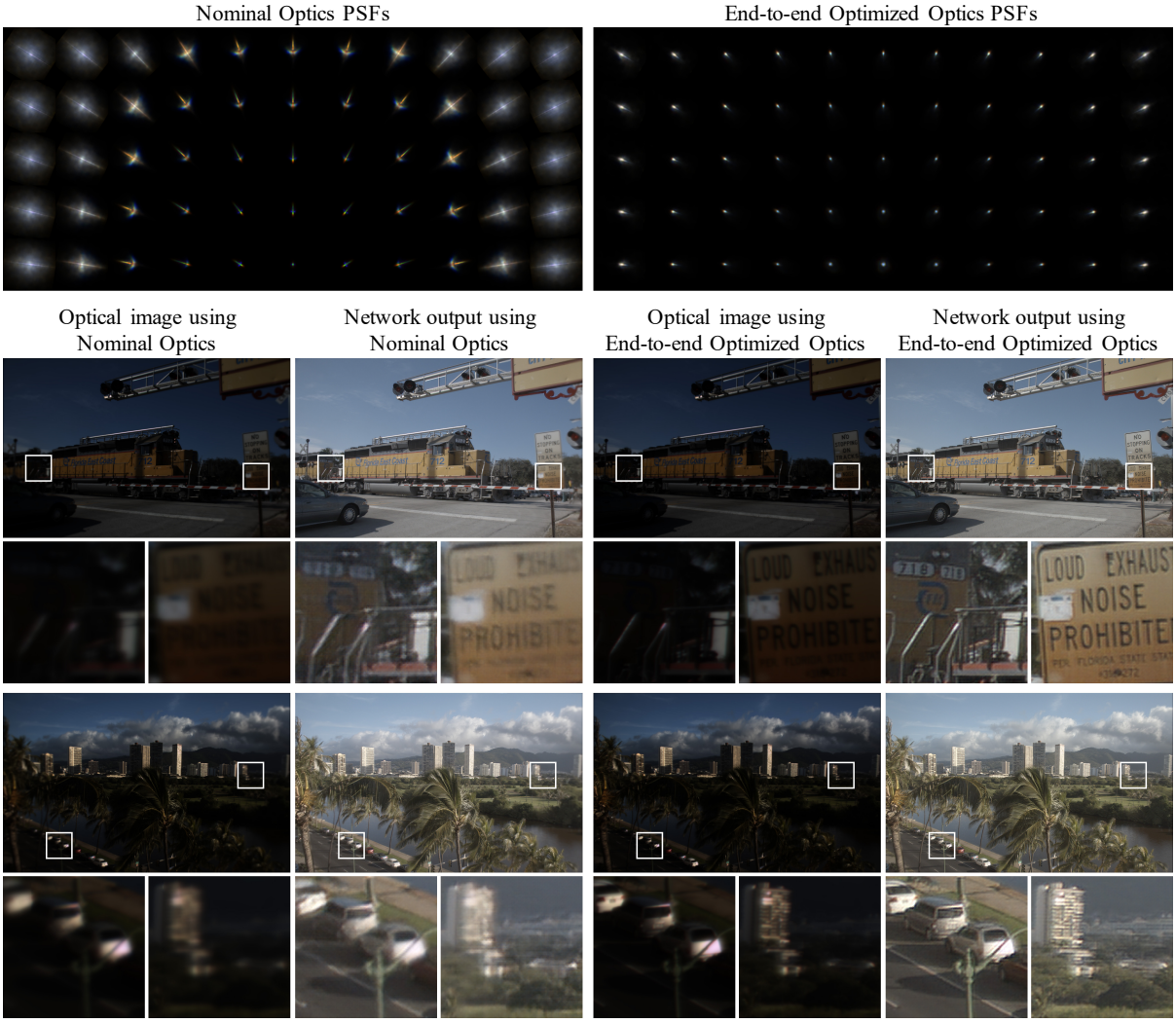



Fig. 10. Image quality with Cooke triplet and software ISP using simulated measurements and a 2× larger FOV than the prototype experiments in the main document. Simulated PSFs from our optics meta-network are shown at the top. Note that the end-to-end optimized PSFs are sharpest not at the center of the FOV, but instead approximately halfway between the center and the periphery of the FOV. We attribute this to the larger FOV, the off-center PSFs have greater impact on image quality than the center PSFs and the proposed end-to-end optimization takes advantage of this.

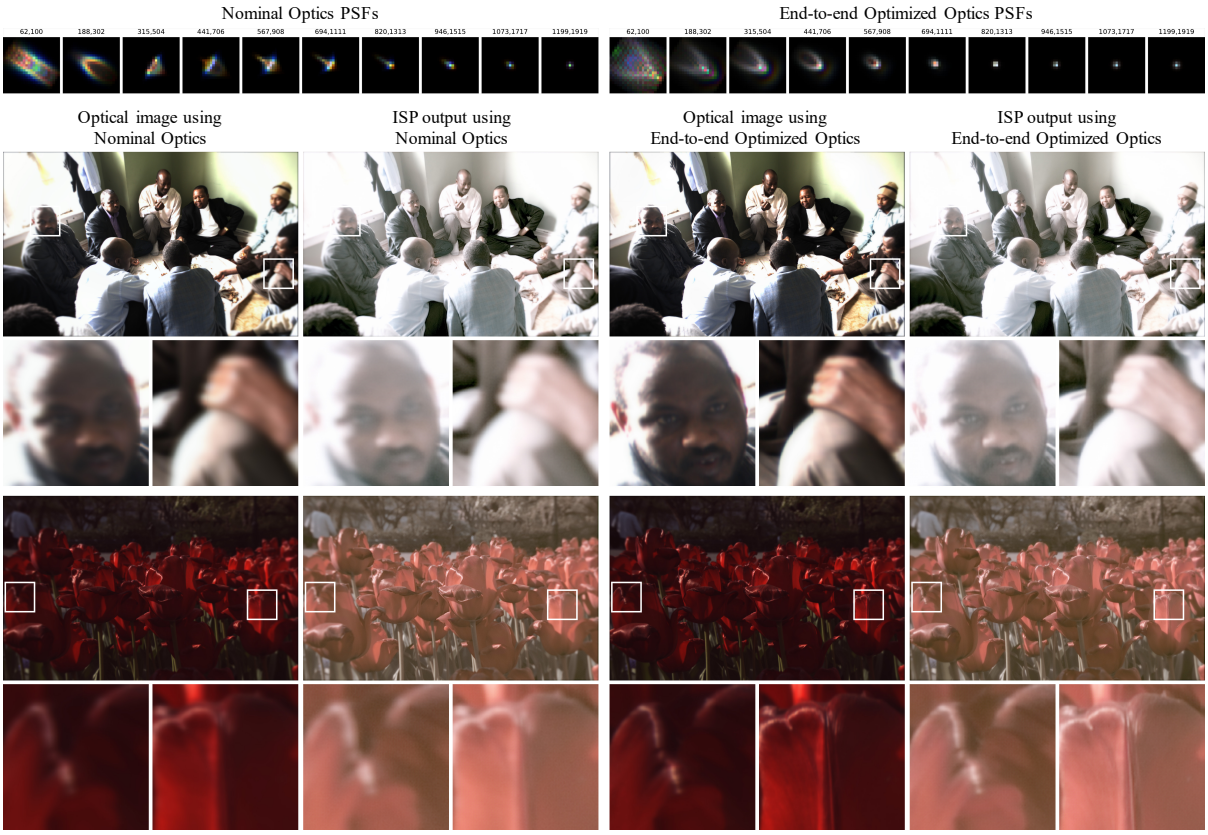



Fig. 11. Image quality with Cooke triplet and hardware ISP using simulated measurements and a 2× larger FOV than the prototype experiments in the main document. Displayed optics PSFs have been resampled for the sensor array. We show the PSFs along the first half of the main diagonal of the $1200 \times 1920$ sensor. The center pixel coordinates (row, column) of the spatial PSFs are indicated above each PSF, where (0, 0) refers to the top-left corner.

## 7 COOKE TRIPLET OPTIMIZATION WITH MANUFACTURING CONSTRAINTS

To supplement our extensive simulations, we also provide physical demonstrations of the proposed end-to-end optimization method in this work. However, lens manufacturing is subject to several constraints: three elements only, middle element fixed and off-the-shelf, and no glass material selection due to machine availability in our case. Moreover, lens optimization is also bound to additional physical constraints such as the back focal distance and distances between different elements. Considering all of these constraints in the traditional lens optimization platforms is not a trivial process.

In our first attempt to obtain a baseline optic for comparison against traditional Zemax optimization, a human error of the experienced optics engineer in considering the manufacturing constraints resulted in a set of over-constrained tight bounds for the lens parameters as shown in Tab. 7. This led what we dub as an "older" Zemax baseline lens design with an RMS spot radius of 30 micron after a two-week process of Hammer optimization and analysis of the design.

We performed the same experiments from Sec. 6.1 of the main document using these tighter manufacturing constraints. Results for end-to-end optimization using the software ISP for image quality are shown in Fig. 12 and Tab. 8 and results for single-image low-light imaging are shown in Fig. 14 and Tab. 9. Sharper optical images obtained using the proposed end-to-end design indicate a better enforcement of good focusing performance across the field of view compared to the older Zemax baseline. Fig. 13 and Tab. 8 show results for end-to-end optics and hardware ISP optimization for image quality. Images obtained through the proposed end-to-end design also indicate a better performance across the field of view as seen in the optical images compared to the old Zemax baseline. Lastly, results for automotive object detection and traffic light detection are shown in Fig. 15 and Tab. 10. These results were evaluated on a separate, smaller dataset of about 1000 validation captures compared to the larger validation set used for the experiments with the final nominal design. Here, we also see that the optimized optic for object detection changes the blur to be more uniform across fields while enhancing light efficiency. Similarly, the optimized optic for traffic light detection also has higher light efficiency in addition to sharper focus in the center and periphery for better acquisition of small traffic lights. The results with the old Zemax baseline demonstrate that our optimization methodology can successfully produce improved end-to-end imaging pipelines using specialized optics for a variety of tasks.

We also manufactured the old Zemax baseline lens and the optimized lenses adhering to the tighter manufacturing constraints, and then we used these prototypes to perform real capture comparisons. Fig. 16 qualitatively shows improved object detection and traffic light detection performance when using our optimized lenses.

Note that the tight parameter ranges favor 0-th order optimization approaches employed in the traditional optimization frameworks, e.g., the Hammer optimization in Zemax. However, our proposed end-to-end optimization produces designs with considerably higher performances compared to the old Zemax baseline lens design. Quantitative improvements are shown in Tabs. 8, 9, 10.

These comparison experiments against our previous Zemax baseline demonstrate that the proposed optics modeling and end-to-end optimization outperform traditional optical design work flows. Nevertheless, we later corrected the human error in considering the manufacturing constraints and obtained a higher quality nominal design with an RMS spot radius of 10 micron. The high quality nominal design is detailed in Sec. 6.1 of the main document and it was used in all other experiments presented in the main and supplemental document. We note that the new design took an experienced designer four man-weeks, bringing the whole design process (including the first baseline) up to *six weeks of Zemax-aided expert-design.*

Table 7. Parameters and their optimization ranges for the three-element Cooke triplet considering tight manufacturing constraints. We follow the optics CAD terminology and denote each lens element by its two surfaces [4]. Accordingly, we refer to the aperture and the imaging plane by surface 5 and surface 8 respectively. The Min/Max constraints are enforced for all lenses optimized under tighter manufacturing constraints and the old Zemax baseline lens.

| Parameter | Min | Max | Units | Description |
|---|---|---|---|---|
| s_1_radius | 10.98 | 14.98 | mm | radius of the 1st surface |
| s_1_conic | -0.49 | 0.29 | - | conic constant of the 1st surface |
| s_2_radius | 10.82 | 14.82 | mm | radius of the 2nd surface |
| l_12 | 4.58 | 7.57 | mm | distance between lens 1 and lens 2 |
| l_2STO | 2.72 | 9.22 | mm | distance between lens 2 and aperture |
| l_STO3 | 0.0 | 9.86 | mm | distance between aperture and lens 3 |
| s_6_radius | 14.17 | 18.17 | mm | radius of the 6th surface |
| s_6_conic | -0.49 | 0.49 | - | conic constant of the 1st surface |
| s_7_radius | -15.00 | -11.50 | mm | radius of the 7th surface |
| s_1_2nd | -4.99e-3 | 4.99e-3 | $mm^{-1}$ | 2nd order coefficient of polynomial fit to 1st surface |
| s_1_4th | -7.91e-5 | -1.45e-5 | $mm^{-3}$ | 4th order coefficient of polynomial fit to 1st surface |
| s_1_6th | -5.10e-7 | 4.88e-7 | $mm^{-5}$ | 6th order coefficient of polynomial fit to 1st surface |
| s_1_8th | -1.58e-8 | -85.27e-11 | $mm^{-7}$ | 8th order coefficient of polynomial fit to 1st surface |
| s_1_10th | -0.91e-10 | 1.08e-10 | $mm^{-9}$ | 10th order coefficient of polynomial fit to 1st surface |
| s_6_2nd | -4.99e-3 | 4.99e-3 | $mm^{-1}$ | 2nd order coefficient of polynomial fit to 6th surface |
| s_6_4th | -2.41e-4 | -1.41e-4 | $mm^{-3}$ | 4th order coefficient of polynomial fit to 6th surface |
| s_6_6th | -0.26e-6 | 1.73e-6 | $mm^{-5}$ | 6th order coefficient of polynomial fit to 6th surface |
| s_6_8th | -4.54e-8 | 0.54e-7 | $mm^{-7}$ | 8th order coefficient of polynomial fit to 6th surface |
| s_6_10th | -0.11e-8 | 8.80e-10 | $mm^{-9}$ | 10th order coefficient of polynomial fit to 6th surface |

Table 8. Quantitative evaluation of end-to-end design and old Zemax baseline design considering tight manufacturing constraints using simulated measurements on an unseen validation set for image quality.

| Methods | 1 - LPIPS | PSNR | SSIM |
|---|---|---|---|
| End-to-end with Neural Network | 0.960 | 36.1 | 0.942 |
| Old Zemax Baseline with Neural Network | 0.926 | 34.2 | 0.914 |
| End-to-End with Hardware ISP | 0.793 | 18.70 | 0.752 |
| Old Zemax Baseline with Hardware ISP | 0.718 | 18.61 | 0.728 |

Table 9. Quantitative evaluation of end-to-end design and old Zemax baseline design considering tight manufacturing constraints using simulated measurements on an unseen validation set for low-light imaging.

| Methods | 1 - LPIPS | PSNR | SSIM |
|---|---|---|---|
| End-to-end with Neural Network | 0.866 | 32.8 | 0.863 |
| Old Zemax Baseline with Neural Network | 0.816 | 31.8 | 0.827 |

Table 10. Mean Average Precision (mAP) for object detection (OD) and traffic light (TL) state detection for our end-to-end optimized pipeline versus the detection pipeline using the old Zemax baseline lens using simulated measurements.

| Methods | OD | TL |
|---|---|---|
| End-to-end with Hardware ISP and FRCNN | 56 | 45 |
| Old Zemax Baseline with Hardware ISP and FRCNN | 37 | 28 |

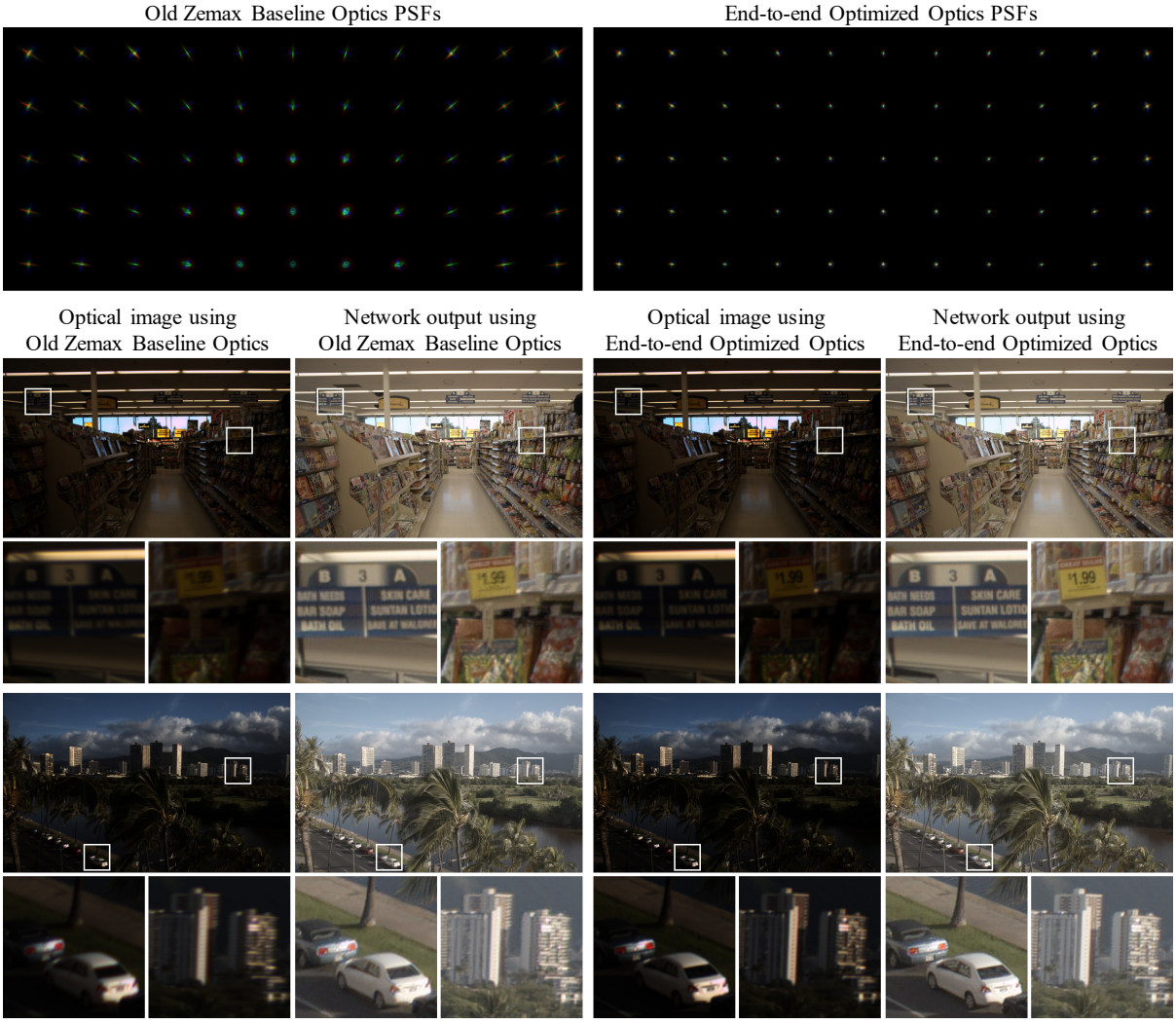



Fig. 12. Image quality with Cooke triplet considering tight manufacturing constraints and software ISP using simulated measurements. Similar to Fig. 7 in the main document, images produced using the old Zemax baseline lens are blurrier and have more color artifacts than images produced using our optimized optics (right). Simulated PSFs from our optics meta-network are shown at the top.

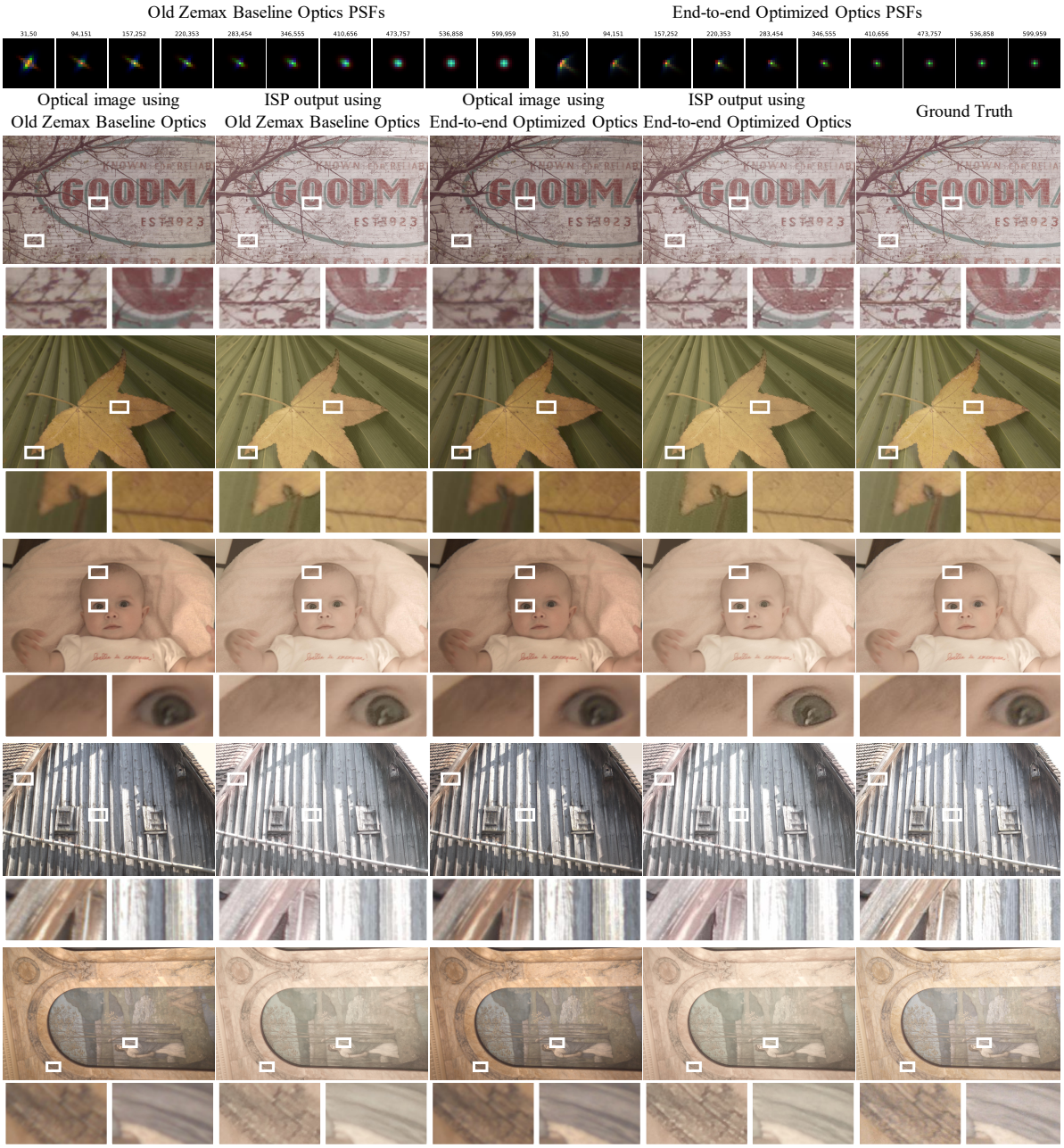



Fig. 13. Image quality with Cooke triplet considering tight manufacturing constraints and hardware ISP using simulated measurements. Similar to Fig. 8 in the main document, images produced using the old Zemax baseline lens are blurrier and thus the ISP tends to overly unsharp mask which generates shadow artifacts and noise. The optical images produced using our optimized optics (right) are much sharper and the ISP output has less artifacts. Displayed optics PSFs have been resampled for the sensor array. We show the PSFs along the first half of the main diagonal of the $1200 \times 1920$ sensor. The center pixel coordinates (row, column) of the spatial PSFs are indicated above each PSF, where (0, 0) refers to the top-left corner.

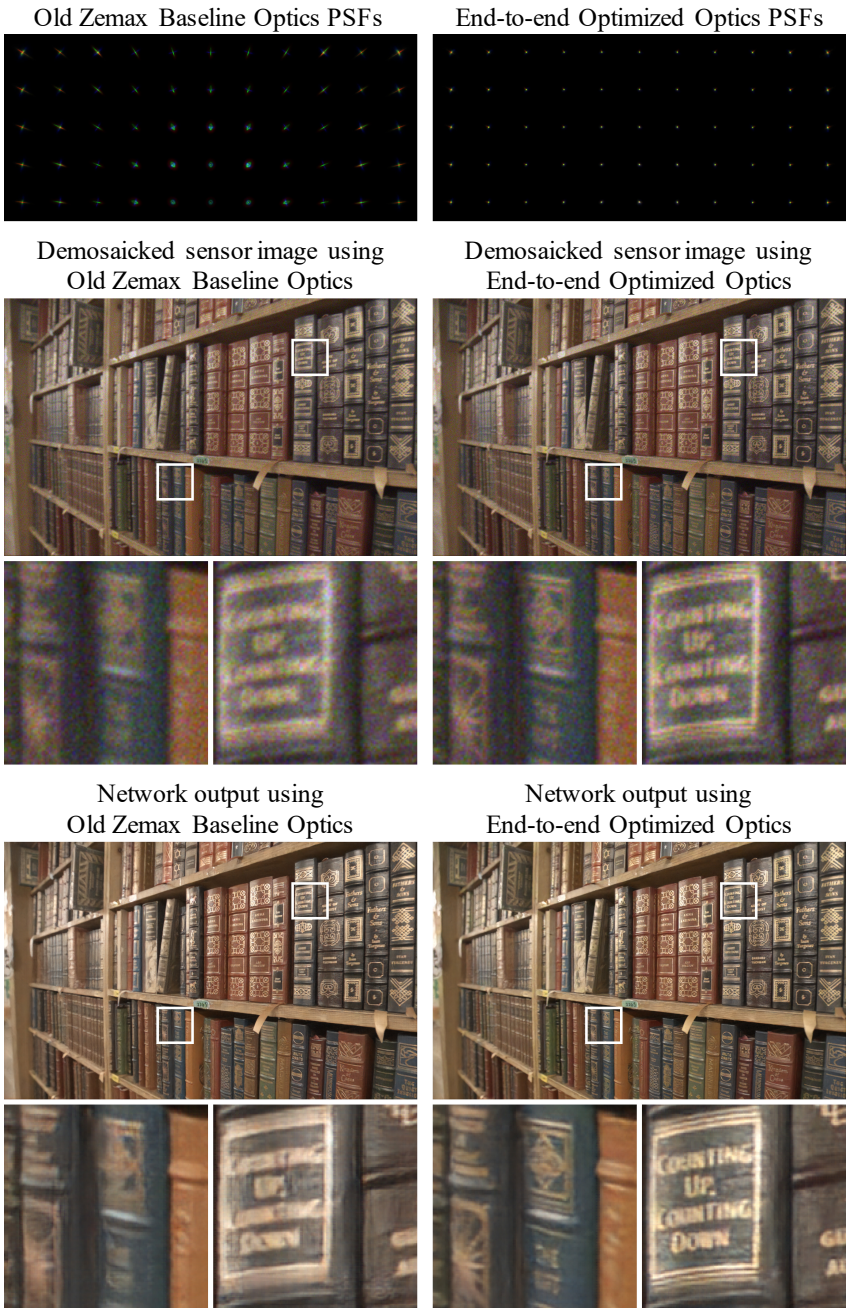



Fig. 14. Low-light imaging with Cooke triplet considering tight manufacturing constraints and software ISP using simulated measurements. Similar to Fig. 9 in the main document, our optimization produces an optic with compact spatial PSFs that assist the neural network ISP in recovering low-light image content. The demosaicked sensor image is shown for both approaches to highlight the noise level in addition to the optical aberrations.




Fig. 15. Automotive detection with Cooke triplet considering tight manufacturing constraints and hardware ISP using simulated measurements. Our joint end-to-end training improves the automotive object detection pipeline over the pipeline using the old ZEMAX baseline lens. Note that although the manufactured optimized lenses have greater light efficiency, the simulated optical images have been adjusted to the same brightness to better display the optical aberrations.




Fig. 16. Real-world prototype captures for automotive detection with Cooke triplet considering tight manufacturing constraints and hardware ISP. Similar to Fig. 10 in the main document, the manufactured prototypes are tested in the wild and demonstrate that our optimization allows for higher accuracy object and traffic light detection and classification over the old Zemax baseline optics pipeline. Note that our traffic light detector is trained to recognize vehicle traffic lights and ignores pedestrian traffic lights.

# 8 ANALYSIS AND SYNTHETIC VALIDATION

## 8.1 Cooke triplet optimization

In this section we present additional qualitative results for the experiments in Sec. 6 of the main document. Fig. 17 shows qualitative results for the "Image quality with software ISP" experiment. Additional results for the "Image quality with hardware ISP" optimization experiment are shown in Fig. 18. More results for end-to-end triplet lens design optimization for "Single-Image Low-light Imaging" are presented in Fig. 19. Qualitative results for "Automotive object detection and traffic light state detection with hardware ISP" using the end-to-end optimized triplet lens are shown in Fig. 20. As described in Sec. 6.1 of the main document, our end-to-end optimization improves object detection performance by maintaining uniform blur across the fields and by improved light efficiency. The optic optimized for traffic light detection is similar by having greater light efficiency while being slightly sharper in the center and periphery to better acquire small traffic lights.

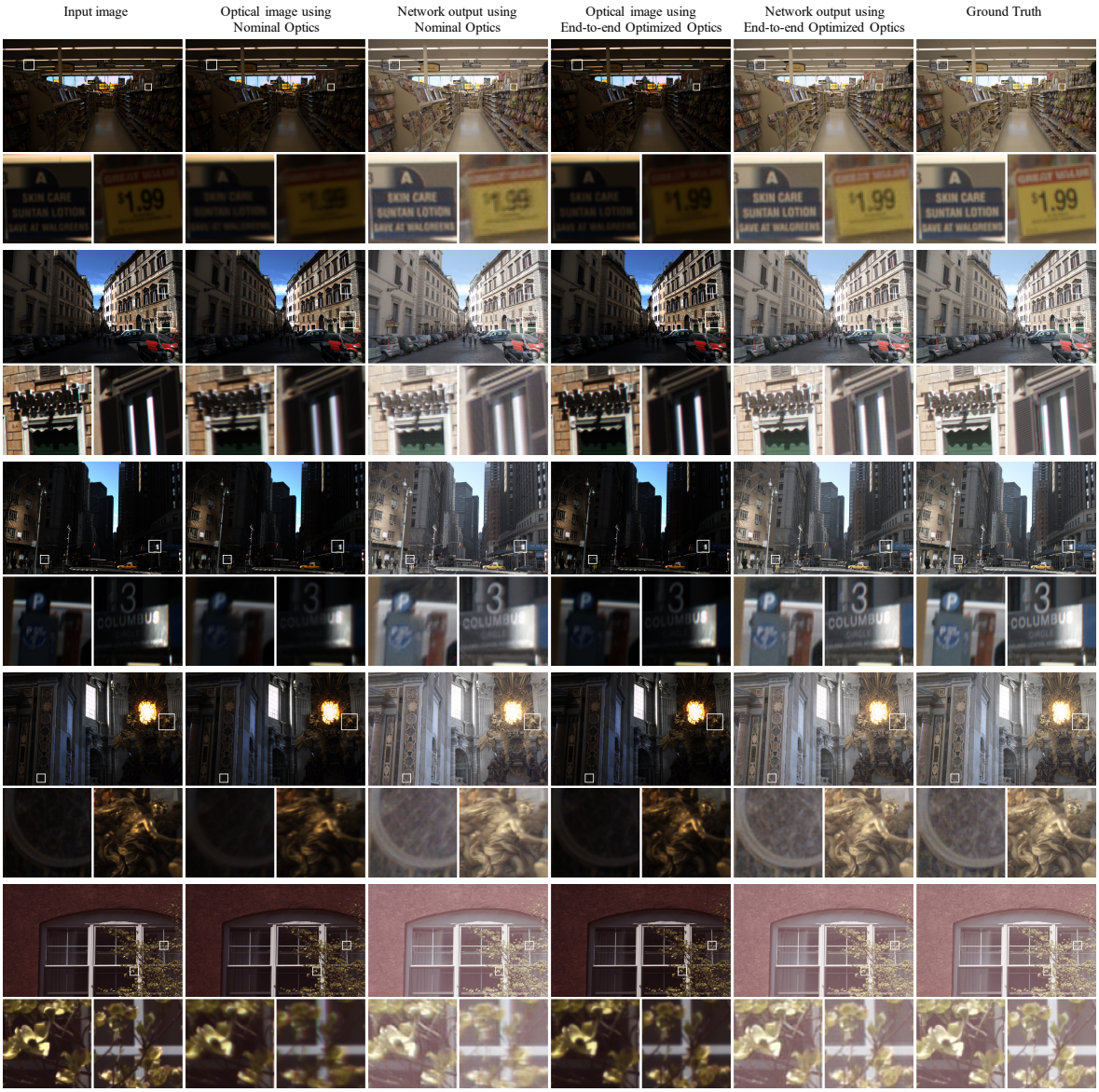



Fig. 17. Image quality with Cooke triplet and software ISP using simulated measurements. These are additional qualitative results for Sec. 6.1 of the main document.

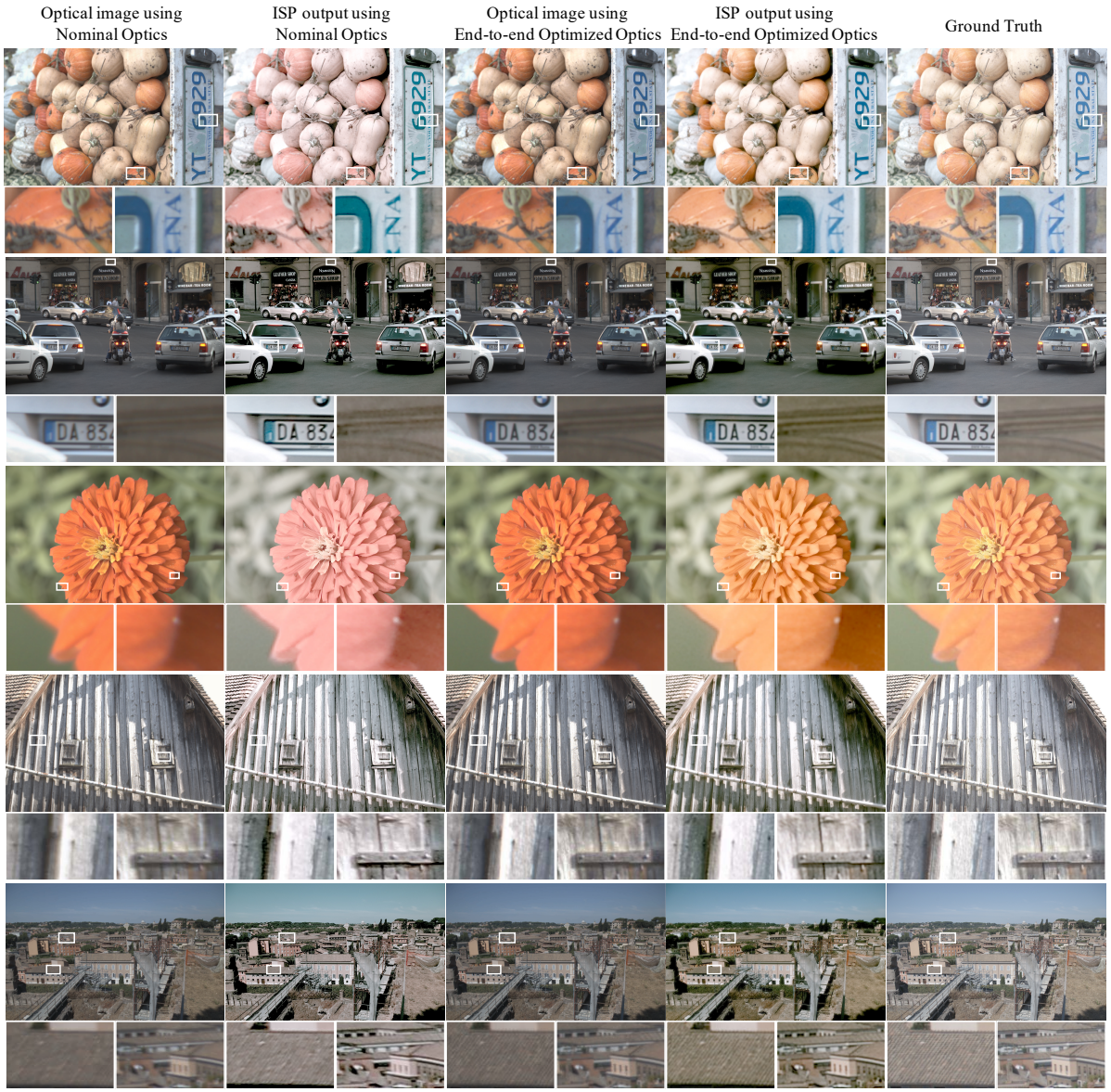



Fig. 18. Image quality with Cooke triplet and hardware ISP using simulated measurements. These are additional qualitative results for Sec. 6.1 of the main document.

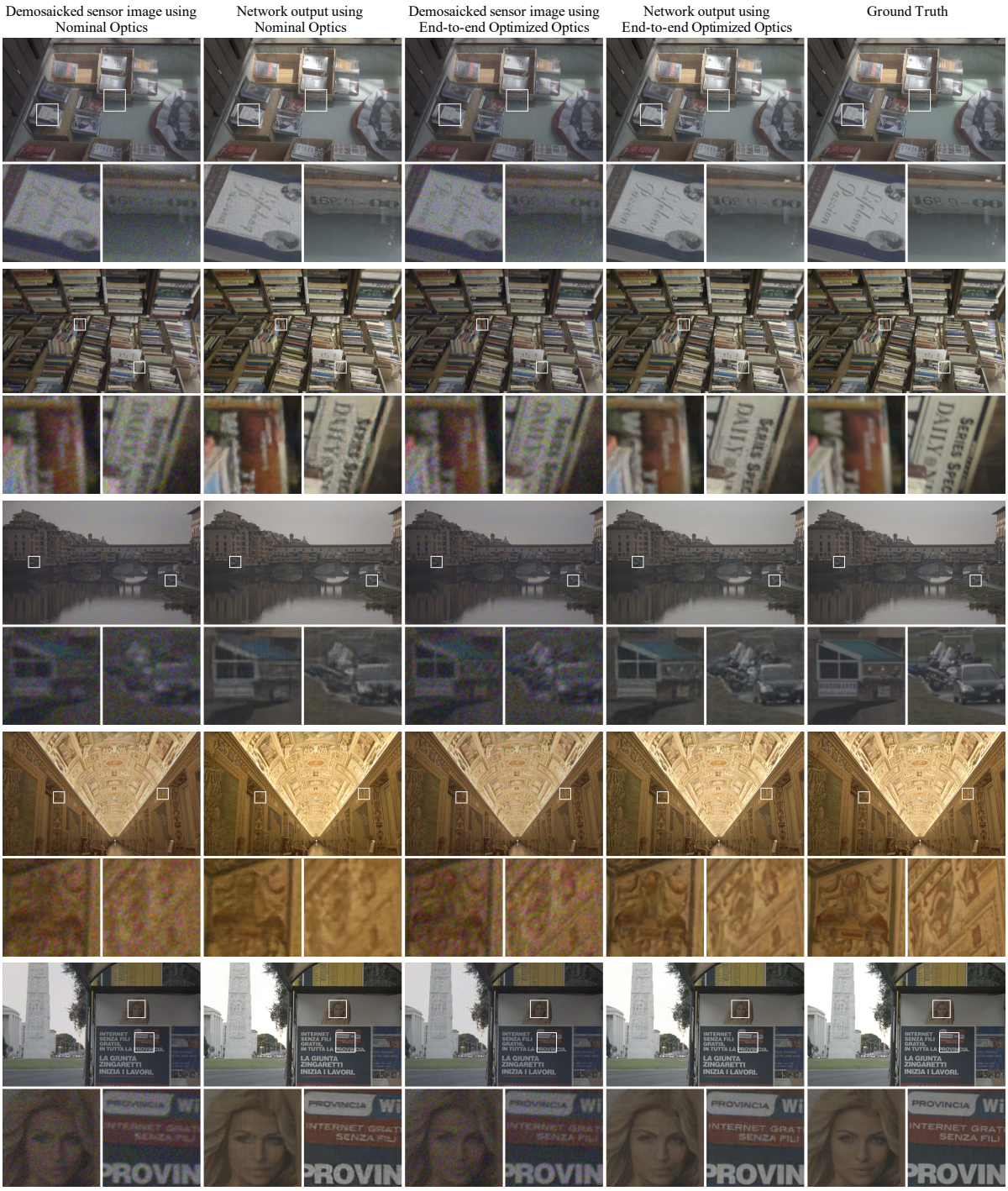



Fig. 19. Single-image low-light imaging with Cooke triplet and software ISP using simulated measurements. These are additional qualitative results for Sec. 6.1 of the main document. The demosaicked sensor image is shown to highlight the noise level in addition to the remaining optical aberrations.

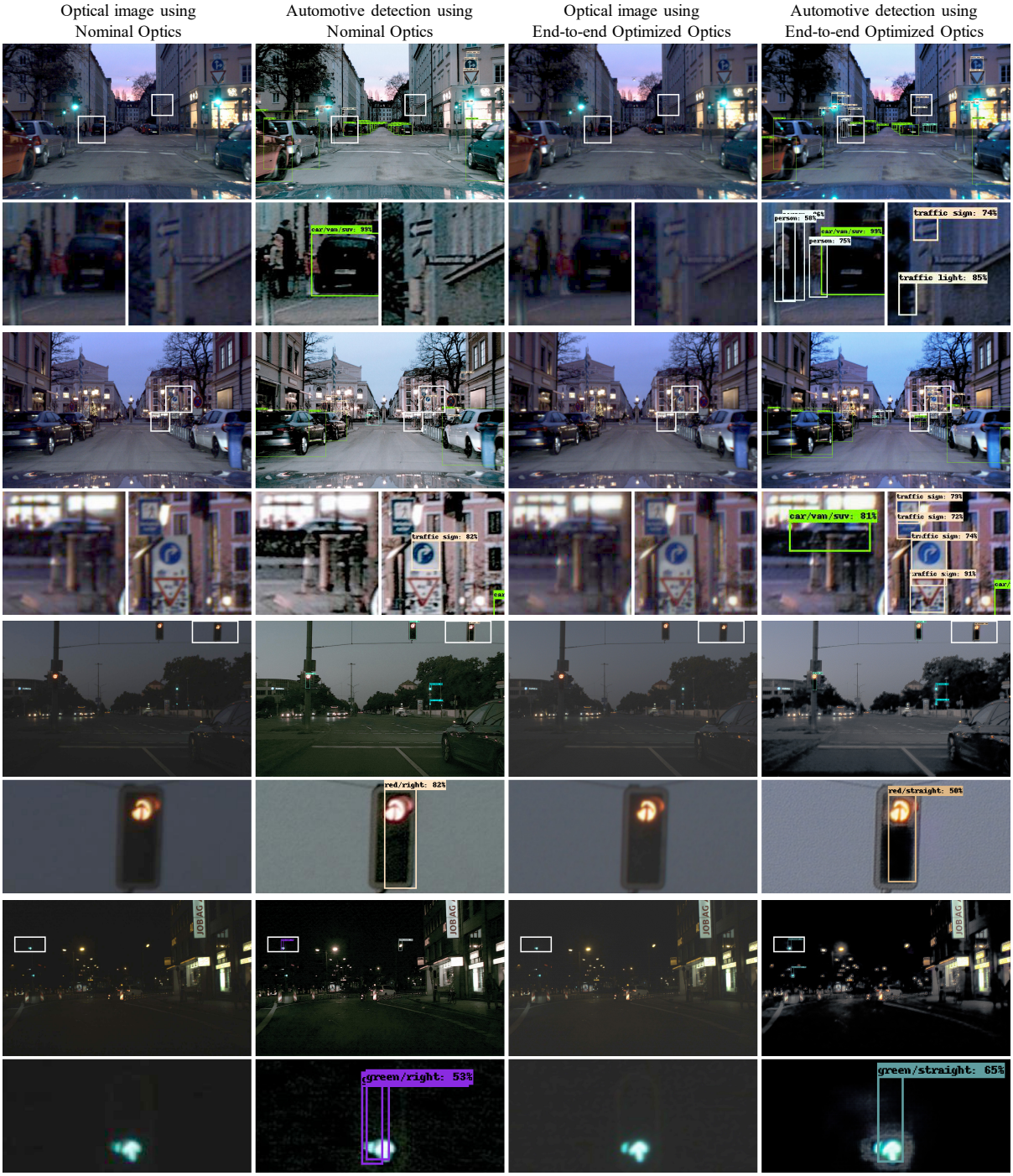
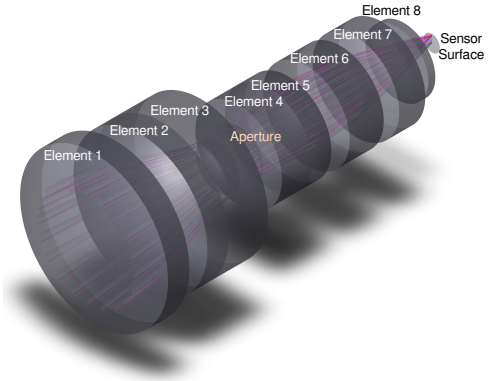



Fig. 20. Automotive detection with Cooke triplet and hardware ISP using simulated measurements. These are qualitative results for Sec. 6.1 of the main document. Our joint end-to-end training improves each block of the detection pipeline: the optical image is more light efficient, the hardware ISP has enhanced contrast and reduced noise, and the object detector has improved accuracy. Note that the simulated optical images have been adjusted to the same brightness to better display the optical aberrations.

## 8.2 Eight-element achromat experiments

The applicability of our method to more complex compound lenses is presented in Sec. 6.2 of the main document. The lens system used in this experiment has eight elements whose parameters are listed in Tab. 11. We optimized this lens design with respect to the 24 parameters for image quality jointly together with the parameters of the ARM Mali-C71 ISP. Our optimization converged towards compact PSFs that match and even slightly reduces the spotsize compared to the nominal PSFs (see PSFs in Fig. 11 in the main document), which demonstrates that the proposed method can handle very complex compound lenses in addition to Cooke triplets. Note that our method was initialized with a random guess, and does not require multiple man-weeks of design iterations that were necessary for the eight-element nominal design. Additional qualitative results are shown in Fig. 21.

Table 11. Schematic and parameters for the eight-element achromatic lens.




| Parameter | Description |
|---|---|
| s_1_radius | radius of the 1st surface |
| s_1_thickness | thickness of the 1st surface |
| s_2_radius | radius of the 2nd surface |
| s_2_thickness | thickness of the 2nd surface |
| s_3_radius | radius of the 3rd surface |
| s_3_thickness | thickness of the 3rd surface |
| s_4_radius | radius of the 4th surface |
| s_4_thickness | thickness of the 4th surface |
| s_5_radius | radius of the 5th surface |
| s_5_thickness | thickness of the 5th surface |
| s_7_radius | radius of the 7th surface |
| s_7_thickness | thickness of the 7th surface |
| s_8_radius | radius of the 8th surface |
| s_8_thickness | thickness of the 8th surface |
| s_9_radius | radius of the 9th surface |
| s_9_thickness | thickness of the 9th surface |
| s_10_radius | radius of the 10th surface |
| s_10_thickness | thickness of the 10th surface |
| s_11_radius | radius of the 11th surface |
| s_11_thickness | thickness of the 11th surface |
| s_12_radius | radius of the 12th surface |
| s_12_thickness | thickness of the 12th surface |
| s_13_radius | radius of the 13th surface |
| s_13_thickness | thickness of the 13th surface |

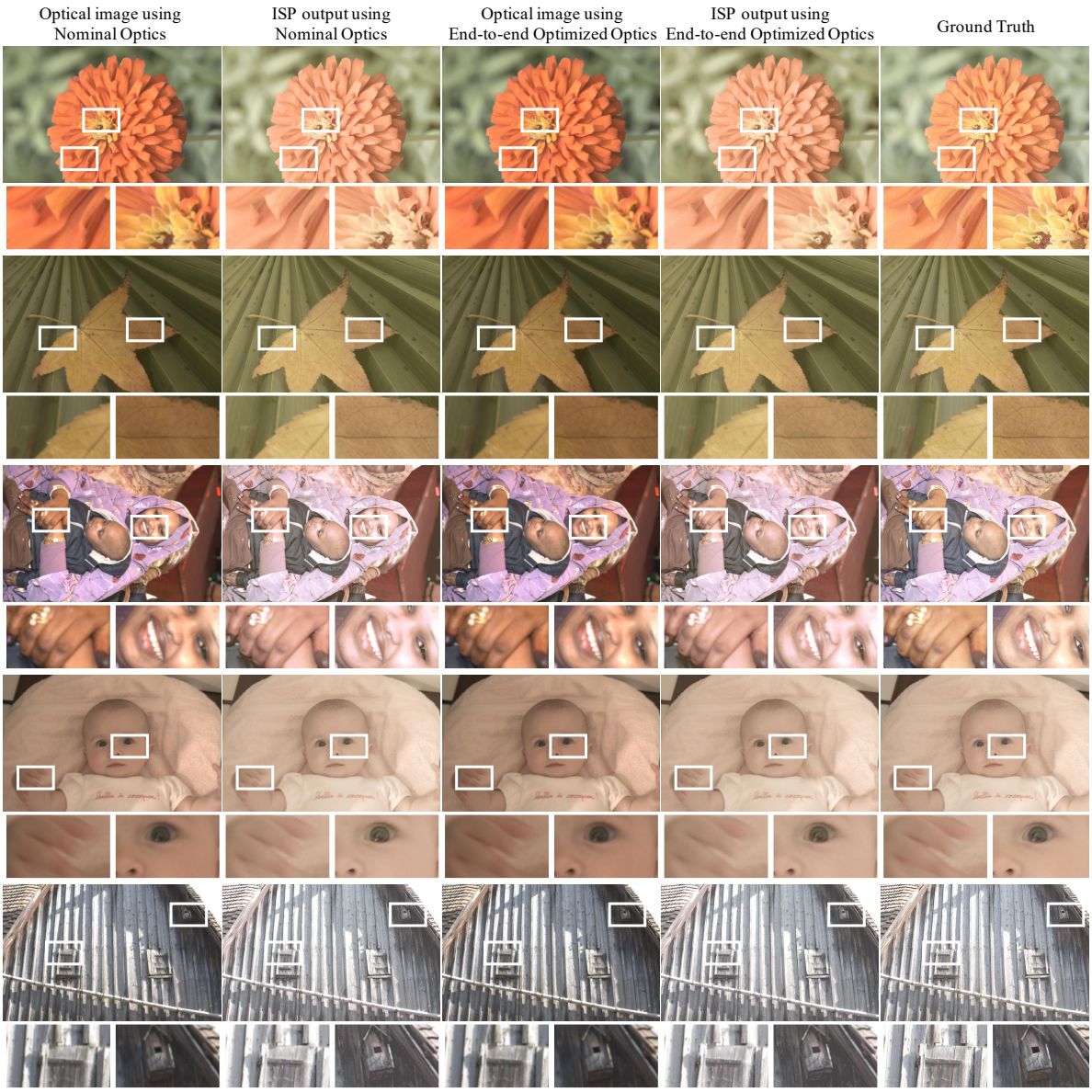



Fig. 21. Image quality with 8-element achromat compound lens and hardware ISP using simulated measurements. These are additional qualitative results for Sec. 6.2 of the main document. The many optical surfaces allow both the nominal and end-to-end optimized optics to remove almost all optical aberrations. Our optimization's ability to match and even slightly improve upon the nominal optic demonstrates that the proposed method generalizes beyond Cooke triplets.

# 9 EXPERIMENTAL VALIDATION

## 9.1 Lens Prototype Manufacturing

Additional captures and qualitative results using the manufactured lenses described in Sec. 7 of the main document are presented in this section. As was done in the main document, we compare the manufactured nominal lens against our optimized lenses for three different tasks: "Image quality with hardware ISP", "Automotive object detection (OD)", "Traffic light state detection (TL)". For all tasks, the manufactured nominal lens and the task-specific optimized lenses are used in a synchronized dual camera setup as shown in Fig. 1 of the main document.

Our optimized lenses demonstrate improved performance across all tasks. For the task of image quality, our optimized optic significantly reduces the aberrations in the periphery while exhibiting similar performance as the nominal optic in the center, see qualitative results in Fig. 22. Our end-to-end imaging pipeline thus produces higher quality images with fewer aberrations across the field of view.

For automotive pedestrian-vehicle object detection, the optimized optic enhances our object detection pipeline in two ways. First, the smaller f-number (f/3.2) of our optimized optic leads to improved light efficiency compared to the nominal lens (f/4.4). In low light regions, the lower light efficiency of the nominal optic results low signal that the hardware ISP is not capable of separating from the sensor noise, which in turn hampers object detection performance. The resulting boosted noise can be seen in Fig. 23. Secondly, our optic maintains uniform blur across the field of view, and this uniformity assists our downstream convolutional object detector. Unlike the optic optimized for image quality, some slight blur is acceptable for the task of object detection and can even enhance performance if the blur is uniform. Refer to Fig. 23 for qualitative results. Note that the nominal detector and ISP, for fairness, were fine-tuned with fixed nominal optics.

For traffic light detection, the optimized optic is similar to the object detection lens with a smaller f-number (f/3.3) for higher light efficiency compared to the slower nominal lens (f/4.4). However, the TL optic is sharper than the optic in the center and exhibits a slightly stronger peak component in the peripheries in order to better capture small traffic lights. These tailored aberrations and their positive impact on traffic light detection performance is validated in Fig. 23. Note again that the nominal detector and ISP, for fairness, were fine-tuned with fixed nominal optics.

For further visual comparison, also performed qualitative comparisons against highly corrected multi-element lenses. For this comparison experiment, we equip a synchronized FLIR BFLY-23S6C-C with a high-quality highly aberration-corrected Fujinon CF12.5HA lens (which comes at 3 times the size and weight compared to the proposed optics). Figs. 24 and 25 show qualitative results for synchronized captures for end-to-end optimized lenses.




Fig. 22. Real-world prototype captures for image quality with Cooke triplet and hardware ISP. The image pipeline using the manufactured optimized optic produces sharper results in the periphery and similar quality results in the center.




Fig. 23. Real-world prototype captures for automotive object and traffic light detection with Cooke triplet and hardware ISP. The optimized optics have greater light efficiency (smaller f-number) and more uniform blur across the field of view than the nominal optic, which leads to greater detection performance. Note that the traffic light detection captures were taken at dusk.




Fig. 24. High-quality scene captures using Fujinon high quality optics and the Cooke triplet prototype lens optimized for automotive object detection.




Fig. 25. High-quality scene captures using Fujinon high quality optics and the Cooke triplet prototype lens optimized for traffic light detection.

## 10 OPTICAL PROPERTIES OF OPTIMIZED LENS DESIGNS

Our novel end-to-end optimization process yielded six distinct lens designs for the four different applications and two sensor sizes. Tab. 12 lists the final optimized lens prescriptions and Tab. 13 lists the general optical properties of the lenses. Lens design layouts are also shown in Fig. 26. Interestingly, no optimized design is close to the nominal design in terms of focal length and f-number (Tab. 12). While the lenses for pedestrian-vehicle detection and traffic light detection favor shorter focal lengths (18.5 mm and 18.8 mm respectively vs. 25 mm) and a faster aperture (3.2 and 3.3, vs. 4.4), the remaining four designs show longer focal lengths between 32.8 mm and 35.4 mm and slower aperture (5.8 to 6.2 vs. 4.4). Because the back focal length of the first two optimized lenses is slightly larger the total track length for all optimized designs is larger than the nominal design (52.2 mm to 60.3 mm vs. 46.6 mm).

The optimization results are intuitive. For object detection in general (both pedestrian-vehicle and traffic light state), shorter focal lengths and faster lenses are likely preferable due to higher contrast (more light being captured for a given physical aperture). The optimization process naturally finds this solution due to the distribution of objects in the visual field. For pedestrian-vehicle detections, objects can uniformly appear anywhere in the visual field, whereas for traffic lights, they occur primarily near the center of the visual field (and in the upper periphery, when close) due to overhanging traffic lights that are visible from far away and that remain longer in the visual field. The downside of shorter focal length is higher distortion near the periphery of the image, but this is acceptable as long as the objects are recognizable. For perceived image quality, the distortion needs to be minimal across the entire field of view. This results in the image quality optimized optics having a longer focal length.

Table 12. Optimized Cooke triplet parameters for the nominal optic and each of the experiments.

| Parameter | Min | Max | Nominal Design | Pedestrian-Vehicle Detection | Traffic Light Detection | Improved Image Quality with Hardware ISP | Improved Image Quality with Software ISP | Improved Image Quality for Larger FOV with Hardware ISP | Improved Image Quality for Larger FOV with Software ISP |
|---|---|---|---|---|---|---|---|---|---|
| s_1_radius | 9.85 | 14.98 | 11.34 | 13.40 | 12.78 | 11.63 | 11.58 | 11.52 | 11.72 |
| s_1_conic | -0.49 | 0.29 | 0.00 | -0.09 | -0.10 | -0.05 | -0.03 | -0.09 | -0.11 |
| s_2_radius | 9.45 | 14.82 | 11.45 | 12.31 | 12.59 | 11.07 | 9.89 | 9.49 | 9.45 |
| l_12 | 4.58 | 10.11 | 8.11 | 6.03 | 6.18 | 8.36 | 7.27 | 6.90 | 8.06 |
| l_2STO | 1.03 | 9.22 | 3.53 | 5.56 | 5.42 | 2.26 | 1.04 | 1.22 | 1.17 |
| l_STO3 | 0.0 | 9.86 | 9.99 | 5.24 | 5.73 | 1.51 | 1.75 | 3.14 | 2.64 |
| s_6_radius | 13.38 | 18.17 | 15.38 | 16.30 | 16.01 | 15.60 | 15.88 | 17.11 | 16.47 |
| s_6_conic | -0.49 | 0.49 | 0.00 | 0.00 | 0.00 | 0.32 | 0.37 | 0.49 | 0.47 |
| s_7_radius | -15.05 | -11.50 | -13.05 | -13.10 | -13.43 | -13.88 | -12.69 | -13.44 | -13.50 |
| s_1_2nd | -4.99e-3 | 4.99e-3 | 0.00 | -0.01e-3 | -0.01e-3 | -2.41e-3 | -0.54e-3 | 0.36e-3 | 0.08e-3 |
| s_1_4th | -9.06e-5 | -1.45e-5 | -4.06e-5 | -4.67e-5 | -4.67e-5 | -2.96e-5 | -3.66e-5 | -1.45e-5 | -1.46e-5 |
| s_1_6th | -5.10e-7 | 6.30e-7 | 1.30e-7 | -0.10e-7 | -0.10e-7 | -2.37e-7 | 0.71e-7 | -1.76e-7 | -5.06e-7 |
| s_1_8th | -1.58e-8 | -2.66e-11 | -5.02e-9 | -8.41e-9 | -8.33e-9 | -2.32e-9 | -4.14e-9 | -2.66e-11 | -2.66e-11 |
| s_1_10th | -1.28e-10 | 1.08e-10 | -0.28e-10 | 0.09e-10 | 0.09e-10 | -1.28e-10 | -1.28e-10 | 1.08e-10 | 1.08-10 |
| s_6_2nd | -4.99e-3 | 4.99e-3 | 0.00 | 0.01e-3 | 0.02e-3 | -4.62e-3 | -3.82e-3 | -4.98e-3 | -4.86e-3 |
| s_6_4th | -2.57e-4 | -1.41e-4 | -2.07e-4 | -1.91e-4 | -1.91e-4 | -1.70e-4 | -1.75e-4 | -1.41e-4 | -1.45e-4 |
| s_6_6th | -1.44e-6 | 1.73e-6 | -0.44e-6 | 0.72e-6 | 0.72e-6 | 0.29e-6 | 0.01e-6 | -1.42e-6 | -1.44e-6 |
| s_6_8th | -4.54e-8 | 4.92e-7 | 4.42e-7 | 0.04e-7 | 0.04e-7 | 4.69e-7 | 3.502e-7 | 4.92e-7 | 4.49e-7 |
| s_6_10th | -2.33e-8 | 8.80e-10 | -2.23e-8 | 0.00 | -0.35e-10 | -2.33e-8 | -1.75e-8 | -2.28e-8 | -2.17e-8 |

Table 13. Optical properties of the seven three-element Cooke triplet designs.

| Optical Property | Nominal Design | Pedestrian-Vehicle Detection | Traffic Light Detection | Improved Image Quality with Hardware ISP | Improved Image Quality with Software ISP | Improved Image Quality for Larger FOV with Hardware ISP | Improved Image Quality for Larger FOV with Software ISP |
|---|---|---|---|---|---|---|---|
| Focal Length (mm) | 25.0 | 18.5 | 18.8 | 33.1 | 32.8 | 35.4 | 32.8 |
| f/# | 4.4 | 3.2 | 3.3 | 5.8 | 5.8 | 6.2 | 5.8 |
| Back Focal Length (mm) | 24.4 | 24.9 | 24.1 | 35.0 | 34.6 | 39.8 | 34.6 |
| Total Track Length (mm) | 46.6 | 52.2 | 52.3 | 56.4 | 53.8 | 60.3 | 53.8 |

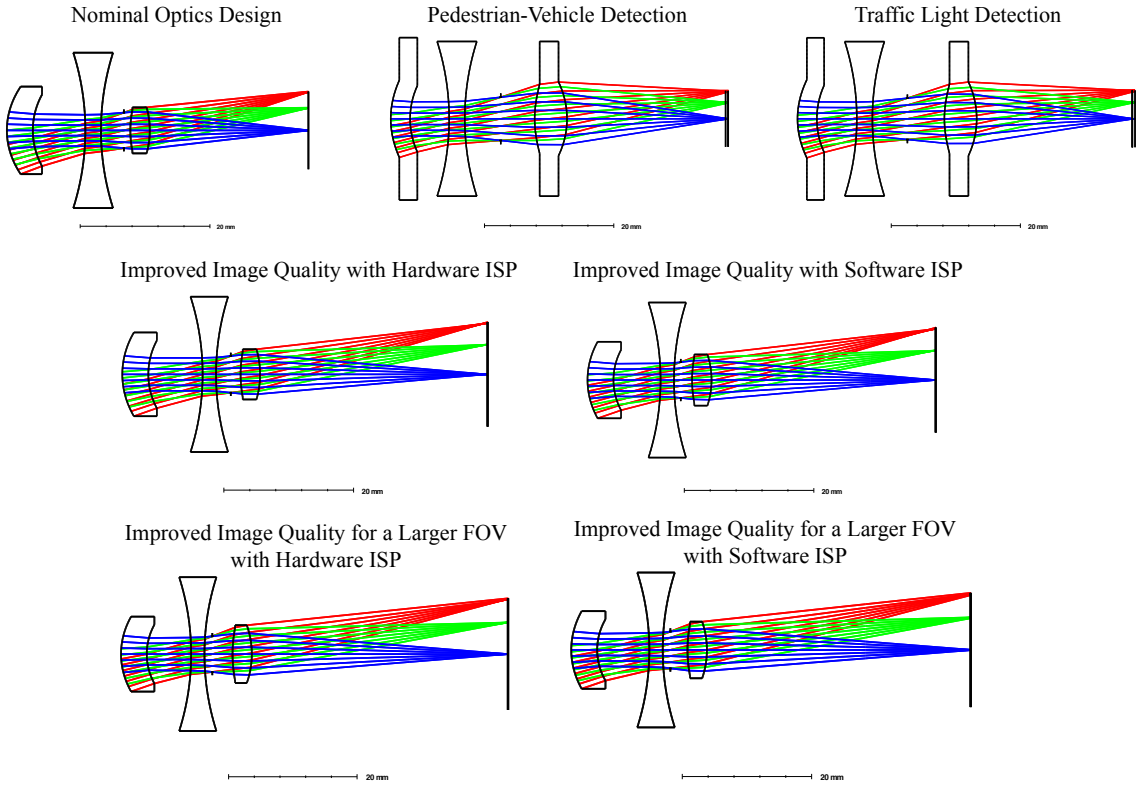



Fig. 26. Design layout of the seven three-element Cooke triplet designs.

Modulation Transfer Function (MTF) is the contrast at a given spatial frequency, where the contrast is determined using a sinusoidal brightness pattern specified in line pairs per mm (lp/mm). In the following analysis we assume a sensor with pixel size of 5.86 µm with Nyquist frequency of 85.32 lp/mm, i.e, $f_{\text{Nyquist}} = 10^3/(2 \cdot 5.86\,\mu\text{m})$. The contrast is normalized to 1, and the different colors refer to the different field positions in the image, which are 0°, 5°, 10°, and 15° for the actual sensor size, and 0°, 5°, 15°, and 25° for the larger field of view (FOV) experiments. The dashed and solid lines indicate the tangential and sagittal orientations of the observed structure.

In the following, we first consider the MTF curves for the actual sensor size in Fig. 27 and 28. The first observation is the quality of the nominal Cooke triplet, as seen in both the tight spot diagrams and the MTF curves. This is the result of several man weeks of optimization using multiple optics design software packages, which is a typical approach even for a lens design of this relatively low complexity. End-to-end designs take into account the stages following the optics which can adjust for different optical aberrations and optimizes for task-specific objectives, such as, detection accuracy and image quality. Hence the final optics can score higher in task-specific metrics (see Tabs. 6 and 7 in the main document) even without ranking high in all traditional optical quality metrics that only consider the optics in isolation from the camera downstream task. Interestingly, the sagittal 5° MTF curves for both hardware and software ISP experiments are distinctly higher than the nominal design and also their own respective center sharpness, pointing to a sensitivity of the image quality end-to-end loss function to radial sharpness. An interesting difference between the end-to-end optimized lenses with software

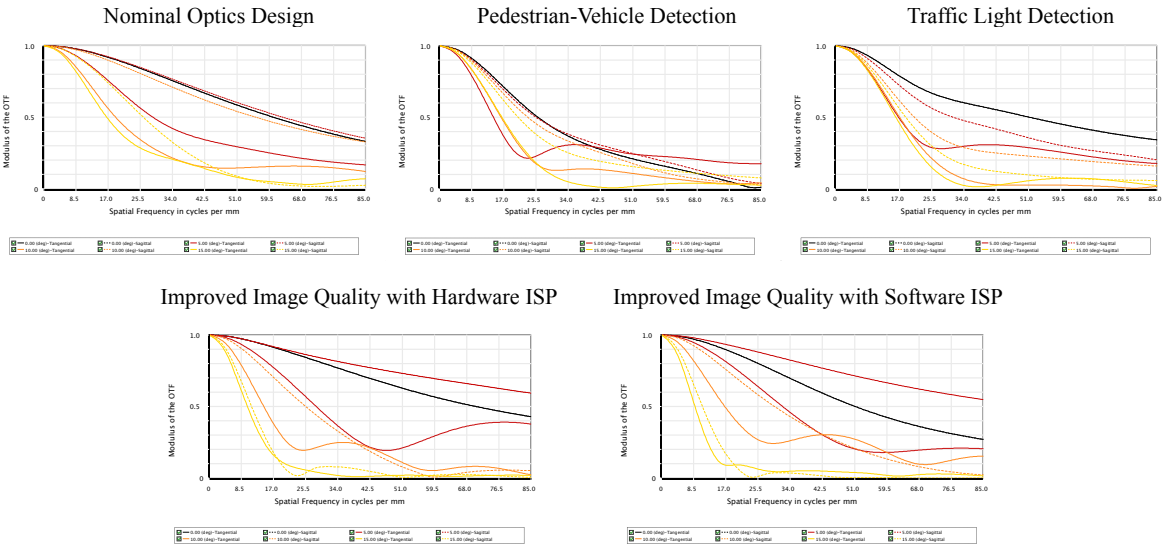



Fig. 27. MTF analyses of the five three-element Cooke triplet designs optimized w.r.t the actual sensor size.

ISP and hardware ISP is the ability of the software ISP to learn a sharpening deconvolution, whereas the hardware ISP can only sharpen the image with an approximately Gaussian kernel. This allows the software ISP to accept less sharpness from the optical system for the center and middle part of the image, because it can rely on its own deconvolution capabilities independently of PSF shape. Here, the hardware ISP has a distinct MTF bump for higher frequencies around 75-80 lp/mm (compare 0° and 5° curves) to achieve the sharpness required for good perceptual image quality, as it cannot compensate as well with the inherently non-optical Gaussian kernel.

For the pedestrian-vehicle lens, typical projection size of objects of interest is at least 20 pixels (e.g., 1 m object at 157 m away will project to 20 pixels for 18.5 mm focal length lens) and as a result fine-grained details are less important than the amount of captured light. The same is true for traffic lights detection where the direction arrows are typically at a higher contrast than other background objects. Moreover, as overhanging traffic lights appear more often near the center of the visual field and in the upper image periphery, the resulting PSFs are sharper in the center and for 5° FOV and slightly improve on the nominal design for high frequencies at the far periphery for 15°, whereas the pedestrian-vehicle detection trades in sharpness in the center to retain a minimum viable sharpness over the whole field. All these results exhibit the influence of the task-specific end-to-end loss function on the design of the optics.

While the MTF curves of the nominal design is almost monotonically decreasing and non-intersecting, the curves for the end-to-end optimized designs exhibit field dependent non-monotonic variations. This is primarily due to the task-specific frequency sensitivity requirements and also taking into account the components such as, hardware/software ISP and task-specific networks that follow the optics model.

The larger FOV optimized optics demonstrate similar behavior as shown in Figs. 29 and 30. These experiments were performed for the image quality metric only. Note that the scale of the spot diagrams were increased to 400 µm from 200 µm for Fig. 28 to accommodate the larger spot sizes for 25°. Nonetheless, in comparison to the spot sizes of the actual sensor size in Fig. 28 the optimization manages to keep the spot sizes in check for both 15° and 25°, best seen for the lens design optimized for improved image quality with software ISP.

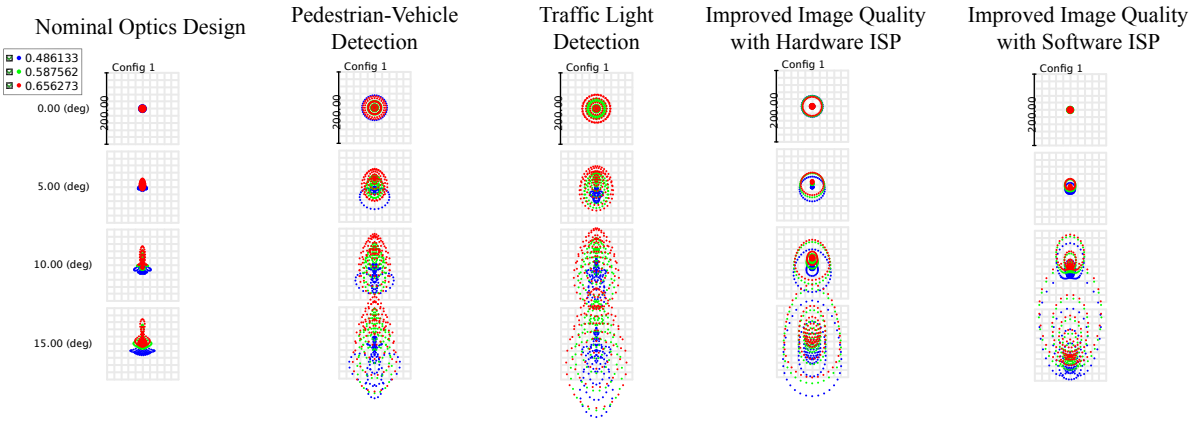



Fig. 28. Spot diagram analyses of the five three-element Cooke triplet designs optimized w.r.t the actual sensor size.

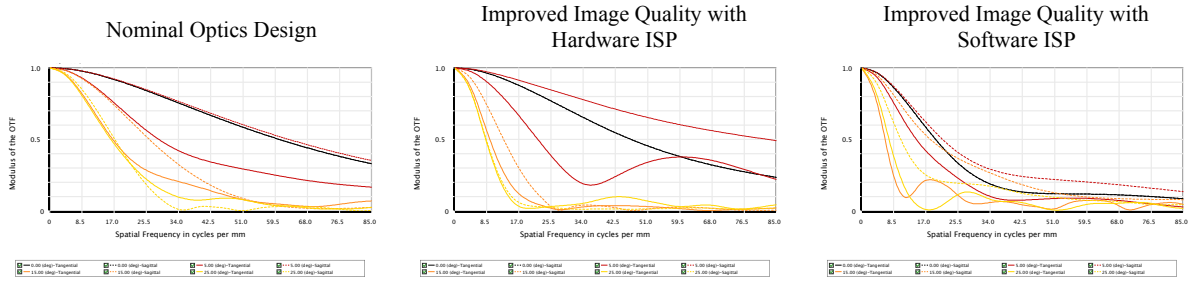



Fig. 29. MTF analyses of the three three-element Cooke triplet designs optimized for a larger field of view.

The requirement of the larger FOV leads to a distinct MTF performance loss for the designs optimized using the hardware ISP and especially the software ISP as the end-to-end optimization deals with the highly difficult task of keeping a sharp image over the entire field for a basic lens design – the Cooke triplet – that is not well suited for larger FOV. Both the MTF curves and the spot diagrams again show the excellent nominal design, which tackles this task very well with MTF values above 20% at 25 lp/mm, even for the 25° field. The optimized lens for the hardware ISP is instead optimized for the center because the hardware ISP is not able to pull high frequency information from low MTF values at higher fields. The optimized lens with software ISP retains higher MTF values for higher frequencies even at the peripheral fields, however, these higher peak MTF values come at the cost of oscillations, with a few valleys where the MTF values become very small. The learned software ISP is able to recover images even with these lost frequency components in a learned deconvolution process, as long as other frequencies are present. Again, the important aspect to note is the influence of the end-to-end design process with its fundamentally different loss function, which is also demonstrated in these examples.

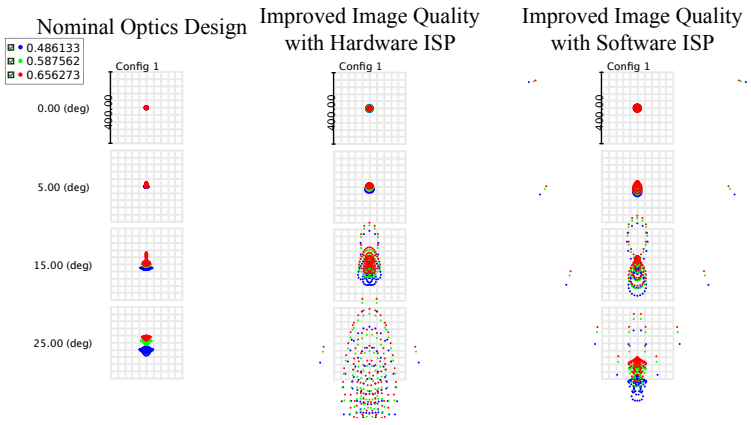



Fig. 30. Spot diagram analyses of the three three-element Cooke triplet designs optimized for a larger field of view.

## 11 PSF CALIBRATION

For every lens, the PSF was measured by imaging a pin-hole light source on the sensor. Fig. 31 shows the full setup for measuring the PSF: on the left is the pin-hole light source and on the right is the camera affixed to a three-axis rotation tripod head. Unfortunately, the Trioptics ImageMaster setup used for the PSF validation captures in Fig. 6 of the main document was not available for this calibration task, resulting in the setup described in the following, which requires careful alignment and demosaicking due to sensor availability.

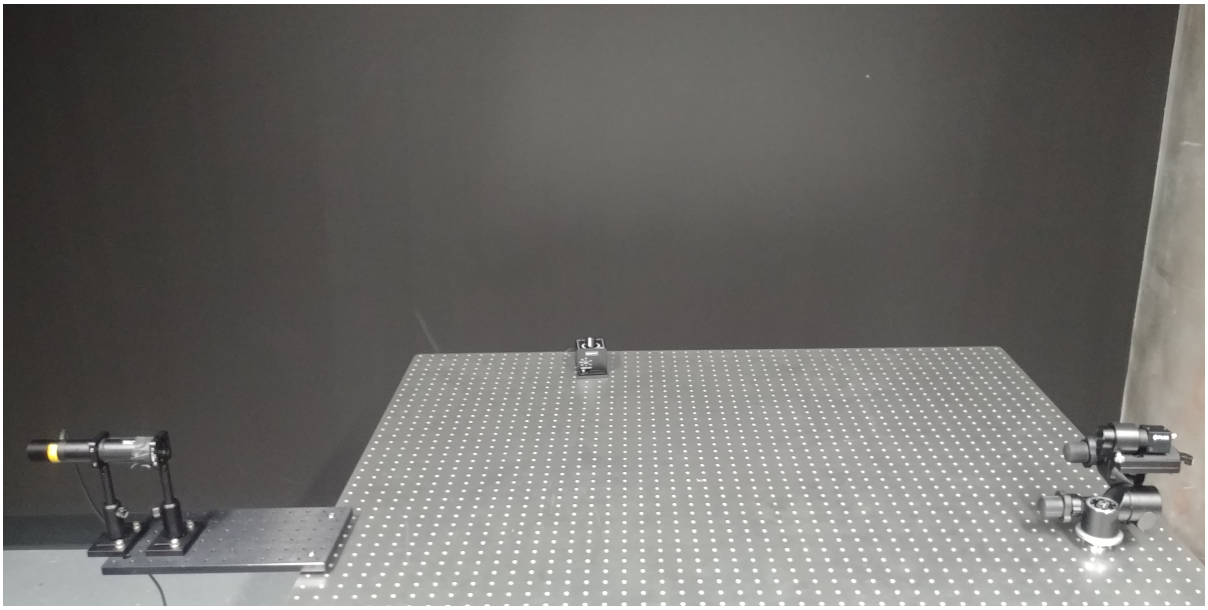


Fig. 31. PSF calibration full setup. The pin-hole light source (left) is captured by the camera (right).

### 11.1 Setup

The pin-hole light source presented in Fig. 32 is composed of a controllable white light emitting diode (4900K) (Thorlabs) and a pin-hole of diameter 75 µm (Thorlabs). The light diode is focused directly on the pin-hole using a converging lens.

The camera is placed at ≈1.5 m from the pin-hole on a three-axis rotation tripod head as shown in Fig. 33. Using an object distance of 1.5 m and an average focal length of 20 mm for the lens, the pin-hole is imaged onto the sensor with a size of ≈1 µm which is below the 5.6 µm pixel size of the camera.

### 11.2 Methodology

All light sources in the laboratory were turned off except the pin-hole light source. The camera was positioned to align the pin-hole light source to the center of the sensor. Acquisitions with the camera were performed with a custom control script displaying a video feed.

Initially, the theoretical PSF was estimated using Zemax at a specific radial distance. Those distances were projected in the sensor space and overlaid on the video display feed to help align the PSF measurement with the theoretical estimation. The camera was aligned with each of the radial distances on the vertical, horizontal, and diagonal axes by moving the camera around the rotational axes of the tripod head.



Fig. 32. Pin-hole light source setup.

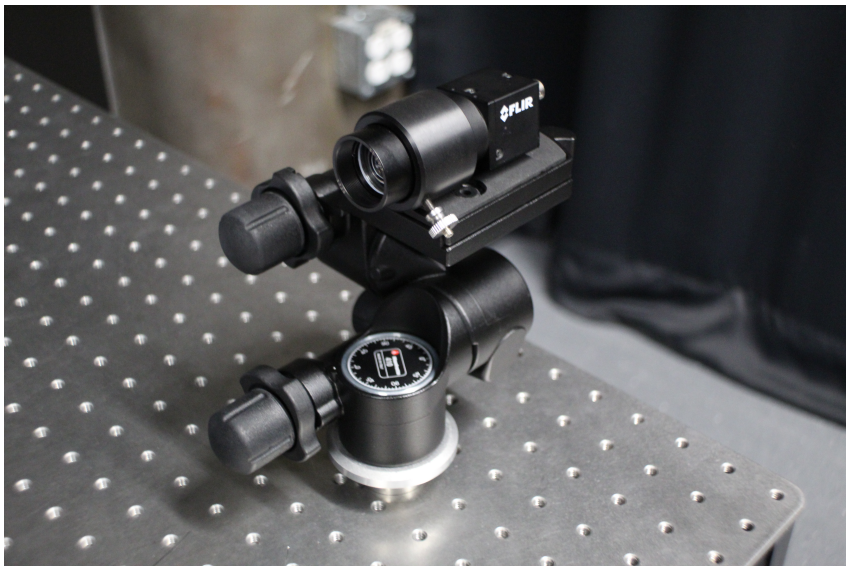



Fig. 33. Camera PSF calibration setup.

For every PSF measurement, 20 acquisitions were taken at 1, 4, and 16 ms exposure. The pin-hole light source's intensity was adjusted to avoid clipping at 16 ms exposure.

For every exposure the 20 acquisitions were averaged and the black level was subtracted. Then the 3 exposures were fed to a custom HDR stitching algorithm. The output was demosaicked to get the PSF measurement for the

R, G, and B channels. While this process is not perfect for small PSFs, we use bilinear demosaicking to minimize interpolation artifacts. For every spatial PSF measurement, the centroid was estimated and a neighborhood of $23 \times 23$ pixels around the centroid was kept as the PSF measurement output.

Fig. 34 shows calibrated lens PSFs along with the simulated PSFs for the three manufactured lens systems, visualized in linear and log scale. The simulated lens PSFs were obtained using ray-tracing in Zemax for 9 different fields. For each field the corresponding calibrated lens PSF was measured as explained above. Each calibrated PSF box in this figure corresponds to an area of $23 \times 23$ pixels on the 2.3 megapixel Sony IMX249 sensor with IR cut-off filter (specifications BFLY-U3-23S6C-C). The difference between the wavelengths of the white light emitting diode used as the pin-hole light source and those used in Zemax's ray-tracing simulation should be noted. Also, the calibrated PSFs are affected by the quantum efficiency of the sensor and the sampling on the color filter array. Hence, the visualizations of calibrated PSFs and the simulated ones are slightly different in their representative colors and shapes of fine structures (due to demosaicking). Nevertheless, these measurements provide additional qualitative validation of the aberrations from fabricated lenses in addition to the captured results provided in the main document and this supplemental document.

## REFERENCES

[1] EMVA 1288 2016. EMVA 1288 Standard for Characterization of Image Sensors and Cameras. https://www.emva.org/wp-content/uploads/EMVA1288-3.1a.pdf

[2] EMVA 1288 PyPI 2018. Reference Implementation of EMVA 1288 Standard for Characterization of Image Sensors and Cameras, Version 0.6.1. https://pypi.org/project/emva1288/

[3] FLIR Specification 14.0 2020. Imaging Performance Specification - FLIR Blackfly GigE Vision - Version 14.0 . https://flir.app.boxcn.net/s/et1u8d9q2u5wpouywowvdl0f5hd5i0c4/file/418608635264

[4] Kenneth Garrard, Thomas Bruegge, Jeff Hoffman, Thomas Dow, and Alex Sohn. 2005. Design tools for freeform optics. In *Current Developments in Lens Design and Optical Engineering VI*, Vol. 5874. International Society for Optics and Photonics, 58740A.

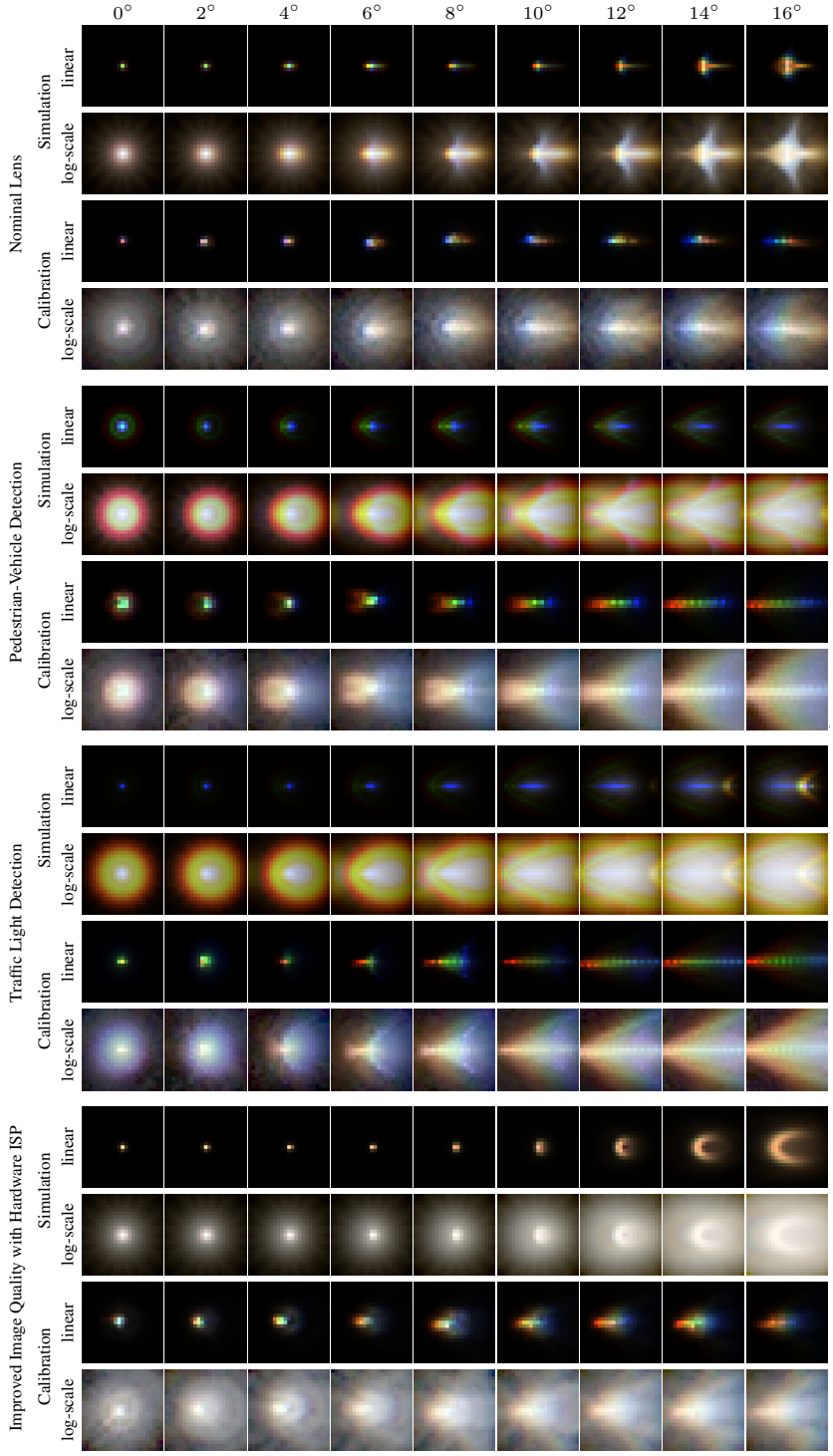



Fig. 34. PSF calibration for the manufactured nominal and end-to-end optimized Cooke triplet lens systems: automotive object detection, traffic light detection, and perceptual image quality with hardware ISP. For better visualization the PSFs are also shown in logarithmic scale.