IMAGE QUALITY IS NOT ALL YOU WANT: TASK-DRIVEN LENS DESIGN

Xinge Yang
KAUST
xinge.yang@kaust.edu.sa

Qiang Fu
KAUST
qiang.fu@kaust.edu.sa

Yunfeng Nie
Vrije Universiteit Brussel
yunfeng.nie@vub.be

Wolfgang Heidrich*
KAUST
wolfgang.heidrich@kaust.edu.sa

March 19, 2024

ABSTRACT

In computer vision, it has long been taken for granted that high-quality images obtained through well-designed camera lenses would lead to superior results. However, we find that this perception is not a "one-size-fits-all" solution and task-driven deep-learned simple optics can actually deliver better performance for computer vision applications. The task-driven lens design idea, which relies solely on a well-trained network model for supervision, is proven to be capable of designing lenses from scratch. The deep-learned lens characteristics are defined by high-level computer vision applications rather than conventional optical objectives. Experimental results on image classification demonstrate the task-driven lens ("TaskLens") exhibits higher accuracy than classical imaging-driven lenses, with even fewer elements. Furthermore, we show that our TaskLens is compatible with various network models while maintaining enhanced classification accuracy. We believe this task-driven idea holds significant potential for the next generation of optical design, particularly when the physical dimensions and cost of lenses are severely constrained.

1 Introduction

Modern deep networks have exhibited exceptional performance in various computer vision tasks, including image classification [1, 2, 3, 4], object detection [5, 6], semantic segmentation [7, 8], and depth estimation [9, 10, 11]. To fully exploit the feature extraction capabilities of deep networks, these works typically rely on sharp, high-quality input images, which are often captured using well-designed precise lenses. However, such lenses can be prohibitively expensive and complex. For example, modern cellphone lenses usually have over five highly aspheric elements [12, 13] and commercial camera lenses usually have six or more precise optical elements [14, 15].

End-to-End lens design [16, 17, 18, 19], an emerging field, simultaneously optimizes camera lenses and deep networks to maximize performance for specific applications. This approach has shown promising results in computational imaging areas such as extended-depth-of-field imaging [19, 17, 20, 21, 22], large-field-of-view imaging [23, 19, 20], hyperspectral imaging [24, 25, 26, 27], high-dynamic-range imaging [28, 29], as well as computer vision tasks such as object detection [30, 18] and depth estimation [31, 32].

However, current End-to-End lens design methods predominantly depend on either pre-corrected lenses as starting points [19, 17, 30] or human expertise as a design prior [33, 18] to find a final solution. As a result, the lenses designed are usually refined versions of conventional imaging lenses, which may not be the optimal solution because classical lens design has usually reached a local minimum. In other words, current End-to-End lens design cannot function without classical lens design, which not only reduces usability but also prevents deep learning methods from discovering novel lens structures for specific computer vision tasks.

To address this limitation, we propose a **task-driven** approach where a well-trained network model is adopted to supervise the lens design from scratch. Unlike conventional End-to-End methods which typically begin with successful lenses and simultaneously optimize both the lens and the network, our approach exclusively optimizes optical parameters while maintaining a fixed downstream network model. This task-driven method simplifies the End-to-End design challenge, facilitating the exploration of more complex optical structures. Additionally, it enhances the explainability of the design process by learning an optical structure that effectively encodes valid features from the object space to the image space.

In this paper, we focus on the image classification task and aim to find the corresponding optimal lens structure. Employing the task-driven approach, we design three image classification lenses ("TaskLens") from scratch solely with network supervision. For comparison, three conventional imaging lenses ("ImagingLens") for each TaskLens are designed by experienced optical engineers. Image classification accuracy evaluation on ImageNet [34] demonstrates that our TaskLenses outperform ImagingLenses with even fewer lens elements. By digging into the optical characteristics, we find the TaskLens exhibits a novel long-tailed point spread function (PSF) which effectively preserves image features in the presence of optical aberrations for image classification purposes. Although this PSF is not desired in conventional imaging-driven lens design, it is beneficial for machine vision and high-level computer vision tasks. Additionally, we demonstrate the practicality of our TaskLens by showing that it is compatible with various network models while maintaining enhanced classification accuracy. The contributions of this paper can be summarized as follows:

- We propose a task-driven lens design method that relies only on a well-trained network model to design lenses from scratch.
- Our TaskLenses achieve higher accuracy in image classification tasks, outperforming conventional imaging lenses while utilizing fewer elements.
- We analyze the optical characteristics of TaskLens and demonstrate its compatibility with various network models, highlighting its ability to maintain or enhance performance across different scenarios.

2 Related works

2.1 Classical lens design.

Classical lens design methods [35, 36, 37] aim to optimize lenses that fulfill specific imaging quality criteria. The lens optimization process typically involves minimizing image quality-based losses, usually measured by the root-mean-square (RMS) spot size. This refers to the RMS radius of all sampled optical rays within a certain field on the image plane. Smaller RMS spot sizes contribute to enhanced image quality. Ray tracing and optimization algorithms are widely explored in this context and have been extensively applied in optical design software such as Zemax [38] and Code V [39].

2.2 End-to-End lens design.

End-to-End optical design [16, 17, 18, 19] jointly optimizes optical systems (including diffractive, refractive, and reflective components) and downstream image processing networks to enhance the overall capabilities for a target application. It has demonstrated remarkable performance in various applications, including hyperspectral imaging [24, 25, 26, 27], extended-depth-of-field imaging [17, 21, 22, 19], large-field-of-view imaging [23, 19], and seeing through obstructions [40]. End-to-End optical design has also enabled the reduction of optical aberrations in compact structures, such as large field of view [23, 19], and achromatic imaging [16, 25, 41]. Furthermore, it has exhibited improved performance in computer vision tasks over classical lenses, including optical character recognition [42], object detection [18, 30, 42], and depth estimation [31, 32]. However, End-to-End optical design presents challenges for convergence, as both lens design and network optimization are highly non-convex problems. The optical gradients back-propagated from the network are indirect and biased compared to classical lens design objectives (*eg*, spot size). To address this challenge, researchers have adopted various strategies, including initiating designs from successful or simple structures [17, 19, 18, 30, 42, 20, 23] or using curriculum learning approaches [21].

2.3 Differentiable image simulation.

In End-to-End optical design, a differentiable optical simulator is necessary to back-propagate the final image loss through the entire pipeline and optimize optical parameters. Existing differentiable optical simulation methods rely on either wave optics or geometric optics. In wave optics, diffractive optical elements (DOEs) [43, 44, 16, 25], metasurfaces [45], and refractive lenses [29, 28, 41, 40, 46] are modeled as phase modulation functions. Refractive

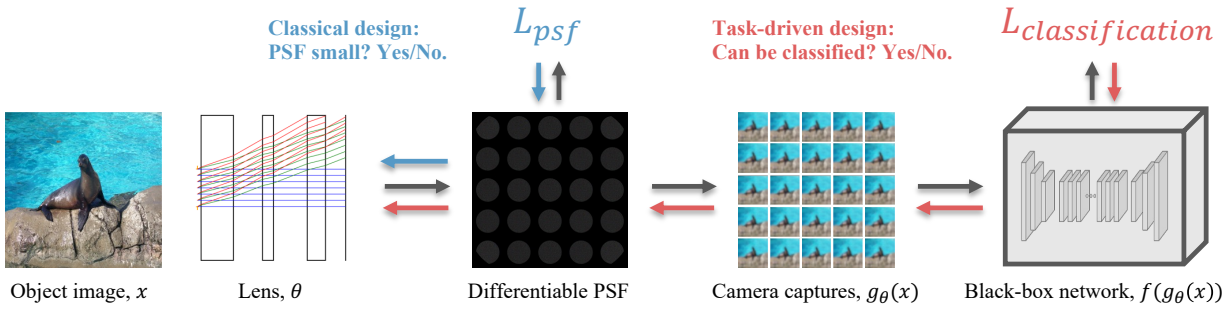

Figure 1: The task-driven lens design pipeline involves the computation of PSF through differentiable ray tracing, followed by convolution with input images to simulate camera-captured images at different fields of view. Subsequently, a well-trained network computes the image classification errors and back-propagates the loss function to optimize the lens parameters. Different from classical lens design which efforts to minimize only the PSF, task-driven lens design focuses on whether the camera-captured images can be successfully classified by the downstream deep network.

lenses are often simplified as thin lenses under the paraxial approximation, which can be inaccurate and fail to represent real lenses. Recent studies in ray tracing-based image simulation [47, 17, 48, 19, 18, 33, 46] have shown promising accuracy in modeling thick, aspherical lenses and even freeform optics. In this approach, images captured by cameras are simulated using either ray tracing-based rendering or point spread function (PSF) convolution [18, 19, 42, 30]. To make the ray tracing process differentiable, researchers either employ auto-differentiation to compute optical gradients [17, 19, 21, 42, 30] or use a network to represent the optical lens [33, 18, 49].

3 Methods

3.1 Task-driven lens design

Illustrated by Fig. 1, the lens design problem for a visual task can be formulated as follows:

$$\begin{aligned} \theta &= \underset{\theta}{\operatorname{argmin}} \|f(g_\theta(x)) - y\| \\ \textbf{s.t.} \quad f &= \underset{f}{\operatorname{argmin}} \|f(x) - y\|, \end{aligned} \tag{1}$$

where x represents the input object image, θ represents the lens parameters, g represents the imaging process, f represents the target visual task, and y is the ground truth for the visual task f. The intuitive solution to Eq. (1) is given by

$$\theta = \underset{\theta}{\operatorname{argmin}} \|g_\theta(x) - x\|, \tag{2}$$

which minimizes the difference between camera-captured image $g_\theta(x)$ and the object image x. This imaging-driven classical lens design philosophy guides optical engineers to minimize the PSF of the lens system, corresponding to the blue arrow in Fig. 1.

However, we propose that Eq. (2) represents only a local minimum for the target visual task, as the solution spaces for the visual task and the best imaging quality are different. In particular, we assume that there are key features in the object images for a visual task, and the object image can be decomposed as

$$x = x_f \oplus x_{bg}, \tag{3}$$

where x_f represents the image features, and x_{bg} represents the background information. They are combined with the relation $\oplus$, but a well-trained network f can effectively extract x_f from the input, and only x_f contributes to the output, formulated as $f(x) = f(x_f)$. Based on this assumption, the optical lens only needs to capture/preserve the image features x_f from the object images:

$$\theta = \underset{\theta}{\operatorname{argmin}} \|g_\theta(x_f) - x_f\|. \tag{4}$$

The philosophy of this task-driven design method is to convert the highly non-convex lens optimization problem into a feature encoding problem. As a result, we can neglect the useless background information x_{bg} and focus on interesting optical features during lens design, rather than minimizing the total aberrations for an intermediate optimal imaging performance as in classical lens design.

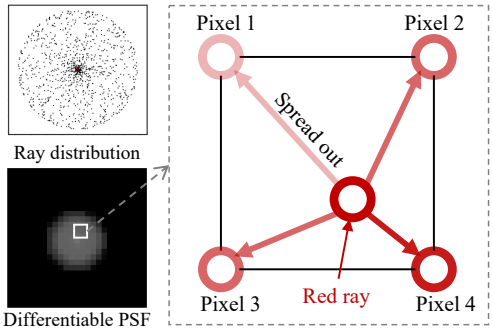

Figure 2: Differentiable PSF computation. Each optical ray is assigned to its neighboring four pixels with a weight term depending on the distance from the ray. The weight term tracks the gradient information and is used to optimize lens parameters during the back-propagation process.

Corresponding to the red arrow in Fig. 1, x_f is determined by a well-trained network f, which operates as a black box to guide the task-driven lens design process. During the optimization, the objective is to generate images that can be successfully classified by the network, without aiming for perfect imaging quality. It is important to note that Eq. (1)–Eq. (4) are not intended to be strict mathematical proofs; instead, they are used to illustrate our task-driven approach.

The task-driven lens design pipeline is similar to the End-to-End optical design pipeline, except that we use a well-trained and intact network during the training. Upon the advantages of End-to-End optical methods, we propose this task-driven lens design approach mainly based on two observations: (1) the visual capabilities of an End-to-End optical system primarily stem from the optical aspects, and (2) achieving convergence in End-to-End optical design is challenging due to the difficulty of obtaining accurate gradients from an undertrained deep network.

3.2 Differentiable point spread function

The PSF characterizes how an optical system blurs a point light source. In image simulation, PSF is convolved with the object image to simulate the camera-captured image. PSF can be computed by ray tracing from a point source to the sensor plane. Subsequently, the optical rays are assigned to their neighboring sensor pixels, as shown in Fig. 2. This process can be formulated as

$$\mathrm{PSF}(\mathbf{o_p}) = \sum_{i=1}^{N} u_i \cdot \sigma(|(\mathbf{o_p} - \mathbf{o_i}) \cdot \hat{\mathbf{e}}_x|/L) \cdot \sigma(|(\mathbf{o_p} - \mathbf{o_i}) \cdot \hat{\mathbf{e}}_y|/L), \tag{5}$$

where $\mathbf{o}_p$ denotes the pixel coordinate, $\mathbf{o}_i$ denotes the intersection position of the ith ray on the sensor plane, N represents the numbers of rays from each point source, u_i denotes the energy, which we assume equals to 1, $\hat{\mathbf{e}}_x$ and $\hat{\mathbf{e}}_y$ are unit vector in the sensor plane, and L denotes the physical width of a sensor pixel. In our experiments, we set $u_i = 1$. The σ function is defined as

$$\sigma(x) = \begin{cases} 1 - x & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

which assesses a ray's impact on its surrounding pixels. As shown in Fig. 2, the total energy of a ray is spread out to its neighboring four pixels. Eq. (5) can be viewed as an inverse bilinear interpolation. By utilizing this sub-pixel information, we can represent the actual light distribution using a limited number of rays. In the backward process, the PSF gradients can guide the rays to move towards the desired pixels as Eq. (5) is differentiable. The gradients can then be back-propagated to adjust the lens surfaces to control the rays.

3.3 Implementation details

In our experiments, we selected a 5×5 PSF grid to represent the optical characteristics of the lens. Instead of simulating full sensor resolution images with this PSF grid, we used a single PSF from the grid to simulate independent images at each field of view. We adopted this setting for two reasons: (1) Commercial camera sensors typically have mega-pixel resolution, while training images usually have much lower resolutions (*eg*, 224×224), making it difficult to obtain full sensor resolution training data; (2) We believe that the learned optical characteristics should be independent of the content distribution of the training images. This approach differs from previous works such as Tseng *et, al* [18] and Côté *et al* [30], which sacrifice edge imaging quality by assuming that objects will not be located there.

Table 1: Accuracy (top) on the validation set of ImageNet achieved by different lenses. PSNR [dB] (middle) of simulated camera captures are adopted to represent the imaging quality of the lens. Avg RMS spot size [μm] (bottom) of the optical lens.

	TaskLens	ImagingLens #1 / #2 / #3
Doublet	70.08% 19.46 27.27	65.63% / 68.54% / 67.01% 22.43 / 23.65 / 22.90 13.61 / 9.85 / 10.99
Triplet	73.40% 23.85 10.41	70.04% / 69.92% / 68.52% 22.77 / 23.14 / 23.51 9.32 / 14.75 / 8.53
Quadruplet	73.61% 23.67 9.97	72.27% / 68.88% / 68.56% 25.00 / 23.63 / 23.98 7.32 / 10.11 / 11.37

We implement the differentiable PSF calculation using an open-source memory-efficient differentiable ray tracer **dO** [19, 21]. The lens surfaces are aspheric, with optimizable parameters of curvature, position, and even polynomial coefficients ranging from α_4 to α_{10}. Lens materials are chosen from the library of commonly used cellphone lens plastics. The image sensor has a diagonal length of 4 mm and a resolution of 1080×1920, corresponding to a pixel size of 1.8 μm. To accommodate lens dispersion, we use three wavelengths (656.3 nm, 589.3 nm, 486.1 nm) to calculate the PSF of each channel. We utilize a 51×51 PSF kernel size to account for significant optical aberrations, also allowing rays to move across a larger region. The AdamW optimizer [50] is employed with a learning rate of $1e^{-4}$ for curvature, position, and α_4 parameters, while a 0.02 learning rate decay is applied to higher-order polynomial coefficients. In accordance with Yang *et, al* [21], we penalize the incident angles between light rays and lens surface distance to prevent self-intersection during the lens design. Considering the rotational symmetry of the aspherical surfaces, we select quarter space for training which corresponds to 9 distinct fields, and we sample 256 rays from each field of view to compute the differentiable PSF.

All training images are resized to 224×224 and convolved with the PSF to simulate the captured images at that field of view. During training, a batch size of 64 is employed, increasing to 576 after concatenating the image batches for each of the 9 different fields of view. We use TrivialAugment Wide [51] for data augmentation to prevent the learned lens from converging to a perfect imaging lens. For the task-driven lens design, we utilize a well-trained ResNet50 network [1] for supervision. The lens is optimized from scratch for 1 epoch on the ImageNet training set, typically achieving convergence. Then we conduct End-to-End training to fine-tune the lens. The optimization process consumes approximately 60 GB of GPU memory for a lens and the previous settings, with the bottleneck being the network size and the number of rays.

After designing the lens, we fine-tune the image classification network for each lens using the AdamW optimizer [50] with a learning rate of $1e^{-5}$ and the CosineAnnealing scheduler [52] with a warm-up scheme. We sample 4096 rays from each field of view during the fine-tuning and testing stages to obtain more accurate PSF. This high sampling rate does not cause memory issues because we do not need to perform differentiable ray tracing at this stage. We fine-tune each network model for an additional 3 epochs, which runs on two 80G A100 GPUs for two days to complete. The classification accuracy is computed on all simulated images at 9 fields of view. Since it would be too expensive if we used the official ImageNet testing set, we split a partition of the training set for validation and use the validation set for testing.

4 Task-driven image classification lens design

With our proposed task-driven approach, we design three image classification lenses from scratch, with two, three, and four lens elements, respectively. Each lens is designed for a target FoV of 68.8°, F/2.8, and paired with an image sensor with a 4 mm diagonal length. We control the aperture to F/2.8 by setting the aperture radius to 0.52 mm. As shown in Fig. 3, the task-driven approach successfully converges optical rays for valid imaging, even though we use a classification network in place of traditional imaging-based objectives.

Baselines. For each image classification lens ("TaskLens"), we design three conventional imaging lenses ("ImagingLens") for comparison. The first ImagingLens (labeled as #1) is optimized using the open-source code [21] by

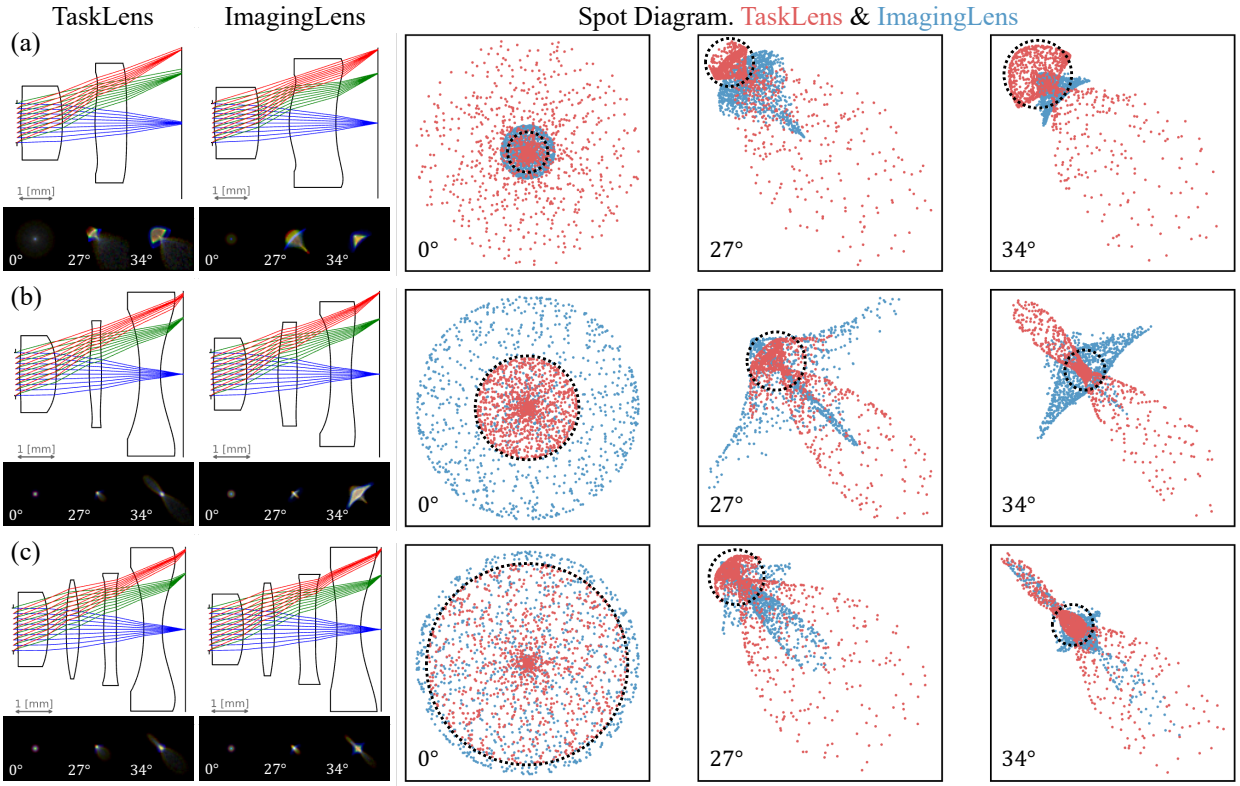

Figure 3: The lens structure, PSF at different fields, and corresponding spot diagram for doublet (a), triplet (b), and quadruplet (c) lenses. Although the total spot diagram of TaskLens is larger than ImagingLens, the majority of the optical rays converge to a small region, resulting in a smaller effective spot diagram (marked by the black circle). This novel spot diagram distribution results in a long-tailed PSF, characterized by a small concentrated center and sparsely populated outer regions. In scenarios where the optical structure cannot fully correct all optical aberrations, this long-tailed PSF proves effective in preserving essential image features from the object images.

minimizing the spot diagram at different fields of view. The other two ImagingLenses (labeled as #2 and #3) are optimized by experienced (10+ years) optical engineers using ZEMAX [38].

Classification accuracy comparisons. Table 1 presents quantitative classification accuracy for all lenses. Our TaskLens demonstrates superior image classification accuracy compared to ImagingLens with the same number of lens elements. The classification accuracy improvement is significant. Due to optical aberrations, no lens can reach the upper bound (75.63%) acquired with the original sharp images. Remarkably, our doublet TaskLens outperforms all triplet ImagingLenses, and the triplet TaskLens outperforms all quadruplet ImagingLenses. These results demonstrate that our TaskLens can achieve enhanced image classification accuracy with fewer lens elements compared to conventional lenses.

Image quality comparisons. In Tab. 1, we also report the imaging performance, quantified by the PSNR metric and average RMS spot size for each lens. Nine distinct fields are selected to calculate these quantitative scores. The results indicate that high imaging quality or a smaller average RMS spot size does not always correlate with higher image classification accuracy. This supports our task-driven design philosophy that a perfect classical optical lens is not necessarily the optimal solution for high-level computational imaging and computer vision applications.

Explanation for the enhanced visual task performance. To investigate the reason behind the improvement in classification accuracy, we evaluate the optical characteristics of the TaskLens and the best-performing ImagingLens. We plot the PSF from the optical axis (0°) to the full FoV, as well as the corresponding spot diagram at 589.3 nm in Fig. 3. The spot diagram illustrates the intersection points of optical rays with the sensor plane. Due to insufficient optical elements to correct all optical aberrations, the spot diagram exhibits noticeable aberrations. An interesting phenomenon was observed: TaskLens converges the majority of optical rays into a small region, disregarding outlier rays.

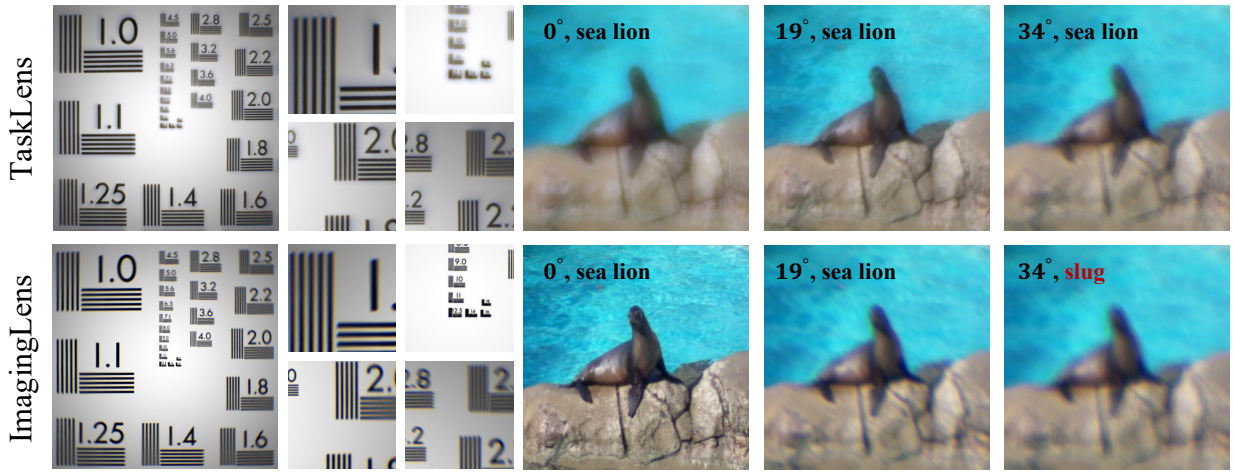

Figure 4: Image simulation results of the doublet TaskLens and ImagingLens (#2). Left: Full-resolution simulation using a resolution chart, with zoomed image patches. For ImagingLens, the center region of the simulated camera capture exhibits well-corrected aberration, while the off-axis regions suffer from significant aberrations. For TaskLens, all regions display a noticeable "haze" pattern caused by the long-tail PSF; however, the structural details are well preserved. Right: Single field image simulation with one PSF function. The simulated camera captures of TaskLens can be correctly recognized by the network despite the haze effect, but those of the ImagingLens fail due to optical aberrations.

In contrast, ImagingLens attempts to converge all optical rays towards the center, aligning with conventional optical design objectives. Although the total spot size of the TaskLens is often larger compared to that of the ImagingLens, the TaskLens demonstrates a smaller effective spot size, as indicated by black dashed circles in Fig. 3.

This novel spot diagram results in a distinctive **long-tailed** PSF at each field of view, particularly for the triplet and quadruplet TaskLenses. Since most light energy effectively converges in the center region, structural details in the object images are well preserved, despite causing a haze effect in simulated camera captures, as illustrated in Fig. 4. Although the haze effect is detrimental to human vision and image quality metrics, it can be compensated for by the network. The network successfully recognizes the sea lion in the inputs from the TaskLens (Fig. 4). During the training process, the optics learn to encode valid image features recognizable by downstream network models. In contrast, the ImagingLens design process, unaware of backend tasks, leads to optical aberrations that blend information in the object image, resulting in blurry camera captures unrecognizable by the network (Fig. 4). These visualization results suggest that, when optical structures are insufficient to correct all aberrations, it is beneficial to selectively tolerate certain aberrations, such as the haze effect in our experiments, to achieve better final visual performance.

5 Ablation study

To demonstrate the effectiveness and usability of our proposed task-driven lens design and the corresponding TaskLens, we conduct ablation experiments with various downstream network models, comparisons with conventional End-to-End lens designs, and evaluations of performance both with and without image post-processing.

5.1 Can TaskLens work with different classification models?

In the Task-Driven lens design process, the lens is designed with a specific network model (ResNet-50). However, in real-world scenarios, it is common to change downstream network models based on practical constraints. For example, using smaller models to accommodate the limited computational power of end devices or employing larger models to achieve better performance. However, optical lenses cannot be modified after design and manufacturing. Therefore, we aim to investigate whether the visual task performance of TaskLenses is compatible with different network models.

We employ three network models with different model sizes and architectures: MobileNetV3-Large [53] ("MobileNetV3-L") with 5.5M parameters, SwinTransformer-Base [2] ("Swin-B") with 88M parameters, and ViT-Large-patch/16 [54] ("ViT-L/16") with 304M parameters. Additionally, the original ResNet50 network contains 26M parameters. ResNet50 and MobileNetV3-L are convolution neural networks, while Swin-B and ViT-L/16 are transformer-based architectures. Table 2 presents the image classification accuracy for different models and our TaskLens still holds the highest accuracy

Table 2: Classification accuracy with different network models. Our designed TaskLens is compatible with different network models while maintaining enhanced performance. The results for the original ResNet50 are reported in Tab. 1. Acc Ref: accuracy achieved on original sharp images.

#params / Acc Ref	MobileNetV3-L [55] 5.4M / 73.96%	Swin-B [2] 88M / 85.59%	ViT-L/16 [54] 304M / 86.70%
TaskLens (Doublet) ImagingLens #1 / #2 / #3	68.22% 64.60% / 67.73% / 66.05%	81.19% 79.03% / 80.87% / 80.23%	81.76% 79.39% / 81.19% / 80.79%
TaskLens (Triplet) ImagingLens #1 / #2 / #3	71.82% 68.36% / 67.94% / 68.01%	82.65% 81.19% / 81.08% / 80.19%	83.46% 81.62% / 81.46% / 81.61%
TaskLens (Quadruplet) ImagingLens #1 / #2 / #3	72.06% 71.00% / 66.78% / 67.19%	82.82% 82.43% / 80.75% / 80.46%	83.62% 82.52% / 81.60% / 81.50%

Table 3: End-to-End design from scratch fails to converge. While starting with a well-designed ImagingLens, End-to-End design fails to discover the optimal classification lens.

	TaskLens	End2End Training	
		ImagingLens	From scratch
Doublet	70.05%	69.55%	✗
Triplet	73.40%	71.94%	✗
Quadruplet	73.61%	73.44%	✗

at the same number of lens elements. The results demonstrate that our TaskLens is compatible with different network models while maintaining enhanced performance, which also implies that the visual task performance of TaskLens comes from its novel optical characteristics.

Moreover, these results provide inspiration for addressing the memory constraints in Task-Driven lens design by utilizing smaller network models for training. And people can switch to larger models in practice after designing the lens.

5.2 Can End-to-End training find the optimal lens?

In this section, we aim to investigate whether the conventional End-to-End optical design approach can also identify the optimal lens for image classification. We consider two initial lens starting points: the best-performing ImagingLens and an all-flat lens. These lenses, along with a well-trained classification network (ResNet50), are jointly optimized using the image classification loss, as is typical in conventional End-to-End optical design approaches. The final classification results are presented in Tab. 3.

When starting from the all-flat optics, the End-to-End training fails to converge. This is primarily attributed to the inherent challenges in achieving convergence directly from scratch when employing differentiable ray tracing-based imaging models. When initiating with a well-designed imaging lens and network, the final performance of the lens designed through this process does not match that of our TaskLens. We hypothesize that this is because the imaging lens starting point has already reached a local minimum for the visual task. As a result, the gradients back-propagated to the lens parameters are too weak to escape from this local minimum. Consequently, the lens designed through the End-to-End process merely fine-tunes the initial configuration (detailed lens structure in Supplemental Material), leading to a classification performance that is inferior compared to our TaskLens.

5.3 Can image restoration bridge the performance gap?

Image restoration is commonly employed to mitigate optical aberrations present in camera captures. These algorithms, especially deep learning-based methods, have exhibited remarkable capabilities in recovering fine structures within images, subsequently enhancing the performance of downstream visual tasks. Given the higher level of optical aberrations in our TaskLens compared to conventional ImagingLens, we want to investigate whether the visual task performance gap between TaskLens and ImagingLens can be bridged through image restoration. Consequently, we apply image restoration to all lenses and compare the resulting classification accuracy of the restored images.

Table 4: Classification accuracy after image restoration. The performance gap between TaskLens and ImagingLens can not be bridged by image restoration.

	PSNR [dB]	Classification Acc
TaskLens (Doublet)	27.24	72.03%
ImagingLens	27.54 / 30.30 / 30.87	68.19% / 71.24% / 71.08%
TaskLens (Triplet)	32.31	74.43%
ImagingLens	29.44 / 30.47 / 29.95	72.42% / 73.35% / 70.90%
TaskLens (Quadruplet)	33.58	74.61%
ImagingLens	34.29 / 29.69 / 29.38	73.98% / 72.16% / 71.52%

We employ NAFNet [56] (width = 32, encoding block number = [1, 1, 1, 8], middle block number = 1, and decoding block number = [1, 1, 1, 1]) to recover the camera captures for each lens. We first train the restoration network on camera-captured images to convergence, then fix the restoration network and fine-tune the image classification network using restored images.

Table 4 presents the image restoration and classification results. The image restoration demonstrates an enhancement in classification accuracy for all lenses. However, TaskLens still outperforms ImagingLens with the same number of lens elements. These results suggest that the gap in classification performance cannot be solely bridged by applying image restoration.

6 Discussion

Application scenarios. In practice, although it is possible to purchase high-quality off-the-shelf lenses, there are instances where desired optical specifications such as focal length and FoV are not readily available. Therefore, there is a need to customize lenses which entails domain-specific expertise, sophisticated manufacturing processes, high costs, and significant leading time. With the advance of automatic lens design techniques [21, 57, 33] and 3D printing technology [58, 59], it has become increasingly possible to customize versatile lenses with significantly less effort. By leveraging our proposed task-driven lens design approach, we can further simplify the optical structure and reduce the number of lens elements. This reduction in complexity offers the additional benefit of minimizing the effort required in lens customization.

Manufacturing. Currently, we are unable to manufacture the designed lenses to validate our results with real-world experiments. However, the ray tracing accuracy of our simulation aligns with the commercial optical design software ZEMAX, which is commonly regarded as the industrial standard. Additionally, we believe the task-driven lens design approach enhances the classical lens design methodology, despite not having the manufacturing process. In the Supplemental Material, we have evaluated the accuracy of our image simulation using off-the-shelf optical lenses. Common manufacturing and assembly errors, which typically reduce imaging and visual performance in real-world scenarios, are also addressed and analyzed by simulation in the Supplemental Material.

Open questions. In our experiments, we observe a novel long-tail PSF that effectively preserves object structures despite optical aberrations. This type of long-tail PSF, although not particularly discussed in our study, can be noted in existing literature [44, 17, 19], and benefits for various applications. Additionally, in real-world scenarios, camera-captured images are often applied in diverse downstream tasks. This observation inspires us to question whether there is an optimal PSF for computational imaging and machine vision. In the Supplemental Material, we assess the performance of our TaskLens in applications like object detection and instance segmentation. Yet, a more comprehensive evaluation of this system is an area for future exploration.

7 Conclusion

In this paper, we introduce a task-driven approach that relies solely on a well-trained network to supervise lens design from scratch. To the best of our knowledge, this is the first attempt to design lenses without using classical lens design methods and knowledge. By applying this approach, we have designed three lenses specifically for image classification, which outperform traditional imaging-oriented lenses, even with fewer elements. We believe that TaskLens has significant potential not only in lens design methodologies but also in practical applications, especially where physical size and cost constraints are critical factors.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[2] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12009–12019, 2022.

[3] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. InternImage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022.

[4] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[5] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Adv. Neural Inf. Process.*, 35:4203–4217, 2022.

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 779–788, 2016.

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, pages 10012–10022, 2021.

[8] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12104–12113, 2022.

[9] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5162–5170, 2015.

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 270–279, 2017.

[11] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4009–4018, 2021.

[12] Gal Shabtay, Ephraim Goldenberg, Michael Dror, Itay Yedid, and Gil Bachar. Folded camera lens designs, February 25 2020. US Patent 10,571,644.

[13] Xuepeng Zhou. Camera optical lens, January 31 2023. US Patent 11,567,301.

[14] Hirofumi Abe. Zoom lens and image pickup apparatus including the same, April 24 2018. US Patent 9,952,446.

[15] Shigenobu Sugita. Zoom lens and image pickup apparatus including same, August 18 2015. US Patent 9,110,278.

[16] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Trans. Graph.*, 37(4):1–13, 2018.

[17] Qilin Sun, Congli Wang, Fu Qiang, Dun Xiong, and Heidrich Wolfgang. End-to-end complex lens design with differentiable ray tracing. *ACM Trans. Graph.*, 40(4):1–13, 2021.

[18] Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-Francois Lalonde, and Felix Heide. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Trans. Graph.*, 40(2):1–19, 2021.

[19] Congli Wang, Ni Chen, and Wolfgang Heidrich. dO: A differentiable engine for deep lens design of computational imaging systems. *IEEE Transactions on Computational Imaging*, 8:905–916, 2022.

[20] Zongling Li, Qingyu Hou, Zhipeng Wang, Fanjiao Tan, Jin Liu, and Wei Zhang. End-to-end learned single lens design using fast differentiable ray tracing. *Opt. Lett.*, 46(21):5453–5456, 2021.

[21] Xinge Yang, Qiang Fu, and Wolfgang Heidrich. Curriculum learning for ab initio deep learned refractive optics. *arXiv preprint arXiv:2302.01089*, 2023.

[22] Yuankun Liu, Chongyang Zhang, Tingdong Kou, Yueyang Li, and Junfei Shen. End-to-end computational optics with a singlet lens for large depth-of-field imaging. *Opt. Express*, 29(18):28530–28548, 2021.

[23] Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. Learned large field-of-view imaging with thin-plate optics. *ACM Trans. Graph.*, 38(6):219–1, 2019.

[24] Daniel S Jeon, Seung-Hwan Baek, Shinyoung Yi, Qiang Fu, Xiong Dun, Wolfgang Heidrich, and Min H Kim. Compact snapshot hyperspectral imaging with diffracted rotation. *ACM Trans. Graph.*, 38(4):1–13, 2019.

[25] Xiong Dun, Hayato Ikoma, Gordon Wetzstein, Zhanshan Wang, Xinbin Cheng, and Yifan Peng. Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging. *Optica*, 7(8):913–922, 2020.

[26] Seung-Hwan Baek, Hayato Ikoma, Daniel S Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H Kim. Single-shot hyperspectral-depth imaging with learned diffractive optics. In *Int. Conf. Comput. Vis.*, pages 2651–2660, 2021.

[27] Lingen Li, Lizhi Wang, Weitao Song, Lei Zhang, Zhiwei Xiong, and Hua Huang. Quantization-aware deep optics for diffractive snapshot hyperspectral imaging. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19780–19789, 2022.

[28] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1375–1385, 2020.

[29] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning Rank-1 diffractive optics for single-shot high dynamic range imaging. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1386–1396, 2020.

[30] Geoffroi Côté, Fahim Mannan, Simon Thibault, Jean-François Lalonde, and Felix Heide. The differentiable lens: Compound lens search over glass surfaces and materials for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20803–20812, 2023.

[31] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3D object detection. In *Int. Conf. Comput. Vis.*, pages 10193–10202, 2019.

[32] Hayato Ikoma, Cindy M Nguyen, Christopher A Metzler, Yifan Peng, and Gordon Wetzstein. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In *Int. Conf. Comput. Photog.*, pages 1–12. IEEE, 2021.

[33] Geoffroi Côté, Jean-François Lalonde, and Simon Thibault. Deep learning-enabled framework for automatic lens design starting point generation. *Opt. Express*, 29(3):3841–3854, 2021.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255. Ieee, 2009.

[35] Michael J Kidger. Fundamental optical design. In *Fundamental optical design*. SPIE-International Society for Optical Engineering, 2001.

[36] Rudolf Kingslake and R Barry Johnson. *Lens design fundamentals*. academic press, 2009.

[37] Warren J Smith. *Modern optical engineering: the design of optical systems*. McGraw-Hill Education, 2008.

[38] Zemax LLC. *Zemax User Manual*, 2021.

[39] Inc. Synopsys. Code V, version 13.0. [Software], 2023.

[40] Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Qiang Fu, Hadi Amata, Xiao Li, Praneeth Chakravarthula, Wolfgang Heidrich, and Felix Heide. Seeing through obstructions with diffractive cloaking. *ACM Trans. Graph.*, 41(4):1–15, 2022.

[41] Samuel Pinilla, Seyyed Reza Miri Rostami, Igor Shevkunov, Vladimir Katkovnik, and Karen Egiazarian. Hybrid diffractive optics design via hardware-in-the-loop methodology for achromatic extended-depth-of-field imaging. *Opt. Express*, 30(18):32633–32649, 2022.

[42] Shiqi Chen, Ting Lin, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Computational optics for mobile terminals in mass production. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[43] Felix Heide, Qiang Fu, Yifan Peng, and Wolfgang Heidrich. Encoded diffractive optics for full-spectrum computational imaging. *Sci. Rep.*, 6(1):33543, 2016.

[44] Yifan Peng, Qiang Fu, Felix Heide, and Wolfgang Heidrich. The diffractive achromat full spectrum computational imaging with diffractive optics. In *SIGGRAPH ASIA 2016 Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments*, pages 1–2. ACM New York, NY, USA, 2016.

[45] Ethan Tseng, Shane Colburn, James Whitehead, Luocheng Huang, Seung-Hwan Baek, Arka Majumdar, and Felix Heide. Neural nano-optics for high-quality thin lens imaging. *Nature communications*, 12(1):6493, 2021.

[46] Yunfeng Nie, Jingang Zhang, Runmu Su, and Heidi Ottevaere. Freeform optical system design with differentiable three-dimensional ray tracing and unsupervised learning. *Opt. Express*, 31(5):7450–7465, 2023.

[47] Craig Kolb, Don Mitchell, and Pat Hanrahan. A realistic camera model for computer graphics. In *Proceedings of the 22nd annual conference on computer graphics and interactive techniques*, pages 317–324, 1995.

[48] Marco Mout, Michael Wick, Florian Bociort, Jörg Petschulat, and Paul Urbach. Simulating multiple diffraction in imaging systems using a path integration method. *Appl. Opt.*, 55(14):3847–3853, 2016.

[49] Xinge Yang, Qiang Fu, Mohammed Elhoseiny, and Wolfgang Heidrich. Aberration-aware depth-from-focus. *arXiv preprint arXiv:2303.04654*, 2023.

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[51] Samuel G Müller and Frank Hutter. TrivialAugment: Tuning-free yet state-of-the-art data augmentation. In *Int. Conf. Comput. Vis.*, pages 774–782, 2021.

[52] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[53] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[54] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[55] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for MobileNetv3. In *Int. Conf. Comput. Vis.*, pages 1314–1324, 2019.

[56] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022.

[57] Libin Sun, Brian Guenter, Neel Joshi, Patrick Therien, and James Hays. Lens factory: Automatic lens generation using off-the-shelf components. *arXiv preprint arXiv:1506.08956*, 2015.

[58] Guangbin Shao, Rihan Hai, and Cheng Sun. 3D printing customized optical lens in minutes. *Advanced Optical Materials*, 8(4):1901646, 2020.

[59] Bisrat G Assefa, Markku Pekkarinen, Henri Partanen, Joris Biskop, Jari Turunen, and Jyrki Saarinen. Imaging-quality 3D-printed centimeter-scale lens. *Opt. Express*, 27(9):12630–12637, 2019.