Minimalist Vision with Freeform Pixels

Jeremy Klotz and Shree K. Nayar

Computer Science Department, Columbia University, New York NY, USA
{jklotz,nayar}@cs.columbia.edu

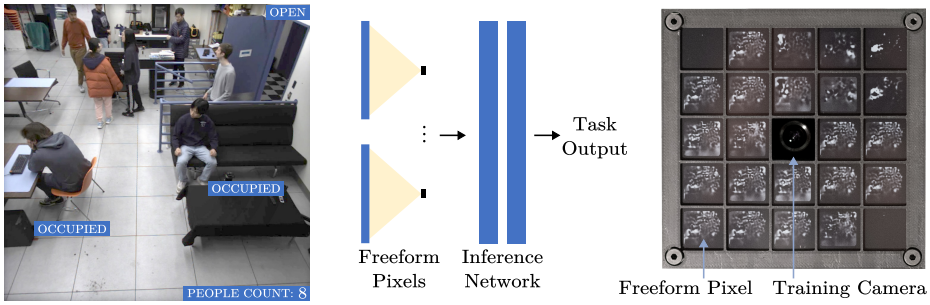
Abstract. A minimalist vision system uses the smallest number of pixels needed to solve a vision task. While traditional cameras use a large grid of square pixels, a minimalist camera uses freeform pixels that can take on arbitrary shapes to increase their information content. We show that the hardware of a minimalist camera can be modeled as the first layer of a neural network, where the subsequent layers are used for inference. Training the network for any given task yields the shapes of the camera's freeform pixels, each of which is implemented using a photodetector and an optical mask. We have designed minimalist cameras for monitoring indoor spaces (with 8 pixels), measuring room lighting (with 8 pixels), and estimating traffic flow (with 8 pixels). The performance demonstrated by these systems is on par with a traditional camera with orders of magnitude more pixels. Minimalist vision has two major advantages. First, it naturally tends to preserve the privacy of individuals in the scene since the captured information is inadequate for extracting visual details. Second, since the number of measurements made by a minimalist camera is very small, we show that it can be fully self-powered, i.e., function without an external power supply or a battery.

Keywords: Freeform Pixels · Minimalist Camera · Lightweight Vision · Self-Powered Camera · Privacy Preservation · Deep Optics · Computational Imaging

1 Why Minimalist Vision?

Today, computer vision plays an indispensable role in our everyday lives. It serves as the backbone in a wide gamut of applications ranging from video surveillance and monitoring to autonomous driving and robotics. Broadly speaking, vision applications can be divided into two categories. In one category, the system seeks to infer detailed information about objects and activities in a scene. Examples include object detection and recognition, optical flow estimation and tracking, and 3D reconstruction. The second category of applications involves high-level inferences about the statistics of objects in a scene and the states of an environment. Examples in this realm include monitoring the occupancy of workspaces, the flow of traffic on highways, and the lighting in an urban environment.

In our work, we are interested in the second category, which we refer to as "lightweight vision." We claim that lightweight tasks can be solved not with traditional images, but rather a very small number of measurements, as long as the measurements are rich in information.



(a) Workspace Monitoring (b) Camera in a Network (c) Minimalist Camera

Fig. 1: Monitoring a workspace with minimalist vision. (a) The task is to count the number of people, track the occupancy of specified zones, and detect when the door is open. A minimalist vision system, composed of a camera and inference network, can perform such lightweight tasks using just a handful of freeform pixels. (b) The entire system can be modeled as a single network. Once this network is trained, the first layer specifies the design of a camera, a prototype of which is shown in (c). This system can count the people in the room (with 2 pixels), track the occupancy of each zone (with 2 pixels, each), and detect when the door is open (with 2 pixels). Given the small number of measurements it makes, a minimalist camera can be completely self-powered.

We introduce minimalist vision as an approach to solve lightweight tasks. In the arts, minimalism is a technique that is used to pare down a piece of work to its essential elements. The goal is to ensure that each element used has a purpose. In our context, traditional cameras that are used in virtually all vision systems today capture far more information than needed to solve a lightweight task. Our work seeks to answer two key questions: (a) *Given a task, what is the smallest number of visual measurements needed to achieve a desired performance?* (b) *How do we construct a camera that produces these measurements?* If we are successful in designing such a minimalist camera, it would have the following two major benefits:

Towards Privacy Preservation: When a traditional camera captures an image, it typically reveals far more information about the scene than necessary for the task. For instance, a single image could reveal a person's identity, location, or even intentions. This is a well-known problem that has made the widespread deployment of cameras highly controversial [38]. Since minimalist vision captures the smallest number of measurements for a given task, it is difficult to extract visual details about the scene such as the biometrics of an individual. Although we cannot guarantee that privacy will be preserved in all applications, we claim that an inherent feature of our approach is that it tends to preserve privacy.

Towards Self-Sustainability: The imaging pipeline of a typical camera involves pixel readout, analog-to-digital (A/D) conversion, signal processing, and transmission. The power consumed by each of these steps, and hence the complete pipeline, is approximately linear in the number of pixels. Since a minimalist system uses an extremely small number of pixels, it consumes orders of magnitude less power than a typical camera. As a result, a minimalist camera can be

designed to function using power harvested from just the light falling upon it, without using an external power supply or a battery. In other words, minimalist cameras can be completely self-sustaining and hence more widely deployed.

To achieve minimalist vision, our key insight is to allow each pixel to have an arbitrary shape. We refer to such a pixel as a "freeform pixel." We show that a freeform pixel performs a linear projection of the scene, allowing us to model a collection of such pixels as a single layer in a neural network. Thus, a minimalist vision system, comprising both the camera and inference network, can be modeled as one network. For a given task, such as monitoring the indoor workspace in Fig. 1(a), we use a video captured from an auxiliary camera to train the network in Fig. 1(b). The trained network reveals both the shapes of the freeform pixels and the weights of the inference network. Then, a camera (Fig. 1(c)) is fabricated, where each freeform pixel is implemented using an optical mask and a photodetector. In Fig. 1(a), the results (people count, door status, and zone occupancy) produced by a camera with only 8 freeform pixels are overlaid on the scene image.

We have conducted extensive synthetic and real experiments that show freeform pixels can solve lightweight tasks using orders of magnitude fewer measurements than a traditional camera. We have used our prototype minimalist camera to demonstrate a variety of tasks: monitoring an indoor space (with 8 pixels), measuring room lighting (with 8 pixels), and estimating traffic flow (with 8 pixels). Finally, we show that our prototype can be powered using just the light falling on it. Under indoor lighting, it can read out and wirelessly transmit the measurements made by 24 freeform pixels at 30 frames per second without the use of an external power supply or a battery.

2 Related Work

Our work is inspired by Pooj *et al.* [28], who introduced the concept of a minimalist camera, where each pixel is a combination of an optical mask and a photodetector. In their work, each mask was handcrafted to solve simple vision tasks such as intrusion detection and object speed estimation. Our work introduces the idea of a freeform pixel that can be automatically designed using training data for any given task. Our key observation is that a camera with freeform pixels can be modeled as the first layer of a neural network. Once the network has been trained, the first layer is used to fabricate the camera, and the rest of the network is used for inference. Furthermore, we show that minimalist cameras can be fully self-powered, making them more easily deployable than traditional cameras.

Our work is closely related to deep optics, an emerging field that jointly designs optics and software using deep learning [40]. Sitzmann *et al.* [30] used this approach to design an optical element for improved image quality. Subsequently, multiple works have used the approach to design optics for image enhancement [13,24,31] and depth estimation [7,15,41]. A similar approach has been taken to design imaging lenses using differentiable ray tracing [9,20,32] and differentiable proxy functions [37]. Tseng *et al.* [36] used this technique to design

a metasurface lens with improved image quality. In each of these works, a differentiable model for the camera's optics is incorporated into a neural network, and the optics is designed by training the network for the specific goal. While we follow a similar approach, our motivation is different. Rather than design optics to enhance image quality or improve task performance, we design cameras that seek to preserve privacy and be self-sustaining by taking the smallest number of measurements.

Prior work has also demonstrated the use of optics in existing network architectures to reduce the computations required during inference. Lin *et al.* [22] fabricated an entire image classification network using layers of diffractive optics. Others have explored hybrid approaches that implement just the first layer of a convolutional network in optics. In [8], angle-sensitive pixels were used to convolve the image with a set of commonly used filters, while in [6], optical phase masks were used to implement learned filters. The goal of our work is different; it is to minimize the number of visual measurements needed for a task, not to reduce the computations in a trained network.

Duarte *et al.* [12] proposed a single pixel camera that captures compressive measurements of a scene. While both the single pixel camera and our minimalist camera capture linear projections of the scene, the single pixel camera uses thousands of measurements, acquired in series using a single detector, to reconstruct an image of the scene. Image and scene reconstruction has also been demonstrated using an image sensor that views the scene through an amplitude mask [2], a phase mask [4], and a diffuser [1]. While all of the above approaches aim to reconstruct an image or 3D scene, the minimalist camera circumvents image reconstruction and seeks to directly solve the task using the smallest number of measurements.

Several works have explored imaging architectures that preserve an individual's privacy while still capturing enough information to perform a task. Some of them attempt to eliminate visual details related to biometrics in captured images by using low-resolution image sensors [10] and time-of-flight sensors [17].[1] Others have approached the problem by introducing optical blur [26, 27], by performing image processing in the analog domain before readout [33], or by designing optical elements that preserve the visual feature of interest while eliminating privacy-related details [16, 34]. Since our approach is to minimize the number of visual measurements, we claim that we implicitly preserve privacy. We demonstrate this via simulations by showing that our freeform pixels are unable to perform face recognition with meaningful accuracy.

It is known that cameras are power-hungry—the image sensor alone can consume hundreds of milliwatts [21]. Nayar *et al.* [25] demonstrated a self-powered camera with 30×40 pixels that harvests energy from the light falling on its sensor to read out full images. The harvested energy, however, was insufficient

[1] Unrelated to privacy preservation, Torralba *et al.* [35] demonstrated image classification using a large dataset of very low resolution (32×32) images. In our experiments, we compare the performance of our minimalist cameras with low-resolution traditional cameras of different resolutions.

for wireless transmission of the images. Since our minimalist approach results in a small number of pixels, our camera is able to both read out and wirelessly transmit the measurements using just the light falling on it. This is an important feature of our approach; we aim to develop a completely self-sustaining camera that does not need to be tethered and hence can be more widely deployed.

3 Freeform Pixels

Since the advent of digital imaging, cameras have used square pixels on a regular grid to record images. While some other pixel tessellations have been suggested in the past [3, 14], square pixels have persisted as the standard sensing element. We posit that there exists a large class of vision tasks for which the square pixel forces the camera to capture significantly more measurements than needed.

Consider the traditional camera model shown in Fig. 2(a), where a single square pixel (detector) receives light from a scene patch. When the pixel is small, there is a good chance that it will measure information that is not relevant to the task. If the pixel is large, its measurement may include information germane to the task, but it may also be corrupted by unrelated information. In short, if we are interested in capturing the smallest number of measurements for a task, square pixels are almost guaranteed to be the wrong choice.

We propose freeform pixels that can take on an arbitrary shape. As shown in Fig. 2(b), a freeform pixel can be implemented by placing an optical mask in front of a photodetector. While we have shown a binary mask in Fig. 2(c), each point on the mask can have an arbitrary transmittance. Let us denote the transmittance function as $M(x, y)$, where $0 \leq M(x, y) \leq 1$. If we assume that the detector is infinitesimally small, then the measurement p produced by it is:

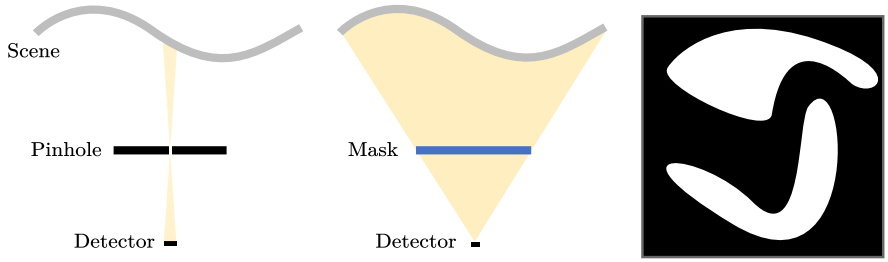
$$p = \iint_{x,y} I(x, y)\, M(x, y)\, dx\, dy\,. \tag{1}$$

Here, $I(x, y)$ is a projection of the 3D scene onto the plane of the mask, where the center of projection is the detector. The above expression shows that, in effect, each freeform pixel performs a linear projection of the scene.

3.1 Minimalist Camera in a Network

Since a freeform pixel performs a linear projection, a set of such pixels can be modeled as a single fully-connected layer in a network, without any bias terms. Based on this observation, we can construct a single network for a task, such as the one in Fig. 1(b), that includes freeform pixels and an inference network. The data for training this network is collected using the training camera shown in Fig. 1(c). Once the network is trained, the learned weights of the first layer are used to fabricate (print out) the masks of the freeform pixels. The smallest set of freeform pixels needed to solve a task constitutes a minimalist camera.

It is important to note that the square pixels found in a traditional camera are a special case of freeform pixels. Thus, if we do not limit the number of pixels used, a minimalist camera can solve any task that a traditional camera can. In general, minimalist vision significantly reduces the number of pixels needed to



(a) Traditional Pixel (b) Freeform Pixel (c) Mask

Fig. 2: A freeform pixel can have an arbitrary shape. (a) A single pixel in a traditional camera is square and captures light from a small patch in the scene. (b) A freeform pixel uses a detector and an optical mask to implement any pixel shape. (c) Example of an optical mask. While this mask is binary, a mask can have any continuous transmittance function.

solve lightweight tasks. As a task becomes more complex, a larger number of freeform pixels would be needed, and the benefits of minimalist vision diminish. For fine-grained tasks (e.g. optical flow or face identification), the number of freeform pixels needed can be expected to approach that of a traditional camera. In short, the use of freeform pixels only serves to dramatically reduce the number of measurements needed to solve a task.

3.2 Sensor Model

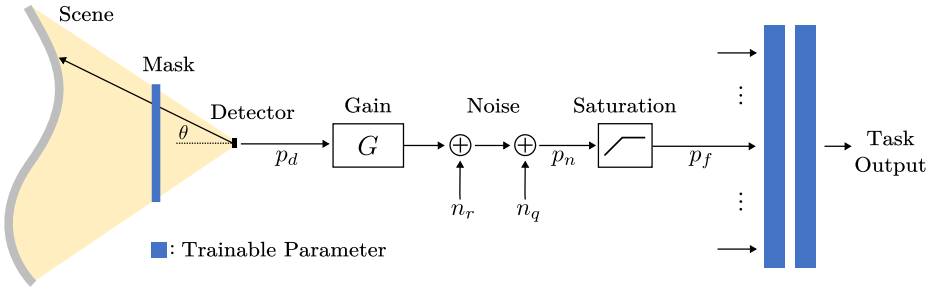
While a freeform pixel gives us flexibility, it is subjected to a set of physical constraints. First, the mask transmittance must be positive and cannot be greater than 1. Second, the detector will have a directional response and a non-zero active area. Finally, the detector will have a limited dynamic range, and its output will include noise. It is important to take all of these factors into consideration when modeling the first layer (freeform pixels) of the network. We now describe the complete sensor model we have developed and how it can be incorporated into the network.

Optics: As shown in Fig. 3, the detector placed behind the mask is expected to have a response that varies with the direction θ of the incoming light. We can represent the directional response as a function $d(x, y)$; it acts like a vignetting function with attenuation that increases with θ. Therefore, the light field received by the detector can be modeled as $I(x, y)\, M(x, y)\, d(x, y)$.

In practice, any detector would have a non-zero active area. The effect of this active area can be modeled by blurring $I(x, y)$ with a kernel $b(x, y)$, the width of which equals that of the active area. The total light energy received by the detector can therefore be expressed as:

$$p_d = \iint_{x,y} \big(I(x, y) * b(x, y)\big)\, M(x, y)\, d(x, y)\, dx\, dy\,. \tag{2}$$

As mentioned before, the value of the mask transmittance function $M(x, y)$ must lie between 0 and 1. When we model a pixel as a part of a network,



(a) Optical Effects (b) Detector Characteristics (c) Inference Network

Fig. 3: A minimalist camera as a part of a network. (a) The optical effects within a freeform pixel include the attenuation due to the mask, the detector's directional response, and its active area. (b) The detector output is amplified by a gain, degraded by readout and quantization noise, and clipped by the finite dynamic range of the detector. (c) The output p_f of the freeform pixel is fed into the inference network, which uses the outputs of all the pixels of the camera to produce the task output.

however, it is desirable to let all the trainable parameters be unbounded. To this end, we define $M_t(x, y)$, such that $M(x, y) = \sigma(M_t(x, y))$, where σ is the sigmoid function. The trainable parameters are now represented by $M_t(x, y)$, and the corresponding $M(x, y)$ is guaranteed to lie between 0 and 1.

Detector: An ideal detector would measure p_d, the total light energy that it receives. A real detector, however, has a gain, noise characteristics, and a finite dynamic range, as illustrated in Fig. 3. First, p_d is amplified by a gain G when the detector converts the incident irradiance to an analog signal. When this analog signal is read out and converted to a digital number, read noise and quantization noise are added, which can be modeled as Gaussian noise ($n_r \sim \mathcal{N}(0, \sigma_r^2)$) and uniform noise ($n_q \sim \mathcal{U}(0,\, p_{lsb})$),[2] respectively. Therefore, the final output of the pixel is:

$$p_n = G\, p_d + n_r + n_q\,. \tag{3}$$

While there are additional sources of noise, such as photon noise and dark current, we only model read and quantization noise since they are the dominant noise sources in our system. Finally, the detector saturates at a maximum measurement p_{max}. Saturation poses a problem during network training because the gradient of a saturated measurement with respect to the trainable parameters in the mask is 0. To avoid such vanishing gradients, we clip the value p_n using a clipping function with a small positive slope in the region of saturation:

$$p_f = \begin{cases} p_n & p_n \le p_{max} \\ \alpha\,(p_n - p_{max}) + p_{max} & p_n > p_{max} \end{cases}, \tag{4}$$

where α is a small, positive value.

[2] p_{lsb} is the brightness value corresponding to the least significant bit of the analog-to-digital converter.

p_f is the final output of a freeform pixel, which serves as an input to the inference network, as shown in Fig. 3. The sensor model described above has a major impact on the final "shape" of a freeform pixel. For instance, a detector's limited dynamic range forces the freeform pixel to "open up" to measure enough light to overcome the measurement noise. In the supplemental material, we show that when the above sensor model is not incorporated into the network during the training process, we obtain freeform pixels that perform poorly.

4 A Toy Example

We begin our empirical evaluation of freeform pixels with a synthetic example. The task is to count the number of patches in an image, which is akin to real tasks that involve counting objects such as people or cars. Figure 4(a) shows one such image with 10 patches. In each generated image, the number of patches can vary from 0 to 10, and each patch is assigned a random position, brightness, and size (within a range). To simulate occlusion effects, the patches are allowed to partially overlap one another. Variations in local illumination are simulated by multiplying each image with a smoothly-varying sinusoid with randomly chosen parameters. We synthesized a set of 1,000,000 images for training, 100,000 images for validation, and 250,000 images for testing.

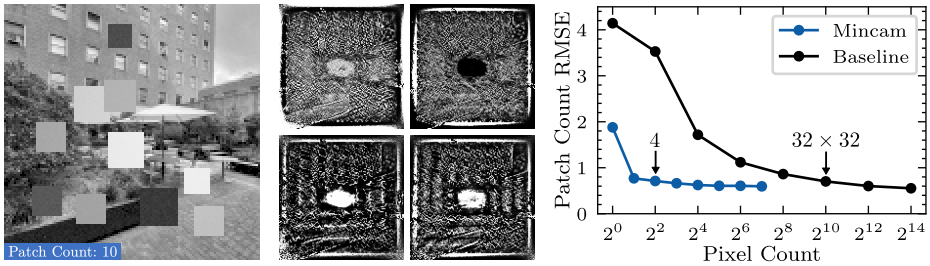
We trained minimalist cameras (mincams) to count the number of patches, starting with 1 freeform pixel up to 128 freeform pixels, incrementing in powers of 2. The parameters of the sensor model were chosen to be similar to that of a real photodetector. The inference network contains 2 hidden layers, each 128 units wide, with a leaky ReLU as the activation function. The masks of each minimalist camera were initialized with uniform noise, $\mathcal{U}(0.08, 0.12)$, and the network was trained by minimizing the cross-entropy loss using the Adam optimizer [19].

We compare the performance of the mincams with a traditional camera, where the output of the camera is used as the input to an inference network that is identical in structure to that of the minimalist camera. We refer to this combination of a traditional camera and inference network as the baseline camera. As we lower the resolution of the baseline camera, each pixel simply integrates the light within a larger square. Put another way, the baseline camera can be viewed as a minimalist camera with *fixed* masks, where each mask is a box function.

Figure 4(b) shows the learned freeform pixels for a minimalist camera with 4 pixels. These freeform pixels achieve 0.71 root-mean-square error (RMSE) in counting patches, on par with the performance of a 32×32 baseline camera (see Fig. 4(c)). This translates to a $256\times$ reduction in pixel count. This toy example demonstrates that with enough training data, freeform pixels can achieve high performance on a lightweight task using orders of magnitude fewer pixels.

5 Camera Architecture

Figure 5(a) shows the prototype of the minimalist camera that we have designed and fabricated. It has a total of 24 freeform pixels. The masks of all 24 freeform pixels are printed on a single sheet of transparency film using an inkjet printer.



(a) Synthetic Image (b) Freeform Pixels (c) Counting Performance

Fig. 4: Reduction in pixel count with freeform pixels. (a) The task is to count the number of patches in an image (up to 10), where the patches have random locations, brightnesses, and sizes. We trained minimalist cameras with an increasing number of freeform pixels (up to 128). (b) The learned freeform pixels for a 4-pixel minimalist camera. (c) The counting performance of a minimalist camera with these 4 freeform pixels is on par with that of a 32×32 baseline camera. This corresponds to a $256\times$ reduction in pixel count. Note that the x-axis is scaled logarithmically.

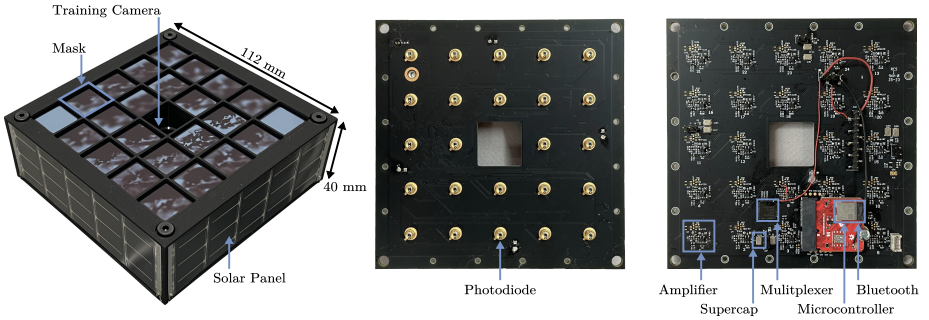
The masks can be interchanged by simply sliding a new transparency into a slot in the camera's chassis.[3] Each mask is $16 \times 16\,\text{mm}^2$ and is placed 11.4 mm above its detector; this corresponds to a $70^\circ \times 70^\circ$ field-of-view for each freeform pixel.

Each detector is a photodiode (Hamamatsu S9119-01), and the array of 24 detectors are arranged on a custom-designed imaging board, the front and back of which are shown in Fig. 5(b) and Fig. 5(c), respectively. The output of each detector is connected to a transimpedance amplifier which converts the detector's photocurrent to a voltage. The voltages of the 24 freeform pixels are passed through a multiplexer to a microcontroller (STM32WB5MMG), which performs A/D conversion and then wirelessly transmits the measurements to a remote receiver using Bluetooth Low Energy (BLE). A traditional camera (Basler daA1920-160uc) with a 3 mm lens is attached to the center of the minimalist camera. This camera is only used to capture videos for training the masks of the minimalist camera and to compare its performance with baseline cameras of different resolutions.

Since the minimalist camera generates just a handful of measurements, it consumes very little power during readout and wireless transmission. This allows us to make our prototype completely self-powered. As seen in Fig. 5(a), a solar panel (PowerFilm MP3-37) is attached to each of the four sides of the camera to harvest energy from the light falling on it. Since the light incident upon the camera, and hence the harvested energy, can vary over time, the solar panels are connected to an 88 mF supercapacitor (see Fig. 5(c)). In an indoor environment, these panels harvest enough energy to power the camera without using a battery or external power supply.

Figure 6 shows the camera operating in fully self-powered mode. In this demonstration, the ambient illumination falling on each of the camera sides is

[3] If a spatial light modulator (SLM) is used in place of the transparency, the masks can be changed via software without any alteration to the hardware.



(a) Minimalist Camera (b) Imaging Board (Top) (c) Imaging Board (Bottom)

Fig. 5: Hardware prototype of a minimalist camera. (a) The masks of the pixels are printed on a single transparency, and the corresponding detectors are arranged on the imaging board in (b). (c) The back of the imaging board shows the key components of the camera, including an amplifier for each pixel, a supercap, a multiplexer, and a microcontroller that is Bluetooth enabled. Attached to each side of the camera is a thin solar panel. The energy harvested from the four panels is sufficient for the camera to function in a fully self-powered mode in an indoor environment (see Fig. 6).

roughly 600 lux. The camera is able to read out and wirelessly transmit the 24 pixel measurements at 30 frames per second. It can continue to function at lower light levels by simply lowering its framerate. It should be mentioned that the current firmware of the camera is far from optimized. It can be made significantly more power-efficient, enabling the camera to function at much lower light levels. In our lightweight vision experiments, we tethered the camera to a benchtop data acquisition system rather than using the self-powered mode, as this configuration made hardware debugging and synchronization with the training camera easier.

6 Lightweight Vision: Experiments

We have used our camera prototype to evaluate the power of freeform pixels in a variety of lightweight vision tasks.

6.1 Workspace Monitoring

In our first application, we use minimalist vision to monitor an indoor space. Consider the workspace shown in Fig. 7(a). In this scenario, people enter/exit the space, move around, and occupy different zones. Our goal is to monitor the room by the counting of number people in it (from 0 to 8), determining which zones are occupied, and detecting when the door is open. As people move around the space, they occlude each other and different parts of the scene, making each of the above tasks more challenging. Furthermore, over time, the lighting of the space can change dramatically. We captured a one-hour video[4] using the training camera to generate a minimalist camera that can solve all of the above tasks.

[4] No information regarding the identities of individuals in the videos was acquired, stored, or used in the experiments. All of the videos were captured after obtaining signed permissions from the participants.

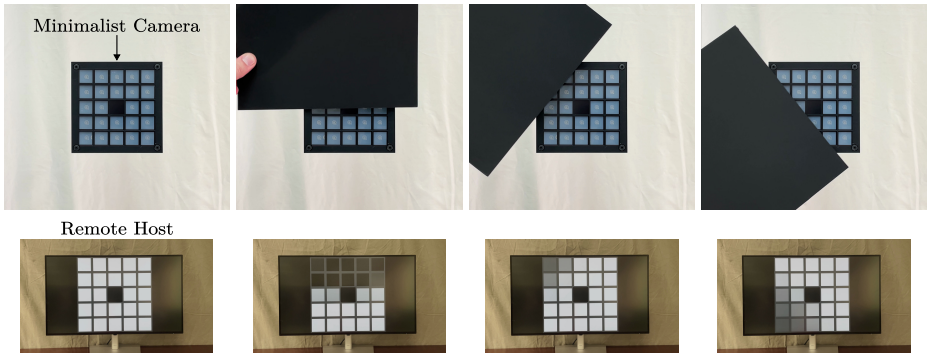
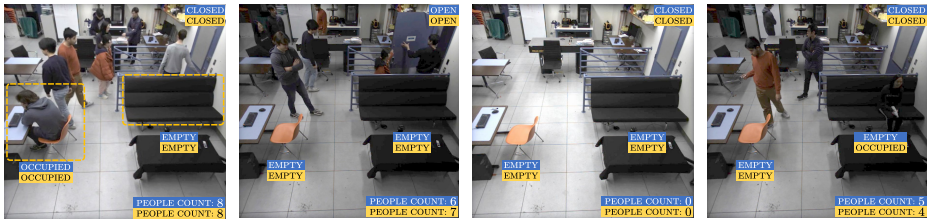


Fig. 6: Minimalist camera in fully self-powered mode. The prototype can be entirely powered by just the light falling on it. In a well-lit indoor environment, it can can read out and wirelessly transmit measurements from 24 pixels at 30 frames per second. In this demonstration, the mask of each pixel is uniform in transmittance. A black sheet is moved over the array of pixels, and the wirelessly received measurements are displayed on a remote host shown below. Please see the supplemental video.

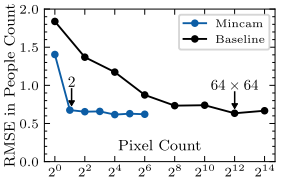
The video is divided into contiguous segments of 40 minutes for training, 10 minutes for validation, and 10 minutes for testing. In each frame of the video, ground truth labels for the tasks are specified.

We generated freeform pixels by training the minimalist camera network, as described in Sec. 3. In Fig. 7(b), we plot the people counting performance of simulated minimalist cameras and baseline cameras with varying pixel counts. A minimalist camera with 2 freeform pixels achieves 0.68 RMSE in the number of people, which is comparable to that of a 64×64 baseline camera. This translates to a $2048\times$ reduction in pixel count. The masks we used (two for each task) to construct the minimalist camera are shown in Fig. 7(c). The performance of this camera is seen in the first four rows of the table in Fig. 7(d). In the four images in Fig. 7(a), the blue boxes show the outputs of the system, and the yellow boxes show the ground truth. Also shown in Fig. 7(d) are people counting performances when we used 4, 8, and 16 freeform pixels. Please see the supplemental material for a video demonstration of workspace monitoring and the post-processing details.

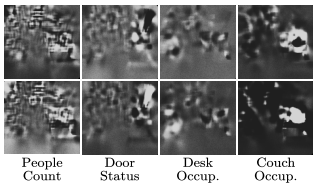
We now illustrate why a typical minimalist camera does not capture enough visual information to recognize faces. State-of-the-art vision systems have attained very high face identification rates (greater than 98%) on traditional images [5, 11, 29, 39]. Using 16 freeform pixels specifically designed for counting people, we retrained the inference network to recognize faces on a subset of the CelebA dataset [23], containing 2751 images of 100 individuals. In this simulation, the faces are scaled to cover the entire field-of-view of the minimalist camera, and each image is augmented with a small amount of noise and a random gain. Once trained, the minimalist camera achieved a recognition rate of 2.0%, suggesting that it is unable to perform meaningful face recognition in any real scenario. While this does not prove that a minimalist camera guarantees



(a) Workspace Monitoring



(b) People Counting Performance



(c) Learned Freeform Pixels

Task	# Pixels	Performance
Door Status	2	95.4%
Desk Occup.	2	99.2%
Couch Occup.	2	84.1%
People Count	2	1.16 RMSE
People Count	4	1.15 RMSE
People Count	8	1.15 RMSE
People Count	16	1.07 RMSE

(d) Prototype Mincam Performance

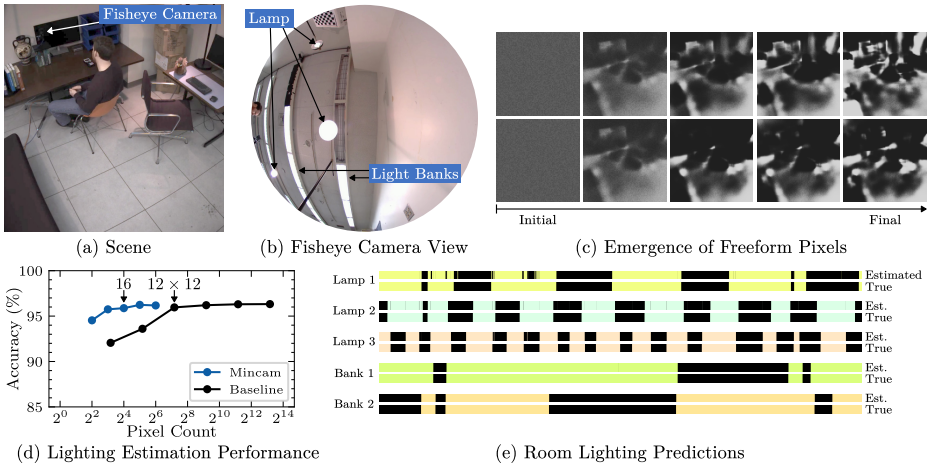
Fig. 7: Workspace monitoring. (a) We use a handful of freeform pixels to count the people in the room, determine which zones are occupied (highlighted in yellow boxes in the left image), and detect when the door is open. The outputs of the prototype camera with 8 freeform pixels are shown in blue, and ground truth is shown in yellow. (b) Minimalist cameras and baseline cameras are trained using a labeled video of the scene, and the people counting performance is plotted as a function of pixel count. For this task, the performance of a 2-pixel minimalist camera is close to that of a 64×64 baseline camera, which corresponds to a $2048\times$ reduction in pixel count. (c) The learned freeform pixels for each task after training the minimalist camera network. (d) The performance of the prototype camera for each of the tasks.

privacy, it strongly supports our conjecture that a person's identity cannot be reliably recovered from the few measurements produced by a minimalist camera.

6.2 Room Lighting Estimation

Modern buildings are moving toward optimized lighting systems to reduce their energy consumption, and hence their carbon footprint. In this context, self-sustaining minimalist cameras can be very effective in estimating the "state" of the light in a room. Coupled with people counting, a lighting-estimation camera can provide exactly the measurements needed to intelligently optimize lighting. Consider the scene shown in Fig. 8(a) with three floor lamps and two banks of overhead lights. Our goal is to use minimalist vision to determine the state (on or off) of each of the five lights as people move in and around the space. The lights are not directly visible to the camera. Therefore, the state of the lighting must be inferred from the shading in the scene, even as people move around and obstruct parts of the space. We captured a 30-minute video of the scene for training and testing. Ground truth labels were obtained using a fisheye camera placed in the scene that directly sees the lights (see Fig. 8(b)).

Using the labeled video, we trained a minimalist camera to determine the state of the room lighting by minimizing the cross-entropy loss for each light. Figure 8(c) shows the evolution of two freeform pixels during training. Each pixel



(a) Scene (b) Fisheye Camera View (c) Emergence of Freeform Pixels

(d) Lighting Estimation Performance (e) Room Lighting Predictions

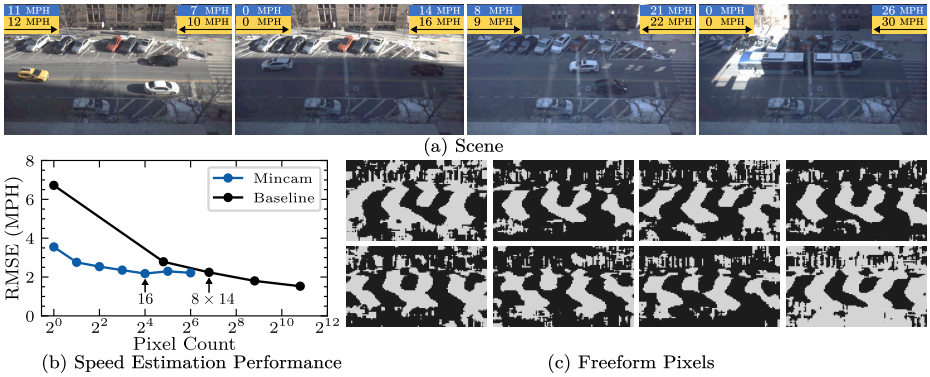
Fig. 8: Room lighting estimation. (a) A room lit by three lamps and two overhead light banks. The task is to determine which lights are turned on. A fisheye camera is placed in the scene. (b) All the lights are visible in the fisheye image, which is used to obtain ground truth labels. We trained a minimalist camera to estimate the room lighting. (c) The evolution during training of two of the freeform pixels that were initialized with random noise. (d) Performance of minimalist cameras and baseline cameras; 16 freeform pixels are sufficient to achieve the performance of a 12×12 baseline camera. (e) The performance of the camera with 8 freeform pixels, compared with ground truth. The black strips correspond to durations for which a light is off.

is initialized to uniform noise, and its shape emerges during the training process. Figure 8(d) compares the performance of minimalist cameras with baseline cameras; a 12×12 baseline camera is needed to achieve the same performance as a minimalist camera with 16 freeform pixels. We fabricated 8 freeform pixels, which could estimate the room lighting (the states of all five lights) with 94.0% accuracy. A comparison between the outputs of the camera and the ground truth is shown in Fig. 8(e). Please see the supplemental video.

6.3 Traffic Monitoring

Minimalist cameras can be attached, without cables or external power, to poles or buildings to monitor traffic. In Fig. 9(a), the task is to estimate the average traffic speed in both directions (left and right). In this system, the minimalist camera uses the temporal history of its pixel measurements over a period of one second to perform the task. The inference network outputs two values: the left and right traffic speeds, in miles per hour. We collected training data by capturing a video of the scene over an entire day and randomly extracted five-minute video clips for validation and testing. The ground truth labels were obtained by applying an off-the-shelf object detector [18] to the captured video to track individual cars. The network was trained by minimizing the mean squared error between the predicted and ground truth traffic speeds.

Figure 9(b) compares the performance of minimalist cameras with that of baseline cameras. We fabricated the 8 freeform pixels shown in Fig. 9(c), which



(a) Scene

(b) Speed Estimation Performance

(c) Freeform Pixels

Fig. 9: Traffic speed estimation. (a) The task is to estimate the average traffic speed in both the left and right directions. A video of the scene captured over an entire day was used to train minimalist cameras. (b) The performance of minimalist and baseline cameras. A minimalist camera with 16 freeform pixels achieves the same performance as an 8×14 baseline camera. (c) We fabricated a minimalist camera with 8 freeform pixels, which can monitor traffic speed with 2.30 RMSE in miles per hour. In (a), the outputs of the camera are shown in blue, and the ground truth in yellow.

were able to estimate the left and right traffic speeds with an RMSE of 2.30 miles per hour. Please see the supplemental material for a video of traffic monitoring and details of the network training and post-processing.

7 Discussion

We have introduced the concept of freeform pixels and shown how they can be effective in solving lightweight vision tasks using just a handful of measurements. There are several directions in which we plan to extend our work. First, in place of the printed transparency we used for our optical masks, a spatial light modulator, such as a liquid-crystal display, can be used to set the shapes of the masks electronically. This would allow us to change the functionality of the minimalist camera as a function of time. This also implies that different tasks can be time-multiplexed, allowing us to use more freeform pixels for any given task. In addition, spatio-temporal control of the masks would allow us to extract more revealing visual features, particularly in the case of dynamic scenes.

While our current notion of a freeform pixel performs a linear projection of the scene, we are interested in generalizing the concept so that it can perform more advanced optical mappings. For instance, by using lenses in addition to a mask, each pixel can be designed to apply a convolution to the scene with a pre-trained kernel. Such a system can also be modeled as a part of a network and has the potential to solve more sophisticated tasks.

With the above enhancements, we believe minimalist cameras can be designed to perform a wider range of vision tasks, while still guaranteeing privacy protection and self-sustainability. Ultimately, our goal is to use minimalist vision to address existing needs in the fields of environment sensing, wildlife monitoring, crowd and traffic analysis, and energy conservation.

Acknowledgements

This work was supported by the Office of Naval Research (ONR) under award number N00014-21-1-2378. The authors are grateful to Behzad Kamgar-Parsi for his support and encouragement. The authors also thank Carl Vondrick for his technical feedback all through the project and Mikhail Fridberg for his help with designing the camera electronics.

References

1. Antipa, N., Kuo, G., Heckel, R., Mildenhall, B., Bostan, E., Ng, R., Waller, L.: DiffuserCam: Lensless single-exposure 3D imaging. Optica **5**(1), 1 (Jan 2018)
2. Asif, M.S., Ayremlou, A., Sankaranarayanan, A., Veeraraghavan, A., Baraniuk, R.G.: FlatCam: Thin, Lensless Cameras Using Coded Aperture and Computation. IEEE Transactions on Computational Imaging **3**(3), 384–397 (Sep 2017)
3. Ben-Ezra, M., Lin, Z., Wilburn, B.: Penrose Pixels Super-Resolution in the Detector Layout Domain. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2007)
4. Boominathan, V., Adams, J.K., Robinson, J.T., Veeraraghavan, A.: PhlatCam: Designed Phase-Mask Based Thin Lensless Camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(7), 1618–1629 (Jul 2020)
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG) (May 2018)
6. Chang, J., Sitzmann, V., Dun, X., Heidrich, W., Wetzstein, G.: Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. Scientific Reports **8**(1) (Aug 2018)
7. Chang, J., Wetzstein, G.: Deep Optics for Monocular Depth Estimation and 3D Object Detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (Oct 2019)
8. Chen, H.G., Jayasuriya, S., Yang, J., Stephen, J., Sivaramakrishnan, S., Veeraraghavan, A., Molnar, A.: ASP Vision: Optically Computing the First Layer of Convolutional Neural Networks Using Angle Sensitive Pixels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2016)
9. Côté, G., Mannan, F., Thibault, S., Lalonde, J.F., Heide, F.: The Differentiable Lens: Compound Lens Search over Glass Surfaces and Materials for Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2023)
10. Dai, J., Wu, J., Saghafi, B., Konrad, J., Ishwar, P.: Towards Privacy-Preserving Activity Recognition Using Extremely Low Temporal and Spatial Resolution Cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (Jun 2015)
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019)
12. Duarte, M.F., Davenport, M.A., Takhar, D., Laska, J.N., Sun, T., Kelly, K.F., Baraniuk, R.G.: Single-Pixel Imaging via Compressive Sampling. IEEE Signal Processing Magazine **25**(2), 83–91 (Mar 2008)

13. Dun, X., Ikoma, H., Wetzstein, G., Wang, Z., Cheng, X., Peng, Y.: Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging. Optica **7**(8), 913–922 (Aug 2020)
14. Grosche, S., Regensky, A., Seiler, J., Kaup, A.: Image Super-Resolution Using T-Tetromino Pixels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2023)
15. Haim, H., Elmalem, S., Giryes, R., Bronstein, A.M., Marom, E.: Depth Estimation From a Single Image Using Deep Learned Phase Coded Mask. IEEE Transactions on Computational Imaging **4**(3), 298–310 (Sep 2018)
16. Hinojosa, C., Niebles, J.C., Arguello, H.: Learning Privacy-Preserving Optics for Human Pose Estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (Oct 2021)
17. Jia, L., Radke, R.J.: Using Time-of-Flight Measurements for Privacy-Preserving Tracking in a Smart Room. IEEE Transactions on Industrial Informatics **10**(1), 689–696 (Feb 2014)
18. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8 (2023)
19. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (May 2015)
20. Li, Z., Hou, Q., Wang, Z., Tan, F., Liu, J., Zhang, W.: End-to-end learned single lens design using fast differentiable ray tracing. Optics Letters **46**(21), 5453–5456 (Nov 2021)
21. LiKamWa, R., Priyantha, B., Philipose, M., Zhong, L., Bahl, P.: Energy Characterization and Optimization of Image Sensing Toward Continuous Mobile Vision. In: Proceeding of the International Conference on Mobile Systems, Applications, and Services (MobiSys). Association for Computing Machinery, New York, NY, USA (Jun 2013)
22. Lin, X., Rivenson, Y., Yardimci, N.T., Veli, M., Luo, Y., Jarrahi, M., Ozcan, A.: All-optical machine learning using diffractive deep neural networks. Science **361**(6406), 1004–1008 (2018)
23. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep Learning Face Attributes in the Wild. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Dec 2015)
24. Metzler, C.A., Ikoma, H., Peng, Y., Wetzstein, G.: Deep Optics for Single-Shot High-Dynamic-Range Imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2020)
25. Nayar, S.K., Sims, D.C., Fridberg, M.: Towards Self-Powered Cameras. In: Proceedings of the IEEE International Conference on Computational Photography (ICCP) (Apr 2015)
26. Pittaluga, F., Koppal, S.J.: Privacy Preserving Optics for Miniature Vision Sensors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2015)
27. Pittaluga, F., Koppal, S.J.: Pre-Capture Privacy for Small Vision Sensors. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(11), 2215–2226 (Nov 2017)
28. Pooj, P., Grossberg, M., Belhumeur, P., Nayar, S.K.: The Minimalist Camera. In: British Machine Vision Conference. BMVA Press, Newcastle, UK (Sep 2018)
29. Rao, Y., Lu, J., Zhou, J.: Attention-Aware Deep Reinforcement Learning for Video Face Recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)

30. Sitzmann, V., Diamond, S., Peng, Y., Dun, X., Boyd, S., Heidrich, W., Heide, F., Wetzstein, G.: End-to-End Optimization of Optics and Image Processing for Achromatic Extended Depth of Field and Super-resolution Imaging. ACM Transactions on Graphics **37**(4) (Aug 2018)
31. Sun, Q., Tseng, E., Fu, Q., Heidrich, W., Heide, F.: Learning Rank-1 Diffractive Optics for Single-Shot High Dynamic Range Imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2020)
32. Sun, Q., Wang, C., Fu, Q., Dun, X., Heidrich, W.: End-to-End Complex Lens Design with Differentiable Ray Tracing. ACM Transactions on Graphics **40**(4) (Jul 2021)
33. Tan, J., Khan, S.S., Boominathan, V., Byrne, J., Baraniuk, R., Mitra, K., Veeraraghavan, A.: CANOPIC: Pre-Digital Privacy-Enhancing Encodings for Computer Vision. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME) (Jul 2020)
34. Tasneem, Z., Milione, G., Tsai, Y.H., Yu, X., Veeraraghavan, A., Chandraker, M., Pittaluga, F.: Learning Phase Mask for Privacy-Preserving Passive Depth Estimation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. LNCS, Springer Nature Switzerland (Oct 2022)
35. Torralba, A., Fergus, R., Freeman, W.T.: 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(11), 1958–1970 (Nov 2008)
36. Tseng, E., Colburn, S., Whitehead, J., Huang, L., Baek, S.H., Majumdar, A., Heide, F.: Neural nano-optics for high-quality thin lens imaging. Nature Communications **12**(1), 6493 (Nov 2021)
37. Tseng, E., Mosleh, A., Mannan, F., St-Arnaud, K., Sharma, A., Peng, Y., Braun, A., Nowrouzezahrai, D., Lalonde, J.F., Heide, F.: Differentiable Compound Optics and Processing Pipeline Optimization for End-to-end Camera Design. ACM Transactions on Graphics **40**(2) (Apr 2021)
38. United Nations High Commissioner for Human Rights: The right to privacy in the digital age (Aug 2022)
39. Wang, M., Deng, W.: Deep face recognition: A survey. Neurocomputing **429**, 215–244 (Mar 2021)
40. Wetzstein, G., Ozcan, A., Gigan, S., Fan, S., Englund, D., Soljačić, M., Denz, C., Miller, D.A.B., Psaltis, D.: Inference in artificial intelligence with deep optics and photonics. Nature **588**(7836), 39–47 (Dec 2020)
41. Wu, Y., Boominathan, V., Chen, H., Sankaranarayanan, A., Veeraraghavan, A.: PhaseCam3D — Learning Phase Masks for Passive Single View Depth Estimation. In: Proceedings of the IEEE International Conference on Computational Photography (ICCP) (May 2019)