Research Article Vol. 28, No. 6 / 16 March 2020 / Optics Express 8619
Optics EXPRESS

Unsupervised deep learning for depth estimation with offset pixels

SAAD IMRAN,[1,3] SIKANDER BIN MUKARRAM,[1,3] MUHAMMAD UMAR KARIM KHAN,[2,*] AND CHONG-MIN KYUNG[1]

[1]*Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, 34141, South Korea*
[2]*Center of Integrated Smart Sensors, Korea Advanced Institute of Science and Technology, Daejeon, 34141, South Korea*
[3]*These authors contributed equally to this work*
[*]*umar@kaist.ac.kr*

Abstract: Offset Pixel Aperture (OPA) camera has been recently proposed to estimate disparity of a scene with a single shot. Disparity is obtained in the image by offsetting the pixels by a fixed distance. Previously, correspondence matching schemes have been used for disparity estimation with OPA. To improve disparity estimation we use a data-oriented approach. Specifically, we use unsupervised deep learning to estimate the disparity in OPA images. We propose a simple modification to the training strategy which solves the vanishing gradients problem with the very small baseline of the OPA camera. Training degenerates to poor disparity maps if the OPA images are used directly for left-right consistency check. By using images obtained from displaced cameras at training, accurate disparity maps are obtained. The performance of the OPA camera is significantly improved compared to previously proposed single-shot cameras and unsupervised disparity estimation methods. The approach provides 8 frames per second on a single Nvidia 1080 GPU with 1024×512 OPA images. Unlike conventional approaches, which are evaluated in controlled environments, our paper shows the utility of deep learning for disparity estimation with real life sensors and low quality images. By combining OPA with deep learning, we obtain a small depth sensor capable of providing accurate disparity at usable frame rates. Also the ideas in this work can be used in small-baseline stereo systems for short-range depth estimation and multi-baseline stereo to increase the depth range.

1. Introduction

Depth from 2D images is a well-studied, yet a challenging problem in computer vision. Depth information is critical for applications such as augmented reality (AR), 3D reconstruction, obstacle avoidance, and self-driving cars. Accurate 3D information is also essential for robotic applications such as object tracking and object grasping.

In the past, active and passive techniques have been used to extract depth of a scene. Active methods such as time-of-flight [1] and structured light [2] require an active light source, making them unsuitable for outdoor environments. On the contrary, LiDAR is an active range sensor, which is well suited for outdoor applications; however, it is typically an expensive choice. Passive methods like structure from motion (SFM) [3], depth from defocus (DFD) [4], and stereo matching [5] directly use images obtained from conventional cameras to estimate depth. As a result, these methods are cheaper, and can be used both indoors and outdoors.

Among different depth estimation schemes, stereo matching has historically most attracted researchers. This can be attributed to the similarity of stereo vision to the human binocular vision. In stereo matching, the disparity between the left and right images is used to indicate the depth of the scene. The disparity estimation across the two images is treated as a correspondence problem. A classic survey of stereo vision is given in [5] where the authors considered different approaches

#385328 https://doi.org/10.1364/OE.385328
Journal © 2020 Received 9 Dec 2019; revised 12 Jan 2020; accepted 22 Jan 2020; published 10 Mar 2020

Optics EXPRESS

used for cost-matching and cost-aggregation. However, most of the approaches discussed in [5] are outdated.

One of the problems with a stereo vision system is that it is quite bulky. A stereo vision system for depth estimation requires at least two cameras. To address this problem, numerous single sensor approaches have been proposed. The Dual Aperture (DA) camera [6] considers the blur across different channels to estimate depth. In Offset Aperture (OA) camera [7], the disparity across channels is used to indicate depth. In Offset Pixel Aperture (OPA) camera, the disparity across individual pixels is used to generate a depth-dependent disparity [8–11]. These approaches require a single camera system; thus, have a smaller footprint compared to the conventional stereo vision systems.

The OPA camera has shown promising performance, yet there is room for improvement. There are numerous challenges for depth estimation using the OPA camera, which include illumination difference across the displaced images and a very small baseline. We strongly believe that overcoming these challenges will significantly improve the utility of the OPA camera.

Deep learning can be used as a tool to improve the performance of the OPA camera. Deep learning-based approaches have shown great performance in a wide array of applications. In fact, many researchers have used deep learning for accurate depth estimation [12–16]. Most of the schemes in the past are supervised, which require the tedious job of developing the ground truth as the training data. Recently, unsupervised methods have been proposed [17–19]. These methods train a generator to create disparity maps that can be used to reconstruct the target images from the reference images. In this paper, we aim to improve disparity estimation with the OPA camera by using unsupervised approaches.

Utilizing deep learning for the OPA camera is of significant importance. By using deep learning for disparity estimation with the OPA camera, we improve the utility of the OPA camera and show that by using the right tools, the OPA camera can show competitive depth estimation performance. Also, we show that deep learning can not only be used for depth estimation with clean and refined images obtained through high PSNR cameras but also with low PSNR OPA images obtained through a small camera. The key contributions of our work are as follows.

- We discover that unsupervised deep learning with a very small baseline fails to generate accurate disparity maps and does not converge to the right solution due to vanishing gradients. We propose a solution to this problem by training with a larger baseline but using the original smaller baseline at inference.
- We obtain accurate disparity maps with the OPA camera despite shading across the matching channels, a very small baseline and low PSNR compared to conventional stereo vision systems.
- We use the stereo approach for unsupervised depth estimation and obtain much-improved disparity maps compared to previous single shot and deep learning approaches.

The rest of the paper is structured as follows. In Section 2, we present a brief literature review. The OPA camera is reviewed in Section 3. We discuss unsupervised disparity extraction with the OPA camera in Section 4. Section 5 provides experimental results with quantitative and qualitative analysis. The paper is concluded in Section 6.

2. Related work

Depth estimation of a scene is a vast field of research and numerous methods have been proposed over the past few decades. In this section, we will only discuss a small set of vision-based approaches related to our work.

Arguably, stereo matching is the depth extraction approach most investigated by researchers. Two cameras, displaced from each other, are used to observe a given scene. The disparity

Optics EXPRESS

between pixels across the two observed images gives a measure of the depth of the scene. Conventional depth estimation with stereo matching of rectified images can be divided into three tasks: obtaining the matching cost for each pixel, aggregating the matching cost of neighboring pixels, and refining the depth. To obtain the matching cost of a pixel, numerous methods have been proposed. The simplest of these is matching a single pixel across the stereo images [20]. This matching cost is highly unreliable and the quality of the disparity is improved by using a dynamic greedy algorithm for cost aggregation in [20]. A common approach for estimating matching costs is to use the sum of squared difference over a local neighborhood for a given pixel [5]. Under illumination differences, normalized cross-correlation (NCC) [5] or its variants [21] are used. For cost aggregation, local methods such as guided filtering [22], semi-global methods (SGM) [23], and global methods such as graph cuts [24] and trees [25] have been used. Local methods are not robust against noise; however, these approaches do not over-smoothen the disparity maps. On the other hand, global approaches are robust against noise at the cost of over-smoothening the disparity map. Disparity maps are refined by ensuring that they do not conflict with the observed intensity images. This is an optional post-processing step. For example, median filter and guided filters [26] have been used for depth refinement in [22] and [7], respectively.

Numerous efforts have been made in the past for monocular depth extraction as well. Classical monocular depth extraction schemes use depth cues to estimate depth. Shape from shading [27] and shape from texture [28] are examples of such approaches. Make3d [29] divides the whole image into color segments called super pixels. They make numerous assumptions about natural scenes based on some common characteristics, and used these assumptions to estimate the depth and 3D structure.

With the advent of deep learning, researchers use data-centric approaches for depth estimation. [12] uses a Siamese neural network for comparing patches to estimate the matching cost of pixels. SGM is used in [12] for cost aggregation. Later, numerous authors proposed different cost aggregation methods over the matching cost of [12], while others proposed end-to-end neural networks for disparity estimation. Authors in [30] use a Markov random field (MRF) with mean-field assumption for cost aggregation and provide an end-to-end implementation of the depth estimation system. In [31], the authors propose two additional layers for semi-global and local cost aggregation. In [32], authors propose a coarse-to-fine convolutional neural network and train it over synthetic data to deal with a variety of image transformations. A multi-scale approach is proposed in [33] for disparity estimation. Another multi-scale approach is proposed in [34], which combines the shallower and deeper features to achieve accurate disparity results. Authors in [35] propose a computationally efficient, accurate and robust network to deal with a dynamic disparity range. [36] uses a dense network to reduce the number of parameters, thereby improving the speed of disparity estimation. [37] proposes using a forest classifier with a neural network to estimate disparity. The method generalizes well and shows good performance across domains that have not been observed in the training data.

Developing the training data for supervised problems is a cumbersome job. It becomes even more difficult for disparity estimation as we have to prepare the ground truth for every pixel [38,39]. Unsupervised disparity estimation schemes can resolve this problem as these do not require ground truth disparity maps during training. Generally, these approaches use a generator to generate a disparity map. The disparity map is used to reconstruct the target image from the reference image by image warping. The generator learns to reduce the difference between the reconstructed and original target images. The seminal work of [17] was the first to propose this approach. There have been numerous enhancements to the method in [17], for both stereo [19] and monocular [18] disparity estimation.

Optics EXPRESS

Some authors have used deep learning for disparity estimation with light-field cameras [40]. Authors in [41] develop a synthetic dataset for light-field cameras. They propose an end-to-end network for mapping the 4D light field to 2D hyper-plane orientations and follow it by a regularization method. In [42], authors use two sequential convolutional neural networks to estimate the color and disparity components of the light-field images. Fixed input views and special data augmentation techniques are used in [43] to estimate disparity. Authors in [44] use dual pixels for disparity estimation. Their method uses Colmap [45,46] to generate ground truth disparities for training. Although these approaches deal with small baselines like OPA, the methods used for disparity estimation are supervised and require training images with respective ground truth.

Unlike stereo imaging, single-shot systems use a single sensor to estimate the depth of the given scene. Binary coded apertures [47] have been used in the past for depth estimation [48]. In [49], the authors try to find the optimal binary aperture pattern for accurate depth estimation. These methods are, however, computationally expensive. The camera's red, green, and blue filters have been shifted in [50,51] and [52] to generate a depth-dependent disparity across the three color channels. [53] explores different configurations of the red, green, and blue filters to maximize the light efficiency of the camera as well as estimate depth.

Displacing the color channels produces unwanted misalignment in the observed image. To circumvent this problem, DA [6] uses two apertures with a four-color sensor. Apart from red, green, and blue, the DA camera also observes near-IR. A smaller aperture is used for the near-IR channel; thus, the near-IR channel is relatively sharp compared to the rest of the channels. The blur of the green channel is compared against that of the near-IR channel to estimate depth. The OA [7] camera uses the same sensor as DA but displaces the near-IR aperture from the rest of the channels using an offset aperture. Thus, the near-IR channel image is displaced from the red, green, and blue color channels. This allows well-aligned red, green, and blue images, and better depth performance compared to DA. However, both DA and OA use the near-IR channel. Allowing near-IR to the sensor corrupts the red, green, and blue images. Furthermore, the observed near-IR is very low; i.e., the PSNR of the near-IR channels is quite poor to be used for depth estimation. To resolve this issue, [8] proposes replacing the two green pixels of the Bayer's pattern with white pixels but with offset. The disparity across the white pixels is used to estimate depth and the green color is estimated by regression over the observed white, red and blue intensities in a local neighborhood. A fast hardware implementation for disparity estimation with OPA is also mentioned in [8].

Despite its numerous advantages, the disparity obtained by the OPA camera still has room for improvement. Unsupervised deep learning has been used for stereo disparity estimation; however, its performance has not been verified for challenging images such as those produced by the OPA camera and has generally been used for high-quality images [38,54]. In this work, we propose modifications to the traditional unsupervised approaches for disparity estimation to significantly improve the performance of the OPA camera.

3. Offset pixel aperture camera

The OPA camera described in [8] has a pixel aperture that is formed by partially covering the pixel diode. A graphical illustration of the structure of the OPA camera is shown in Fig. 1. The CMOS image sensor array of the OPA camera, shown in Fig. 2, consists of red, blue, and white color filters. The size of pixel and pixel aperture is 2.8 μm $\times$ 2.8 μm and 1.3 μm $\times$ 2.8 μm, respectively. For every odd row, the right side of the white pixels is covered (LW in Fig. 2) while for every even row, the left side of the white pixels is covered (RW in Fig. 2). This arrangement creates an offset across the pixels of the white channel, which is analogous to the baseline of a stereo camera. Without the offset, all pixels are aligned. However, the metal covering over the LW and RW pixels displaces their centers by 0.65 μm approximately. Thus, a horizontal

Optics EXPRESS

offset of approximately 1.3 μm between the pixel apertures results in left and right images with a depth-dependent disparity. The left and right images are obtained by skipping the odd and even rows of white channel image taken from the sensor, respectively. Due to the small offset, the disparity between the left and right images is quite small compared to conventional stereo cameras.

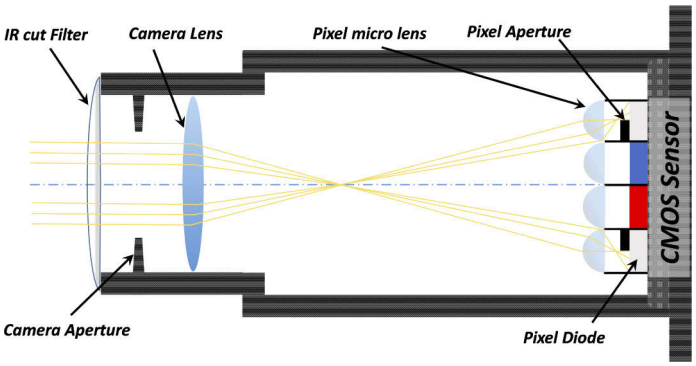


Fig. 1. Illustration of OPA camera. Apertures are offset at pixel level.

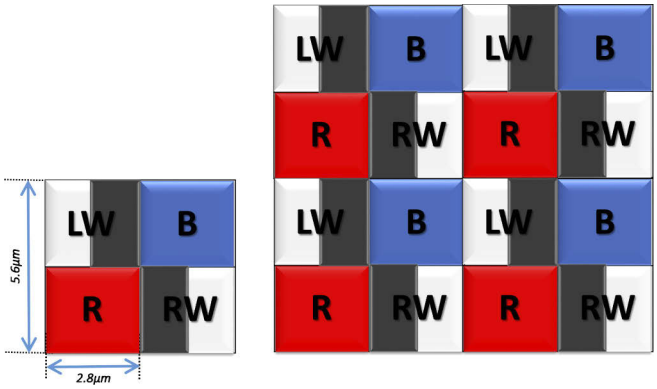


Fig. 2. Color Filter Array of the CMOS sensor. Green pixels in Bayer CFA are replaced by white pixels with offset apertures where LW and RW indicate left and right white images, respectively.

OPA has numerous advantages over competing technologies. It uses disparity instead of blur, which generally yields better depth. Also, the IR signal is not present in OPA that corrupts the RGB image in [6][7]. On the downside, the OPA sensor has disparity-dependent blur in the images, shading and low PSNR. To circumvent these problems, [8] employs a pre-processing stage. To minimize the effect of blur, first-order gradients are used for correspondence matching, i.e.,

$$I_g(x, y) = I(x + 1, y) - I(x - 1, y), \tag{1}$$

where $I(x, y)$ and $I_g(x, y)$ are the pixel intensity and the one-dimensional gradient at the position (x, y), respectively. Local normalization is used in [8] to counter the effect of shading as

$$I_N(x, y) = \frac{I_g(x, y) - \mu_\Omega(x, y)}{\sigma_\Omega(x, y)}, \tag{2}$$

Optics EXPRESS

where $\mu_\Omega(x, y)$ and $\sigma_\Omega(x, y)$ are the mean and standard deviation of pixel gradients in the $N \times N$ neighborhood centered at (x, y). For noise reduction, simple mean filtering was applied to the normalized images.

To find the pixel-to-pixel correspondence between left and right images for the cost volume, [8] uses Sum of Absolute Difference (SAD) as the cost, i.e.,

$$C(x, y, d) = \sum_{x,y \in \Pi} |I_N^{(L)}(x, y) - I_N^{(R)}(x + d, y)|, \tag{3}$$

where $I_N^{(L)}$ and $I_N^{(R)}$ are the left and right images after pre-processing, respectively, and Π is the window centered at (x, y).

SGM [23] is used for cost aggregation in [8] to improve the cost volume. After cost aggregation, disparity at each pixel is chosen by the winner-takes-all approach. A filter based on Poisson assumption over image noise is used for disparity-map refinement. A dedicated hardware implementation is also proposed to generate disparity maps from OPA images.

4. Unsupervised disparity estimation using OPA

In this section, we propose using deep learning for disparity estimation with OPA owing to the challenging nature of OPA images.

As discussed in Section II, unsupervised disparity estimation is based on the principle of reconstructing one of the stereo images from a generated disparity map. Both of the stereo images are input to the generator and the generator is trained to generate disparity maps. The parameters of the generator are adjusted such that the reconstruction loss of the original and generated image(s) is minimized.

4.1. OPA disparity estimation based on stereo

Since two images are obtained from the OPA camera, the left white channel image I_L and the right white channel image I_R, it is possible to use the reconstruction error with both images rather than one. A cyclic Generative Adversarial Network (GAN) [55] allows us to consider the reconstruction loss of both images. Cyclic GANs have been used for domain translation in the past. However, in this work we use cyclic GANs to generate an image of a scene from a different viewpoint. In more detail, one half of the cyclic GAN tries to reconstruct the right white image, I'_R, using I_R and I_L, while the other half tries to reconstruct the left white image, I'_L, using I'_R and I_L. At inference, the two disparities obtained by reconstructing the two images are fused to get a final disparity.

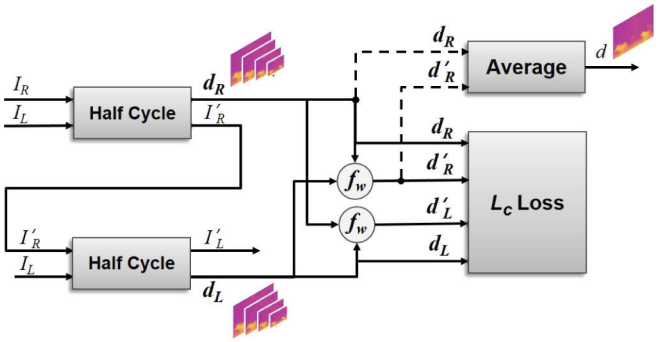


Fig. 3. Top-level block diagram of the OPA-Stereo approach. The dashed lines show the data path at inference. Bold letters indicate multi-scale results.

Optics EXPRESS

A top-level block diagram of the disparity estimation network with the stereo approach is shown in Fig. 3. There are two half-cycles in the network, each generating the disparity map at multiple scales. For comparison across viewpoints, the left disparity maps $\boldsymbol{d_L}$ are warped with the right disparity maps $\boldsymbol{d_R}$ at respective scales as in [19]. The warping function f_w used to warp an image I with the disparity map d is defined as

$$f_w(I; d) = I(x + d(x, y), y), \ \forall x, y. \tag{4}$$

In detail, warping is used to change the viewpoint of a given scene across two views with a given disparity map. For example, if I_L is the left white image and d_R is the disparity between I_L and I_R with the right white image taken as reference then $f_w(I_L; d_R)$ should be equal to I_R. Warping is always followed by bilinear interpolation as described in [56]. In the remaining text, f_w will indicate warping followed by bilinear sampling. Note that f_w is differentiable [56]. The loss between the right disparity maps $\boldsymbol{d_R}$ generated by the first half cycle and the warped disparity maps $\boldsymbol{d'_R} = f_w(\boldsymbol{d_L}; \boldsymbol{d_R})$ from the second half cycle is used to train the network prarameters. At inference, right disparity map d_R and the warped left disparity map $d'_R = f_w(d_L; d_R)$ at finest resolution are averaged to obtain the final result, i.e.,

$$d = \frac{d_R + f_w(d_L; d_R)}{2}. \tag{5}$$

Note that in $f_w(d_L; d_R)$, the disparity map d_R is used as a spatial transformation that is applied to the disparity map d_L. In simplistic terms, the disparity values at every pixel in d_L disparity map are displaced by the corresponding disparity value in d_R. For further details of applying spatial transformations in a neural network, we refer the reader to the seminal work of [56].

The half cycle used in the network is shown in detail in Fig. 4. The CNN is a GAN-based encoder-decoder neural network trained to generate disparity maps at multiple scales. The multi-scale disparities are obtained by extracting the output of the decoder at different layers. This is a common approach used in the past as well [57]. Reconstructed right images are obtained by warping multi-scale left images with the disparities obtained by the CNN. The appearance loss between the reconstructed and actual right images of the OPA camera is minimized to train the network. A discriminator is also trained to distinguish between the reconstructed I'_R and the real right images I_R of the OPA camera. The operations of the CNN are shown in Fig. 5. The outputs of the CNN are multiplied by a constant value to match the expected dynamic range of the disparity map. This constant remains fixed in all the experiments and is set to five percent of

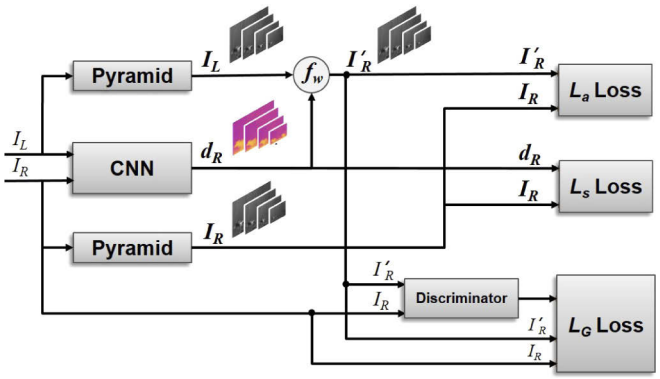


Fig. 4. A half cycle used for generating disparities in OPA-Stereo. Pyramid block subsamples images to 1/2, 1/4 and 1/8 of the original resolution while also returning the original image. Bold letters indicate multi-scale results.

Optics EXPRESS

Algorithm 1 Training Stage

Input ← Dataset, hyper parameters, network architecture, initial parameters
Output ← Updated parameters
for ∀ epochs **do**
  for ∀ batches **do**
    Forward Pass
    Read the left I_L and right I_R images
    Reconstruct $\boldsymbol{I'_R}$ with G_R
    Compute losses $L_a^{(R)}$ and $L_s^{(R)}$ using $\boldsymbol{I'_R}$, $\boldsymbol{I_R}$ and $\boldsymbol{d_R}$
    Input I'_R and I_R to D_R to compute $L_D^{(R)}$
    Input I'_R to D_R to compute $L_G^{(R)}$
    Reconstruct $\boldsymbol{I'_L}$ with G_L
    Compute losses $L_a^{(L)}$ and $L_s^{(L)}$ using $\boldsymbol{I'_L}$, $\boldsymbol{I_L}$ and $\boldsymbol{d_L}$
    Input I'_L and I_L to D_L to compute $L_D^{(L)}$
    Input I'_L to D_L to compute $L_G^{(L)}$
    Reconstruct the right disparity map $\boldsymbol{d'_R}$ from the left disparity map $\boldsymbol{d'_L}$ using $\boldsymbol{d_R}$
    Reconstruct the left disparity map $\boldsymbol{d'_L}$ from the left disparity map $\boldsymbol{d'_R}$ using $\boldsymbol{d_L}$
    Compute the loss L_c using $\boldsymbol{d'_R}$, $\boldsymbol{d_R}$ and $\boldsymbol{d'_L}$, $\boldsymbol{d_L}$
    Backward Pass
    Update the weights of G_R to minimize $L_c + L_a^{(R)} + L_s^{(R)} + L_G^{(R)}$
    Update the weights of G_L to minimize $L_c + L_a^{(L)} + L_s^{(L)} + L_G^{(L)}$
    Update the weights of D_R to minimize $L_D^{(R)}$
    Update the weights of D_L to minimize $L_D^{(L)}$
  end for
end for

the width of the OPA frames. The ResNet blocks are described in [58]. The discriminator is a simple feed-forward neural network shown in Fig. 6. Note that the whole network shown in Fig. 3 is jointly trained.

There are numerous loss functions used in this approach. Except for the GAN loss, L_G, the rest of the losses are computed and minimized at multiple scales. To ensure similarity between the reconstructed and original images, we use the weighted sum of the SSIM loss [59] and the L_1 loss as the appearance loss L_a, i.e.,

$$L_a = \frac{1}{N}\sum_{i,j}\left(\alpha\frac{1 - S_\Omega(I_{ij}, I'_{ij})}{2} + (1-\alpha)\left|I_{ij} - I'_{ij}\right|\right), \tag{6}$$

where α is the weighting factor, $S_\Omega(I_{i,j}, I'_{i,j})$ is the SSIM difference between the original and reconstructed images in the local neighborhood Ω centered at (i,j), and N is the number of pixels in the image. We used $\alpha = 0.85$ based on experiments. This loss is different from the loss used in [19] to counter the severe illumination difference between right and left OPA images. By using SSIM, the neural network is somewhat capable of identifying matching regions even under illumination differences. This is because SSIM takes into account the appearance pattern of the inputs while computing the matching cost.

The neural network of Fig. 3 produces two disparity maps. The system is self-supervised by ensuring similarity between the two disparity maps. Thus, we train the generators to minimize

Optics EXPRESS

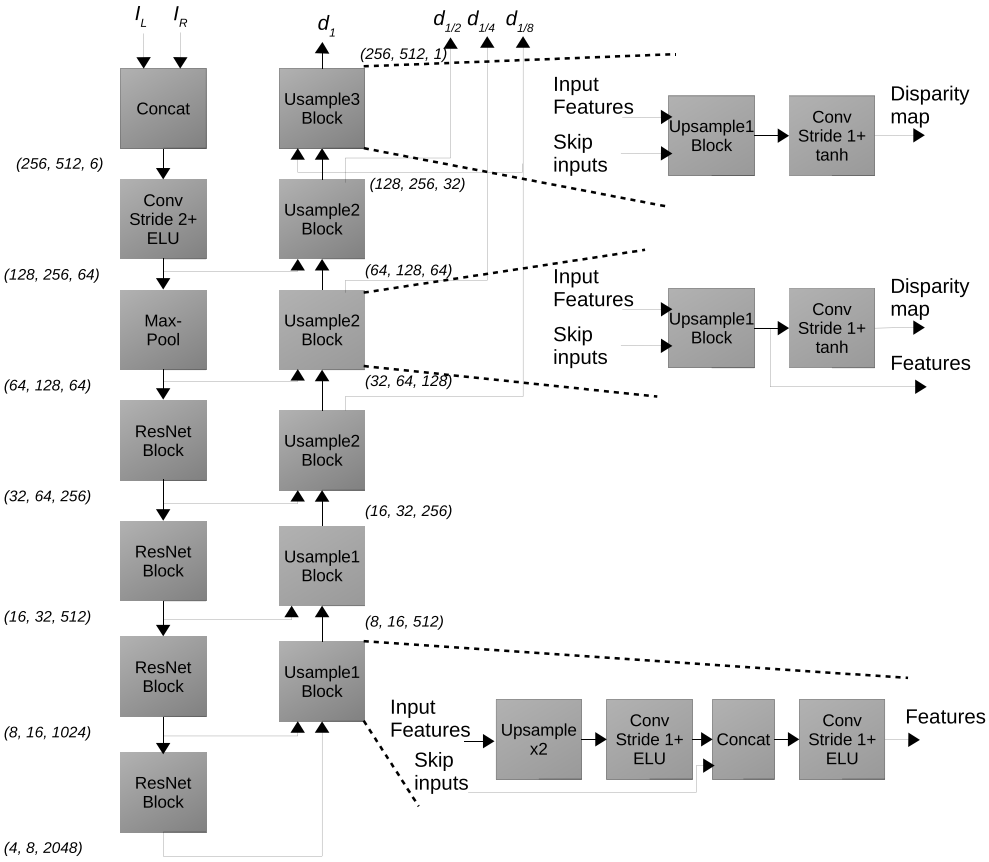


Fig. 5. Detailed operations of the CNN used in the half cycle.

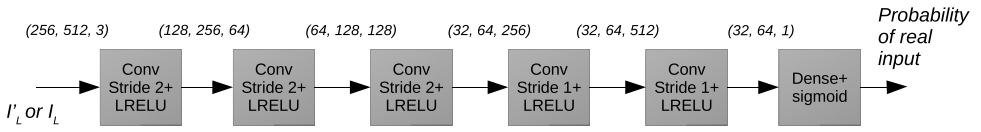


Fig. 6. The discriminator used in each half cycle.

the loss across the two disparity maps. The left-right consistency loss L_c is defined as

$$L_c = |d_L - f_w(d_R; d_L)| + |d_R - f_w(d_L; d_R)|. \tag{7}$$

With these losses, we also use the disparity smoothness loss L_s given in [18].

The adversarial loss or the GAN loss is only computed at the original resolution. For the first half cycle, the parameters of the discriminator are adjusted to minimize the loss

$$\begin{aligned} L_D^{(R)}(G_L, D_R, I_L, I_R) = &-\mathbb{E}_{I_R \sim p(I_R)}[\log D_R(I_R)] \\ &- \mathbb{E}_{I_L \sim p(I_L)}[\log(1 - D_R(f_w(I_L; d_R)))], \end{aligned} \tag{8}$$

where the generator G_L is the CNN that generates the disparity d_R while the discriminator, D_R, discriminates between the original image I_R and the reconstructed image $I'_R = f_w(I_L; d_R)$. Minimizing L_G with the discriminator requires maximizing both the terms on the RHS of (8).

Optics EXPRESS

Maximizing the first term on the RHS of (8) ensures that the output of the discriminator is close to one for the real images, whereas maximizing the second term on the RHS ensures that the output of the discriminator is close to zero for the reconstructed images. On the contrary, the weights of the CNN in Fig. 4 are updated so that the output of the discriminator is close to one with the reconstructed images. More specifically, the weights of the CNN are updated to minimize

$$L_G^{(R)}(G_L, D_R, I_L, I_R) = -\mathbb{E}_{I_L \sim p(I_L)}[\log D_R(f_w(I_L; d_R))]. \tag{9}$$

Training is performed in two alternating steps. In the first step, the weights of the CNN are updated to minimize all the losses except the loss in (8). In the next step, the weights of the discriminator are updated to minimize the loss in (8). More details on training a GAN can be found in [60]. The equations for the second half cycle are straightforward. The procedure for training is shown in Algorithm 1. We have also included Table 1 with the algorithm. Please note that in the rest of the paper, the approach of Fig. 3 will be termed as OPA-Stereo.

Table 1. Symbols used in Algorithm 1.

Symbol	Description
G_L, G_R	Image reconstruction neural networks for the first and second half-cycles
D_L, D_R	Discriminators for the GAN framework for the first and second half-cycles
I_L, I_R	Input left and right images
I'_L, I'_R	Reconstructed left and right images
$\boldsymbol{I'_L}$, $\boldsymbol{I'_R}$	Reconstructed multi-scale left and right images
d_L, d_R	Disparities obtained from the first and second half-cycles
$\boldsymbol{d_L}$, $\boldsymbol{d_R}$	Multi-scale disparities obtained from the first and second half-cycles
$\boldsymbol{d'_L}$, $\boldsymbol{d'_L}$	Multi-scale disparity maps obtained by warping the $\boldsymbol{d_R}$ and $\boldsymbol{d_L}$
L_G, L_D	GAN loss for the generator and the discriminator
L_a, L_c, L_s	Appearance, left-right consistency and smoothness losses

Detailed operations of the CNN used in the half cycle are shown in Fig. 5. The CNN has an architecture similar to the U-Net [61]. The encoder is a ResNet-50 [58]. The decoder repeatedly upsamples the encoded input to generate the disparity maps. Note that multiple disparity maps of different resolutions are generated by the CNN.

Occlusion has always been a significant problem with stereo matching and video-based depth estimation. Numerous classical methods proposed rule-based approaches for dealing with disparity, which generally revolve around the left-right consistency check. To our knowledge, occlusion has not been dealt with in unsupervised disparity estimation from stereo. Some researchers have made efforts to deal with the occlusion problem for video-based depth estimation [62–64]. However, it should be noted that occlusion is not a serious problem with the OPA camera due to its apparently small baseline. Therefore, we do not propose any special methods to deal with occlusion.

4.2. Challenges with the small baseline

The apparent baseline of the OPA camera is very small, as discussed in Section III. The OPA camera offsets the pixels of the camera by a very small value. This results in disparity typically not larger than 8 pixels.

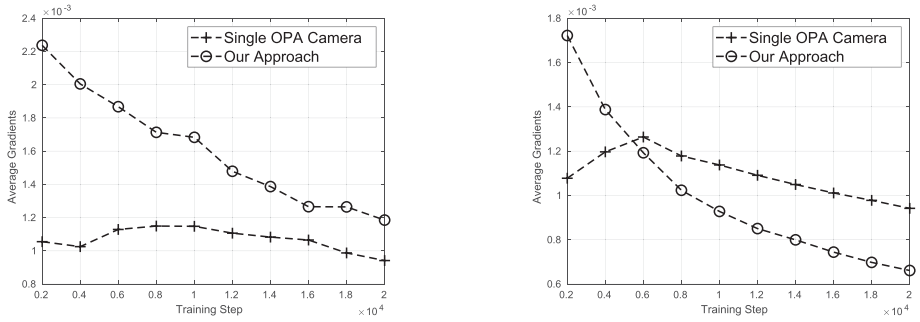
During our experiments, we note that at the initial iterations in training each generator acts as an autoencoder. In other words, if the left white image is input to the generator then it recreates the input left white image rather than the right white image at the beginning of training. Generator can easily act as an autoencoder for two reasons. First, the generator only has to act as an identity

Optics EXPRESS

function. Second, the difference between the left and right white images is generally quite small due to the small baseline. By acting as an autoencoder, the generators achieve a very small reconstruction loss. It is only after further training that the generators learn to generate the other images, and, thereby, the disparity maps.

The above situation results in vanishing gradients. Once the generator acts as an autoencoder, its reconstruction loss is already quite small. As a result, the gradients at the deeper layers are quite small which become even smaller at the shallower layers due to the multiplication of gradients in back-propagation. After a while the network stops learning, limiting the overall disparity estimation performance. Although our models use skip connections to subside the vanishing gradients problem as suggested in [58], the problem still persists.

In this work, we present a unique solution to train unsupervised neural networks for disparity estimation with a small baseline. We do not use the left and right images from the same OPA camera at training. We propose using two OPA cameras at training, and use the left and the right white images of the first and second cameras, respectively, to train the neural networks. Both OPA cameras are displaced by a larger distance compared to the pixel offset. Thus, the difference between the left and right white images is quite large, allowing the network to learn to properly reconstruct corresponding images. The training images not only contain large disparities but also small disparities. Thus, the model trained in this fashion can be directly used with the original OPA images (with a small baseline) as the model learns to generate objects at a small disparity from the input images. The issue is graphically depicted in Fig. 7, which shows the average of the gradients of first and third layers of the neural networks used for disparity estimation, respectively. For these shallow layers in the network, it is seen that the gradients of the parameters are initially small and never decrease as training proceeds if we use the left and right white images from a single OPA camera. In contrast, with our approach the gradients of the parameters are relatively large at the beginning of training and decrease afterwards showing that the shallow layers are updated as expected while learning.



(a) (b)

Fig. 7. Comparison of gradients during training if the images are taken from a single OPA camera and our approach at the (a) first and (b) third layers of the neural networks.

Previous unsupervised disparity estimation methods, such as [19], are incapable of estimating the disparity in OPA images. Although [19] has a similar neural network to ours, the loss function and the training method are different from the proposed approach. These differences allow proper estimation of disparity in OPA images. On the other hand, adapting unsupervised stereo [19] or monocular [18] schemes produces erratic results, as seen in the next section.

Optics EXPRESS

5. Experimental results

The experimental results show some interesting insights into the proposed strategy for disparity estimation with OPA. We extract qualitative and quantitative results with the proposed approach, and compare against recently proposed single shot, single sensor depth extraction cameras such as DA [6] and OA [7]. Also, we compare against the originally proposed method for depth extraction with OPA [8]. To see the effectiveness of our approach, we compare against previously proposed deep learning approaches for unsupervised stereo [19] and unsupervised monocular disparity estimation [18]. In the remaining text, we will denote the direct adaptation of the [19] and [18] for OPA images as Unsup-Stereo and Unsup-Monocular, respectively. The results show that the proposed deep learning approaches can be used to significantly improve disparity estimation with special sensors.

The dataset was gathered with the OPA camera using the method described in the previous paragraph. The dataset was divided into three categories: flat surfaces, gestures, and small objects. We used a total of 2000, 1050 and 17000 images for flat surfaces, gestures and small objects, respectively. The dataset has been gathered considering that the OPA camera estimates depth in a short range [65]. All the gathered images are of 1024×512 resolution. Around 10% of the images were used for evaluation while the rest were used for training the networks, with a separate model for a category. For quantitative evaluation, ground truth was prepared for eight images shown in Fig. 9. This ground truth was prepared using Colmap [66] following [44]. We took on average 10 images of the scene with the OPA camera from varying viewpoints. Out of the 10, one was chosen as the reference image. A 3D model of the scene was developed using the set of images and thus, the ground truth depth corresponding to the reference image was obtained. Depth at pixels where the geometric and photometric depths did not conform were excluded from the evaluation. The depths were converted to disparity corresponding to the camera system using typical methods [67]. For a given disparity map d, ground truth disparity map d^*, and total number of pixels N, the following metrics were used to evaluate disparity estimation performance:

the absolute relative difference

$$M_{AR} = \frac{1}{N}\sum_{x,y}\frac{|d(x,y) - d^*(x,y)|}{d^*(x,y)}; \tag{10}$$

the squared relative difference

$$M_{SR} = \frac{1}{N}\sum_{x,y}\frac{||d(x,y) - d^*(x,y)||^2}{d^*(x,y)}; \tag{11}$$

and the root-mean-squared error (RMSE)

$$M_{RMSE} = \sqrt{\frac{1}{N}\sum_{x,y}||d(x,y) - d^*(x,y)||^2}. \tag{12}$$

We have used the percent of bad depth (PBD) as well, which is defined as the percent of pixels for which the difference of disparity from the ground truth is greater than one pixel. Furthermore, we have also used the following threshold-based accuracy metric evaluate the performance.

$$A(t) = \frac{1}{N}\sum_{x,y} 1\left\{max\left(\frac{d^*(x,y)}{d(x,y)}, \frac{d(x,y)}{d^*(x,y)}\right) > t\right\}, \tag{13}$$

where $1\{\cdot\}$ shows the indicator function which returns 1 if the condition is true and 0 otherwise.

Optics EXPRESS

We require two images of the same scenes with the OPA camera for training, as described previously. Rather than using two OPA cameras to extract the training images, we used a single OPA camera and observed a given static scene from two predefined camera positions. The two positions were controlled by using the A-LSQ600D linearized motion controller from Zaber. We displaced the camera to different positions and obtained images at different baselines. We trained and evaluated OPA-Stereo for each baseline separately. The results are shown in Fig. 8. From the figure, it is seen that the best performance is obtained at 1 *mm*, which effectively becomes the stereo baseline during training. It would have been convenient to use two cameras for training rather than displacing a single camera. However, it is not possible to place cameras with conventional lens systems such that their lens centers are 1 *mm* apart due to the size of the cameras.

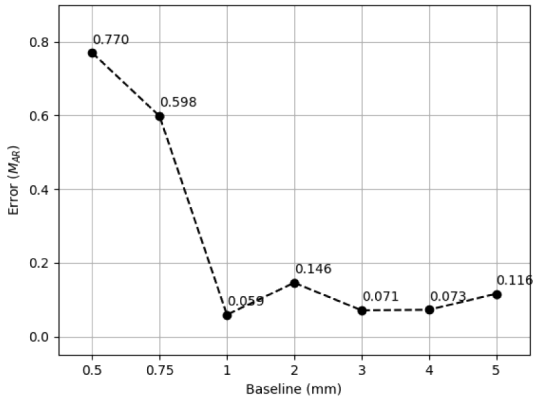


Fig. 8. Performance of OPA-Stereo with different baselines.

The experiments were conducted on a system with an Intel Core i5 processor running at a clock speed of 3.4GHz with 16GB RAM. Training was performed on a single Nvidia 1080 GPU with a memory of 8GB. The tensorflow library [68] was used for training and testing the neural networks. We used the Adam optimizer with a learning rate of 10^{-5}. The momentum and weight decay parameters were set to 0.9 and 2×10^{-4}, respectively.

5.1. Ablative analysis

OPA-Stereo combine different methods to produce disparity maps from OPA images. It is interesting to see the effect of these methods on the overall disparity estimation performance. These results are summarized in Tables 2.

Table 2. RMSE of OPA-Stereo.

Overall	0.384
Single Scale	0.489
Half cycle only	2.267
Without GAN loss	0.434
Without SSIM loss	0.63

From the results, it is seen that using the GAN loss slightly improves the overall RMSE of OPA-Stereo. Not using multiple scales to compute the loss during training also leads to degraded performance. The effect of using the SSIM loss to train the network parameters is clearly seen in the results. It is seen that depending on a single half-cycle to produce the disparity map leads to

Optics EXPRESS

extremely poor performance. Also, the exclusion of left-right consistency check produces very poor results.

5.2. Comparison with other methods

In this section, we compare our results with similar approaches proposed in the past. We compare the OPA-Stereo approach against DA [6], OA [7], previous disparity estimation system for OPA [8], Unsup-Stereo [19] and Unsup-Monocular [18]. Both qualitative and quantitative comparisons are discussed including depth sensitivity analysis.

Quantitative results for depth estimation with different approaches are shown in Table 3. To obtain these results, we used the manually labeled ground-truth shown in Fig. 9. For OA, we have used the dataset in [7] for evaluation. Note that the scene structure in the dataset used in [7] for quantitative evaluation and ours (Fig. 9) is very similar so that results can be compared across the two datasets. Also, we scaled the OA disparity maps to the same range as the OPA disparity maps for valid comparison. From the results, it is seen that the results obtained by using the methods described in this paper produce much superior results compared to OA. It is interesting to note that the PBD of the previously proposed approach for disparity estimation with OPA was higher compared to OA; however, by using deep learning the performance is significantly improved and is relatively better. Also, from the table it is seen the OPA-Stereo greatly improves the results of Unsup-Stereo and Unsup-Monocular.

Table 3. Disparity estimation performance of different methods.

Approach	M_{AR}	M_{SR}	M_{RMSE}	PBD	$A(1.25)$	$A(1.25^2)$	$A(1.25^3)$
OA [7]	0.085	0.311	0.557	3.7	N.A.	N.A.	N.A.
OPA [8]	0.161	0.764	2.874	8.262	0.752	0.906	0.966
Ref. [19]	0.81	0.972	1.01	79.02	0.077	0.149	0.213
Ref. [18]	0.273	0.672	2.297	57.73	0.283	0.358	0.471
OPA-Stereo	0.0594	0.0397	0.384	1.699	0.973	0.988	0.993

For further insight, we provide the qualitative results of different approaches in Fig. 10. The images obtained from DA, OA, and OPA are different from each other as these are different cameras with different intrinsic and spectral characteristics. However, the results are comparable as we have observed similar scenes with each of the cameras. For the hand-gesture images, the results of DA are inaccurate. OA shows slightly better results compared to DA. Further improvement is seen with the previous approach used with the OPA camera. However, it is seen that the disparity maps produced by OPA-Stereo are comparatively much better than original OPA (without deep learning). For the flat surfaces, the results of DA and OA are somewhat better for the closer surface. However, the results of DA for the farther surfaces are quite erratic and OA cannot distinguish between the two farther surfaces. Legacy approach with OPA generates erratic disparity for the closer surface. It is seen that the results of OPA-Stereo are much better compared to the rest of the methods. DA and OA produce slightly inaccurate results for the small objects scene. The results with OPA are accurate but quite blurred, so much so that it is difficult to comprehend the shape of the objects from the disparity map. On the contrary, it is seen that OPA-Stereo produces crisp and accurate disparity maps. Again we note that Unsup-Stereo and Unsup-Monocualr completely fail to extract a disparity map for any of the test images. Some more results of the OPA-Stereo are shown in Fig. 11.

Unsupervised disparity estimation methods generally show degraded performance with heterogeneous datasets, i.e., datasets with different scenes. Researchers report results with the KITTI [39] and Cityscapes [69] datasets to show the performance of their methods across different datasets. However, both KITTI and Cityscapes observe a road with a camera system installed on

Optics EXPRESS

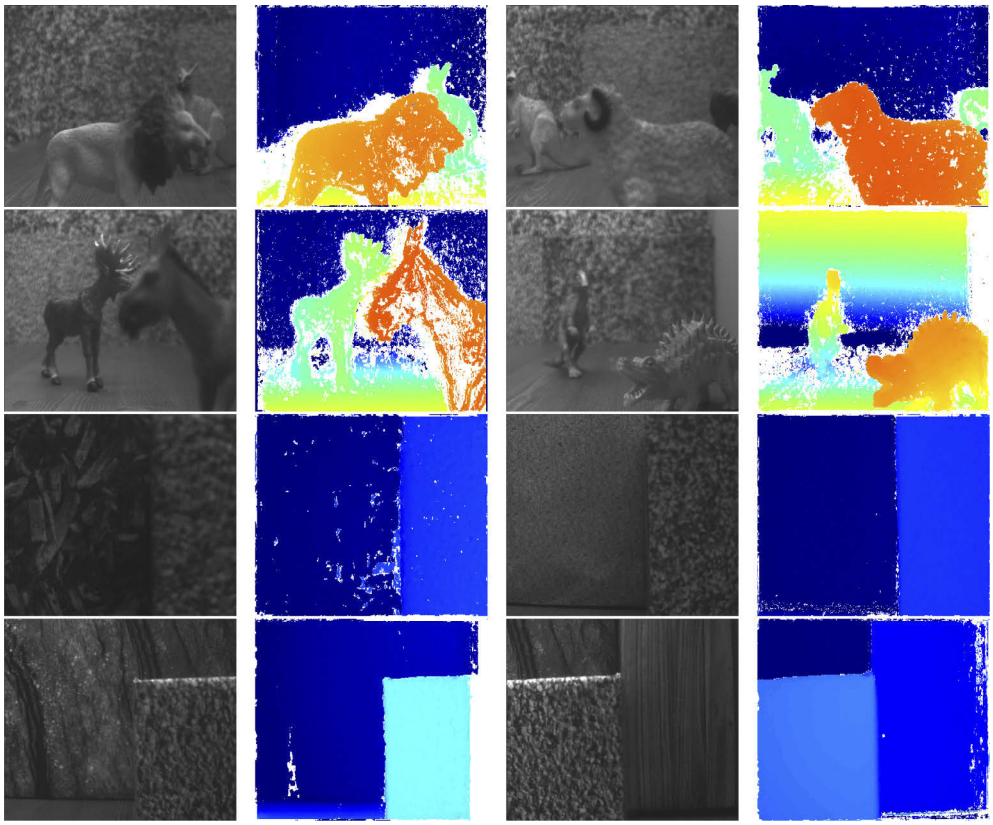


Fig. 9. Images and respective ground truths used in quantitative evaluation with OPA images. Warmer colors show closer objects. White pixels are excluded from the evaluation. (Best seen in color).

a car. In other words, both datasets have very similar scenes. Authors in [70] have tried to deal with heterogeneous datasets for unsupervised depth estimation but they use monocular videos in an online setting. The RMSE of OPA-Stereo is increased to 0.459 if training is performed over a heterogeneous dataset composed of small objects and flat surfaces. This degraded result, however, is still better than the results of [19] and [18] with homogeneous datasets shown in Table 3. Furthermore, the RMSE of [19] and [18] is degraded to 3.29 and 4.91 with the heterogeneous dataset, respectively.

One important parameter of depth extraction cameras is their depth resolution or depth sensitivity. To evaluate the depth sensitivity of OPA-Stereo, we place a flat surface before the camera at varying distances and observe the depth of the flat surface. For an ideal depth camera, the disparity values for the whole flat surface should remain the same at a given depth and should change equally for all pixels as the depth changes. The change in the disparity values should be observable as well. Therefore, we define the depth sensitivity metric, which takes into account both the variation of depth across a flat surface as well as the change in the depth metric. For n images arranged in order of their distance, the depth sensitivity is defined as

$$S_d = \frac{\sum_{i=1}^{n-1} |\mu_i - \mu_{i-1}|}{\sum_{i=0}^{n-1} \sigma_i^2}, \tag{14}$$

where μ_i and σ_i^2 are the average and variance of disparity or depth index of the i-th image.

Optics EXPRESS

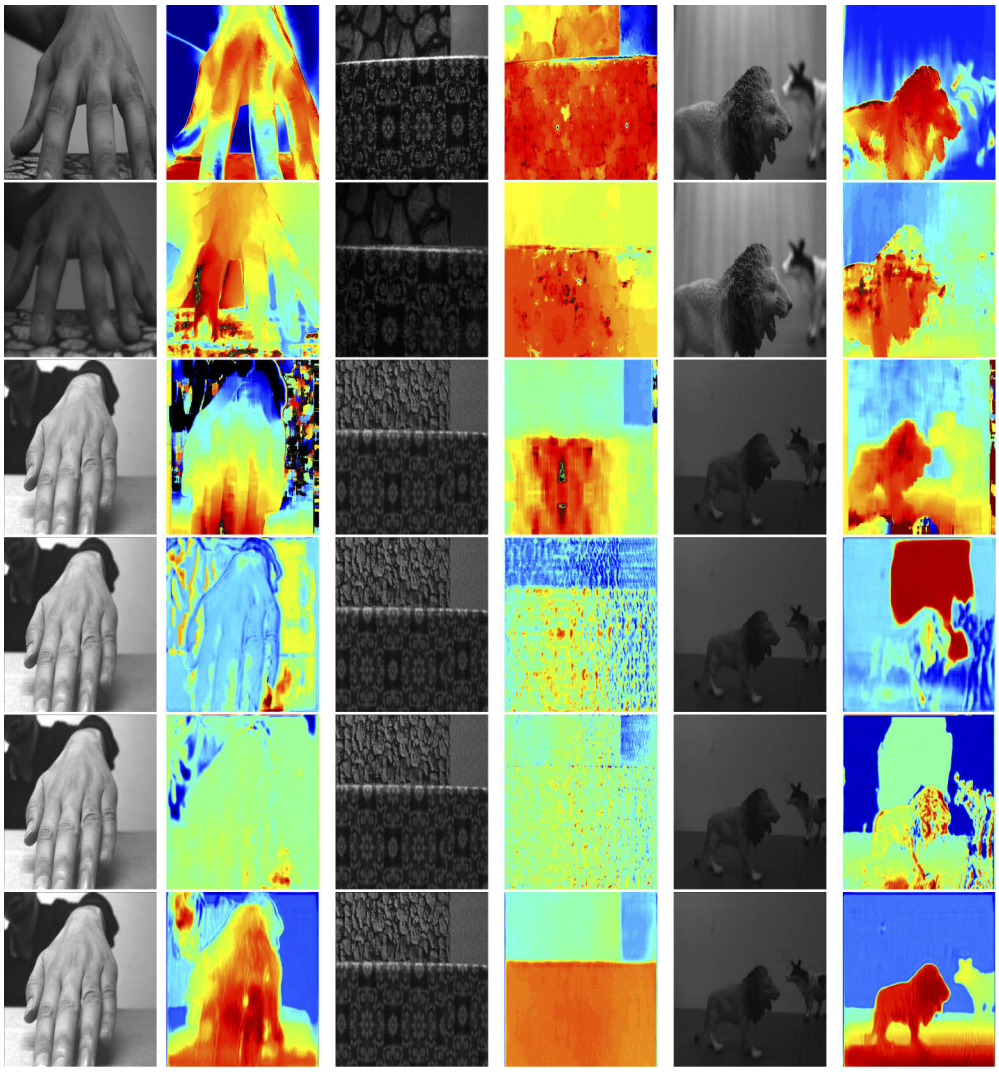


Fig. 10. Qualitative comparison of different methods where the rows from top to bottom show the results of DA [6], OA [7], OPA [8], Unsup-Stereo [19], Unsup-Monocular [18] and OPA-Stereo methods. Warmer colors show closer distance (Best seen in color).

The results of depth sensitivity for different cameras are shown in Fig. 12. It is seen that the depth sensitivity is significantly improved by using our OPA-Stereo approach. It should be noted that the DA camera does not provide disparity, rather it provides a depth index at every pixel based on the blur difference between different frequency channels. The DA camera does not produce reliable results with depth. The results of the OA camera are somewhat better compared to DA. The previously proposed approach for OPA also produces decent results for depth sensitivity. However, the best depth sensitivity results are seen by using the OPA-Stereo approach. The results are shown in Table 4.

5.3. Time complexity

For our experiments, we have used the Nvidia GTX 1080 GPU. Training for 100 epochs with a training dataset of 2000 images takes 16 hours for OPA-Stereo. The number of trainable

Optics EXPRESS

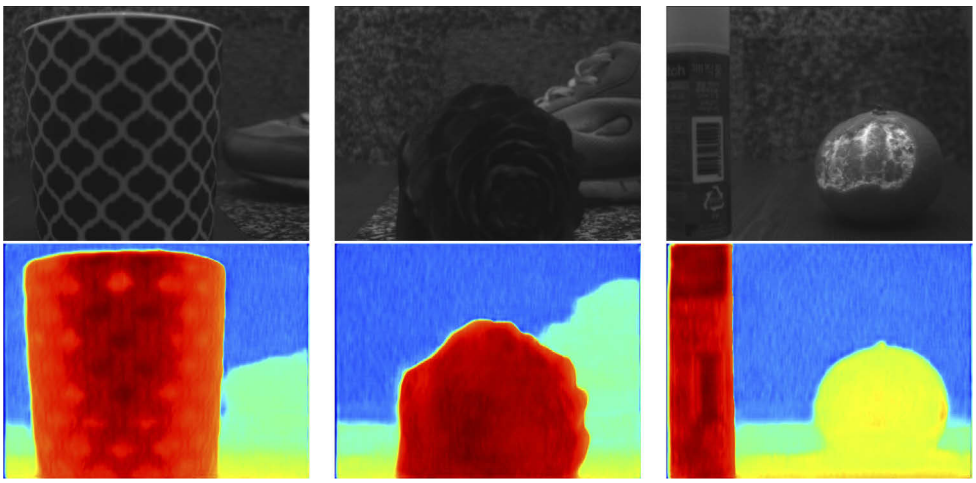
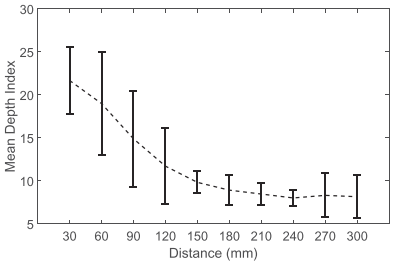
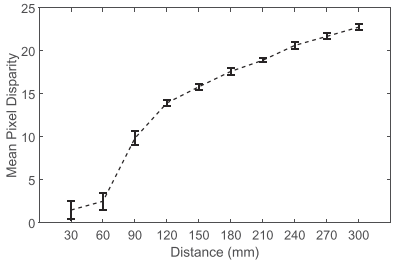


Fig. 11. Qualitative results with OPA-Stereo. Warmer colors show closer distance (Best seen in color).

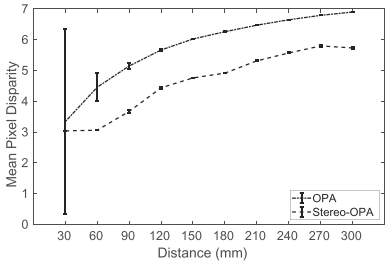
(a)



(b)



(c)



(d)

Fig. 12. (a) Depth sensitivity analysis of different approaches, the error plot of evaluated depth for (b) DA (c) OA and (d) OPA camera with different approaches for a flat surface placed at varying distances from the camera.

Table 4. Depth sensitivity performance.

Approach	Depth Sensitivity
DA [6]	0.1164
OA [7]	6.3510
OPA [8]	1.3550
OPA-Stereo	14.7483

Optics EXPRESS

parameters for OPA-Stereo are approximately 125.1 million. OPA-Stereo provides 8 frames per second. However, it should be noted that these results are with a relatively smaller Nvidia 1080 GPU with 8GB of memory. The inference speed will be much higher if a larger GPU is used. A CPU implementation of OPA-Stereo on an Intel Core i5 processor with 8GB RAM provides 0.5fps. For comparison, the speed of competing methods is relatively lower compared to the proposed approaches. For example, on an Intel Core i5 processor with 8GB RAM, MATLAB implementations of DA [6], OA [7] and previous OPA [8] take 151s, 51s and 4s, respectively. Note that dedicated hardware implementations for disparity estimation have been proposed for OA [65] and OPA [8] cameras, which are much faster. However, here we limit ourselves to software implementations only. The neural networks cannot accept images of arbitrary sizes; therefore, all images input to all the neural networks are resized to 512×256 as shown in Fig. 5. The unsupervised method proposed in [19] also provides 8 frames per second on the Nvidia 1080 GPU. The speed of [18], however, is much higher at 20 frames per second. Despite the better speed, experimental results clearly demonstrate that [18] cannot be used for disparity estimation with the OPA camera.

6. Conclusion

In this paper, we propose using deep learning for disparity estimation with the Offset Pixel Aperture (OPA) camera. Specifically, we use an unsupervised approach based on stereo principles. By using the unsupervised approach, we avoid pixel-level labeling of ground truth disparity maps for training. Furthermore, we discover that a very small baseline does not allow the neural networks to estimate disparity due to vanishing gradients. We propose using two OPA cameras for training, which results in improved disparity estimation. Our work practically takes deep learning into the wild for disparity estimation as the OPA images have low PSNR, small baseline and other common imaging problems compared to the environments and systems where disparity estimation is generally performed in research. By using the OPA sensor with deep learning, our work is in line with the recent paradigm of using sensors on the edge with artificial intelligence.

Funding

Ministry of Science, ICT and Future Planning (CISS-2013073718).

Disclosures

The authors declare no conflicts of interest.

References

1. J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2008), pp. 1–8.
2. D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1 (IEEE, 2003), p. I.
3. P. Sturm and B. Triggs, "A factorization based algorithm for multi-image projective structure and motion," in *European conference on computer vision*, (Springer, 1996), pp. 709–720.
4. A. P. Pentland, "A new sense for depth of field," IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-9**(4), 523–531 (1987).
5. D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," Int. J. Comput. Vision **47**(1/3), 7–42 (2002).
6. M. Martinello, A. Wajs, S. Quan, H. Lee, C. Lim, T. Woo, W. Lee, S.-S. Kim, and D. Lee, "Dual aperture photography: Image and depth from a mobile camera," in *Computational Photography (ICCP), 2015 IEEE International Conference on*, (IEEE, 2015), pp. 1–10.
7. M. U. K. Khan, A. Khan, J. Lim, S. Hamidov, W.-S. Choi, W. Yun, Y. Lee, Y.-G. Kim, H.-S. Park, and C.-M. Kyung, "Offset aperture: A passive single-lens camera for depth sensing," IEEE Transactions on Circuits and Systems for Video Technology **29**(5), 1380–1393 (2019).
8. W. Yun, Y. Kim, Y. Lee, J. Lim, H. Kim, M. Khan, S. Chang, H. Park, and C.-M. Kyung, "Depth extraction with offset pixels," Opt. Express **26**(12), 15825–15841 (2018).

Optics EXPRESS

9. B.-S. Choi, J. Lee, S.-H. Kim, S. Chang, J. Park, S.-J. Lee, and J.-K. Shin, "Analysis of disparity information for depth extraction using cmos image sensor with offset pixel aperture technique," Sensors **19**(3), 472 (2019).
10. J. Lee, B.-S. Choi, S.-H. Kim, J. Lee, J. Lee, S. Chang, J. Park, S.-J. Lee, and J.-K. Shin, "Effects of offset pixel aperture width on the performances of monochrome cmos image sensors for depth extraction," Sensors **19**(8), 1823 (2019).
11. B.-S. Choi, M. Bae, S.-H. Kim, J. Lee, C.-W. Oh, S. Chang, J. Park, S.-J. Lee, and J.-K. Shin, "Cmos image sensor for extracting depth information using offset pixel aperture technique," in *Novel Optical Systems Design and Optimization XX*, vol. 10376 (International Society for Optics and Photonics, 2017), p. 103760Y.
12. J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), pp. 1592–1599.
13. D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, (2014), pp. 2366–2374.
14. F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), pp. 5162–5170.
15. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), pp. 4040–4048.
16. A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), pp. 66–75.
17. R. Garg, B. G. Vijay Kumar, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*, (Springer, 2016), pp. 740–756.
18. C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), pp. 270–279.
19. A. Pilzer, D. Xu, M. Puscas, E. Ricci, and N. Sebe, "Unsupervised adversarial depth estimation using cycled generative networks," in *2018 International Conference on 3D Vision (3DV)*, (IEEE, 2018), pp. 587–595.
20. S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," Int. J. Comput. Vision **35**(3), 269–293 (1999).
21. A. Khan, M. U. K. Khan, and C.-M. Kyung, "Intensity guided cost metric for fast stereo matching under radiometric variations," Opt. Express **26**(4), 4096–4111 (2018).
22. A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," IEEE Trans. Pattern Anal. Machine Intell. **35**(2), 504–511 (2013).
23. H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2 (IEEE, 2005), pp. 807–814.
24. M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters," in *null*, (IEEE, 2003), p. 900.
25. Q. Yang, "A non-local cost aggregation method for stereo matching," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2012), pp. 1402–1409.
26. K. He, J. Sun, and X. Tang, "Guided image filtering," in *European conference on computer vision*, (Springer, 2010), pp. 1–14.
27. R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: a survey," IEEE Trans. Pattern Anal. Machine Intell. **21**(8), 690–706 (1999).
28. D. Blostein and N. Ahuja, "Shape from texture: Integrating texture-element extraction and surface estimation," IEEE Trans. Pattern Anal. Machine Intell. **11**(12), 1233–1251 (1989).
29. A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," IEEE Trans. Pattern Anal. Machine Intell. **31**(5), 824–840 (2009).
30. P. Knobelreiter, C. Reinbacher, A. Shekhovtsov, and T. Pock, "End-to-end training of hybrid cnn-crf models for stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), pp. 2339–2348.
31. F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," arXiv preprint arXiv:1904.06587 (2019).
32. I. Melekhov, A. Tiulpin, T. Sattler, M. Pollefeys, E. Rahtu, and J. Kannala, "Dgc-net: Dense geometric correspondence network," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (IEEE, 2019), pp. 1034–1042.
33. P. Yadati and A. M. Namboodiri, "Multiscale two-view stereo using convolutional neural networks for unrectified images," in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, (IEEE, 2017), pp. 346–349.
34. F. Tosi, M. Poggi, A. Benincasa, and S. Mattoccia, "Beyond local reasoning for stereo confidence estimation with deep learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), pp. 319–334.
35. S. Tulyakov, A. Ivanov, and F. Fleuret, "Practical deep stereo (pds): Toward applications-friendly deep stereo matching," in *Advances in Neural Information Processing Systems*, (2018), pp. 5871–5881.
36. R. Atienza, "Fast disparity estimation using dense networks," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, (IEEE, 2018), pp. 3207–3212.

Optics EXPRESS

37. K. Batsos, C. Cai, and P. Mordohai, "Cbmv: A coalesced bidirectional matching volume for disparity estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), pp. 2060–2069.
38. D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition*, (Springer, 2014), pp. 31–42.
39. A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," The International Journal of Robotics Research **32**(11), 1231–1237 (2013).
40. R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Computer Science Technical Report CSTR **2**, 1–11 (2005).
41. S. Heber and T. Pock, "Convolutional networks for shape from light field," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), pp. 3746–3754.
42. N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," ACM Trans. Graph. **35**(6), 1–10 (2016).
43. C. Shin, H.-G. Jeon, Y. Yoon, I. So Kweon, and S. Joo Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), pp. 4748–4757.
44. R. Garg, N. Wadhwa, S. Ansari, and J. T. Barron, "Learning single camera depth estimation using dual-pixels," arXiv preprint arXiv:1904.05822 (2019).
45. J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016).
46. J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, (2016).
47. M. J. Cieślak, K. A. Gamage, and R. Glover, "Coded-aperture imaging systems: Past, present and future development–a review," Radiat. Meas. **92**, 59–71 (2016).
48. A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," ACM Trans. Graph. **26**(3), 70 (2007).
49. C. Zhou, S. Lin, and S. Nayar, "Coded aperture pairs for depth from defocus," in *2009 IEEE 12th International Conference on Computer Vision*, (IEEE, 2009), pp. 325–332.
50. Y. Amari and E. Adelson, "Single-eye range estimation by using displaced apertures with color filters," in *Industrial Electronics, Control, Instrumentation, and Automation, 1992. Power Electronics and Motion Control., Proceedings of the 1992 International Conference on*, (IEEE, 1992), pp. 1588–1592.
51. Y. Bando, B.-Y. Chen, and T. Nishita, "Extracting depth and matte using a color-filtered aperture," in *ACM Transactions on Graphics (TOG)*, vol. 27 (ACM, 2008), p. 134.
52. E. Lee, W. Kang, S. Kim, and J. Paik, "Color shift model-based image enhancement for digital multifocusing based on a multiple color-filter aperture camera," IEEE Transactions on Consumer Electronics **56**(2), 317–323 (2010).
53. V. Paramonov, I. Panchenko, V. Bucha, A. Drogolyub, and S. Zagoruyko, "Depth camera based on color-coded aperture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (2016), pp. 1–9.
54. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (2012).
55. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, (2017), pp. 2223–2232.
56. M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in neural information processing systems*, (2015), pp. 2017–2025.
57. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, (Springer, 2016), pp. 21–37.
58. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp. 770–778.
59. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2 (Ieee, 2003), pp. 1398–1402.
60. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, (2014), pp. 2672–2680.
61. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), pp. 1125–1134.
62. Q. Teng, Y. Chen, and C. Huang, "Occlusion-aware unsupervised learning of monocular depth, optical flow and camera pose with geometric constraints," Future Internet **10**(10), 92 (2018).
63. A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," arXiv preprint arXiv:1904.04998 (2019).
64. C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, (2019), pp. 3828–3838.
65. W. Yun, Y.-G. Kim, Y. Lee, J. Lim, W. Choi, M. U. K. Khan, A. Khan, S. Homidov, P. Kareem, H. S. Park, and C.-M. Kyung, "Offset aperture based hardware architecture for real-time depth extraction," in *2017 IEEE International Conference on Image Processing (ICIP)*, (IEEE, 2017), pp. 4392–4396.

Optics EXPRESS

66. J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision*, (Springer, 2016), pp. 501–518.
67. R. Szeliski, *Computer vision: algorithms and applications* (Springer Science & Business Media, 2010).
68. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, (2016), pp. 265–283.
69. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp. 3213–3223.
70. V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33 (2019), pp. 8001–8008.