Deep End-to-End Time-of-Flight Imaging

Shuochen Su
UBC

Felix Heide
Stanford University

Gordon Wetzstein
Stanford University

Wolfgang Heidrich
KAUST

Abstract

We present an end-to-end image processing framework for time-of-flight (ToF) cameras. Existing ToF image processing pipelines consist of a sequence of operations including modulated exposures, denoising, phase unwrapping and multipath interference correction. While this cascaded modular design offers several benefits, such as closed-form solutions and power-efficient processing, it also suffers from error accumulation and information loss as each module can only observe the output from its direct predecessor, resulting in erroneous depth estimates. We depart from a conventional pipeline model and propose a deep convolutional neural network architecture that recovers scene depth directly from dual-frequency, raw ToF correlation measurements. To train this network, we simulate ToF images for a variety of scenes using a time-resolved renderer, devise depth-specific losses, and apply normalization and augmentation strategies to generalize this model to real captures. We demonstrate that the proposed network can efficiently exploit the spatio-temporal structures of ToF frequency measurements, and validate the performance of the joint multipath removal, denoising and phase unwrapping method on a wide range of challenging scenes.

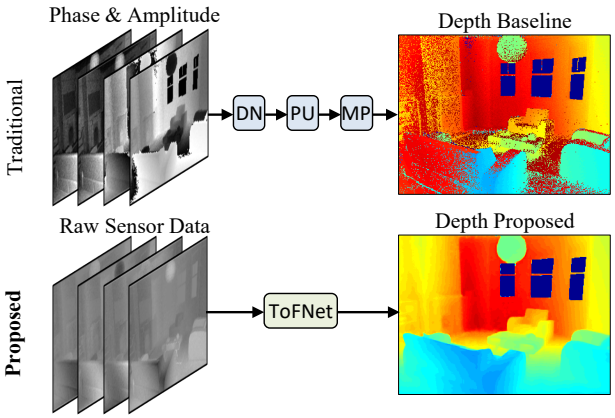


Figure 1: Top: given phase and amplitude images from dual-frequency measurements, traditional ToF cameras apply a sequence of techniques for depth map generation, such as denoising (DN), phase unwrapping (PU) and multipath correction (MP). This often leads to inaccurate depth estimation as low frequency phases are particularly prone to global illumination [19] and various types of sensor noise; Bottom: we train a deep convolutional network to predict scene depth directly from a ToF camera's raw correlation measurements. The proposed method is substantially more robust to noise and MPI, and runs in real-time.

1. Introduction

Recently, amplitude-modulated continuous wave (AMCW) time-of-flight cameras such as Microsoft's Kinect One have not only become widely adopted in interactive commercial applications, but have also emerged as an exciting imaging modality in computer vision [15, 22, 45]. Combined with conventional color cameras, RGB-D data allows for high-fidelity scene reconstruction [26], enabling the collection of large 3D datasets which drive 3D deep learning [9] for scene understanding [47, 24], action recognition [40], and facial and pose tracking [31]. Beyond enabling these core computer vision applications, RGB-D cameras have wide-spread applications in human-computer interaction, robotics, and for tracking in emerging augmented or virtual reality applications [35]. Due to low power requirements, low-cost CMOS sensor technology, and small sensor-illumination baseline [20], AMCW time-of-flight cameras have the potential to become a cornerstone imaging technology. For brevity, we will in the following refer to AMCW time-of-flight cameras simply as ToF cameras, with the implicit understanding that they are distinct from other time-of-flight imaging technologies, such as direct temporal sampling with SPADs (e.g. [49]).

ToF cameras measure depth by illuminating a scene with periodic amplitude-modulated flood-light, which is reflected back to the camera along direct as well as indirect light paths. The camera then measures the phase shift of the incident signal with respect to the illumination signal. To extract depth from these raw phase measurements, a number of challenging reconstruction problems must be solved. For a single diffuse reflector in the scene, the phase measurements encode depth unambiguously only up to an integer phase wrapping, which is addressed by phase unwrapping

methods [20]. In the presence of global illumination, multiple light paths interfere along direct and indirect paths, leading to severe multipath interference (MPI) distortion of the depth maps. Finally, raw ToF measurements are affected by severe noise due to the low absorption depth of the IR modulation, and immature sensor technology [31] compared to RGB CMOS image sensors.

Conventionally, these three reconstruction problems, phase unwrapping, MPI reduction, and denoising, are solved in a pipeline approach where each step addresses an individual subproblem in isolation, as in [11, 14, 34, 39]. While this design facilitates divide-and-conquer algorithms, it ignores the coupling between individual sub-modules and introduces cumulative error and information loss in the reconstruction pipeline. For example, established multi-frequency unwrapping methods [12] become inaccurate in the presence of MPI or noise, leading to noticeable unwrapping errors and subsequently inaccurate shape recovery.

Instead of building a reconstruction pipeline, or relying on additional hardware, we present a data-driven approach that generates a depth map directly from the raw modulated exposures of the ToF camera (see Fig. 1). Specifically, we make the following contributions:

- We propose a learning-based approach for end-to-end time-of-flight imaging by jointly solving phase unwrapping, MPI compensation and denoising from the raw correlation measurements. The proposed architecture significantly outperforms conventional depth image pipelines, while being highly efficient with interactive framerates on modern GPUs.
- We validate that the proposed reconstruction approach effectively removes MPI, phase wrapping and sensor noise, both in simulation and on experimentally acquired raw-dual frequency measurements.
- We introduce a large-scale raw correlation time-of-flight dataset with known ground truth depth labels for every pixel. The dataset and architecture will be published for full reproducibility of the proposed method.

2. Related Work

Phase unwrapping. An established method for resolving phase ambiguity acquires measurements at two different modulation frequencies [11], preserving long distance range by unwrapping high-frequency phases with their lower-frequency counterpart. While effective for direct-only scenes, this dual-frequency acquisition approach becomes inaccurate in the presence of MPI. When multi-frequency measurements are not accessible, statistical priors, such as the amplitude smoothness [20] and surface normal constraints [10, 12], can be leveraged. Our method, however, is not built on such hand-crafted priors that only model a subset of the rich statistics of natural scenes. Instead we learn the spatial prior directly from a large corpus of training data.

MPI correction. MPI distortions are commonly reduced in a post-processing step. A large body of work explores either analytic solutions to the simplified two-path or diffuse-only problems [17, 14], or attempts to solve MPI in isolation as a computationally costly optimization problem [13, 29, 7] with strong assumptions on the scene sparsity. Although MPI, phase unwrapping and denoising are coupled, none of the existing reconstruction methods address them in a joint and computationally efficient manner.

Alternative acquisition. Recent alternate approaches attempt to resolve ambiguities in the capture process. Gupta et al. [19] propose to separate indirect illumination using high-frequency modulations in the GHz range, which remains theoretical due to the limitations of existing hundred-MHz-range CMOS techniques. A number of works have proposed hybrid structured light-ToF systems [42, 39, 5], requiring coded and carefully synchronized illumination with a significantly enlarged footprint due to the projector-camera baseline, thus removing many of the inherent benefits of ToF technology.

Learned post-processing. We are not the first to apply deep learning to resolve ambiguities in ToF depth reconstruction. Son et al. [46] use a robotic arm to collect ToF range images with the corresponding ground truth labels from a structured light sensor, and train a feedforward neural network to remove multipath distortions. Concurrent to our work, Marco et al. [37] train an encoder-decoder network that takes ToF range images as input and predicts the multipath-corrected version. Both of these approaches, however, are not end-to-end, as they post-process depth from a specific type of camera's pipeline output. Much of the information presented in the raw ToF images has already been destroyed in the depth images that serve as input to these methods. By ignoring the coupled nature of the many subproblems in ToF reconstruction they artificially limit the depth imaging performance, as we show in this work.

Deep image generation. Deep convolutional neural networks have enabled great advances in supervised image reconstruction problems, including deblurring [52], denoising/inpainting [51], and super-resolution [28]. While such feedforward architectures work well for local operations on natural images, a large receptive field is typically desired for non-local inverse problems, such as MPI removal.

Recently, conditional generative adversarial networks (cGAN) have shown high-quality image translation results under supervised [25] and unsupervised [53] settings. Unlike traditional GANs [18], in cGANs both the generator G and discriminator D observe an input image. By combining

a GAN loss with traditional pixel losses, one can then learn to penalize *structured* differences between output and target images, without relying on domain knowledge [4]. We adopt these successful cGAN strategies to train our depth generation network, and combine them with pixel loss and smoothness terms on the depth maps and their gradients.

3. Time-of-Flight Imaging

In this section, we review ToF depth imaging and its core challenges.

Depth acquisition. A ToF camera measures the modulated exposure of the scene with periodic flood-light illumination and sensor demodulation signals. Following Lange [32], we model a raw correlation measurement for integration time T as

$$b_{\omega,\psi} = \int_0^T E(t) f_\omega \left(t - \psi/\omega\right) dt, \tag{1}$$

where E is the irradiance, and f is a programmable reference signal with angular frequency ω and phase offset ψ. Typically, f is zero-mean so that the measurement is robust to ambient light.

In the ideal case where indirect light paths are not present, one can reliably recover the scene depth[1] at each pixel by capturing a pair of modulated exposures with the same ω but different ψ as

$$d = c\phi/2\omega, \tag{2}$$

$$\text{where} \quad \phi = \text{atan2}(b_{\omega,\pi/2}, b_{\omega,0}) \tag{3}$$

is the measured phase, and c denotes the speed of light. While computationally cheap, applying Eq. 2 to real-world data often leads to poor depth estimates. This is not only because of sensor noise, but also because the measurements from Eq. 3 are inherently ambiguous due to phase wrapping and MPI which requires solving the ill-posed reconstruction problems described in the following.

Dual-frequency phase unwrapping. Due to the periodic nature of the phase measurements in Eq. 3, the depth estimate also "wraps around", and is only unambiguous for distances smaller than a half of the modulation wave length, i.e. in the $[0, c\pi/\omega]$ range. Dual-frequency methods disambiguate the true depth from other, phase wrapped candidates by measuring b at two different frequencies, ω_1 and ω_2 [11]. This effectively extends the maximum unambiguous depth range to $d_{max} = c\pi/\text{GCD}(\omega_1, \omega_2)$, where GCD denotes the greatest common divisor of the two frequencies. To recover an unknown depth d^* one can create a lookup table $\mathcal{T}_\omega(\hat{d})$ between candidate depth $\hat{d} \in [0, ..., d_{max}]$ and phase observations $\Phi = [\phi_1, \phi_2]$, and solve the following 1D search problem [19],

$$d^* = \arg\min_{\hat{d}} ||\mathcal{T}_{[\omega_1,\omega_2]}(\hat{d}) - \Phi||_2. \tag{4}$$

However, in the presence of noise, this idealized per-pixel method often fails due to the lack of spatial priors. Recently, Lawin et al. [33] proposed kernel density function estimates as a hand-crafted prior for more robust phase unwrapping.

[1] Converting distance to depth is trivial given camera intrinsic matrix. We use depth, distance, and path length interchangeably in this work.

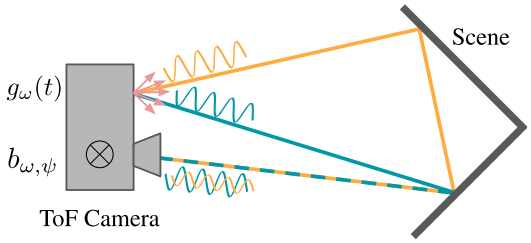


Figure 2: A ToF sensor integrates a mixture of direct (green) and indirect (orange) reflectance from the corner scene.

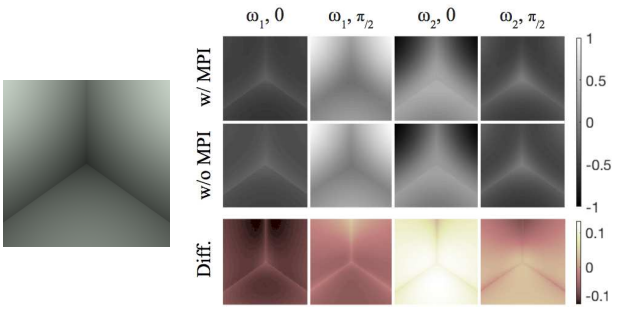
Multi-path interference. The second core challenge when applying Eq. 2 to real world scenarios is that instead of a single directly-reflected path, scenes with different geometric and material properties can cause multiple light paths to be linearly combined in a single sensor pixel, illustrated in Fig. 2. For common sinusoidal modulation, these path mixtures lead to measurements that are identical to the ones from longer direct paths, resulting in inherently ambiguous measurements. Formalizing the intensity-modulated light source in homodyne mode as $g_\omega(t)$, $E(t)$ becomes a superposition of many attenuated and phase-shifted copies of $g_\omega(t)$, along all possible paths of equal travel time τ:

$$E(t) = \int_0^{\tau_{\max}} \alpha(\tau) g_\omega(t - \tau) d\tau. \tag{5}$$

When $E(t)$ is substituted in Eq. 1, we model the correlation integral as in [23],

$$b_{\omega,\psi} = \int_0^{\tau_{\max}} \alpha(\tau) \cdot \rho\left(\omega, \psi/\omega + \tau\right) d\tau, \tag{6}$$

where the scene-independent functions f_ω and g_ω have been folded into ρ which is only dependent on the imaging device, and can be calibrated in advance. Essentially, Eq. 6 probes the latent, scene-dependent temporal point spread function (TPSF, its first peak indicates true depth), $\alpha(\tau)$, to the sensor observations $b_{\omega,\psi}$. This motivates us to devise the learning framework, which will be described in the next section.



(a) Scene. (b) Correlation images.

Figure 3: Illustration of dual-frequency correlation images of a corner scene synthesized with and without multipath interference. MPI introduces scene- and frequency-dependent offsets to ToF data (bottom right), which is in turn treated as features by our method. Images are simulated by the method in Sec. 5.1, and are normalized for visualization.

4. Learning Time-of-Flight Imaging

In this section, we describe the proposed reconstruction architecture and learning loss functions that allow us to directly estimate depth from raw ToF measurements. To build intuition for this end-to-end approach, we synthesize and analyze the correlation images of a corner scene with Eq. 6 in Fig. 3. MPI introduces a per-pixel phase offset, depending not only on scene-specific properties such as distance, geometry and material [16], but also the modulation signals. We demonstrate that the inverse mapping from correlation images to depth maps can be learned by leveraging the spatio-temporal structures of raw ToF measurements in a large corpus of synthetic training data. Specifically, we treat depth generation as a multi-channel image fusion problem, where a desired depth map is the weighted combination of the same scene measured at multiple $[\omega_i, \psi_j]$ illumination-sensor configurations. Our reconstruction network is trained to fuse these spatio-temporal structures, jointly performing MPI removal, denoising and phase unwrapping, while penalizing artifacts in the resulting depth maps via a novel loss function.

4.1. Depth Estimation Network

The proposed depth generation network architecture takes the correlated nature of the raw ToF measurements into account. In contrast to conventional RGB or grayscale intensity images, the pixel values in $B_{\omega,\psi}$ are more sensitive to scene and camera settings, e.g. the frequency, phase offset and power of illumination signals. An ideal network should therefore learn cross channel correlations, as well as spatial features that are invariant to albedo, amplitude and scale variations.

Moreover, the input correlation measurements and output depth images should both be consistent with the underlying scene geometry. While the two should share depth gradients, albedo gradients do not necessarily align with depth edges and should be rejected.

With these motivations, we design a multi-scale network, ToFNet, following an encoder-decoder network architecture with skip connections [44] and ResNet [21] bottleneck layers (see Fig. 4). Specifically, the network takes a stack of modulated exposures $[B_{\omega_i,\psi_j}], i, j = [1, 2]$ as input to generate a phase-unwrapped and MPI-compensated distance image. We then convert the network output to depth map with calibrated camera intrinsics.

The encoder (F1_1 to D2) of the generator G spatially compresses the input up to $1/4$ of its original resolution, while generating feature maps with increasing receptive field. The ResNet blocks at the bottleneck maintain the number of features, while refining their residuals across multiple channels so that they can reconstruct a finer, cleaner depth after upsampling. We also design symmetrically connected skip layers between F1_2-U2 and F2-U1 by element-wise summation. These skip connections are designed around the notion that scene structures should be shared between inputs and outputs [25]. The discriminator network D consists of 3 down-convolutional layers, classifying G's prediction in overlapping patches. During training, we also randomly augment the scale of input images by a number of coarse and fine levels to learn scale-robust features.

We propose a number of input/output configurations as well as data normalization and augmentation strategies to accompany the network design. Specifically, instead of relying on the network to learn amplitude invariant features, we apply pixel-wise normalization to correlation inputs with their corresponding amplitudes. This effectively improves the model's robustness to illumination power and scene albedo, thereby reducing the required training time as amplitude augmentation becomes unnecessary. One drawback with the normalization scheme is that the input may contain significantly amplified noise in regions where reflectivity is low or distances are too large due to the inverse square law. To this end we introduce an edge-aware smoothness term to leverage the unused amplitude information, by feeding the amplitude maps into the TV regularization layer described in the following section.

4.2. Loss Functions

Due to the vastly different image statistics of depth and RGB image data, traditional ℓ_1/ℓ_2-norm pixel losses that work well in RGB generation tasks lead to poor depth reconstruction performance with blurry image outputs. In the following we devise domain-specific criteria tailored to depth image statistics.

L_1 loss. We minimize the mean absolute error between the generator's output depth d and target depth $\tilde{d}$ due to its ro-

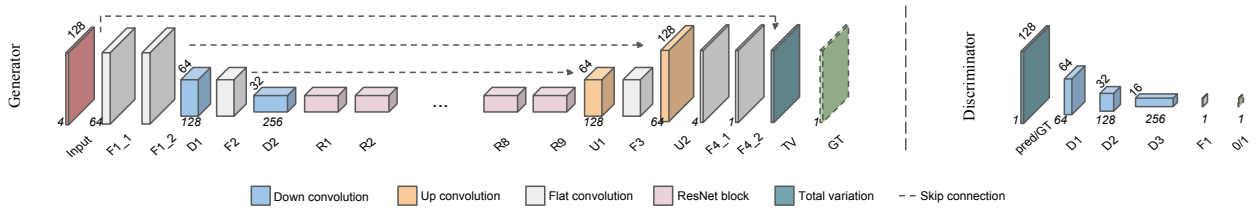


Figure 4: The proposed ToFNet architectures, consisting of, top: a symmetrically skip-connected encoder-decoder generator network G, and bottom: a patchGAN discriminator network D. We implement $\mathcal{L}_{smooth}$ as a regularization layer, denoted as TV (total variation) here. Please refer to the supplemental material for detailed layer specifications.

bustness to outliers,

$$\mathcal{L}_{L_1} = \frac{1}{N}\sum_{i,j} |d_{ij} - \tilde{d_{ij}}|. \tag{7}$$

Depth gradient loss. To enforce locally-smooth depth maps, we introduce an L_1 penalty term on depth gradients, i.e. a total variation loss, which is further weighted by image gradients in an edge-aware fashion [4]. Denoting w as the amplitude of correlation inputs $[b_{\omega_i,0}, b_{\omega_i,\pi/2}]$, we have

$$\mathcal{L}_{smooth} = \frac{1}{N}\sum_{i,j} |\partial_x d_{ij}| e^{-|\partial_x w_{ij}|} + |\partial_y d_{ij}| e^{-|\partial_y w_{ij}|}. \tag{8}$$

Adversarial loss. To further adapt to depth-statistics, we introduce a patch-level conditional adversarial loss [53], minimizing the structural gap between a model-generated depth d and ground-truth depth $\tilde{d}$. We adopt the least square GAN [36] to stabilize the training process,

$$\begin{aligned}\mathcal{L}_{adv} &= \frac{1}{2}\mathbb{E}_{y\sim p_{\text{depth}}(y)}[(D(y)-1)^2] \\ &+ \frac{1}{2}\mathbb{E}_{x\sim p_{\text{corr}}(x)}[(D(G(x)))^2].\end{aligned} \tag{9}$$

Overall loss. Our final loss is a weighted combination of

$$\mathcal{L}_{total} = \mathcal{L}_{L_1} + \lambda_s \mathcal{L}_{smooth} + \lambda_a \mathcal{L}_{adv}. \tag{10}$$

During training, G and D are optimized alternatingly, such that G gradually refines the depth it generates to convince D to assume the result to be correct (label 1), while D gets better and better at distinguishing correct and incorrect depth estimates by minimizing the squared distance in Eq. 9.

4.3. Training and Implementation

Both G and D are trained on 128×128 patches. We first randomly downsample the original 240×320 images within a $[0.6, 1]$ scaling range and apply random cropping. This multiscale strategy effectively increases the receptive field and improves the model's robustness to spatial scales. Each convolution block in Fig. 4 contains spatial convolution and ReLU/Leaky ReLU (in D) nonlinearity layers, omitting batch normalization to preserve cross-channel correlations. In all of our experiments, we set the loss weights in Eq. 10 to be $\lambda_s = 0.0001$ and $\lambda_a = 0.1$. We train our model using the ADAM optimizer with an initial learning rate of 0.00005 for the first 50 epochs, before linearly decaying it to 0 over another 100 epochs. The training takes 40 hours to complete on a single Titan X GPU.

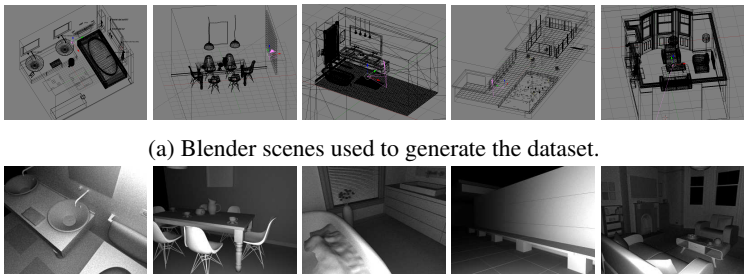
5. Datasets

Because large raw ToF datasets with ground truth depth do not exist, we simulate synthetic measurements with known ground truth to train the proposed architecture. To validate that the synthetic training results map to real camera sensors, we evaluate on experimental measurements acquired with a ToF development board with raw data access.

5.1. Synthetic Dataset

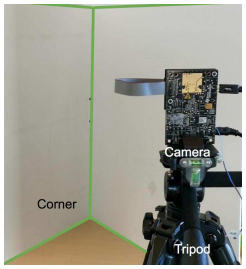
To simulate realistic ToF measurements, we have extended pbrt-v3 [43] for time-resolved rendering. Specifically, we perform bidirectional path tracing [41] with histogram binning according to the path-length of the sampled path. For each scene model and camera-light configuration, our renderer synthesizes a sequence of transient images consisting of a discretized TPSF at every pixel. The raw ToF images can then be simulated by correlating the transient pixels with the frequency-dependent correlation matrix ρ (see Eq. 6). During training we randomly apply additive Gaussian noise to the raw images, which generalizes well to real ToF data of various noise levels due to the fact that both Poisson and Skellam [8] noise are well approximated by Gaussian noise at high photon counts.

We select a number of publicly available indoor and outdoor scene models [2], which include a diverse set of geometric structures at real-world 3D scales (see Fig. 5a and 5b). The ground truth depth maps are generated using Blender's Z pass renderer. Each scene is observed by flying the virtual camera across multiple viewing points and angles that lie along physically plausible paths [38]. To generalize our model to real-world reflectivity variations, we additionally augment the surface albedo of each object for training. In total, our synthetic ToF dataset contains 100,000 correlation-depth image pairs of size 320×240, including 5 scenes with 10 reflectivity variations observed from 250 viewing points and 8 sensor mir-

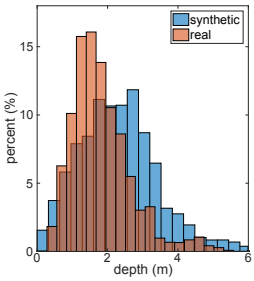


(a) Blender scenes used to generate the dataset.

(b) Exemplar intensity images.



(c) Real data collection.



(d) Depth statistics.

Figure 5: We synthesize transient/correlation images by "animating" a virtual camera along physically-plausible paths in the publicly available blender scenes: BATHROOM, BREAKFAST, CONTEMPORARY-BATHROOM, PAVILION, and WHITE-ROOM [2]. Reasonable alignment can be observed between depth distributions of synthetic and real datasets.

roring/orientations.

We further validate our synthetic dataset by comparing the depth-range distribution between synthetic and real datasets. Our synthetic dataset has a mean depth of 2.35m as a reasonable range for indoor scenes, and it matches the measured empirical depth distribution (see Fig. 5d).

5.2. Real Dataset

We capture the physical validation ToF measurements using an off-the-shelf Texas Instrument OPT8241-CDK-EVM camera, shown in Fig. 5c, which operates at 48MHz by the default. We modify the frequency setting by adjusting the corresponding on-board register via the VoxelSDK [3]. We select 40 and 70MHz as the modulation frequencies for both real and synthesized measurements as our camera prototype achieves a high modulation contrast within this range. Note that the proposed architecture itself is not limited to this range and our network can generalize to any pair/set of modulation frequencies. We also calibrate the phase non-linearity [1] for the two frequencies, after which we treat the measured signal as sinusoidal.

We evaluate the proposed framework on a diverse set of scenes collected under both controlled and in-the-wild conditions, including wall corners, concave objects, as well as every-day environments such as an office, bedroom, bathroom, living room, and kitchen. See Fig. 8 for examples. Note that the real scenes are much more cluttered, consisting of skins, cloths, fabric and mirrors with irregular shape and complex reflectance not presented during training.

6. Experiments and Results

In this section, we present an ablation study to validate the proposed architecture design, and present synthetic and physical experiments that verify the reconstruction performance compared to existing approaches. Tab. 1 and Fig. 7 show synthetic results on a test set containing 9,400 synthetic correlation-depth images sampled from *unseen* scene-reflectivity-view configurations. Fig. 8 shows physical results on raw ToF measurements. We follow Adam et al. [6] to categorize the pixel-wise multipath ratio into low, average, and strong levels, which allows us to understand the performance of each method when performed on direct illumination, e.g. a planar wall, and difficult global illumination cases, e.g. a concave corner. In the following, we quantify depth error with the mean absolute error (MAE) and the *structural* similarity (SSIM) [50] of predicted depth map compared to the ground truth.

6.1. Ablation Study

We evaluate the contribution of individual architecture component to the overall reconstruction performance by designing a series of ablation experiments with truncated architectures and varying input configurations.

Effect of architecture components. Tab. 1 compares the performance of the proposed network architecture, denoted as COMBINED, against four ablated variants, namely

- BASELINE: where we remove the skip connections from G and only minimize the pixel loss $\mathcal{L}_{L_1}$;
- SKIPCONN: same as BASELINE except that G now includes skip connections which encourage structural similarity between input and output;
- TV: same as SKIPCONN except that two loss functions are used for training: $\mathcal{L}_{L_1}$ and $\mathcal{L}_{smooth}$;
- ADV: same as SKIPCONN except that both $\mathcal{L}_{L_1}$ and $\mathcal{L}_{adv}$ are minimized.

Corresponding corner scene scanlines are also shown in Fig. 6. The BASELINE and SKIPCONN networks achieve an overall 3.1cm and 3.0cm depth error which already outperforms traditional pipeline approaches by a substantial margin. However, the generated depth maps suffer from noticeable reconstruction noise in flat areas. By introducing total variation regularization during training, the TV network generates outputs without such artifacts, however still containing global depth offsets. Introducing the adversarial loss in ADV network, which learns a depth-specific structural loss, this global offset is reduced. We also find that

Network	Input	Low MPI	Avg. MPI	Strong MPI	Overall	Speed
EMPTY	N/A	1.205 / 0.0000	2.412 / 0.0000	2.453 / 0.0000	2.190 / 0.0000	N/A
BASELINE	corr.	0.028 / 0.9994	0.030 / 0.9945	0.110 / 0.9959	0.031 / 0.9613	415.5
SKIPCONN	corr.	0.029 / 0.9993	0.030 / 0.9930	0.109 / 0.9949	0.030 / 0.9565	421.0
TV	corr.	0.026 / 0.9995	0.028 / 0.9956	0.109 / 0.9957	0.030 / 0.9625	418.4
ADV	corr.	0.026 / 0.9994	**0.027** / 0.9937	**0.107** / 0.9953	**0.028** / 0.9593	418.8
COMBINED	corr.	0.025 / **0.9996**	0.028 / **0.9957**	**0.107** / **0.9958**	0.029 / **0.9631**	418.8
COMBINED	phase	0.034 / 0.9987	0.051 / 0.9888	0.143 / 0.9938	0.055 / 0.9395	521.4
COMBINED [37]	depth	0.061 / 0.9960	0.060 / 0.9633	0.171 / 0.9815	0.064 / 0.8291	**529.8**
PHASOR [19]	phase	**0.011** / 0.9975	0.102 / 0.9523	1.500 / 0.8869	0.347 / 0.6898	5.2*
SRA [13]	corr.	0.193 / 0.9739	0.479 / 0.8171	0.815 / 0.8822	0.463 / 0.6005	32.3*

Table 1: Quantitative ablation studies on the proposed network and its performance against traditional sequential approaches. EMPTY serves as reference for the mean depth range of the test set. We report MAE and SSIM for each scenario, with MAE measured in meters. In the rightmost column, runtime is reported in FPS (*CPU implementation).

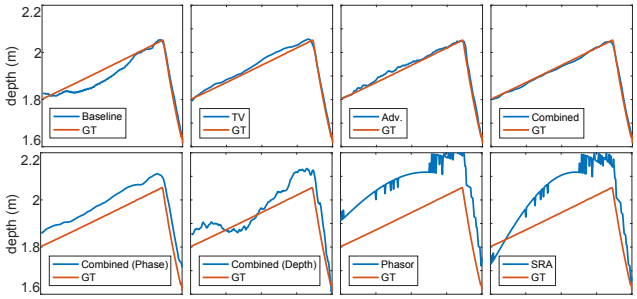


Figure 6: Comparions and ablation study on a corner scene.

the adversarial network generates much sharper depth maps with fewer "flying pixel" artifacts around depth edges. Finally, with skip connections, TV, and adversarial combined, our proposed network achieves the best balance between accuracy, smoothness and processing speed.

Effect of inputs. Although raw correlation images are the natural input choice for the proposed end-to-end architecture, it is also possible to post-process the phase or depth estimation from existing methods' output. Specifically, we evaluate the following input configurations

- CORR., where the input to the network is a stack of raw dual frequency correlation images as presented before;
- PHASE, where we convert the ToF data into two phase maps using Eq. 3; and
- DEPTH, where similar to [37] we first apply phase unwrapping (Eq. 4) to obtain raw depth, and relying on TOFNET to remove noise and MPI.

To this end, we modify the number of input channels at F1_1 layer of G and retrain the weights. All other layers and hyperparameters are kept the same.

As shown in Tab. 1, the COMBINED+PHASE network achieves an overall 5.5cm depth error, which is closest to the COMBINED+CORR. variant. Different from the smooth, correlation inputs, the COMBINED+PHASE network must learn to disambiguate edges caused by phase wrapping from those as a result of depth boundaries, thus becomes less confident when assigning depth values.

The COMBINED+DEPTH network, on the other hand, takes the phase unwrapped depth as input, but must learn to remove the newly introduced depth errors from the previous step as well as correcting for MPI. Consequently, it generates depth maps that are much noisier than COMBINED+PHASE, yet still quantitatively superior to pipeline approaches. Note that this observation matches that in [37], the code of which is unavailable at the time of submission.

6.2. Comparison to Sequential Approaches

Next, we compare the proposed direct reconstruction network to representative sequential pipeline approaches. Specifically, we compare with a ToF pipeline consisting of raw bilateral denoising [48], lookup-table phase unwrapping [20, 19], and non-linearity correction as first three blocks. We will denote the depth map generated from this sub-pipeline as PHASOR. To compensate for MPI we also apply the state-of-the-art sparse reflections analysis technique [13] as the last stage, indicated as SRA. We note that other works on MPI and phase unwrapping [11, 14, 17, 27, 30] either share similar image formation models, or require tailored acquisition strategies, e.g. a larger number of phase or frequency measurements than our approach, making it difficult to draw direct comparisons.

Quantitative results on synthetic dataset. In Fig. 7 we compare our proposed end-to-end solution against PHASOR, SRA, and our depth post-processing variant COMBINED+DEPTH, denoted as DEPTH2DEPTH here [37], on two representative scenes from the test set. As expected, PHASOR generates the most noise among all of the methods, due to the lack of MPI modeling in its image formations. SRA better suppresses the sensor and multipath noise, however its does not significantly compensate for MPI distortions in our experiments, possibly due to the violation of the sparsity assumption [13] in our synthesized backscattering $\alpha(\tau)$ (Eq. 6) which contains strong indirect decay. The DEPTH2DEPTH variant performs inconsistently and are particularly prone to input depth quality. Finally, our method consistently generates depth that is much closer to the ground truth in terms of noise suppression and detail preservation. Please refer to the supplemental material for an extensive set of additional scenes.

Qualitative results on real data. To validate that TOFNET generalizes to real camera data, we conduct qualitative experiments in challenging environments, shown in Fig. 8. Particularly, we evaluate on everyday scenes such as CONCAVEWALL, KITCHEN, LIVINGROOM, OFFICE and PERSON, where traditional pipeline methods commonly fail in the presence of noise, low reflectivity, long range and MPI. While the pipeline methods either partially or overly

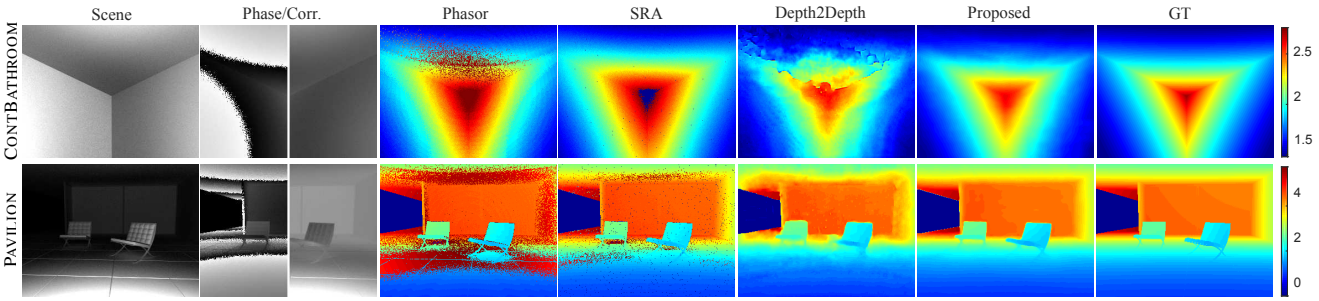


Figure 7: Results on synthetic dataset. Top: Reduction of MPI in a corner scene from Cont-Bathroom. Bottom: Challenging long range scene from Pavilion where denoising, phase unwrapping and MPI are jointly solved by our approach.

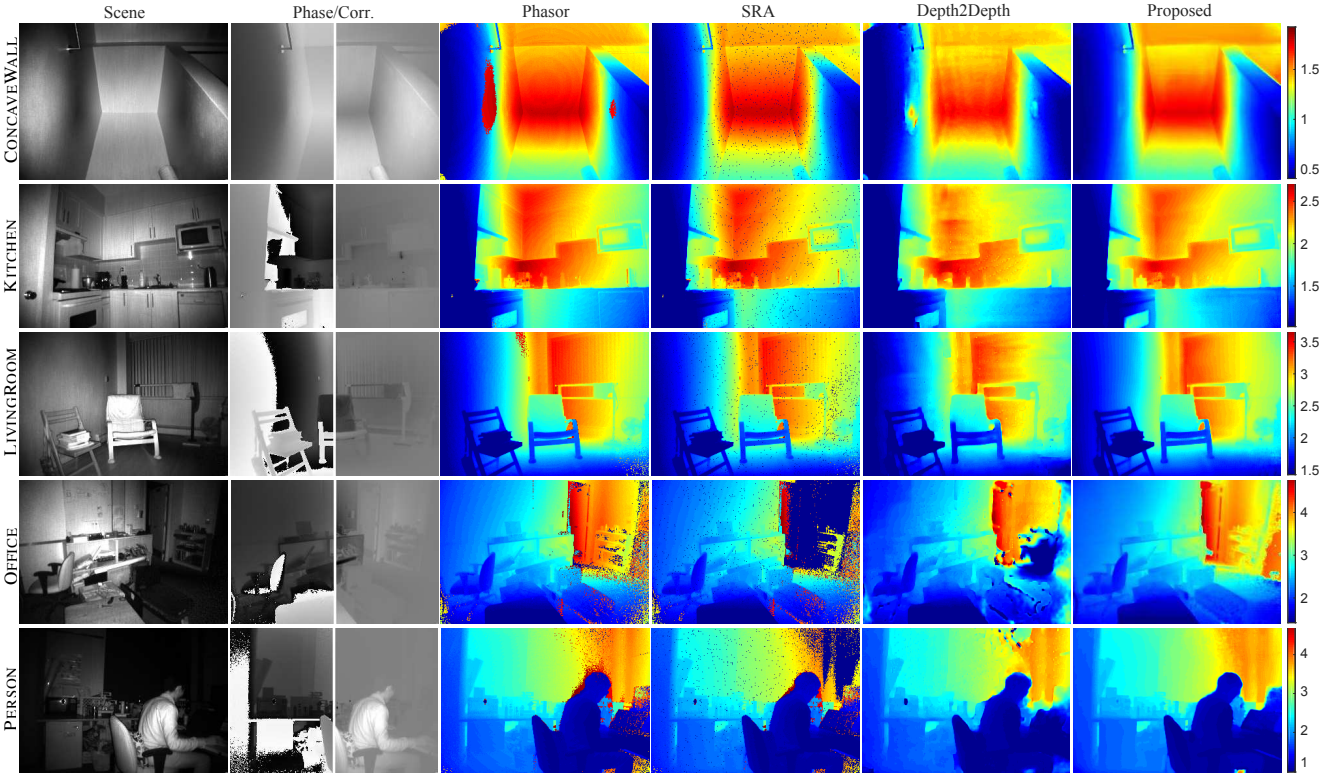


Figure 8: Results on real indoor scenes, where the coupled sensor noise, depth discontinuity (see wrapped edges in phase images) and multipath ambiguity must be addressed in a joint end-to-end manner. Our approach faithfully reconstructs cleaner depth with reduced multipath distortions (see Fig. 6 and supplemental for scanline comparisons). Notice the elimination of "flying" regions in our end-to-end recovered depth compared to the ToF depth as a result of isolated pipeline steps.

compensate MPI and introduce high frequency artifacts, the proposed method consistently generates piece-wise smooth depth maps with reasonable shapes, proving the effectiveness of the learned spatial-correlation features.

Failure cases. ToFNet gracefully fails when the measurement contains saturation, inadequately modeled materials, low reflectivity and finer geometric structures. Nevertheless, due to the depth-dependent prior architecture, our model will estimate the unreliable regions adaptively based on the local neighborhood, achieving a more stable performance than traditional techniques.

7. Conclusion and Future Work

We have presented a learning framework for end-to-end ToF imaging and validated its effectiveness on joint denoising, phase unwrapping and MPI correction for both synthesized and experimentally captured ToF measurements. In the future, we plan to apply our framework to more types of ToF cameras, including impulse-based SPAD detectors. We are also exploring the co-design of modulation function and reconstruction method with our framework, potentially enabling imaging modalities beyond the capabilities of current ToF depth cameras, such as imaging in scattering media.

References

[1] OPT8241-CDK-EVM: Voxel Viewer User Guide. http://www.ti.com/lit/ug/sbou157/sbou157.pdf. Accessed: 2017-10-29. 6

[2] Scenes for pbrt-v3. http://www.pbrt.org/scenes-v3.html. Accessed: 2017-10-29. 5, 6

[3] VoxelSDK: an SDK supporting TI's 3D Time of Flight cameras. https://github.com/3dtof/voxelsdk. Accessed: 2017-10-29. 6

[4] Unsupervised monocular depth estimation with left-right consistency. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, July 2017. 3, 5

[5] S. Achar, J. R. Bartels, W. L. Whittaker, K. N. Kutulakos, and S. G. Narasimhan. Epipolar time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 36(4):37, 2017. 2

[6] A. Adam, C. Dann, O. Yair, S. Mazor, and S. Nowozin. Bayesian time-of-flight for realtime shape, illumination and albedo. *IEEE transactions on pattern analysis and machine intelligence*, 39(5):851–864, 2017. 6

[7] A. Bhandari, A. Kadambi, R. Whyte, C. Barsi, M. Feigin, A. Dorrington, and R. Raskar. Resolving multipath interference in time-of-flight imaging via modulation frequency diversity and sparse regularization. *Optics letters*, 39(6):1705–1708, 2014. 2

[8] C. Callenberg, F. Heide, G. Wetzstein, and M. Hullin. Snapshot difference imaging using time-of-flight sensors. *ACM Transactions on Graphics (ToG)*, 36(6):220:1–220:11, 2017. To appear. 5

[9] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[10] R. Crabb and R. Manduchi. Fast single-frequency time-of-flight range imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 58–65, 2015. 2

[11] A. A. Dorrington, J. P. Godbaz, M. J. Cree, A. D. Payne, and L. V. Streeter. Separating true range measurements from multi-path and scattering interference in commercial range cameras. In *Conference on the Three-Dimensional Imaging, Interaction, and Measurement*, volume 7864, pages 1–1. SPIE–The International Society for Optical Engineering, 2011. 2, 3, 7

[12] D. Droeschel, D. Holz, and S. Behnke. Multi-frequency phase unwrapping for time-of-flight cameras. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 1463–1469. IEEE, 2010. 2

[13] D. Freedman, Y. Smolin, E. Krupka, I. Leichter, and M. Schmidt. Sra: Fast removal of general multipath for tof sensors. In *European Conference on Computer Vision*, pages 234–249. Springer, 2014. 2, 7

[14] S. Fuchs, M. Suppa, and O. Hellwich. Compensation for multipath in tof camera measurements supported by photometric calibration and environment integration. In *International Conference on Computer Vision Systems*, pages 31–41. Springer, 2013. 2, 7

[15] J. Gall, H. Ho, S. Izadi, P. Kohli, X. Ren, and R. Yang. Towards solving real-world vision problems with rgb-d cameras. In *The IEEE Conference on Computer Vision and Pattern Recognition Tutorial*, 2014. 1

[16] I. Gkioulekas, A. Levin, F. Durand, and T. Zickler. Micron-scale light transport decomposition using interferometry. *ACM Transactions on Graphics (ToG)*, 34(4):37, 2015. 4

[17] J. P. Godbaz, M. J. Cree, and A. A. Dorrington. Closed-form inverses for the mixed pixel/multipath interference problem in amcw lidar. In *Conference on Computational Imaging X*, volume 8296, pages 1–15. SPIE, 2012. 2, 7

[18] 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*, pages 2672–2680, 2014. 2

[19] M. Gupta, S. K. Nayar, M. B. Hullin, and J. Martin. Phasor imaging: A generalization of correlation-based time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 34(5):156, 2015. 1, 2, 3, 7

[20] M. Hansard, S. Lee, O. Choi, and R. P. Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012. 1, 2, 7

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[22] F. Heide, W. Heidrich, M. Hullin, and G. Wetzstein. Doppler time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 34(4):36, 2015. 1

[23] F. Heide, M. B. Hullin, J. Gregson, and W. Heidrich. Low-budget transient imaging using photonic mixer devices. *ACM Transactions on Graphics (ToG)*, 32(4):45, 2013. 3

[24] S. Hickson, S. Birchfield, I. Essa, and H. Christensen. Efficient hierarchical graph-based segmentation of rgbd videos. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 344–351, 2014. 1

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *The IEEE conference on computer vision and pattern recognition*, 2017. 2, 4

[26] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011. 1

[27] D. Jimenez, D. Pizarro, M. Mazo, and S. Palazuelos. Modelling and correction of multipath interference in time of flight cameras. In *The IEEE conference on Computer Vision and Pattern Recognition*, pages 893–900. IEEE, 2012. 7

[28] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 2

[29] A. Kadambi, R. Whyte, A. Bhandari, L. Streeter, C. Barsi, A. Dorrington, and R. Raskar. Coded time of flight cameras: sparse deconvolution to address multipath interference and

recover time profiles. *ACM Transactions on Graphics (ToG)*, 32(6):167, 2013. 2

[30] A. Kirmani, A. Benedetti, and P. A. Chou. Spumic: Simultaneous phase unwrapping and multipath interference cancellation in time-of-flight cameras using spectral methods. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013. 7

[31] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight cameras in computer graphics. In *Computer Graphics Forum*, volume 29, pages 141–159. Wiley Online Library, 2010. 1, 2

[32] R. Lange. 3d time-of-flight distance measurement with custom solid-state image sensors in cmos/ccd-technology. 2000. 3

[33] F. J. Lawin, P.-E. Forssén, and H. Ovrén. Efficient multi-frequency phase unwrapping using kernel density estimation. In *European Conference on Computer Vision*, pages 170–185. Springer, 2016. 3

[34] F. Lenzen, K. I. Kim, H. Schäfer, R. Nair, S. Meister, F. Becker, C. S. Garbe, and C. Theobalt. Denoising strategies for time-of-flight data. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pages 25–45. Springer, 2013. 2

[35] H. Li, L. Trutoiu, K. Olszewski, L. Wei, T. Trutna, P.-L. Hsieh, A. Nicholls, and C. Ma. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG)*, 34(4):47, 2015. 1

[36] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *The IEEE International Conference on Computer Vision*, 2017. 5

[37] J. Marco, Q. Hernandez, A. Muoz, Y. Dong, A. Jarabo, M. Kim, X. Tong, and D. Gutierrez. Deeptof: Off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 36(6), 2017. to appear. 2, 7

[38] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 2678–2687, 2017. 5

[39] N. Naik, A. Kadambi, C. Rhemann, S. Izadi, R. Raskar, and S. B. Kang. A light transport model for mitigating multipath interference in time-of-flight sensors. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–81. IEEE, June 2015. 2

[40] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013. 1

[41] M. O'Toole, F. Heide, D. B. Lindell, K. Zang, S. Diamond, and G. Wetzstein. Reconstructing transient images from single-photon sensors. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1539–1547, 2017. 5

[42] M. O'Toole, F. Heide, L. Xiao, M. B. Hullin, W. Heidrich, and K. N. Kutulakos. Temporal frequency probing for 5d transient analysis of global light transport. *ACM Transactions on Graphics (ToG)*, 33(4):87, 2014. 2

[43] M. Pharr, W. Jakob, and G. Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. 5

[44] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 4

[45] S. Shrestha, F. Heide, W. Heidrich, and G. Wetzstein. Computational imaging with multi-camera time-of-flight systems. *ACM Transactions on Graphics (ToG)*, 35(4):33, 2016. 1

[46] K. Son, M.-Y. Liu, and Y. Taguchi. Automatic learning to remove multipath distortions in time-of-flight range images for a robotic arm setup. In *IEEE International Conference on Robotics and Automation*, 2016. 2

[47] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *The IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1

[48] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *The IEEE International Conference on Computer Vision*, pages 839–846. IEEE, 1998. 7

[49] F. Villa, B. Markovic, S. Bellisai, D. Bronzi, A. Tosi, F. Zappa, S. Tisa, D. Durini, S. Weyers, U. Paschen, et al. Spad smart pixel for time-of-flight and time-correlated single-photon counting measurements. *IEEE Photonics Journal*, 4(3):795–804, 2012. 1

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[51] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, pages 341–349, 2012. 2

[52] L. Xu, J. S. Ren, C. Liu, and J. Jia. Deep convolutional neural network for image deconvolution. In *Advances in Neural Information Processing Systems*, pages 1790–1798, 2014. 2

[53] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision*, 2017. 2, 5