Deep Learning for Multi-path Error Removal in ToF Sensors

Gianluca Agresti and Pietro Zanuttigh(✉)

Department of Information Engineering, University of Padova, Padova, Italy
{gianluca.agresti,zanuttigh}@dei.unipd.it

Abstract. The removal of Multi-Path Interference (MPI) is one of the major open challenges in depth estimation with Time-of-Flight (ToF) cameras. In this paper we propose a novel method for MPI removal and depth refinement exploiting an ad-hoc deep learning architecture working on data from a multi-frequency ToF camera. In order to estimate the MPI we use a Convolutional Neural Network (CNN) made of two sub-networks: a coarse network analyzing the global structure of the data at a lower resolution and a fine one exploiting the output of the coarse network in order to remove the MPI while preserving the small details. The critical issue of the lack of ToF data with ground truth is solved by training the CNN with synthetic information. Finally, the residual zero-mean error is removed with an adaptive bilateral filter guided from a noise model for the camera. Experimental results prove the effectiveness of the proposed approach on both synthetic and real data.

Keywords: ToF sensors · Denoising · Multi-path interference · Depth acquisition · Convolutional Neural Networks

1 Introduction

Time-of-Flight (ToF) cameras are active range imaging systems able to estimate the depth of a scene by illuminating it with a periodic amplitude modulated light signal and measuring the phase displacement between the transmitted and received signal [1]. These sensors achieved a wide popularity thanks to their ability to acquire reliable 3D data at video frame rate. In this paper we propose a method for ToF data denoising that focus on the removal of the Multi-Path Interference (MPI) corruption and of the zero-mean error caused by photon shot and sensor thermal noise. ToF acquisitions rely on the key assumption that the received light signal has been reflected only once inside the scene. Unfortunately this is not true in practice and the projected light can be reflected multiple times before going back to the ToF sensor: this issue is called Multi-Path Interference

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-11015-4_30) contains supplementary material, which is available to authorized users.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11131, pp. 410–426, 2019.
https://doi.org/10.1007/978-3-030-11015-4_30

(MPI) and it is one of the main sources of error in ToF data acquisition. The MPI leads to a depth overestimation and this phenomenon is scene dependent, indeed it is related to both the geometry and the properties of the materials inside the scene. The removal of MPI is an ill posed problem with standard single frequency ToF acquisitions but since MPI depth distortion is related to the modulation frequency of the ToF signal, multi-frequency ToF (MF-ToF) sensors can be used for MPI estimation. Some approaches [2–4] following this rationale have been proposed even if with not completely satisfactory results.

Convolutional Neural Networks (CNN) have been widely employed for tasks like denoising and super resolution of image and video data. The application of CNNs to data acquired with ToF cameras for denoising and MPI removal has been investigated only in few works [5–9] due to the difficulty of acquiring the depth ground truth information needed for supervised learning. Here we exploit the information acquired with MF-ToF sensors as input for a deep network architecture able to estimate the unknown MPI corruption. We designed an ad-hoc CNN for this task made of two parts, a coarse one able to understand the global structure of the scene and to globally locate MPI, and a fine one that takes in input the output of the coarse network and allows us to remove the MPI while preserving the small details. Furthermore, an ad-hoc pre-processing step combines the information about the depth and amplitude at multiple frequencies into novel representations that allows the network to learn key clues to estimate the MPI. The critical task of training the deep network has been solved by constructing a synthetic dataset and generating the MF-ToF data using a ToF simulator. The MPI corruption is then removed by subtracting the CNN estimation of the interference. Finally an adaptive bilateral filter guided by an estimation of the ToF noise is used to remove also the zero-mean error. The experimental evaluation has been done on both synthetic and real data, proving how the training on synthetic data can generalize to real world scenes.

2 Related Works

Many different approaches for MPI removal in continuous wave ToF systems employing sinusoidal waveforms have been proposed [2,3,10,11] and an extensive review can be found in [12]. This task is particularly complex since it is an ill-posed problem regarding the retrieval of the sinusoidal light waves related to the shortest paths linking the ToF pixels to the scene points. This is due to various reasons: first of all, the light rays which are interfering are sinusoidal waves at the same modulation frequency and the MPI effects can not be directly detected only by looking at the received waveform in the single frequency case. Moreover since the MPI is scene dependent, the scene geometry is needed to solve the problem but the MPI needs to be removed to estimate the geometry thus creating a chicken-and-egg problem. There are four main families of approaches for MPI correction: methods that use *single frequency ToF data and scene geometry*, methods based on *ray separation*, methods based on *direct and global light separation* and those based on *machine learning* approaches.

The methods which use single frequency ToF data exploit some reflection models in order to estimate the geometry of the scene and correct MPI as done by Fuchs et al. in [13], where reflections with a maximum of 2 bounces are considered. This method is further extended in [14] where multiple albedo and reflections are taken in account. Jimenez et al. [15] proposed a radiometric model to simulate ToF acquisitions and the reflection phenomenon and then correct MPI through non linear optimization.

In methods based on *ray separation*, the light is described as a summation of single sinusoidal waves which are interfering one another in case of MPI. The ray with the shortest path is assumed to be the one carrying the correct depth information (direct light). The method proposed by Freedman et al. in [2] uses 3 modulation frequencies and exploits MPI *frequency diversity* and an L_1 *optimization* to find the light backscattering vector that is assumed to be sparse. In [3], a closed form solution for MPI removal using multi-frequency ToF data is proposed. A method based on the backscattering vector estimation by using random on-off codes instead of standard sinusoidal waveforms for light modulation is proposed in [10].

In the third family of approaches the light is described as the summation of only two sinusoidal waves, one related to the direct component while the other groups together all the sinusoidal waves related to global light (the summation of all the interfering rays). Gupta et al. [11] proposed to use an high modulation frequency to cancel out the sinusoidal global component of the light. The methods proposed by Naik et al. [16], Whyte et al. [17] and Agresti et al. [18] are inspired by the work presented by Nayar in [19]. These methods use an external projector to illuminate the scene with spatial high frequency patterns modulated by the ToF sinusoidal signal to separate the global and direct component of the light and correct MPI.

Only recently, methods based on deep learning have been used on ToF data for denoising purpose. This is due to the fact that depth ground truth is difficult to collect. In [5], ToF data is acquired from a robotic arm setup and the depth ground truth is estimated with a Structured Light system. In [6], an auto-encoder CNN is used with a 2 phase training, in the first phase real depth data without ground truth is used, then the encoder part is kept fix and the decoder part is trained with a synthetic dataset in order to learn how to correct MPI. These methods have in input data taken from single-frequency ToF cameras while our proposal rely on data acquired with multi-frequency ToF cameras. We will show that this choice will improve the MPI correction performance as also done very recently by Su et al. in [8] and Guo et al. in [9]. The first presents an end-to-end deep learning approach to directly estimate a denoised depth map from raw correlation frames. In the second the motion blur is taken in account.

It is possible to find in literature many applications of CNN for 3D estimation from monocular and stereo RGB images [20–22]. An example is the monocular estimation method proposed by Eigen in [21], that exploits a Coarse-Fine network, and its improvement proposed in [22] where a multi-scale CNN is used.

3 Proposed Method

The task of the proposed method is to obtain accurate ToF depth data by removing MPI corruption and reducing zero-mean error related to shot noise. The MPI is estimated by exploiting a CNN whose input are data extracted from a MF-ToF camera, while for the zero-mean error reduction instead we use an adaptive bilateral filter guided by the noise statistic estimated on the input data.

ToF camera pixels are able to compute the correlation function between the received light sinusoidal waves and a reference signal: the computed correlation function appears to be a sinusoidal wave that can be modeled as:

$$c(\theta_i) = B + A\cos\Big(\theta_i - \frac{4\pi f_m \cdot d}{c_l}\Big) = B + A\cos\Big(\theta_i - Kd\Big) \tag{1}$$

where $\theta_i \in [0; 2\pi)$ is the phase sample of the correlation function that is captured by the ToF pixels (nowadays ToF cameras use 4 samples), f_m is the modulation frequency of the light signal, B and A are respectively proportional to the intensity and the amplitude of the received signal, c_l is the speed of light. The depth d of the observed scene point can then be estimated from the 4 correlation samples [1]. This model is correct if each camera pixel receives only one ray from the scene, the direct one. If the pixels receive the summation of light rays reflected multiple times, we have the MPI phenomenon. In this case the ToF correlation function can be modeled as

$$c(\theta_i) = B + A_d\cos\Big(\theta_i - Kd_d\Big) + \int_{d_d+\epsilon}^{\infty} A'_x \cos\Big(\theta_i - Kx\Big)dx = B + A_{FF}\cos(\theta_i - Kd_{FF}) \tag{2}$$

where A_d and d_d are respectively the amplitude and the depth related to the direct light, instead the integral models the global light as the superposition of the rays that are reflected more than once inside the scene. Each interfering ray has its own phase offset and amplitude. In case of MPI, i.e., when $\exists x : A'_x \neq 0$, the depth estimated from Eq. 1, d_{FF}, will be bigger than the correct depth d_d. The phase offsets of the interfering rays are frequency dependent and by changing f_m also the estimated depth d_{FF} and the amplitude A_{FF} will be different. This *frequency diversity* can be used to understand if MPI is acting on MF-ToF cameras and can give us cues for its correction as discussed in [2,3].

In this paper we are going to use data from a ToF camera that captures the scene using the modulation frequencies of 20, 50 and 60 MHz. We will extract some features that are meaningful for MPI analysis directly exploiting the *frequency diversity* on the acquired depth and amplitude images. Moreover, we are going to use also information about the geometry of the scene to estimate the MPI as done in some approaches using single frequency data as [6,13,14]. We devised a CNN for the prediction of the MPI corruption to use all these aspects together. The general architecture of the proposed approach for ToF depth denoising is shown in Fig. 1.

The data acquired by the MF-ToF system is first pre-processed in order to extract a representation that contains relevant information about the MPI

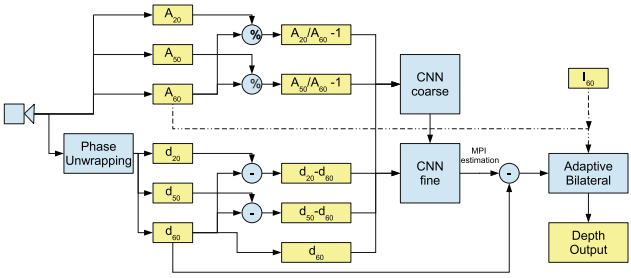

Fig. 1. Architecture of the proposed approach.

presence and strength. As detailed in Sect. 4, where also the motivation for the selection of each input source is presented, the deep network has 5 different input channels containing the ToF depth extracted from the phase at 60 MHz, the difference between the depth maps at different frequencies and the ratio of the amplitudes also at different frequencies.

The CNN architecture of Sect. 5 is made of two main blocks, a coarse network that takes in input the 5 representations and estimates the MPI at low resolution and a fine network that takes in input the 5 representations and the output of the coarse network in order to estimate the MPI interference at full resolution. The estimated multi-path error is then directly subtracted from the ToF depth map (at 60 MHz), thus obtaining a depth map free from multi-path distortion (but still affected by other zero-mean error sources).

The resulting depth map is first filtered with a 3×3 median filter in order to remove depth outliers, then the final output of the proposed method is obtained by further filtering it with an adaptive version of the bilateral filter [23] because of its capability of reducing noise while preserving edges. Bilateral filters have been already used on ToF data [24,25], specially to denoise and upsample the depth map using information from a standard video camera. In our implementation the bilateral filter is guided by the noise information estimated from the received signal amplitude and intensity from which the error variance related to shot noise can be estimated. As suggested in [26], we fixed the spatial smoothing parameter σ_d to a constant value, while the range parameter σ_r is taken proportional to the level of noise. We made the bilateral filter adaptive by using a per pixel noise model for σ_r. In particular we took $\sigma_r = c_r \cdot \tilde{\sigma}_n$, where $\tilde{\sigma}_n$ is an estimate of the depth noise standard deviation due to shot noise. This can be estimated from the amplitude A_{ph} and the intensity I_{ph} of the received light signal [27]:

$$\sigma_n(k,h) = \frac{c_l}{4\sqrt{2}\pi f_m} \frac{\sqrt{I_{ph}(k,h)}}{A_{ph}(k,h)}. \tag{3}$$

A_{ph} and I_{ph} are proportional to the intensity and the amplitude of the correlation function estimated by the ToF pixels. In our experiments we computed $\tilde{\sigma}_n$ from the correlation data and we optimized the values of σ_d and c_r on a subset of

the synthetic training dataset. Then we used the selected values ($\sigma_d = 3$ and $c_r = 3.5$) in the evaluation phase.

4 ToF Data Representation

As mentioned before, we used a CNN to estimate the MPI corruption on the ToF depth map at 60 MHz that is phase unwrapped by using the 20 MHz and 50 MHz ToF data. Notice that these frequency values have been selected since they resemble the ones used in real world ToF cameras. We also investigated the possibility of performing the phase unwrapping using the CNN of Sect. 5, but the disambiguation using the MF data proved to be reliable and the deep network optimization is more stable if already phase unwrapped data is fed to it. A critical aspect is the selection of input data that should be informative about the MPI phenomenon. We decided to use as input the following elements:

- The first input $C_1 = d_{60}$ is the ToF depth map at 60 MHz. It is required not only because it is the corrupted input that needs to be denoised but also because the geometry of the scene influences the MPI error and the ToF depth represents the best estimate of the geometry available before the MPI removal process. We selected the depth captured at 60 MHz since the higher the modulation frequency, the more accurate the depth estimation.
- The difference between the depth maps estimated at the different modulation frequencies, used since the MPI corruption changes with the frequency (generally the higher the modulation frequency, the smaller is MPI [11]). We used the differences between the depths at 20 Mhz and 60 Mhz, and between the ones at 50 Mhz and 60 Mhz, i.e., $C_2 = d_{20} - d_{60}$ and $C_3 = d_{50} - d_{60}$.
- The ratio of the amplitudes of the received light signal at different modulation frequencies. In presence of MPI the light waves experiences destructive interferences and in ToF data acquired in presence of MPI the higher the modulation frequency, the lower the resulting amplitude. For this reason, comparing the amplitudes at different frequencies gives us a hint about the MPI presence and strength. We used the ratios between the amplitudes at 20 Mhz and 60 Mhz, and between the ones at 50 Mhz and 60 Mhz, i.e., $C_4 = (A_{20}/A_{60}) - 1$ and $C_5 = (A_{50}/A_{60}) - 1$ (the "-1" term has been introduced to center the data around 0 in case of MPI absence).

The proposed CNN aims at estimating the MPI corruption on the 60 MHz depth map: the targets for the training procedure have been computed by taking a filtered version of the difference between d_{60} and the ground truth depth d_{GT} (the filtering is used to remove the zero mean error, notice that MPI is a low frequency noise). We decided to use this set of inputs for the proposed *Coarse-Fine CNN* since depth and amplitude are data which are generally accessible from commercial ToF cameras. Moreover, by taking the ratio between the amplitudes we are canceling out the gain of the sensor, that can be different for different sensors, making the method more robust to hardware changes. We have tried

to use subsets of the input data, but this reduced the performance in MPI estimation. Notice that other techniques based on multi-frequency approaches as [2,3] use a per pixel model based on the sparsity of the backscattering vector A'_x of Eq. 2, while in our proposal we are implementing a data-driven model that will suit the diffuse reflection case and thanks to the CNN receptive fields we are capturing the geometrical structure of the scene in addition to the *frequency diversity*. We decided to pre-filter the CNN inputs with a 5×5 median filter to obtain a more stable input and reduce their zero-mean variation.

As aforementioned, it is difficult to collect a real world dataset big enough for CNN training with ToF data and the related depth ground truth. For this reason, we decided to exploit a dataset composed by synthetic scenes, for which the true depth is known. The ToF acquisitions have been performed with the *ToF Explorer* simulator realized by Sony EuTEC starting from the work of Meister et al. [28] which is able to faithfully reproduce ToF acquisition issues like the shot and thermal noise, the read-out noise, artifacts due to lens effects, mixed pixels and specially the multi-path interference. This simulator uses as input the scene data generated by *Blender* [29] and *LuxRender* [30]. In order to build the synthetic dataset we started from the set of *Blender* scenes used in [7]. We used 40 scenes for the training set, while the other 14 different scenes have been used for testing. Each scene has been rendered from a virtual viewpoint with the ToF simulator in order to acquire the ToF raw data (amplitude, intensity and depth image) at the modulation frequencies of 20, 50 and 60 MHz. We also used the Blender rendering engine to acquire the scene depth ground truth. The dataset is publicly available at http://lttm.dei.unipd.it/paper_data/MPI_CNN. The various scenes contain surfaces and objects of different shapes and texture and correspond to very different environments. They also have a very wide depth acquisition range from about 50 cm to 10 m and various corners and structures in which the multi-path phenomenon is critical.

We collected also a set of real scenes with related depth ground truth in order to validate the proposed method for ToF depth refinement. We used a *SoftKinetic* ToF camera in combination with an active stereo system for the acquisitions. The stereo system and the ToF camera have been jointly calibrated and ground truth depth estimated with the active stereo system has been reprojected on the ToF sensor viewpoint. We set up a wooden box (see Fig. 2) that is about 1.5 m wide and 1 m high and composed by a 90° and a 135° angle. The real world dataset is composed by 8 scenes, each captured by looking at the box from different viewpoints and by placing in it objects made of wood, polystyrene and ceramic. Since the acquired dataset is quite small we used it only for testing purposes, while we used the synthetic data for the training of the network. By looking at the scene in Fig. 2, it is possible to see how some critical situations for MPI are present, e.g., the surface angles that are the typical cases where MPI corruption can be clearly observed.

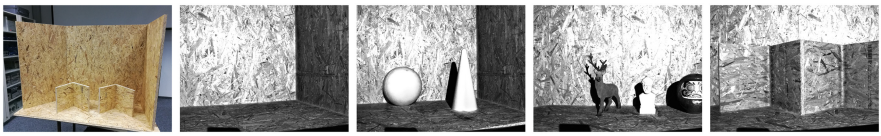

Fig. 2. Wooden box used for the real world acquisition and examples of amplitude images of the acquired scenes. The wooden box has been captured from different viewpoints and with different objects inside.

5 Proposed Deep Learning Architecture

The architecture of the proposed *Coarse-Fine CNN* is shown in Fig. 3: the network is made of two main parts, a coarse sub-network and a fine one.

Since the MPI phenomenon depends on reflections happening in different locations, a proper estimation of its presence needs a relatively wide receptive field of the CNN in order to understand the geometrical structure of the scene. Following this rationale, the coarse network performs an analysis of the input data by applying downsampling with pooling layers increasing the receptive field as a consequence. The coarse network takes in input the 5 data channels described in Sect. 4 and is made of a stack of 5 convolutional layers each followed by a ReLU with the exception of the last one. The first 2 convolutional layers are also followed by a max-pooling stage reducing the resolution of a factor of 2. All the layers perform 3×3 pixels convolutions and have 32 filters, except the last one that has a single filter, producing as output a low resolution estimate of the MPI. The estimated MPI error is finally upsampled of a factor of 4 using a bilinear interpolation in order to bring it back to the original input resolution. This network allows us to obtain a reliable estimate of the regions affected by MPI but, mostly due to the pooling operations, the localization of the interference is not precise and directly subtracting the output of this network to the acquired data would lead to artifacts specially in proximity of the edges. For this reason, we used a second network working at full resolution to obtain a more precise localization of the error. This second network also has 5 convolutional layers with 3×3 convolutions and ReLU activation functions (except the last as before). It has instead 64 filters for each layer and no pooling blocks. The input of the first layer is the same of the previous network but the fourth layer takes as input not only the output of the third layer but also the upsampled output of the coarse network. This allows us to combine the low resolution estimation with a wide receptive field of the previous network with the more detailed but local estimation done by the fine network and to obtain an MPI estimation that captures both the scene global structure and the fine details.

The network has been trained using the synthetic dataset of Sect. 4. Even if it is one of the largest ToF dataset with multi-frequency data and ground truth information, its size is still quite small if compared to datasets typically used for CNNs training. In order to deal with this issue and avoid over-fitting we applied data augmentation techniques on the training data as random sampling

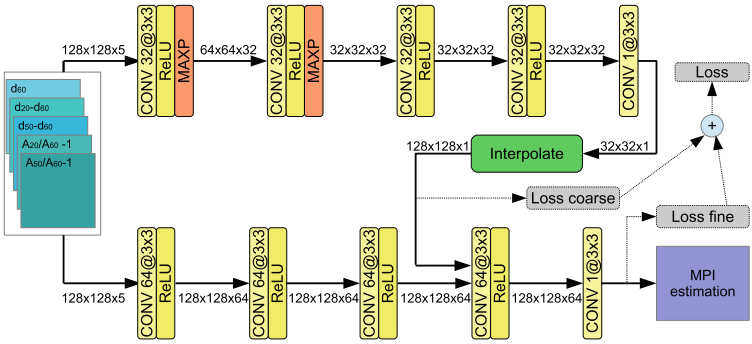

Fig. 3. Architecture of the Coarse-Fine CNN used for MPI estimation in ToF data

of patches, rotation and flipping operations. We extracted 10 random patches of size 128×128 pixels from each of the 40 scenes, then we applied to each of them a rotation of $\pm 5°$ and horizontal and vertical flipping. This leads to a total of about $40 \times 10 \times 5 = 2000$ patches (invalid patches with non complete covering on rotated images have been excluded), that represents a good amount of data for the training of the proposed deep network. The number of patches could be increased by using smaller patches, but this would weaken the ability of the network to understand the geometrical structures of the scenes and to retrieve the MPI corruption.

Due to the small amount of data we have used *K-fold validation* with $K = 5$ on the training set to validate the hyper-parameters of the CNN and of the training procedure as the architecture of the network, the number and depth of the layers, the learning rate and the regularization constant. We have selected the CNN hyper-parameters in order to avoid overfitting and obtain the minimum mean validation MAE among the 5 folds. Once the hyper-parameters have been selected, the CNN has been trained on the whole training set.

For the training we minimized a combined loss made by the sum of two loss functions, one computed on the interpolated output of the coarse network and the other computed on the output of the fine network. This approach allowed to obtain better performances than the separate training of the two sub-networks. Each of the two loss functions is the l_1 *norm* of the difference between the MPI error estimated by the corresponding network and the MPI error computed by comparing the ToF depth at 60 MHz with true depth as described in Sect. 4. The l_1 *norm* is more robust to outliers in the training process if compared with the l_2 *norm* and had more stable results in the validation of the network hyper-parameters. Furthermore the use of l_1 *norm* proved to be more efficient for image denoising [31]. During the training, we exploited the *ADAM* optimizer [32] and a batch size of 16. We started the training with an initial set of weight values derived with Xavier's procedure [33], a learning rate of 10^{-4} and a l_2 regularization with a weighting factor of 10^{-4} for the *norm* of the CNN weights. Figure 4 shows the mean training and validation error across all the epochs of

the *K-fold validation*: we trained the network for 150 epochs, that in our case proved to be enough for the validation error to stabilize. The network has been implemented using the *TensorFlow* framework and the training took about 30 minutes on a desktop PC with an Intel i7-4790 CPU and an *NVIDIA Titan X (Pascal)* GPU. The evaluation of a single frame with the proposed network takes instead just 9.5 ms.

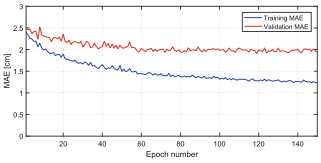

Fig. 4. Mean training (blue) and validation (red) error at each epoch of the *K-fold validation*. (Color figure online)

6 Experimental Results

In order to evaluate the proposed approach we used the two different datasets presented in Sect. 4. The first is the synthetic dataset, that has been used both for the training of the CNN and for the evaluation of the performances of the proposed method. The second one is a smaller real world dataset that has been used only for evaluation purposes due to its limited size.

6.1 Results on Synthetic Data

As already pointed out, we kept 14 synthetic scenes for evaluation purposes only. The scenes used for testing are shown in Fig. 5, notice how they include various types of settings with different sizes, types of textures and several situations where the multi-path error can arise. Figure 6 shows the results of the application of the proposed approach on a subset of the scenes used for testing. It shows the input depth map from the ToF camera at 60 MHz (with phase unwrapping), the depth map after the application of the adaptive bilateral filter and the final result of the proposed approach with their related errors map and the depth ground truth information. By looking at the third and fourth columns it is possible to notice how the adaptive bilateral filter is able to reduce the zero-mean error by preserving the fine details in the scenes, e.g., the small moon in the *castle* is preserved by the filtering process, but the depth overestimation due to MPI is still present. From the fifth and sixth columns it is possible to see how both the multi-path error and the zero-mean noise have been widely reduced by the complete version of the proposed approach. For example in the first 3 scenes there is a very strong multi-path distortion on the walls in the back that has been almost completely removed by the proposed approach for MPI correction. The multi-path estimation is very accurate on all the main surfaces of the scenes, even if the task proved to be more challenging on some small details like the top

of the pots in row 1 or the stairs in row 2. However notice that thanks to the usage of the Coarse-Fine network the small details of the various scenes are preserved and there is no blurring of the edges. This can be seen for example another time from the details of the castle (e.g., the moon shape) in row 3. The box scene (row 4) is another example of the MPI removal capabilities. Notice how the multi-path on the edges between the floor and the walls is correctly removed. Also the error on the slope in the middle of the box (that is more challenging due to bounces from locations farther away) is greatly reduced even if not completely removed. This evaluation is confirmed also by numerical results, the Mean Absolute Error (MAE) is reduced from 156 mm on the input data to 70 mm.

Fig. 5. Synthetic test set used for evaluating the proposed approach. The figure shows a color view for each scene in the dataset. (Color figure online)

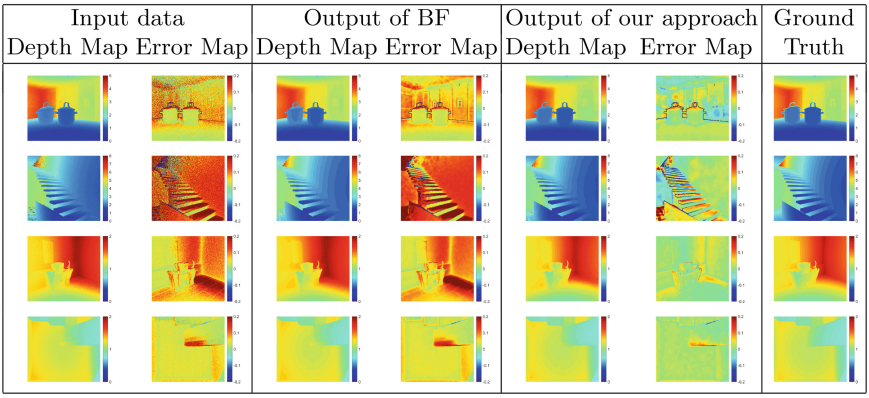

Fig. 6. Input depth map at 60 MHz, output of the adaptive bilateral filter (BF) and output of the proposed approach (with MPI correction) on same sample synthetic scenes with the corresponding error maps. All the values are measured in meters.

Figure 7 compares the error at 60 MHz with its estimation made by the proposed CNN architecture. The second column shows the estimation taken from the interpolated output of the coarse network: notice how the general distribution of the MPI is correctly estimated but the edges and details (e.g., the moon over the castle) are lost in this estimation due to the pooling operations that reduce the resolution. The last column shows instead the output of the Coarse-Fine architecture and it is possible to notice how the general MPI distribution is maintained but there is a much higher precision on boundaries and small details.

The second row shows the same data for the *stairs* scene, also in this case notice how the general structure is the same but the estimation follows more accurately the shape of the stairs in the Coarse-Fine output.

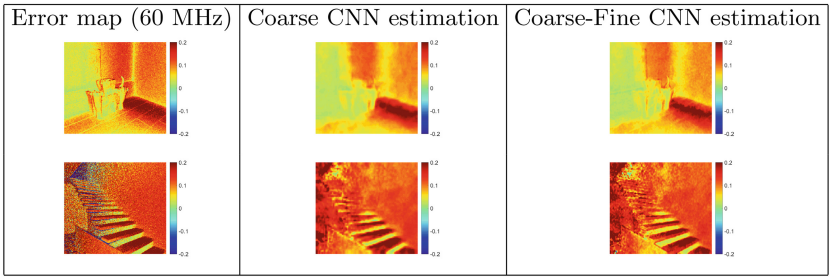

Fig. 7. Estimation of the MPI performed by the proposed approach using only the coarse network or the complete Coarse-Fine architecture.

We also compared the proposed approach with some competing approaches from the literature. In particular we considered the MF-ToF MPI correction scheme proposed by Freedman [2] and the approach based on deep learning presented by Marco in [6] that takes in input the depth map at 20 MHz to remove MPI. The method proposed by Freedman was adapted to use the same triple of frequencies used by the proposed approach. The first column of Table 1 shows the MAE obtained by comparing the output of the 3 methods with the ground truth data on the synthetic dataset. Our approach is able to reduce the error from 156 to 70 mm, reducing it to less than half of the original error. It outperforms with a wide margin both competing approaches. The Freedman method [2] is able to remove only about 10% of the error in the source data obtaining an accuracy of 140 mm. The method of [2] works under the hypothesis that the light backscattering vector is sparse and this is not true in scenes where diffuse reflections are predominant as the considered ones. For this reason, its effectiveness is limited. The method of [6] works under the assumption that the reflections are diffuse and it achieves better results removing about 20% of the original error, but it is still far from the performances of the proposed approach. This is due to the fact that the CNN proposed in [6] uses single frequency ToF data, instead we have shown that the multi-frequency approach can achieve much higher performance using a less complex CNN. The additional material contains also a detailed analysis of the different methods behavior in proximity of corners.

6.2 Results on Real World Data

After evaluating the proposed approach on synthetic data we preformed also some experiments on real world data. For this evaluation we used the real test set introduced in Sect. 4 that is composed by 8 scenes. It has a more limited

Table 1. Mean MAE for competing schemes from the literature and for the proposed approach on synthetic and real world data. The table shows the MAE in millimeters and the relative error between the output of the various methods and the error on input data. Our approach and [2] are multi-frequency methods and are compared with the highest employed frequency (60 MHz) for the relative error, instead [6] (*) is compared with the only frequency it uses (20 MHz).

	Synthetic data		Real World data	
	MAE (*mm*)	Relative Err. (%)	MAE (*mm*)	Relative Err. (%)
ToF input (60 Mhz)	167.3	-	54.3	-
ToF input (20 Mhz)	327.8	-	72.8	-
Freedman et al. [2]	149.8	89.5%	51.1	94.1%
Marco et al. [6]	260.9*	79.6%	51.3*	70.5%
Our approach	**74.9**	**44.8%**	**31.9**	**58.7%**

variety of settings with respect to the synthetic data but still the scenes contain objects of different sizes, types of material and surfaces with different orientations where the MPI can arise. The Coarse-Fine CNN was trained on the synthetic dataset that is composed by scenes with ideal properties, e.g., the reflections are perfectly diffuse, and due to some limitations of the simulator, the synthetic data, even if quite accurate, does not exactly model all the issues of real data.

Figure 8 shows the results of the application of the proposed approach to the set of real world scenes. As before, it shows the input depth map from the ToF camera at 60 MHz and the depth map resulting after the application of the proposed approach with their corresponding error maps and ground truth information. By looking at the images, it is possible to see how the MPI interference is reduced by the application of the proposed approach even if some MPI error remains in the scenes. It is possible to notice how the MPI is almost completely removed on the vertical walls, in particular in proximity of edges between facing surfaces. The reduction is strong also on the small objects like the sphere, the cone or the deer even if some multi-path in proximity of boundaries remains on these objects. On the other side the MPI error is under-estimated on surfaces with a strong inclination, in particular the floor in the various scenes, where the approach is able to reduce only part of the multi-path. By comparing Figs. 6 and 8 it is possible to notice how the strong MPI on these surfaces (e.g., the floors) is not present in the synthetic scenes. This is probably due to the fact that reflections happening on the considered real materials are not ideally diffuse when the light rays are strongly inclined and the ToF simulator does not model this phenomenon. Our approach, as any other machine learning scheme, learns from the training data and is not able to correct issues not present in the training examples.

We compared our approach with [2] and [6] also on the real world data. The results are in the third and fourth column of Table 1. On real data our approach is able to reduce the error from 54.3 to 31.9 mm, i.e., to 58.7% of

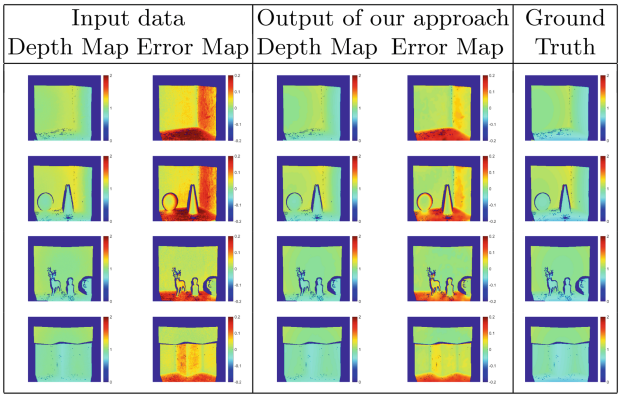

Fig. 8. Input depth map at 60 MHz and output of the proposed approach on same sample real world scenes with the corresponding error maps.

the original error, a very good performance outperforming both the compared approaches even if lower than the one achieved on synthetic data. In particular the proposed method was able to improve the accuracy of the depth estimation on all the considered scenes: in the worst case scene the error reduction is around 29%. Recall that the training is done on synthetic information only, as pointed out in the visual evaluation the issues on the floor reduce the performances. The error removal capability of [2] is limited also in this case, it removes about 6.5% of the error. The method of [6] removes about 30% of the error and gets a bit closer to ours in this experiment, but there is still a gap of more than 10%.

7 Conclusions

In this paper we proposed a novel method for MPI removal and denoising in ToF data. We extracted from MF-ToF data multiple clues based on depth differences and amplitude ratios that proved to be very informative about the MPI presence. Furthermore, by using a Coarse-Fine deep network we were able both to capture the general structure of the MPI interference and to preserve the small details and edges of the scene. Finally, we dealt with the critical issue of ground truth for training data by using synthetic information. Experimental results demonstrated how the proposed approach is able to remove the MPI interference and to remove the ToF data noise without introducing artifacts on both synthetic and real world scenes. Results are impressive on synthetic data and good on real data, but in the second case there are still some limitations due to the differences between the simulated training data and real world acquisitions. For this reason further research will be devoted at improving the results on real data both by improving the realism of the synthetic data and by trying multi-stage training procedures using together synthetic and real data. Semi-supervised learning strategies and Generative Adversarial Networks (GANs) will also be investigated.

Acknowledgment. We would like to thank the Computational Imaging Group at the Sony European Technology Center (EuTEC) for allowing us to use their *ToF Explorer* simulator and Oliver Erdler, Markus Kamm and Henrik Schaefer for their precious comments and insights. We also thank prof. Calvagno for his support and gratefully acknowledge NVIDIA Corporation for the donation of the GPUs used for this research.

References

1. Zanuttigh, P., Marin, G., Dal Mutto, C., Dominio, F., Minto, L., Cortelazzo, G.M.: Time-of-Flight and Structured Light Depth Cameras. Springer, Switzerland (2016). https://doi.org/10.1007/978-3-319-30973-6
2. Freedman, D., Smolin, Y., Krupka, E., Leichter, I., Schmidt, M.: SRA: fast removal of general multipath for ToF sensors. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 234–249. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_16
3. Bhandari, A., et al.: Resolving multipath interference in time-of-flight imaging via modulation frequency diversity and sparse regularization. Optics Lett. **39**(6), 1705–1708 (2014)
4. Peters, C., Klein, J., Hullin, M.B., Klein, R.: Solving trigonometric moment problems for fast transient imaging. ACM Trans. Graph. (TOG) **34**(6), 220 (2015)
5. Son, K., Liu, M.Y., Taguchi, Y.: Learning to remove multipath distortions in time-of-flight range images for a robotic arm setup. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), pp. 3390–3397 (2016)
6. Marco, J., et al.: Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. ACM Trans. Graph. (TOG) **36**(6), 219 (2017)
7. Agresti, G., Minto, L., Marin, G., Zanuttigh, P.: Deep learning for confidence information in stereo and ToF data fusion. In: Geometry Meets Deep Learning ICCV Workshop, pp. 697–705 (2017)
8. Su, S., Heide, F., Wetzstein, G., Heidrich, W.: Deep end-to-end time-of-flight imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6383–6392 (2018)
9. Guo, Q., Frosio, I., Gallo, O., Zickler, T., Kautz, J.: Tackling 3D ToF artifacts through learning and the FLAT dataset. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 381–396. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_23
10. Kadambi, A., et al.: Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. ACM Trans. Graph. (TOG) **32**(6), 167 (2013)
11. Gupta, M., Nayar, S.K., Hullin, M.B., Martin, J.: Phasor imaging: a generalization of correlation-based time-of-flight imaging. ACM Trans. Graph. (TOG) **34**(5), 156 (2015)
12. Whyte, R., Streeter, L., Cree, M.J., Dorrington, A.A.: Review of methods for resolving multi-path interference in time-of-flight range cameras. In: IEEE Sensors, pp. 629–632. IEEE (2014)
13. Fuchs, S.: Multipath interference compensation in time-of-flight camera images. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 3583–3586. IEEE (2010)
14. Fuchs, S., Suppa, M., Hellwich, O.: Compensation for multipath in ToF camera measurements supported by photometric calibration and environment integration.

In: Chen, M., Leibe, B., Neumann, B. (eds.) ICVS 2013. LNCS, vol. 7963, pp. 31–41. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39402-7_4
15. Jiménez, D., Pizarro, D., Mazo, M., Palazuelos, S.: Modeling and correction of multipath interference in time of flight cameras. Image Vis. Comput. **32**(1), 1–13 (2014)
16. Naik, N., Kadambi, A., Rhemann, C., Izadi, S., Raskar, R., Bing Kang, S.: A light transport model for mitigating multipath interference in time-of-flight sensors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 73–81 (2015)
17. Whyte, R., Streeter, L., Cree, M.J., Dorrington, A.A.: Resolving multiple propagation paths in time of flight range cameras using direct and global separation methods. Opt. Eng. **54**(11), 113109 (2015)
18. Agresti, G., Zanuttigh, P.: Combination of spatially-modulated ToF and structured light for MPI-free depth estimation. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018 Workshops. LNCS, vol. 11129, pp. 355–371. Springer, Cham (2018)
19. Nayar, S.K., Krishnan, G., Grossberg, M.D., Raskar, R.: Fast separation of direct and global components of a scene using high frequency illumination. ACM Trans. Graph. (TOG) **25**(3), 935–944 (2006)
20. Luo, W., Schwing, A.G., Urtasun, R.: Efficient deep learning for stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5695–5703 (2016)
21. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems, pp. 2366–2374 (2014)
22. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2650–2658 (2015)
23. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of International Conference on Computer Vision (ICCV), pp. 839–846. IEEE (1998)
24. Chan, D., Buisman, H., Theobalt, C., Thrun, S.: A noise-aware filter for real-time depth upsampling. In: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008 (2008)
25. Marin, G., Zanuttigh, P., Mattoccia, S.: Reliable fusion of ToF and stereo depth driven by confidence measures. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 386–401. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_24
26. Zhang, M., Gunturk, B.K.: Multiresolution bilateral filtering for image denoising. IEEE Trans. Image Process. **17**(12), 2324–2333 (2008)
27. Lange, R., Seitz, P., Biber, A., Lauxtermann, S.C.: Demodulation pixels in CCD and CMOS technologies for time-of-flight ranging. In: Sensors and Camera Systems for Scientific, Industrial, and Digital Photography Applications, vol. 3965, pp. 177–189. International Society for Optics and Photonics (2000)
28. Meister, S., Nair, R., Kondermann, D.: Simulation of time-of-flight sensors using global illumination. In: Bronstein, M., Favre, J., Hormann, K. (eds.) Vision, Modeling and Visualization. The Eurographics Association, Goslar (2013)
29. The Blender Foundation: Blender website. https://www.blender.org/. Accessed 14 Mar 2018
30. The LuxRender Project: Luxrender website. http://www.luxrender.net. Accessed 14 Mar 2018

31. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Trans. Comput. Imaging **3**(1), 47–57 (2017)
32. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
33. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)