Multi-Frame Depth Super-Resolution for ToF Sensor With Total Variation Regularized L1 Function

JONGHYUN KIM, JAEDUK HAN, AND MOON GI KANG
School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea
Corresponding author: Moon Gi Kang (mkang@yonsei.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C2002167).

ABSTRACT In this paper, we propose a multi-frame depth super-resolution (SR) method based on $L_1$ data fidelity with the total variation regularization (TV-$L_1$) model. The majority of time-of-flight (ToF) sensors exhibit limited spatial resolution compared to RGB sensors and the improvement of the depth image resolution is an inherently ill-posed problem. To overcome this under-determined problem, the solution space is limited by the regularization term through prior knowledge and the data fidelity term using statistical information of the noise. Firstly, the statistical characteristics of ToF depth images are analyzed to specify the appropriate observation model. Thereafter, the objective function for multi-frame depth SR based on the TV-$L_1$ model is designed by considering the prior knowledge of the depth images. This approach enables the sharpness of the edges to be preserved and the noise amplification to be suppressed simultaneously. Furthermore, an efficient solver based on half-quadratic splitting is proposed. The algorithm minimizes the objective function for the multi-frame SR problem consisting of the TV regularization term and $L_1$ data fidelity term. The proposed method is verified on a synthetic dataset and real-world data acquired from a ToF sensor. The experimental results demonstrate that the proposed method can substantially reconstruct high-resolution depth images in terms of preserving sharp depth discontinuities, without any obvious artifacts, and can increase robustness to noise.

INDEX TERMS Depth super-resolution, time-of-flight sensor, multi-frame super-resolution, total variation regularization, $L_1$ data fidelity.

I. INTRODUCTION

Accurate depth (distance) measurement is an essential requirement in image processing and computer vision. In recent years, the demand for accurate depth information has increased in numerous applications, such as image enhancement [1], virtual reality [2], robotics [3], and autonomous driving [4]. In particular, face recognition technology that utilizes depth information [5] has been adopted in traditional business models to provide access control in various fields (e.g., finance systems or identity verification). Although the significance of depth images is rapidly increasing, several artifacts are still frequently observed. For example, integrated detectors or sensors are usually disturbed by visible light from the background and certain devices suffer from interferences between waves during depth measurements. Moreover, depth images exhibit the additional problem of low quality because the depth measurement requires high costs in terms of complexity or hardware implementation.

Various methods for depth sensing have been developed over the past several decades. The structured light scanner can measure the distances to objects using projected light patterns [6]. Light scanners can measure accurate distances in certain environments, such as those with static objects. Nonetheless, these scanners generally capture distorted depth images because the projection onto the surfaces of moving objects is typically biased. A method often referred to as stereo photography uses a stereo camera, which is a type of camera with two or more lenses with a separate image sensor, as well as independent optical systems to enable the camera to imitate the human binocular vision system. However, stereo
cameras require exhaustive computational costs to estimate the point of inconsistency between two independent sensors. Furthermore, although the performance of the stereo matching algorithm has been improved, stereo cameras may exhibit limitations in environments under low-light conditions or with severe noise.

The time-of-flight (ToF) is the measurement of the time taken by an electromagnetic wave to travel a specific distance to an object. ToF cameras employ the ToF technique to resolve the distance between the camera and objects for each point of the depth image. These cameras measure the time required by near-infrared (NIR) to hit the object and then return to the camera [7]. The transmitter transmits NIR with a specific frequency and the receiver detects a phase change with respect to the time. That is, the phase difference between the input and the reflected signals can easily be measured to calculate the flight time. Although the time can be calculated based on simple phase detection in this manner, the majority of ToF cameras have a limited spatial resolution because ToF sensors experience difficulties in scaling the sensor architecture to large arrays owing to the size of the on-chip time-to-digital converters required [8].

Numerous attempts have been made to overcome the physical limitations of ToF sensors, such as their limited spatial resolution. Algorithms for improving the depth image resolution of ToF sensors can be categorized into three groups, which are described as follows: The first approach is single-frame depth upsampling, which aims to recover a high-resolution (HR) image from a single low-resolution (LR) image [9], [10]. Such methods generally use the statistical characteristics inside the depth image or those of the training images. In recent years, numerous test images have been employed to improve single-frame depth upsampling. For example, deep learning methods, which are used extensively to model the mapping from LR patches to HR patches using a convolutional neural network (CNN), have achieved remarkable success in natural image processing [11]. However, single-frame depth upsampling methods exhibit limited performance owing to issues such as a loss of details and jagging artifacts compared to the methods in the other two categories, because they do not employ additional hardware, extra sensors, or alignment.

The second approach employs an intensity image (e.g., an RGB image), the field of view of which is the same as the ToF camera, as additional information. The additional input of these methods stems from the correlation between the depth and intensity discontinuities. These techniques assume that the edges of the intensity image coincide with the edges (or discontinuities) of the depth image [12]–[17]. As the additional intensity image is typically captured at a higher resolution, its discontinuities are used to locate the depth discontinuities at a higher resolution than the LR depth image. However, in practice, this approach tends to introduce serious artifacts such as inaccurate depth, depth bleeding, and texture copying artifacts, because the intensity and depth discontinuities do not always coincide.

To avoid artifacts from the intensity image, methods for increasing the spatial resolution only from the depth data have been proposed. The third approach restores the high-frequency information from multiple LR depth images. This approach originates from the classical super-resolution (SR) theories in image processing, and it is assumed that aliasing is always present in the LR images [18]. Hardie et al. have formulated a joint approach for reconstruction and registration using Tikhonov regularization [19]. Farsiu et al. have introduced a fast and robust scheme based on bilateral total variation (BTV) regularization with $L_1$ data fidelity [20]. An algorithm to solve the nonlinear total variation (TV) based SR problem for digital video has been proposed in [21]. Babacan et al. have utilized the quadratic approximation of the TV prior based on a variational Bayesian formulation [22]. A combination of the sparse and non-sparse priors has been introduced in [23].

In terms of increasing the resolution of the ToF sensor, a technique using multi-frame depth images has been developed by extending the classical SR method for existing cameras [24]. Recently, a method based on the TV prior model has been proposed to preserve the overall sharpness of depth images [25]. However, the multi-frame depth SR techniques for ToF sensors presented in [24] and [25] do not reflect the depth image characteristics, and do not consider problems such as oversmoothing and noise amplification. Therefore, we propose a method based on the $L_1$ data fidelity with the TV regularization (TV-$L_1$) model to overcome the limitations of other methods in terms of preserving the sharp depth discontinuities and suppressing the noise amplification.

In this paper, a multi-frame depth SR approach is proposed to improve the spatial resolution of the depth image. Our approach can be classified into the third category: multiple LR depth images are used to reconstruct a single HR depth image. The main contributions of this study are as follows:

- We analyze the statistical characteristics of ToF depth images.
- We design an objective function for multi-frame depth SR based on the TV-$L_1$ model which is robust to noise and produces sharp depth discontinuities. The experimental results verify that the TV-$L_1$ model is more robust than the $L_2$ data fidelity with the Tikhonov regularization ($\text{TV}+\text{Tikh}$) and the $L_2$ data fidelity with the TV regularization (TV-$L_2$) models.
- We propose an efficient solver based on half-quadratic splitting. The experimental results show that our method is more reliable than previous methods using a similar objective function as well as other non-similar ones.

The remainder of this paper is organized as follows. Section II presents related works on multi-frame depth SR and TV-$L_1$ approach for deconvolution and denoising. In Section III, the analysis of the characteristics of ToF images is discussed and the proposed method based on the TV-$L_1$ model is described in detail. Section IV presents the experimental results. Finally, conclusions are outlined in Section V.
II. RELATED WORK

In this section, we present an observation model for LR depth images. Firstly, we briefly review multi-frame depth SR methods, which are related to our method. Thereafter, we introduce an image restoration algorithm based on the TV-L1 approach. Observed LR depth images are employed to restore the HR depth image.

A. IMAGE OBSERVATION MODEL

SR image reconstruction is the process of estimating an original image from multiple degraded images, with the aim of overcoming the physical limitations of optical and sensor systems. During the process of obtaining depth images, a natural loss of spatial resolution occurs as a result of aliasing owing to the limited spatial resolution of the sensor, optical blur owing to the physical limitations of the optical system, and sensor noise. The observation model can be described as follows:

\[ y_k = DBM_k x + n_k = W_k x + n_k, \]  

(1)

where \( y_k \) is the \( k \)th observed LR depth image, \( D \) is a downsampling matrix, \( B \) is a blurring matrix, \( M \) is a warping matrix, \( x \) is the HR depth image, and \( n_k \) is the corresponding additive noise [18]. These series of degradations are summarized as the single matrix \( W_k \) in this study.

In the observation model of Eq. (1), the task of restoring \( x \) from \( y_k \) is an inverse problem that is solved by an optimization process, as follows:

\[ \hat{x} = \arg\min_x \left\{ \sum_{k=1}^{K} \frac{1}{\lambda_k} D(y_k - W_k x) + R(x) \right\}, \]  

(2)

where \( D \) is the data fidelity term, \( R \) is the regularization term for incorporating prior information, and \( \lambda_k \) is the regularization parameter controlling the trade-off between the two terms. In this formulation, \( K \) denotes the number of LR images in the reconstruction. The process of Eq. (2), which is also known as multi-frame SR and is illustrated in Fig. 1, simultaneously performs inverse filtering to obtain the HR image.

B. MULTI-FRAME DEPTH SR

The data fidelity term is often derived from the negative log-likelihood; that is, \( D(y_k - W_k x) = -\log P(y_k | W_k x) \), and it is usually associated with the noise model. In the general case, where \( n_k \) is assumed to be additive white Gaussian noise, the data fidelity term utilizes the \( L_2 \)-norm, and the corresponding inverse problem is expressed as follows:

\[ \hat{x} = \arg\min_x \left\{ \sum_{k=1}^{K} \frac{1}{\lambda_k} ||y_k - W_k x||_2^2 + R(x) \right\}. \]  

(3)

The data fidelity term with the \( L_2 \)-norm is frequently used as a standard approach in regression analysis, which is known as the method of least squares.

The regularization term is adopted to stabilize the inverse problem by confining the set of feasible solutions. In early years, Tikhonov regularization was representative of smoothing constraints, which assume that the distribution of the image gradients is smooth [18]. This regularization is constructed by employing the \( L_2 \)-norm and is represented as follows:

\[ R(x) = ||\Gamma x||_2^2, \]  

(4)

where \( \Gamma \) denotes the Tikhonov matrix, which is usually constructed by a high-pass filter (e.g., the Laplacian) in image restoration. This regularization strategy, also known as \( L_2 \)-regularization [26], has been applied extensively because fast transforms can be developed with low computational costs [27].

The objective function, in which both the data fidelity and regularization terms employ \( L_2 \)-regularization, can be formulated as:

\[ \hat{x} = \arg\min_x \left\{ \sum_{k=1}^{K} \frac{1}{\lambda_k} ||y_k - W_k x||_2^2 + ||C x||_2^2 \right\}. \]  

(5)
where \( C \) denotes the discrete Laplacian operator. The Tikhonov-L2 model has become popular in various applications. For example, Hardie et al. formulated the multi-frame SR reconstruction method by minimizing the Tikhonov-L2 objective function [19]. In terms of increasing the depth resolution, Schuon et al. proposed multi-frame based depth SR with Eq. (5) as an optimization framework [24].

One of the most successful regularization techniques for image restoration is TV, in which the gradient magnitude of the image is assumed to be sparse. Osher et al. introduced a TV model for regularization, and proposed a denoising and deblurring algorithm based on the regularization strategy [28], [29]. Wang et al. proposed a minimization algorithm using the TV strategy [30]. TV regularization has become a common regularization strategy in image restoration because of its special characteristics. One major problem in image restoration is the ringing artifacts that appear near sharp transitions. Tikhonov regularization suffers from ringing artifacts owing to its assumption of smooth gradients. Unlike in Tikhonov regularization, edge information can be preserved more effectively in TV regularization, and the regularization term based on TV is expressed as follows:

\[
R(x) = ||\nabla x||_1,
\]

where \( \nabla \) denotes the first derivative linear operator. The norm \( || \cdot || \) denotes the \( L_2 \)-norm if the TV regularization is isotropic and it denotes the \( L_1 \)-norm if the TV regularization is anisotropic. However, we only deal with the anisotropic case, \( || \cdot || = || \cdot ||_1 \), because the treatment of the isotropic case is very similar.

The TV regularization strategy is generally combined with the data fidelity term using the \( L_2 \)-norm, which is usually denoted as the TV-L2 model. The objective function, in which the regularization terms employ TV-regularization, can be formulated as follows:

\[
\hat{x} = \arg\min_x \left\{ \sum_{k=1}^{K} \frac{1}{\lambda_k} ||y_k - W_k x||_2^2 + ||\nabla x||_1 \right\}.
\]

Ng et al. formulated the multi-frame SR reconstruction method by minimizing the TV-L2 objective function [21], which was recently applied to multi-frame depth SR [25]. This method is effective in suppressing ringing artifacts and preventing sharp edges from oversmoothing.

C. TV-L1 APPROACH FOR DECONVOLUTION AND DENOISING

The theoretical approach for image restoration begins with acquiring an HR image from a single input LR image. This approach mainly focuses on recovering high-frequency information by inverse filtering of a point spread function (PSF) of the optical system and is known as deconvolution. In deconvolution problems, the blurring matrix \( B \) is defined by a certain PSF and the noise term \( n_k \) is modeled as Gaussian noise. The majority of deconvolution methods use a certain type of regularization to overcome the issue of noise amplification in the final solution, and consider the problem of various noise types by applying various norms to the data fidelity term. The \( L_1 \)-norm can also be adopted in data fidelity to increase the noise immunity of the optimization process.

We refer to the objective function with the \( L_1 \)-norm based data fidelity term and TV regularization as the TV-L1 model. TV-L1 model has been employed to several applications. Yang et al. proposed the TV-L1 model for deblurring images corrupted by random-valued impulse noise [31]. Micchelli et al. presented proximity algorithms for TV-L1 image denoising [32]. This approach can be modeled as follows:

\[
\hat{x} = \arg\min_x \left\{ \frac{1}{2} ||y - Bx||_1 + ||\nabla x||_1 \right\}.
\]

The \( L_2 \) data fidelity term assumes Gaussian noise, which causes side effects in other noise types. In real imaging systems, LR observations are often degraded by various sensor noise types. Therefore, it is not appropriate to assume that the noise follows a Gaussian distribution. However, in the approach of Eq. (8), the data fidelity term with the \( L_1 \)-norm is more effective and robust than that of the \( L_2 \)-norm when the images contain non-Gaussian noise. Furthermore, the optimization process based on Eq. (8) conserves strong edges and generates few ringing artifacts. To preserve sharp depth discontinuities and improve robustness to noise, we design an objective function for multi-frame depth SR based on the TV-L1 model and propose an efficient solver based on half-quadratic splitting.

III. PROPOSED METHOD

In this section, we first analyze the statistical characteristics of ToF depth images, and subsequently design the objective function to restrict the solution space by means of prior knowledge through the regularization term and statistical information of the noise through the data fidelity term. The problem of depth image SR being corrupted by various sources is considered to improve the robustness to noise. Thereafter, we propose an efficient solver based on half-quadratic splitting and describe the proposed method in detail.

A. TV-L1 FUNCTION FOR MULTI-FRAME DEPTH SR

Depth SR is an inherently ill-posed problem; that is, many solutions exist for any given LR depth images. To overcome this under-determined problem, the development of effective image regularization and data fidelity has been the topic of substantial research over the years. Under the Bayesian framework, the maximum a posteriori (MAP) estimation of the HR image \( x \) with the observed LR images \( y_1, y_2, \ldots, y_K \) is given by

\[
\hat{x} = \arg\max_x P(x|y_1, y_2, \ldots, y_K).
\]

Using Bayes’ rule, Eq. (9) can alternatively be expressed as

\[
\hat{x} = \arg\max_x \frac{P(y_1, y_2, \ldots, y_K|x)P(x)}{P(y_1, y_2, \ldots, y_K)}.
\]
where \( P(x) \) describes the prior probability of the HR image, \( P(y_1, y_2, \ldots, y_K) \) represents the prior probability of the LR images, and \( P(y_1, y_2, \ldots, y_K | x) \) represents the conditional probability of the LR images given the HR image. As \( P(y_1, y_2, \ldots, y_K) \) has no influence on the maximization probability function, the above MAP estimation can be rewritten as

\[
\hat{x} = \arg \max_x P(y_1, y_2, \ldots, y_K | x) P(x). \tag{11}
\]

It is convenient and equivalent to minimize the minus log of Eq. (11), which yields

\[
\hat{x} = \arg \min_x \left\{ -\log P(y_1, y_2, \ldots, y_K | x) - \log P(x) \right\}. \tag{12}
\]

In general, the observation of each LR image is independent and identically distributed (i.i.d.). Therefore, Eq. (12) can be simplified as

\[
\hat{x} = \arg \min_x \left\{ -\log \prod_{k=1}^{K} P(y_k | x) - \log P(x) \right\}, \tag{13}
\]

where \( P(y_k | x) \) represents the conditional probability of the LR image given the HR image. Therefore, the prior probability and conditional probability concerning prior knowledge based on the statistical information of the HR image and noise must be specified.

Firstly, let us consider the prior probability involved in Eq. (13). In general, an image is assumed to be globally smooth, which is incorporated into the reconstruction process through a Gaussian prior [18]. The prior probability is expressed as follows:

\[
P(x) = \frac{1}{Z_g} \exp \left( -\frac{1}{2} ||\Gamma x||_2^2 \right). \tag{14}
\]

where \( Z_g \) is a simplified normalizing constant. This approach is referred to as Tikhonov regularization, which is the representative smoothing constraint. The approach in Eq. (14) smooths the restored image by penalizing the high-frequency component and has been applied extensively owing to its low computational costs. However, strategies that use this regularization inevitably smooth the sharp edges and detailed information. Recent research in natural image statistics has demonstrated that real-world scenes exhibit significantly heavy-tailed or hyper-Laplacian distributions in their gradients [33]. Image processing based on such priors has yielded state-of-the-art results in image restoration [34].

As explained previously, an edge-preserving HR depth image can be obtained if the prior probability of the desired depth images is modeled. Fig. 2a presents the logarithmic first-order derivative using 20 depth images from a ToF sensor in the real-world. The empirical distribution of the gradient magnitude exhibits a higher correlation with the Laplace distribution than the Gaussian distribution. As a result, the prior of the depth images takes the form of the logarithm of a Laplace prior, and is represented as follows:

\[
P(x) = \frac{1}{Z_l} \exp \left( -\frac{1}{2} ||Gx||_1 \right). \tag{15}
\]

where \( Z_l \) is a simplified normalizing constant and \( G \) denotes the derivative matrix corresponding to the first derivative linear operator. This model is associated with TV regularization, which assumes that the gradient magnitude of the depth image is sparse.

Based on the MAP theory, the problem of multiple depth SR with the TV-\( L_2 \) model can be transformed into a minimization problem, as follows:

\[
\hat{x} = \arg \min_x \left\{ \sum_{k=1}^{K} \frac{1}{\lambda_k} ||y_k - W_k x||_2^2 + ||Gx||_1 \right\}. \tag{16}
\]

where \( ||Gx||_1 \) denotes the TV regularization.

Secondly, the distribution of \( P(y_k | x) \) is derived from the observation model. In general, the noise in the observation model is assumed to follow a Gaussian distribution. The conditional probability is represented as follows:

\[
P(y_k | x) = \frac{1}{\sigma^2_\theta \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2_\theta} ||y_k - W_k x||_2^2 \right), \tag{17}
\]

where \( \sigma_\theta \) denotes the standard deviation of the Gaussian distribution. However, the result obtained by the strategy based on the data fidelity term with the \( L_2 \)-norm is only optimal when the depth images contain Gaussian noise.

In practical applications, the noises that exist in the LR depth images are usually mixed. Because an image can be degraded by noise from various sources, these are typically non-Gaussian distributed. In ToF sensors, a noise source is correlated with the light source (shot noise) and an additional additive noise source (dark current noise) [35]. Furthermore, ToF sensors contain defective pixels that cause bright or dark spots in the depth images. Fig. 2b presents the logarithmic probability density of various noise types. The mixed noise is similar to the Laplace distribution, rather than the Gaussian distribution, in that it exhibits a heavy-tailed distribution. Therefore, the conditional probability takes the form of the logarithm of a Laplace distribution.
Laplace distribution, and it is represented as follows:

\[
P(y_k|x) = \frac{1}{\sqrt{2\sigma_l}} \exp\left(-\frac{\sqrt{2}}{\sigma_l}||y_k - W_kx||_1\right).
\]  

(18)

where \(\sigma_l\) denotes the standard deviation of the Laplace distribution. This approach is associated with the \(L_1\)-norm data fidelity term, which is robust to noise regardless of the noise type. As the median geometry is defined as a point in Euclidean space that minimizes the sum of the distances, it is related to the \(L_1\)-norm [36]. Thus, the \(L_1\)-norm data fidelity term exhibits the characteristics of a median operator. Consequently, the data fidelity term with the \(L_1\)-norm is more effective and robust than the \(L_2\)-norm when the images contain non-Gaussian noise such as random-valued Laplace noise and Salt-and-Pepper noise.

According to the MAP theory, the problem of multiple depth SR with the TV-\(L_1\) model can be transformed into a minimization problem, as follows:

\[
\hat{x} = \arg\min_x \left\{ \sum_{k=1}^{K} \frac{1}{\lambda_k} ||y_k - W_kx||_1 + ||Gx||_1 \right\}.
\]  

(19)

In Eq. (19), the \(L_1\)-norm is used for both the data fidelity and regularization terms. The objective function based on the TV-\(L_1\) model is known as the TV-\(L_1\) function.

Fig. 3 compares the SR results obtained by the Tikhonov-L\(_2\), TV-L\(_2\), and TV-L\(_1\) models. The observed LR depth image is degraded by downsampling with a factor of 4 and adding mixed noises containing Gaussian noise with \(\sigma_g = 5\), random-valued Laplace noise with \(\sigma_l = 5\), and Salt-and-Pepper noise with probability \(p = 0.05\). In the SR result restored by the Tikhonov-L\(_2\) model, severe artifacts such as smooth edges and overshooting on the boundaries are unavoidable owing to the smoothing constraints. Moreover, noise amplification appears during the restoration process. In the SR result generated by the TV-L\(_2\) model, the noise amplification problem is suppressed to a certain extent by adjusting the regularization parameters. However, the sharp depth discontinuities are not preserved. However, the TV-L\(_1\) model outperforms the others in terms of sharpness, without any obvious artifacts, and suppresses the noise amplification.

Fig. 3b presents the one-dimensional (1D) sectional view of the dashed line in Fig. 3a. We further investigate the performance of the TV-L\(_1\) model in suppressing the noise in Fig. 3c. The histogram of the green square box in Fig. 3a for the TV-L\(_1\) model exhibits the most significant spike compared to the other models. More specifically, the TV-L\(_1\) model produces a lower standard deviation, which indicates the level of noise in a flat area. It is important to measure the accurate distance through the depth image. As the proposed method exhibits the characteristics of a median operator, the mean value is preserved in a depth image that is corrupted by mixed noise, as indicated in Fig. 3c. Therefore, the proposed method based on the TV-L\(_1\) model overcomes the limitations of other models in terms of preserving the sharp depth discontinuities and suppressing the noise amplification.

![Fig. 3. Comparison of Tikhonov-L\(_2\), TV-L\(_2\), and TV-L\(_1\) models at \(x\times4\) upsampling rate for test image Art corrupted by mixed noises (Gaussian noise with \(\sigma_g = 5\), random-valued Laplace noise with \(\sigma_l = 5\), and Salt-and-Pepper noise with \(p = 0.05\)): (a) upsampled depth images, enlarged parts, and difference maps; (b) 1D sectional view of dashed line in upsampled depth images; (c) histogram of green square box in upsampled depth images. The proposed model produces sharp depth discontinuities without overshooting artifacts and suppresses the noise amplification compared to the other models.](image1)

![Fig. 4. Objective performance comparison of Tikhonov-L\(_2\), TV-L\(_2\), and TV-L\(_1\) models at \(x\times4\) upsampling rate for test image Art according to varying noise levels (Gaussian noise with \(\sigma_g = 0\) to 10, random-valued Laplace noise with \(\sigma_l = 0\) to 10, and Salt-and-Pepper noise with \(p = 0\) to 0.1): (a) PSNR values and (b) SSIM values. The proposed method based on the TV-L\(_1\) model achieves the highest PSNR and SSIM performance in all cases.](image2)
with probability $p = 0$ to $0.1$). Moreover, as the noise level increases, the performance difference between the models increases significantly.

### B. OPTIMIZATION

In this section, we explain how to solve the optimization problem of Eq. (19), which is computationally difficult owing to the non-differentiability and non-linearity of the TV-L1 function. To recover the HR depth image $x$ efficiently, we propose an efficient technique based on half-quadratic splitting for TV-L1 minimization [31]. In this strategy, the auxiliary variables $v_k$ and $u$ are introduced to transfer $y_k - W_k x$ and $Gx$ out of the non-differentiable term. The objective function can be rewritten as follows:

$$
\hat{x} = \arg \min_{x,v_k,u} \left\{ \sum_{k=1}^{K} \frac{1}{2} \beta_k \|y_k - W_k x - v_k\|^2_2 + \sum_{k=1}^{K} \frac{1}{\lambda_k} \|v_k\|_1 + \frac{1}{2\theta} \|Gx - u\|^2_2 + \|u\|_1 \right\},
$$

(20)

where $v_k$ and $u$ are the auxiliary variables, and $\beta_k$ and $\theta$ are the regularization parameters. It is well known that the solution of Eq. (20) converges to that of Eq. (19) as $\beta_k \to 0$ and $\theta \to 0$. The objective function of Eq. (20) is strictly convex and differentiable. Furthermore, if two of the three variables $v_k$, $u$, and $x$ are fixed, the minimization of the objective function with respect to the others has a closed-form solution.

As $v_k$ and $u$ are decoupled for a given $x$, the objective function is numerically easy to minimize iteratively. For a fixed $x$, the objective function of Eq. (20) is separable with respect to $v_k$ and $u$. The objective function can be efficiently minimized although alternatively minimizing the following two sub-problems. According to the alternative minimization method [30], the problem can be transformed into two separate objective functions, as follows:

$$
\hat{u} = \arg \min_u \left\{ \frac{1}{2} \|u - Gx\|^2_2 + \theta \|u\|_1 \right\},
$$

(21)

$$
\hat{v}_k = \arg \min_{v_k} \left\{ \frac{1}{2} \|v_k - (y_k - W_k x)\|^2_2 + \frac{\beta_k}{\lambda_k} \|v_k\|_1 \right\}.
$$

(22)

The objective functions in Eq. (21) and Eq. (22) are each a single-variable optimization problem. Firstly, the solution of Eq. (21) is expressed by the shrinkage formula, as follows:

$$
\hat{u} = \frac{Gx}{\|Gx\|_1} \max(\|Gx\|_1 - \theta, 0),
$$

(23)

where the convention $0 \cdot (0/0) = 0$ is followed. Each component of $u$ becomes

$$
\hat{u}_x = \text{sgn}(\partial_x x) \max(\|\partial_x x\|_1 - \theta, 0),
$$

$$
\hat{u}_y = \text{sgn}(\partial_y x) \max(\|\partial_y x\|_1 - \theta, 0),
$$

(24)

where $\text{sgn}(\cdot)$ is the signum function and all operations are performed component-wise. Secondly, the minimization of the objective function in Eq. (22) is given by the 1D shrinkage, as follows:

$$
\hat{v}_k = \frac{y_k - W_k x}{\|y_k - W_k x\|_1} \max(\|y_k - W_k x\|_1 - \frac{\beta_k}{\lambda_k}, 0)
$$

$$
= \text{sgn}(y_k - W_k x) \max(\|y_k - W_k x\|_1 - \frac{\beta_k}{\lambda_k}, 0),
$$

(25)

where all operations involve component-wise processing.
For fixed $v_k$ and $u$ variables, the minimization with respect to $x$ in each iteration is a least-squares problem, as follows:

$$\hat{x} = \arg\min_x \left\{ \sum_{k=1}^{K} \|y_k - W_k x - v_k\|_2^2 + \sum_{k=1}^{K} \frac{\beta_k}{\theta} \|Gx - u\|_2^2 \right\},$$

(26)

and the objective function is convex and differentiable with the use of quadratic data fidelity and regularization terms. Because the gradient of the objective function with respect to $x$ is equal to 0, Eq. (26) for the optimal solution becomes

$$\sum_{k=1}^{K} \left[ W_k^T W_k + \frac{\beta_k}{\theta} G^T G \right] x = \sum_{k=1}^{K} W_k^T (y_k - v_k) + \sum_{k=1}^{K} \frac{\beta_k}{\theta} G^T u,$$

(27)

and the unique solution can be obtained by deterministic techniques.

The proposed multi-frame depth SR process with the TV-L1 function is summarized in Algorithm 1. A continuation scheme on the regularization parameters $\beta_k$ and $\theta$ is employed to accelerate the overall convergence. Small values of $\beta_k$ and $\theta$ are initially set and gradually incremented in iterations. The solution of Eq. (20) converges to that of Eq. (19) as the values of $\beta_k$ and $\theta$ move towards 0. In terms of convergence, the cost function in Eq. (19) is convex with respect to all variables.

**IV. EXPERIMENTAL RESULTS**

We verified the performance of the proposed method both quantitatively and qualitatively through various experiments. The experiments were conducted using three degradation types: downsampling degradation (noise-free observation), ToF-like degradation (sub-sampling with mixed noises), and real-world ToF sensor data. In this section, we first describe the datasets used in the experiments. Thereafter, the conventional methods that are compared with the proposed method are briefly outlined. Subsequently, the comparative results on the datasets in terms of the performance metrics and visual quality are presented. Finally, the influence of the parameters
and convergence analysis are discussed. In the implementation of our proposed method, the following were used in each LR image: $\lambda_k = 0.005$ for the downsampling degradation, $\lambda_k = 0.1$ for the ToF-like degradation, and $\lambda_k = 0.01$ for the real-world ToF data. The same regularization parameters $\lambda_k$ were applied to each LR image. The criterion $||x_{n+1} - x_n||^2/||x_n||^2 < 10^{-6}$ was used to terminate the iteration. The PSNR, SSIM [38], and percentage of bad matched pixels (BMP) [39] were used as objective performance metrics. The PSNR was employed to evaluate the gray value similarity, whereas the SSIM was mainly used to evaluate the structural similarity. To evaluate the depth accuracy, the percentage of BMP between the upsampled depth image $\hat{x}(i,j)$ and ground truth depth image $x(i,j)$ was measured, which is defined as

$$\text{BMP} = \frac{1}{N} \sum_{(i,j)} (|\hat{x}(i,j) - x(i,j)| > \delta_d),$$

(28)

where $\delta_d$ is the depth error tolerance and $N$ is the total number of pixels. We used $\delta_d = 1.0$ for the downsampling degradation and $\delta_d = 255 \times 0.05$ for the ToF-like degradation. Moreover, the cross-correlation method [40] was applied to estimate the sub-pixel registration between the reference LR frame and other LR frames. The sub-pixel registration is represented by the warping matrix $M_k$ in Eq. (1).

A. DATASETS

We used a synthetic dataset and a real-world ToF dataset to demonstrate the performance of the proposed method, as illustrated in Fig. 5. The Middlebury dataset [37] and ToFMark dataset [13], which include HR color images and ground truth HR depth images, were used for evaluation. The former and latter comprise six test images, namely Art, Book, Dolls, Laundry, Moebius, and Reindeer, as well as three test images, namely Books, Devil, and Shark, respectively. As these datasets include ground truth depth images, quantitative analysis was possible. To synthesize the LR depth images $y_k$ for the downsampling degradation, we translated the HR image $x$ and sub-sampled it by a certain upsampling factor in each dimension using the same bicubic kernel. We generated 4 or 16 LR depth images from one HR depth image at $\times 2$ or $\times 4$ upsampling factors, respectively. The sub-pixel registration of every LR image was simulated as a global motion randomly changing from $-1$ to $1$ in the horizontal and vertical directions. We assumed that the blurring matrix was...
FIGURE 9. Visual comparison of \( \times 4 \) upsampled depth images, enlarged parts, difference maps between the ground truth and the corresponding result from downsampling degradation and ToF-like degradation (sub-sampling with mixed noises): (a) ground truth, (b) bicubic interpolation, (c) SRCNN [11], (d) FGI [15], (e) RCG [16], (f) RSR [20], (g) MTSR [25], (h) proposed method, and (i) to (p) are the same methods as (a) to (h). The first and second rows present the results of Books and Devil in the ToFMark dataset, respectively.

an identity matrix (\( B_k = I \)) to focus on the downsampling and noise in the SR problem in all experiments.

For the ToF-like degradation, a combination of Gaussian noise, random-valued Laplace noise, and Salt-and-Pepper noise was added to the LR depth images simultaneously as the mixed noises. The noisy observation for each pixel \((i, j)\) of mixed noises could be described as

\[
y_{i,j}^n = \begin{cases} 
  d_{\min}, & \text{with probability } p/2 \\
  d_{\max}, & \text{with probability } p/2 \\
  v_{i,j} + n_{i,j}, & \text{with probability } 1 - p, 
\end{cases}
\]

where \( n_{i,j} \) is a sample of i.i.d. zero-mean Gaussian distribution with standard deviation \( \sigma_g \), and \( v_{i,j} \) is a sample of i.i.d. zero-mean Laplace distribution with standard deviation \( \sigma_l \). Within the given dynamic range \([d_{\min}, d_{\max}]\), \( d_{\min} \) and \( d_{\max} \) denote the minimum and maximum values of an image pixel, respectively. An LR image was corrupted by Salt-and-Pepper noise when \( v_{i,j} \) was either \( d_{\min} \) or \( d_{\max} \) with an equal probability of \( p/2 \). For the mixed noises, a pixel was corrupted by Gaussian and Laplace noise with a probability of \( 1 - p \). Fig. 6 presents a noisy LR depth image corrupted with four levels of mixed noises. We considered four values of the Gaussian noise parameter, namely \( \sigma_g = 2.5, 5.0, 7.5, \text{ and } 10.0 \); four values of the Laplace noise parameter, namely \( \sigma_l = 2.5, 5.0, 7.5, \text{ and } 10.0 \); and four values of the Salt-and-Pepper noise parameter, namely \( p = 0.025, 0.05, 0.075, \text{ and } 0.01 \).

Moreover, we acquired multiple depth images from the real-world ToF sensor data, which included \( 172 \times 224 \) dense depth measurements, as indicated in the six images on the right side of Fig. 5. As multi-frame SR algorithms require LR depth images with sub-pixel registration, global translations are introduced into real-world ToF data by arbitrarily manipulating the ToF imaging system during acquisition. Moreover, it is possible to obtain LR depth images using multiple ToF camera arrays. Regardless of which method is used to obtain the LR depth images, the new information contained in each LR image can be exploited to obtain the HR depth image if the sub-pixel registrations of the LR depth images differ from one another and if aliasing is present.

B. COMPARED METHODS

We compared the proposed method with the following three method categories: 1) the single-frame depth image upsampling approach including general image upsampling methods: bicubic interpolation, self-guided residual interpolation (SG) [9], edge-guided upsampling (EG) [10], and
FIGURE 10. Visual comparison of \( \times 4 \) upsampled depth images, enlarged parts, gradient maps, and 1D sectional view of purple dashed ellipse from ToF sensor in real-world: (a) bicubic interpolation, (b) SG [9], (c) SRCNN [11], (d) RSR [20], (e) CSR [23], (f) MTSR [25], (g) proposed method, and (h) to (n) are the same methods as (a) to (g).

SRCNN [11]; 2) the intensity-guided depth upsampling approach: guided image filtering (GF) [12], fast global image smoothing (FGI) [15], robust color guided restoration (RCG) [16], and joint local structure and nonlocal low-rank regularization (LN) [17]; and 3) the multi-frame depth image SR approach: robust SR (RSR) [20] based on the BTV model, combination SR (CSR) using sparse and non-sparse priors [23], and multi-frame ToF SR (MTSR) [25] based on the TV model. The implementations from the publicly available codes provided by the authors and appropriate parameters were used in our experiments. However, the source code of LN [17] was executed only for the downsampling degradation on the Middlebury dataset. Experiments were conducted on the Tikhonov-\( L_2 \) and TV-\( L_2 \) models implemented with our scheme (denoted by “ours”), as proposed in Lidarboost [24] and MTSR [25]. We selected the regularization parameters that generated the highest PSNR values for all of the testing images.

C. EVALUATION ON DOWNSampling DEGRADATION

For the evaluation on the downsampling degradation, the LR images were generated by sub-sampling the synthetic datasets at upsampling rates of \( \times 2 \) and \( \times 4 \). In this case, the degradation matrix \( W_k \) represented the operation that first translated the HR depth image and then downsampled it to the desired resolution. The noise \( n_k \) was 0. The quantitative evaluation of the upsampled depth images in terms of the PSNR, SSIM, and BMP values is summarized in Table 1 and Table 3. It can be observed that the proposed method achieved the highest PSNR and SSIM values as well as the lowest BMP value among all method types, including the single-frame and intensity-guided approaches.

Fig. 7 and the first row of Fig. 9 presents the results of three approach types: single-frame depth upsampling (bicubic interpolation and SRCNN [11]), intensity-guided depth upsampling (FGI [15] and RCG [16]), and multi-frame depth SR (RSR [20], MTSR [25], and the proposed method). The bicubic interpolation produced jagging artifacts and oversmoothing. The SRCNN [11] method generated sharper edges than the bicubic interpolation. However, the single-frame depth upsampling approach exhibited low performance compared to those of the intensity-guided upsampling and multi-frame SR approaches. The intensity-guided upsampling methods yielded sharper depth edges than the single-frame upsampling methods, as the intensity guidance could aid in maintaining the depth edge. However, these methods suffered from depth bleeding or texture copy artifacts owing to the inconsistency between the color edges and depth discontinuities. In particular, FGI [15] resulted in more severe artifacts than RCG [16]. In contrast, no texture...
transfer artifacts existed in the results of the multi-frame SR methods. Note that RSR [20], MTSR [25], and the proposed method used multiple LR depth images. Although RSR [20] is based on the BTV-\(L_1\) model which uses gradient descent, a problem similar to the ringing artifact occurred during the restoration process. MTSR [25] based on the TV-\(L_2\) model exhibited relatively good results in a noise-free environment. However, as can be observed, the proposed method based on the TV-\(L_1\) model produced much sharper edges than the other methods, without any obvious artifacts. In particular, the proposed method successfully suppressed the ringing artifacts and outperformed the other methods in terms of the sharpness around the depth discontinuity region. The visual comparison demonstrates that the proposed method intuitively achieved low errors through the difference images.

### TABLE 1. Objective Performance Comparison of Downsampling Degradation on Middlebury Dataset in PSNR (dB), SSIM, and BMP (%).

| Method          | Art \(x2\) | Art \(x4\) | Book \(x2\) | Book \(x4\) | Dolls \(x2\) | Dolls \(x4\) | Laundry \(x2\) | Laundry \(x4\) | Mobius \(x2\) | Mobius \(x4\) | Render \(x2\) | Render \(x4\) | Average \(x2\) | Average \(x4\) |
|-----------------|------------|------------|------------|------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Bioptic         | 40.15      | 35.73      | 48.34      | 43.46      | 49.11       | 45.04       | 44.21       | 39.68       | 49.22       | 44.65       | 42.62       | 38.41       | 45.61       | 41.16       |
| SG [9]          | 39.37      | 35.32      | 46.88      | 40.95      | 47.51       | 42.25       | 43.35       | 37.21       | 47.28       | 41.30       | 41.55       | 35.38       | 44.32       | 38.34       |
| EG [10]         | 38.72      | 34.84      | 46.98      | 41.91      | 47.08       | 44.08       | 43.61       | 39.42       | 47.43       | 43.78       | 41.07       | 37.16       | 44.15       | 40.20       |
| SRGN [11]       | 42.67      | 39.83      | 50.47      | 46.98      | 50.02       | 47.05       | 47.46       | 44.05       | 50.35       | 47.15       | 45.53       | 42.61       | 47.75       | 44.61       |
| GF [12]         | 38.64      | 32.83      | 47.54      | 40.35      | 47.01       | 42.27       | 42.36       | 36.64       | 46.70       | 41.07       | 40.58       | 34.53       | 43.67       | 37.95       |
| FGI [15]        | 36.87      | 35.79      | 45.92      | 42.49      | 45.91       | 42.82       | 42.99       | 39.72       | 46.14       | 43.18       | 41.67       | 38.56       | 43.55       | 40.43       |
| RCG [16]        | 41.19      | 36.56      | 46.39      | 40.09      | 47.01       | 41.66       | 44.64       | 40.47       | 48.05       | 43.41       | 43.18       | 39.02       | 45.08       | 41.18       |
| LN [17]         | 41.35      | 38.19      | 49.06      | 45.65      | 49.22       | 46.79       | 45.58       | 42.61       | 49.68       | 46.40       | 43.76       | 40.79       | 46.44       | 43.40       |
| RSR (TVL-1) [20]| 38.48      | 38.48      | 44.51      | 45.41      | 45.16       | 45.17       | 41.90       | 41.90       | 44.93       | 44.93       | 40.48       | 40.48       | 42.58       | 42.58       |
| CSR (SRAS+TVL-2) [23] | 38.78     | 32.20      | 46.69      | 39.29      | 47.44       | 41.16       | 42.80       | 35.92       | 47.05       | 40.24       | 41.10       | 34.56       | 43.94       | 37.23       |
| MTSR (TVL-2) [25]| 38.78      | 32.19      | 46.69      | 39.29      | 47.43       | 41.16       | 42.79       | 35.91       | 47.05       | 40.24       | 41.10       | 34.55       | 43.93       | 37.22       |
| ours (Tik_L2)   | 45.86      | 39.72      | 54.64      | 47.77      | 54.66       | 48.79      | 50.43       | 43.86      | 52.02       | 48.67       | 48.44       | 42.40       | 51.54       | 45.20       |
| ours (Tik_L2)   | 50.27      | 43.12      | 59.46      | 49.64      | 56.34       | 49.70      | 55.58       | 45.77      | 57.32       | 50.53      | 53.84       | 44.17       | 55.39       | 46.81       |
| Proposed (TVL-1) | 58.04      | 50.42      | 67.98      | 51.78      | 64.71       | 50.83      | 64.37       | 47.75      | 66.13       | 51.62       | 60.98       | 45.66       | 63.70       | 48.44       |

### D. EVALUATION ON ToF-LIKE DEGRADATION

In this case, the degradation matrix \(W_k\) was the same as that of the downsampling degradation. However, the noise \(n_k\) represented mixed noises. To verify the robustness of the proposed method to noise, mixed noises including Gaussian noise with \(\sigma_g = 5\), random-valued Laplace noise with \(\sigma_l = 5\), and Salt-and-Pepper noise with \(p = 0.05\) were added to the LR depth images. The comparison results in Table 2 and Table 3 indicate that the proposed method achieved the highest PSNR performance in all cases. Moreover, the proposed method exhibited the highest SSIM performance when the SSIM was used as the performance metric. The reconstructed depth images had high structural similarity to the ground truth depth image. Furthermore, because the proposed method yielded the lowest BMP value among several comparison methods, the highest depth accuracy was achieved.
TABLE 2. Objective Performance Comparison of ToF-like Degradation (sub-sampling with mixed noises) on Middlebury Dataset in PSNR (dB), SSIM, and BMP (%).

| Method              | Art x2  | Book x2  | Dolls x2  | Laundry x2  | Meubies x2  | Reindeer x2  | Average x2  |
|---------------------|---------|----------|-----------|-------------|-------------|--------------|-------------|
| SRCNN [11]          | 18.04   | 18.48    | 18.04     | 18.04       | 18.04       | 18.04        | 18.04       |
| CSR [SAR+TV-L1] [23]| 18.04   | 18.48    | 18.04     | 18.04       | 18.04       | 18.04        | 18.04       |
| MTSR (TV-L1) [25]   | 18.04   | 18.48    | 18.04     | 18.04       | 18.04       | 18.04        | 18.04       |

A visual comparison of the upsampling evaluation is presented in Fig. 8 and the second row of Fig. 9. The majority of single-frame upsampling methods had limitations in noisy depth images produced by the ToF sensor. Although FGI [15] and RCG [16] could reduce the noise, several severe outliers such as bleeding artifacts were propagated. RSR [20] removed most of the noise but yielded an unnatural overall result. In MTSR [25], the boundaries of the edge were broken, and the level of the overall depth varied. In contrast, the proposed method based on TV-L1 model successfully suppressed the noise amplification while preserving the sharpness of the edges.

E. EVALUATION ON REAL-WORLD TOF SENSOR DATA

We also performed comparisons on real-world ToF sensor data. The depth images captured with the real-world ToF camera were degraded by various noise types, such as Gaussian, Poisson, and Speckle noise. A total of 10 LR depth images with different sub-pixel registrations were used in this experiment. We compared the single-frame upsampling methods (bicubic interpolation, SG [9], and SRCNN [11]) and multi-frame SR methods (RSR [20], CSR [23], MTSR [25], and the proposed method). As no ground truth HR depth images existed, only a subjective comparison of the upsampled depth images could be performed. The gradient maps, as opposed to difference maps, are presented in Fig. 10, which express the absolute magnitude of the gradient at each pixel. The noise in the flat areas could also produce local maxima in the gradient map, and the sharpness of the edges.

A visual comparison of the upsampling evaluation is presented in Fig. 8 and the second row of Fig. 9. The majority of single-frame upsampling methods had limitations in noisy depth images produced by the ToF sensor. Although FGI [15] and RCG [16] could reduce the noise, several severe outliers such as bleeding artifacts were propagated. RSR [20] removed most of the noise but yielded an unnatural overall result. In MTSR [25], the boundaries of the edge were broken, and the level of the overall depth varied. In contrast, the proposed method based on TV-L1 model successfully suppressed the noise amplification while preserving the sharpness of the edges.

A visual comparison of the upsampling evaluation is presented in Fig. 8 and the second row of Fig. 9. The majority of single-frame upsampling methods had limitations in noisy depth images produced by the ToF sensor. Although FGI [15] and RCG [16] could reduce the noise, several severe outliers such as bleeding artifacts were propagated. RSR [20] removed most of the noise but yielded an unnatural overall result. In MTSR [25], the boundaries of the edge were broken, and the level of the overall depth varied. In contrast, the proposed method based on TV-L1 model successfully suppressed the noise amplification while preserving the sharpness of the edges.

A visual comparison of the upsampling evaluation is presented in Fig. 8 and the second row of Fig. 9. The majority of single-frame upsampling methods had limitations in noisy depth images produced by the ToF sensor. Although FGI [15] and RCG [16] could reduce the noise, several severe outliers such as bleeding artifacts were propagated. RSR [20] removed most of the noise but yielded an unnatural overall result. In MTSR [25], the boundaries of the edge were broken, and the level of the overall depth varied. In contrast, the proposed method based on TV-L1 model successfully suppressed the noise amplification while preserving the sharpness of the edges.

A visual comparison of the upsampling evaluation is presented in Fig. 8 and the second row of Fig. 9. The majority of single-frame upsampling methods had limitations in noisy depth images produced by the ToF sensor. Although FGI [15] and RCG [16] could reduce the noise, several severe outliers such as bleeding artifacts were propagated. RSR [20] removed most of the noise but yielded an unnatural overall result. In MTSR [25], the boundaries of the edge were broken, and the level of the overall depth varied. In contrast, the proposed method based on TV-L1 model successfully suppressed the noise amplification while preserving the sharpness of the edges.
could be observed with the gradient map. Moreover, the 1D sectional view indicates overshooting artifacts and sharp edge preservation.

Several examples of this comparison with an upsampling factor of ×4 are presented in Fig. 10. The results of the bicubic interpolation contained considerable outliers such as blurred edges and the SG [9] suffered from similar artifacts. However, the noise reduction effect could be expected because this method was based on the guided filter [12]. SRCNN [11] generated much sharper edges, but artifacts appeared on the flat areas because the noise was not considered during the training procedure. In general, the methods based on a single-frame depth image lacked sharpness in the high-frequency region. The multi-frame SR methods produced sharper edges than the single-frame upsampling methods. RSR [20] produced periodic artifacts and yielded unnatural results overall. CSR [23] blurred the edges owing to the use of a combination of sparse and non-sparse priors. Although MTSR [25] suppressed the noise amplification, zipper artifacts could be observed at the edges. In contrast, the proposed method based on the TV-L1 model exhibited superior performance without any obvious artifacts.

| TABLE 3. Objective Performance Comparison of Downsampling Degradation and ToF-like Degradation on ToFMark Dataset in PSNR (dB), SSIM, and BMP (%) |

| Method            | Downsampling Degradation | ToF-like Degradation |
|-------------------|--------------------------|----------------------|
|                   | Books ×2 | Devil ×2 | Shark ×2 | Books ×2 | Devil ×2 | Shark ×2 |
| Bicubic           | 31.15    | 27.18   | 32.59    | 28.49    | 30.29    | 26.30    | 19.14    | 18.44   | 19.06    | 18.58    | 19.18    | 18.71    |
| SG [9]            | 30.83    | 25.20   | 31.93    | 26.41    | 30.00    | 24.64    | 21.02    | 21.11   | 20.96    | 21.29    | 20.81    | 21.11    |
| EG [10]           | 29.63    | 26.29   | 31.46    | 27.77    | 28.90    | 25.63    | 18.43    | 18.34   | 18.38    | 18.02    | 19.63    | 19.16    |
| SRCNN [11]        | 30.81    | 25.31   | 32.41    | 26.42    | 30.51    | 24.74    | 16.49    | 16.02   | 16.42    | 16.21    | 16.44    | 16.19    |
| GF [12]           | 32.07    | 27.72   | 33.03    | 28.97    | 30.11    | 26.09    | 27.61    | 25.77   | 25.86    | 26.57    | 26.28    | 24.37    |
| FGI [15]          | 27.55    | 24.50   | 29.40    | 26.40    | 25.69    | 22.37    | 26.25    | 23.61   | 24.50    | 25.03    | 24.34    | 21.96    |
| RCG [16]          | 31.50    | 27.41   | 32.88    | 26.79    | 30.56    | 25.62    | 19.59    | 18.80   | 19.47    | 18.94    | 19.63    | 19.09    |
| LN [17]           | -        | -       | -        | -        | -        | -        | -        | -      | -        | -        | -        | -        |
| RSR (BTv-L1) [20] | 29.46    | 29.46   | 30.64    | 30.64    | 28.78    | 28.78    | 28.99    | 29.03   | 30.12    | 30.11    | 28.21    | 28.21    |
| CSR ([SAR+TV]-L2) [23] | 29.45  | 23.40   | 30.65    | 24.56    | 28.77    | 22.60    | 25.77    | 24.67   | 26.61    | 25.56    | 24.87    | 23.93    |
| MTSR (TV-L2) [25] | 29.45    | 23.40   | 30.65    | 24.56    | 28.77    | 22.60    | 25.33    | 24.87   | 26.61    | 25.06    | 26.25    | 24.26    |
| ours (TvL-L2)     | 36.71    | 30.60   | 38.14    | 32.18    | 35.94    | 29.83    | 25.54    | 25.29   | 25.63    | 25.79    | 25.47    | 25.42    |
| ours (TV-L2)      | 43.45    | 32.04   | 44.75    | 33.66    | 42.83    | 31.45    | 29.72    | 29.07   | 30.02    | 29.79    | 29.42    | 28.32    |
| Proposed (TV-L1)  | 55.57    | 33.13   | 55.89    | 34.83    | 53.23    | 32.96    | 35.56    | 31.85   | 36.79    | 33.18    | 35.11    | 30.92    |

| Method            | SSIM     | BNM (%)  |
|-------------------|----------|----------|
| Bicubic           | 0.9747   | 0.9404   |
| SG [9]            | 0.9740   | 0.9404   |
| EG [10]           | 0.9700   | 0.9430   |
| SRCNN [11]        | 0.9665   | 0.9175   |
| GF [12]           | 0.9708   | 0.9443   |
| FGI [15]          | 0.9383   | 0.9086   |
| RCG [16]          | 0.9736   | 0.9442   |
| LN [17]           | -        | -        |
| RSR (BTv-L1) [20] | 0.9628   | 0.9628   |
| CSR ([SAR+TV]-L2) [23] | 0.9747 | 0.9187   |
| MTSR (TV-L2) [25] | 0.9747   | 0.9187   |
| ours (TvL-L2)     | 0.9802   | 0.9459   |
| ours (TV-L2)      | 0.9967   | 0.9767   |
| Proposed (TV-L1)  | 0.9991   | 0.9811   |

| Method            | DMP (%)  |
|-------------------|----------|
| Bicubic           | 7.29     |
| SG [9]            | 5.40     |
| EG [10]           | 7.90     |
| SRCNN [11]        | 10.25    |
| GF [12]           | 2.49     |
| FGI [15]          | 14.38    |
| RCG [16]          | 9.67     |
| LN [17]           | -        |
| RSR (BTv-L1) [20] | 11.33    |
| CSR ([SAR+TV]-L2) [23] | 4.25  |
| MTSR (TV-L2) [25] | 4.24     |
| ours (TvL-L2)     | 12.89    |
| ours (TV-L2)      | 2.51     |
| Proposed (TV-L1)  | 0.68     |
by reconstructing detailed information and generating sharp depth discontinuities. Therefore, the proposed method provided more reliable results than previous methods using both a similar objective function and other non-similar ones.

F. INFLUENCE OF PARAMETERS

The proposed method requires regularization parameters $\lambda$, which control the trade-off between the data fidelity and regularization terms. Small values of $\lambda$ tend to yield sharper results, but the noise will be amplified. Large values of $\lambda$ tend to yield less noisy results, but the restored image may be smoothed. Fig. 11 presents an example of the influence of the regularization parameters in the case of mixed noises. Because the choice of $\lambda$ is not known prior to solving the minimization, $\lambda$ should be selected empirically. An appropriate selection of $\lambda$ according to the noise level can significantly improve the objective performance.

Fig. 12 presents a comparison of the objective performance with an upsampling factor of $\times 4$ according to $\lambda$ with various noise levels ($\sigma_g = 0$ to 10, $\sigma_l = 0$ to 10, and $p = 0$ to 0.1). Through this experiment, appropriate values of $\lambda$ were established according to the noise level. Moreover, the available information contained in each LR image was dependent on the image quality, such as the noise level and high-frequency information. Therefore, if different regularization parameters $\lambda$ are considered in each LR image, improved results can be expected compared to when using the same regularization parameter.

G. CONVERGENCE ANALYSIS

Finally, to demonstrate the convergence behavior of the multi-frame depth SR with TV-$L_1$ model, the cost function in Eq. (19) versus the iteration number is plotted in Fig. 13 for various conditions. In this test, the initial estimate $x_0 = 0$, upsampling rate of $\times 4$, and test image Art without noise were set as the default values and only one variable was changed to conduct the experiment. For the initial estimates $x_0$, we experimented with a blank image ($x_0 = 0$) or an interpolated version of the first LR image ($x_0 = D^2 y_1$). Furthermore, experiments were conducted with various noise levels ($\sigma_g = 0$ to 10, $\sigma_l = 0$ to 10, and $p = 0$ to 0.1). As illustrated in Fig. 13, the objective function of the proposed model converged for various variables and the graphs indicate that the cost function rapidly decreased in several iterations.

V. CONCLUSION

The multi-frame depth SR method based on the TV-regularized $L_1$ function has been presented. The statistical characteristics of ToF depth images were analyzed and an objective function based on the TV-$L_1$ model for multi-frame depth SR was designed. The multi-frame depth SR based on the TV-$L_1$ model is robust to noise and produces sharp depth discontinuities without obvious artifacts compared to the Tikhonov-$L_2$ and TV-$L_1$ models. Furthermore, an efficient solver based on half-quadratic splitting has been presented. The cost function decreases within only a few iterations for various variables. The proposed method was compared with existing state-of-the-art methods using synthetic and real-world ToF data. In both the quantitative and qualitative evaluations, the proposed method outperformed existing state-of-the-art methods on the synthetic and real-world ToF data.

FIGURE 11. Influence of regularization parameter at $\times 4$ upsampling for test image Art corrupted by mixed noises (Gaussian noise with $\sigma_g = 5$, random-valued Laplace noise with $\sigma_l = 5$, and Salt-and-Pepper noise with $p = 0.05$): (a) $\lambda = 0.05$, (b) $\lambda = 0.1$, (c) $\lambda = 0.3$, and (d) $\lambda = 1$.

FIGURE 12. Objective performance comparison at $\times 4$ upsampling rate for test image Art according to regularization parameter $\lambda$ with varying mixed noises: (a) PSNR values and (b) SSIM values.

FIGURE 13. Convergence of cost functions at $\times 4$ upsampling rate for test image Art: (a) value of cost function versus iteration number of various initial conditions and (b) value of cost function versus iteration number of various noise levels.
REFERENCES

[1] Q. Zhu, J. Mai, and L. Shao, “A fast single image haze removal algorithm using color attenuation prior,” IEEE Trans. Image Process., vol. 24, no. 11, pp. 3522–3533, Nov. 2015.

[2] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, “Scanning 3D full human bodies using kinects,” IEEE Trans. Vis. Comput. Graphics, vol. 18, no. 4, pp. 643–650, Apr. 2012.

[3] J. Engel, T. Schonberger, and D. Cremers, “LSD-SLAM: Large-scale direct monocular slam,” in Computer Vision—ECCV, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 834–849.

[4] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3D object detection network for autonomous driving,” in Tech. Rep., IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1907–1915.

[5] R. Min, N. Kose, and J.-L. Dugelay, “KinectFaceDB: A Kinect database for face recognition,” IEEE Trans. Syst., Man., Cybern., Syst., vol. 44, no. 11, pp. 1534–1548, Nov. 2014.

[6] D. Scharstein and R. Szeliski, “High-accuracy stereo depth maps using structured light,” in Proc. IEEE Comput. Soc. Conf. Vis. Pattern Recognit., Jun. 2003, p. 1.

[7] R. Lange and P. Seitz, “Solid-state time-of-flight range camera,” IEEE J. Quantum Electron., vol. 37, no. 3, pp. 390–397, Mar. 2001.

[8] D. Stoppa, N. Massari, L. Pancheri, M. Malfatti, M. Peronzeni, and L. Gonzo, “A range image sensor based on 10-µm lock-in pixels in 0.18-µm CMOS imaging technology,” IEEE J. Solid-State Circuits, vol. 46, no. 1, pp. 248–258, Jan. 2011.

[9] Y. Li, D. Min, M. N. Do, and J. Lu, “Fast guided global interpolation,” in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), Dec. 2016, pp. 1394–1399.

[10] J. Xie, R. S. Feris, and M.-T. Sun, “Edge-guided single depth image super resolution,” IEEE Trans. Image Process., vol. 25, no. 1, pp. 428–438, Jan. 2016.

[11] C. Dong, C. L. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, Feb. 2016.

[12] K. He, J. Sun, and X. Tang, “Guided image filtering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[13] D. Fersstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, “Depth map upsampling by self-guided residual interpolation,” in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), Dec. 2016, pp. 1394–1399.

[14] B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, Computer Vision—ECCV, vol. 19, no. 2, pp. 293–301, Feb. 2017.

[15] J. Engel, T. Schoeps, and D. Cremers, “LSD-SLAM: Large-scale direct monocular slam,” in Computer Vision—ECCV, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 834–849.

[16] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3D object detection network for autonomous driving,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1907–1915.

[17] A. Y. Ng, “Feature selection, l1 vs. l2 regularization, and rotational invariance,” in Proc. 21st Int. Conf. Mach. Learn. ICML, 2004, p. 78, doi: 10.1145/1015330.1015435.

[18] S. C. Park, M. K. Park, and M. G. Kang, “Non-lipschitz ℓp-regularization and box constrained model for image restoration,” IEEE Trans. Image Process., vol. 21, no. 12, pp. 4709–4721, Dec. 2012.

[19] J. Kim, “Total variation based image restoration with free local constraints,” in Proc. IEEE Int. Conf. Image Process. (ICIP), 1st Vol. 1, Nov. 1994, pp. 31–35.

[20] L. I. Rudin and S. Osher, “Total variation based image restoration with free local constraints,” in Proc. IEEE Int. Conf. Image Process. (ICIP), 1st Vol. 1, Nov. 1994, pp. 31–35.

[21] J. Y. Yang, J. Yang, W. Yin, and Y. Zhang, “A new alternating minimization algorithm for total variation image reconstruction,” SIAM J. Imag. Sci., vol. 1, no. 3, pp. 248–272, Jan. 2008, doi: 10.1137/080724265.

[22] J. Y. Yang, Y. Zhang, and W. Yin, “An efficient TVL1 algorithm for deblurring multichannel images corrupted by impulse noise,” SIAM J. Sci. Comput., vol. 31, no. 4, pp. 2842–2865, Jan. 2009, doi: 10.1137/080732894.

[23] C. A. Michelli, L. Shen, Y. Xu, and X. Zeng, “Proximity algorithms for the L1-TV image denoising model,” Adv. Comput. Math., vol. 38, no. 2, pp. 401–426, Feb. 2013, doi: 10.1007/s10444-011-9234-y.

[24] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, “LidarBoost: Depth superresolution for ToF 3D shape scanning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 343–350.

[25] S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Multi-frame super-resolution for Time-of-flight imaging,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2011.

[26] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, “Fast and robust image deconvolution,” IEEE Trans. Image Process., vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[27] M. K. Ng, H. Shen, E. Y. Lam, and L. Zhang, “A total variation regularization based super-resolution reconstruction algorithm for digital video,” EURASIP J. Adv. Signal Process., vol. 2007, no. 1, pp. 1–16, Dec. 2007.

[28] B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, Computer Vision—ECCV, vol. 19, no. 2, pp. 293–301, Feb. 2017.

[29] S. C. Park, M. K. Park, and M. G. Kang, “Super-resolution image reconstruction: A technical overview,” IEEE Signal Process. Mag., vol. 20, no. 3, pp. 21–36, May 2003.

[30] R. Hardie, K. J. Barnard, and E. E. Armstrong, “Joint MAP registration and high-resolution image estimation using a sequence of undersampled images,” IEEE Trans. Image Process., vol. 6, no. 12, pp. 1621–1633, Dec. 1997.

[31] S. Farsiu, D. M. Robinson, M. Elad, and P. Milanfar, “Fast and robust multiframe super resolution,” IEEE Trans. Image Process., vol. 13, no. 10, pp. 3522–3533, Nov. 2004.

[32] S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Variational Bayesian super resolution,” IEEE Trans. Image Process., vol. 20, no. 4, pp. 984–999, Apr. 2011.

[33] S. Villena, M. Vega, S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Bayesian combination of sparse and non-sparse priors in image super resolution,” Digit. Signal Process., vol. 23, no. 2, pp. 530–541, Mar. 2013, [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051500412002400

JONGHYUN KIM received the B.S. degree in electrical and electronic engineering from Yonsei University, South Korea, in 2012, where he is currently pursuing the joint M.S. and Ph.D. degrees with the Department of Electrical and Electronic Engineering.

His current research interests include super-resolution image processing, image restoration, and color interpolation.
JAEDUK HAN received the B.S. degree in electrical and electronic engineering from Yonsei University, South Korea, in 2014, where he is currently pursuing the joint M.S. and Ph.D. degrees with the Department of Electrical and Electronic Engineering.

His current research interests include image restoration and inverse problems in image and video processing.

MOON GI KANG received the B.S. and M.S. degrees in electronics engineering from Seoul National University, Seoul, South Korea, in 1986 and 1988, respectively, and the Ph.D. degree in electrical engineering from Northwestern University, Evanston, IL, USA, in 1994.

He was an Assistant Professor with the University of Minnesota, Duluth, MN, USA, from 1994 to 1997. Since 1997, he has been with the Department of Electronic Engineering, Yonsei University, Seoul, where he is currently a Professor. His current research interests include image and video filtering, restoration, enhancement, and superresolution reconstruction. He has authored more than 100 technical articles in his areas of expertise. He was a recipient of 2006, 2007, and 2012 Yonsei Outstanding Research Achievement Awards (Technology Transfer), the 2002 HaeDong Foundation Best Paper Award, and the 2000, 2009, and 2010 Awards of Teaching Excellence at Yonsei University. He has served as the Editorial Board Member for IEEE Signal Processing Magazine, the Editor for SPIE Milestone Series Volume (CCD and CMOS imagers), and the Guest Editor for the IEEE Signal Processing Magazine Special Issue on Superresolution Image Reconstruction (May 2003). He has served in the technical program and steering committees for several international conferences. He has also served as the Associate Editor for the EURASIP Journal of Advances in Signal Processing and the Digital Signal Processing (Elsevier) journal.