IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 2, 2020, accepted May 8, 2020, date of publication May 22, 2020, date of current version June 5, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2996689*

# A General Framework for Depth Compression and Multi-Sensor Fusion in Asymmetric View-Plus-Depth 3D Representation

**MIHAIL GEORGIEV[1], (Member, IEEE), EVGENY BELYAEV[2], AND ATANAS GOTCHEV[3], (Member, IEEE)**

[1]Panasonic Automotive Systems GmbH, 63225 Langen, Germany
[2]International Laboratory ‘‘Computer Technologies’’, ITMO University, 197101 Saint Petersburg, Russia
[3]Faculty of Information Technology and Communication Sciences, Tampere University, 33720 Tampere, Finland

Corresponding author: Mihail Georgiev (mihail.georgiev@ext.eu.panasonic.com)

**ABSTRACT** We present a general framework which can handle different processing stages of the three-dimensional (3D) scene representation referred to as ‘‘view-plus-depth’’ (V+Z). The main component of the framework is the relation between the depth map and the super-pixel segmentation of the color image. We propose a hierarchical super-pixel segmentation which keeps the same boundaries between hierarchical segmentation layers. Such segmentation allows for a corresponding depth segmentation, decimation and reconstruction with varying quality and is instrumental in tasks such as depth compression and 3D data fusion. For the latter we utilize a cross-modality reconstruction filter which is adaptive to the size of the refining super-pixel segments. We propose a novel depth encoding scheme, which includes specific arithmetic encoder and handles misalignment outliers. We demonstrate that our scheme is especially applicable for low bit-rate depth encoding and for fusing color and depth data, where the latter is noisy and with lower spatial resolution.

**INDEX TERMS** 3D, 3-D depth, fusion, compression, super-pixel, time-of-flight, ToF, view-plus-depth, V+Z, V+D.

## I. INTRODUCTION

Representation and processing of real-world three-dimensional (3D) visual scenes has been of increasing interest recently in the light of new forms of immersive visualization achieved by the advancement of 3D display technology. The geometrical information about scenery can be sensed into an intensity image-like representation referred to as ‘‘depth map’’. Each pixel of a depth map represents the distance to a particular point in 3D space as seen from a particular view perspective. Depth maps are combined with confocal captures of 2D color images to form a 3D representation, referred to as ‘‘View-plus-depth’’ (V+Z) [1], [2], where both images have the same size and are pixel-to-pixel aligned to augment each color pixel with its position in space. V+Z can be used for various applications, such as virtual view synthesis by Depth-Image Based Rendering (DIBR) [3], computational photography effects of refocusing, vertigo or synthetic aperture [4], and mixed reality [5] The format has been standardized in 3D video compression standards (3DVC) [1]. Figure 1 illustrates the color and depth modalities in blended transparent combination (i.e. the actual color is shown on the upper left corner and depth is shown pseudo-color coded in the lower right corner). As seen in the figure, the depth modality is a piece-wise smooth function, where edges are formed by objects situated in different distances. The blended transparency reveals that there is a certain alignment congruency between edges of both modalities (i.e. scene objects are at a certain depth).

Depth maps of real scenes are captured and estimated by, generally, two groups of techniques, referred to as passive or active sensing. The ‘‘*structure-from-stereo*’’ estimates depth by matching similar (corresponding) pixels between two or more images captured from different perspectives. Dedicated (i.e. active) range sensors employ Time-of-Flight (ToF) principles to directly capture depth [6], [7]. In all cases, depth estimation or measurement usually come degraded by various artifacts. For example, in passive sensing, degradation

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

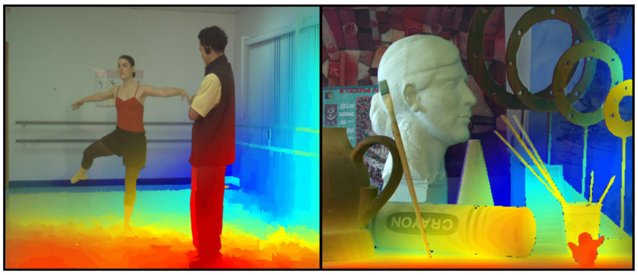


FIGURE 1. View-plus-Depth edge congruency examples (low-right parts show depth modality in pseudo colors) for (a) "*Ballet*", (b) "*Art*" [8] data sets.

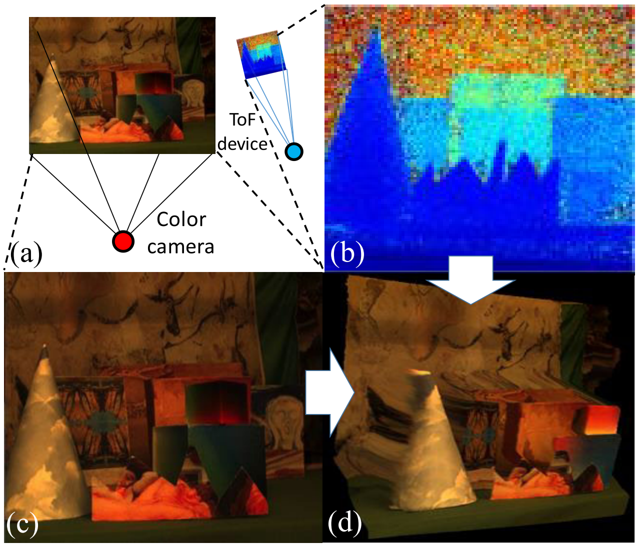


FIGURE 2. Example of 3D sensing by (a) non-confocal asymmetric camera setup, where (b) HD color sensor, (c) sensed depth map zoomed to emphasize noise, (d) de-noised, aligned, and fused output for virtual DIBR view synthesis.

is caused by ambiguity in texture-less areas or repetitive patterns. Furthermore, depth resolution is degraded by the non-linear conversion (quantification) of matched disparities [8]. In ToF approaches, depth data is limited by the low sensor resolution, e.g. 120 × 160 [9]. It is constrained by the requirement of the photo-elements to work in high-sensitivity conditions, which is ensured by increasing the sensing element area. ToF sensing elements typically have plate size of 150 $\mu m$, compared to the size of modern color sensors which is about 2 $\mu m$ [7]. Otherwise, ToF sensors provide better depth resolution quality, however they are usually non-confocally located with respect to the companion color sensors. A 3D data fusion is required to mix the modalities into a *confocal* representation. Such processing stage includes projection alignment, non-uniform data resampling, denoising, and depth enhancement filtering [10], [11]. Figure 2 illustrates the fusion process for a non-confocal asymmetric V+Z setup.

In this work, we focus on the problem of optimally representing the V+Z data. Our inspiration is based on the fact that the depth is a piecewise smooth function aligned to scene object edges, which open possibilities for its sparse representation. We consider two cases. First, we consider an already aligned V+Z representation where depth and color maps are with the same resolution and we target the smallest decimated depth map representation which would ensure a faithful full-resolution depth reconstruction. Such approach is instrumental for depth compression and streaming in the form of auxiliary data. Second, we consider a case, where the depth comes as low-resolution, noise-degraded map and the task is to restore it to its full resolution. Such case is instrumental in non-confocal ToF/color data fusion systems.

### A. DEPTH AND VIEW-PLUS-DEPTH COMPRESSION

Depth compression schemes can be roughly separated into two categories regarding whether the depth maps are compressed independently from or jointly with the aligned color images [12]–[23]. Methods for direct depth map compression include decomposition techniques for effective prediction of the underlying piecewise-smooth function [12]–[15] or techniques for representing and compressing depth contours [16]–[18]. The inter-relation between the V and Z modalities has been explored in several works utilizing different cross-segmentation approaches [18]–[20]. Other works have considered block partitioning and "*wedgelet*" edge modeling of non-rectangular intra-block segments [2], sometime combined with inter-component prediction [1]. Some of the tools in image/video compression standards such as "JPEG/JPEG2000" [21], [22] or "H.264/HEVC/AVC" [23] are also effectively applicable for depth compression.

### B. 3D FUSION OF ASYMMETRIC VIEW-PLUS-DEPTH DATA

3D data fusion problem has been considered in different research settings aiming at aligning the edges of the two modalities while enforcing piecewise smoothness of the depth. A layered Markov Random Field (MRF) model in [24] with the purpose to correlate a continuous smooth surface to the given samples of depth data. The MRF formalization have been further advanced in [25], [26] and [27]. In [28], the problem has been cast as in a dissipated heat anisotropic diffusion network, where the heat sources are the available data samples. Simultaneous surface fit and denoising have been considered in a number of works, employing either joint-geodesic distance [29], or moving least squares [30], or multi-point regression [31]. Cross-modality filters such as bilateral [32] and non-local [33], [34] have been implemented as to utilize the high-resolution color map as a guiding modality is the depth reconstruction process. Solutions based on bilateral filtering have been proposed in [35]–[38], and solutions based on non-local filtering have been proposed in [39] and [40]. Other forms of edge-preserving guided filtering have been proposed as well [41]. A method based on total generalized variation (TGV) for optimization of anisotropic diffusion tensor structure has been proposed in [42]. The article provides also a benchmark data set for 3D fusion resampling quality evaluation for real-case data of asymmetric V+Z capturing setup, where depth maps are obtained by noisy ToF sensor.

### C. RELATION WITH PREVIOUS WORK

Previously, we have proposed techniques for depth resampling and 3D fusion for the case of an asymmetric non-confocal V+Z camera setup, where the depth is sensed in low-sensing conditions [43], [44], as well as techniques for near-lossless depth encoding [45], [46]. In the present work, we present a general framework, which addresses both cases.

We further extend the technical stages of super-pixel (SP) segmentation, resampling, regularization, encoding, and 3D fusion. More specifically, we modify the segmentation clustering stage proposed in [45], [44] to ensure border congruency at hierarchical refinement levels and seed the SP clusters for non-uniform data samples to serve the case of projected data. Furthermore, we address the problem of possible misalignment between V and Z modalities caused by sensing artifacts. Such misalignment produces edge outliers that concentrate high amount of errors in the global cost metrics and thus mislead the error optimization in the coding process. To this end, we propose an efficient encoding scheme of such outliers in so-called "*yield-flow*" protocol. A modification of the adaptive regularized reconstruction is proposed as well.

The article is organized as follows: *Section II* provides some preliminaries and notation conventions along with description of basic super-pixel clustering, *Section III* describes the proposed general framework, *Section IV* describes application realizations for depth encoding and 3D fusion of asymmetric V+Z sensor data utilizing a proposed multi-layer congruent super-pixel clustering mechanism, *Section V* provides experimental results, and the manuscript is finalized in *Section VI* for some conclusive remarks.

## II. PRELIMINARIES

### A. DEPTH AND VIEW-PLUS-DEPTH COMPRESSION

Consider a color image is some three-component color space, for example CIELAB [47]. Each pixel with index $j$ is a three-component vector $\mathbf{V}_j = [l, a, b]_j$, $j = (1, .., J)$. When needed, the pixel is given with its coordinates related to the camera projective system $\mathbf{x} = (x, y)$, $\{x, y\} \in \mathbb{R}^2$ [48]. The associated depth value is denoted by $Z_j$.

When sensed by active sensors, the depth map relates with the range data $D$, which represents distances from pixels to scene points [42]. When estimated from stereo, the depth values relate with disparity values $d$ showing the shifts between corresponding pixels [8]. In many encoding applications, depth is quantized as "*inverse depth*" [7]

$$z_n = 1\Big/\frac{n}{N}\left(\frac{1}{Z_{MIN}} - \frac{1}{Z_{MAX}}\right) + Z_{MAX}, \qquad (1)$$

where $\{z_{MIN}, z_{MAX}\}$ are the minimum (MIN) and the maximum (MAX) sensed values of the depth in the scene, and $N$ is the number of quantization levels. Usually, disparity and depth are represented as 8-bit integers, $Z = (0, .., 2^8 - 1)$.

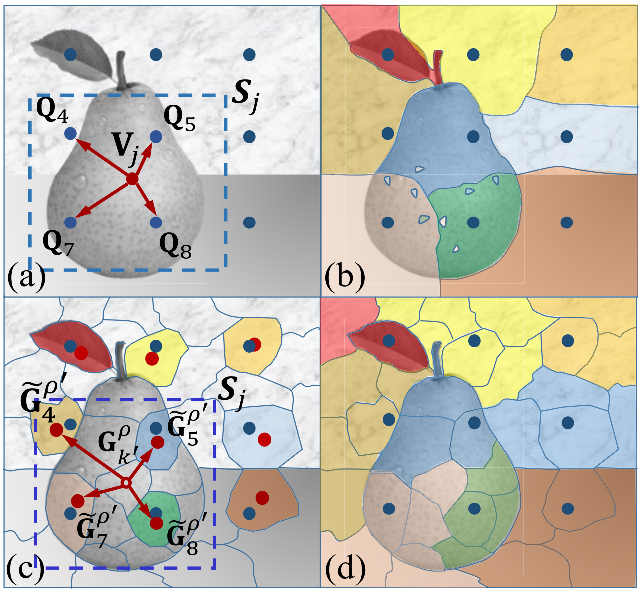


**FIGURE 3.** Super-pixel clustering: (a) Simple Linear Iterative Clustering (SLIC) method [49], [51], (b) its possible output, (c) proposed modification, and (d) its possible output.

When sensed by some active range sensor, depth maps are non-confocal to the color maps and can come with lower spatial resolution and floating-point higher range, e.g. $Z_l$, $l = (1, .., L), L \ll J, Z \in \left[0, 2^{16}\right]$. In such case, the output of the V+ Z representation is calculated by projective alignment and depth resampling, referred to as "*3D fusion*".

### B. SUPER-PIXEL CLUSTERING

Super-pixel (SP) based segmentation plays an essential role in the proposed framework. Super-pixels are segments that have near-isotropic and compact representation with low-computational overhead. A typical super-pixel behaves as a raster pixel on a low-resolution near-regular grid. Perceptually, SP areas are homogeneous in terms of color and texture. Two main approaches for generating super-pixels can be cited, namely: Simple Linear Iterative Clustering (SLIC) [49], [50] and Super-pixels Extracted via Energy-Driven Sampling (SEEDS) [51]. Hereafter we adopt the SLIC approach.

An elegant feature of the super-pixel segmentation is that it takes the desired number of SPs as an input parameter and that for this number it is reproducible in terms of same SP areas (clusters) and indexing that follow the edge shape between color textures. For that reason, SP segmentation is instrumental for finding objects shapes in a scene, see the pear example in Figure 3 (b). The SP clustering is initialized by defining $K$ seed locations of color points $\mathbf{Q}_k, k = (1, .., K)$. Those points are chosen to be equidistantly sampled in image coordinates $\mathbf{x}_k = \{x, y\}_k$ for roughly calculated sampling shifts [50]

$$s_{\{\mathcal{H},\mathcal{W}\}} = (\mathcal{H}\mathcal{W})\big/K, \qquad (2)$$

where $\mathcal{H}$ and $\mathcal{W}$ are the pixel dimensions of the sensor (c.f. blue dots in Figure 3 (a)). Pixels $\mathbf{V}_j$ are clustered to SP segments $C_k$, where each segment $C_k$ span $N_k$ pixels, as follows. For each image pixel $\mathbf{V}_j$, a neighborhood $S_j$ (i.e. seeding support region) is associated. The neighborhood seeding support region spans a rectangular area of dimensions $2s_{\{R,C\}}$ around $\mathbf{V}_j$. The closest similarity of $\mathbf{V}_j$ to seeding points $\mathbf{Q}_k$ within $S_j$ is found by applying e.g. a bilateral cost, which assigns $\mathbf{V}_j$ to segment $C_\kappa$

$$k = \arg \underset{k}{\mathrm{MIN}} \left\{ \sqrt{\lambda_\rho \left\| \mathbf{x}_k - \mathbf{x}_n \right\|_2^2 + \lambda_C \left\| \mathbf{Q}_k - \mathbf{P}_j \right\|_2^2} \right\} \quad (3)$$

where $\{\lambda_\rho, \lambda_C\}$ are weighting constants. The clustering is iterated by updating the seeding points $\mathbf{Q}_k$ with the arithmetic mean for the pixels assigned to the associated cluster $C_k$

$$\mathbf{Q}_k = \frac{1}{N_k} \left\{ \sum\nolimits_{j \in C_k} \mathbf{V}_j \right\}. \quad (4)$$

A polishing step that enforces connectivity of points of each segment is applied at the end [50].

## III. PROPOSED GENERAL FRAMEWORK FOR V+Z RESAMPLING AND FUSION

### A. DEPTH RESAMPLING SCHEME

We propose a general depth resampling scheme (DRS) to be used as a building block in various applications. The aim is to find an optimal representation of the depth map, for either compression or depth reconstruction. The block diagram of the proposed scheme is given in Figure 4. It takes as input the color image $\mathbf{V}$, a set of initial seeding points $\mathbf{Q}$, and a depth map $Z$, which might be or might be not with the same resolution as the color image. The color image is segmented by a SP clustering operator $\Xi$,

$$C = \Xi\left(\mathbf{V}, \mathbf{Q}\right). \quad (5)$$

A masking operator $\mathcal{M}$ fills each segment $C_k$ with constant depth values $\bar{Z}_k$, thus generating a depth map with the same resolution as the color image $\mathbf{V}$

$$\bar{Z} = \mathcal{M}\left(Z, C\right). \quad (6)$$

The values $\bar{Z}_k$ are selected or calculated depending on the application. A cross-modality adaptive reconstruction filter $B$ reconstructs an estimate of the depth map

$$\hat{Z} = \mathcal{B}\left(\bar{Z}, \mathbf{V}, C\right), \quad (7)$$

Furthermore, a depth down-sampling operator $D$ turns either $\bar{Z}$ or $\hat{Z}$ into low-resolution depth map $Z$

$$\mathcal{Z} = \mathfrak{D}\left(\{Z, \hat{Z}\}, C\right), \quad (8)$$

The scheme is general and can be integrated in other techniques requiring depth resampling and refinement. We develop two such techniques, one related with near-lossless depth encoding and one related with asymmetric V+Z data fusion. However, we first propose a modification of the SP segmentation which would better serve the targeted applications.

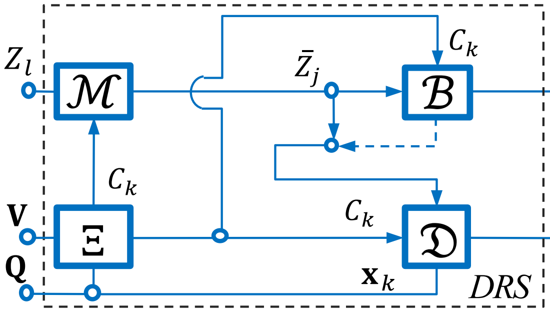


**FIGURE 4.** Proposed depth resampling scheme (DRS).

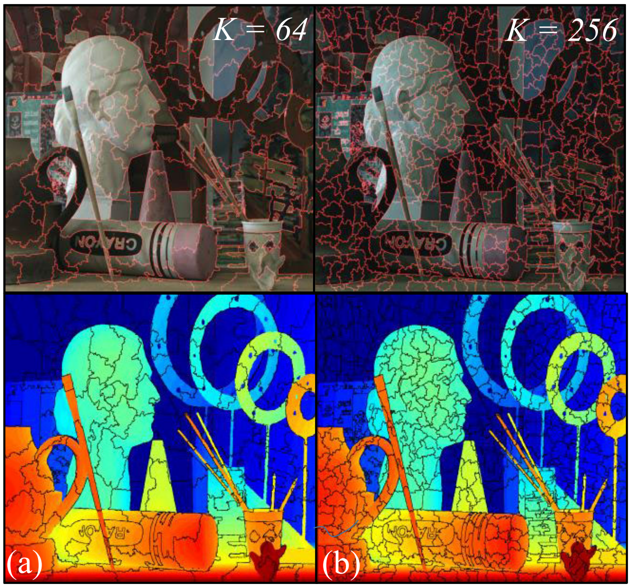


**FIGURE 5.** Example of edge congruency of proposed super-pixel partitioning scheme applied on color map of V+Z data set *"Art"* for number of segments $K$ = (a) 64, (b) 256; boundaries visualized on color (up) and depth (down) modalities.

### B. MULTI-LAYER CONGRUENT SUPER-PIXEL CLUSTERING

In order to facilitate the operations in the DRS, we propose a novel multi-layer SP clustering to serve as the operator $\Xi$ (5). It is based on the SLIC method [50] and aims at finding a segmentation that has contour congruency among different refinement levels in a sense that a refinement level with smaller number of segments has segment boundaries of SPs that are in union to those of a refinement level with higher number of segments (c.f. Figure 5 (a, b)).

In the proposed solution, the clustering for some desired number $K$ of SPs is done by several refinement stages $\rho$, starting from an initial very fine mosaic $\rho = 0$. Assume the initial number of segments $K^0$ and the corresponding seeds $\mathbf{Q}_k^0$ are selected in a way that only a few points $M$ define each cluster $C_k^0$(e.g. $M_k^0 = 4$). For each iterative step $\rho > 0$, the number of SPs is chosen to be smaller (e.g. decreased by two in each iteration)

$$K^\rho = \left((\mathcal{HW}) \Big/ M^0\right) \Big/ 2^\rho. \quad (9)$$

The clustering process for $\rho > 0$ combines segments of SP cluster $C^{\rho'}$ obtained in previous iteration $\rho' = \rho - 1$. Denote the actual seeding points at iteration $\rho'$ by $\mathbf{G}_{k'}^{\rho'}$. In the general case, these are at non-uniform locations $\mathbf{x}_{k'}$, $k' = 1, 2, \ldots, K^{\rho'}$. In order to find the new seeders for iteration $\rho$, one first sets a coarser uniform grid with steps $s_{\{\mathcal{H},\mathcal{W}\}}^{\rho} = (\mathcal{H}\mathcal{W})\big/K^{\rho}$(see the blue points in Figure 3 (c)). Seeders $\tilde{\mathbf{G}}_{k}^{\rho'}$ being closest to this grid are considered as *attractors*, as they are meant to attract other super-pixel centroids to form the new segments $C_k^{\rho}$ (see the red points in Figure 3 (c)). This is done by calculating the bilateral distances (3) between each $\mathbf{G}_{k'}^{\rho'}$ and $\tilde{\mathbf{G}}_{k}^{\rho'}$ within the neighborhood $2s_{\{H,W\}}^{\rho}$ (see the blue dashed rectangle in Figure 3 (c)), and appending super-pixels from iteration $\rho'$ to corresponding attractor super-pixels. This operation gives the new segment support $C_k^{\rho}$. The new seeding points $\mathbf{G}_k^{\rho}$ are then updated both for their positions and values. The seeding point location $\mathbf{x}_k$, $k = (1, 2, \ldots, K^{\rho})$ and its intensity value is calculated as the arithmetic mean of segment points

$$\mathbf{x}_k = \frac{1}{M_k^{\rho}}\left\{\sum\nolimits_{j\in C_k^{\rho}} \mathbf{x}_j\right\}, \quad \mathbf{G}_k^{\rho} = \frac{1}{M_k^{\rho}}\left\{\sum\nolimits_{j\in C_k^{\rho}} \mathbf{V}_j\right\}, \qquad (10)$$

Essentially, the operation is repeating the basic SLIC but working at each layer with super-pixels instead of pixels and maintaining non-uniform seeding positions to better describe the properties of the embedded super-pixels. It is important to mention that the resulted clustered segments $C_k^{\rho}$ combine pixels from sub-segments $C^{\rho'}$(c.f. Figure 3 (d)), thus ensuring border congruency. The iterations end upon reaching a desired number of super-pixel segments $K^{\rho}$.

The proposed modification of the SP clustering brings a few benefits. First, it leads to a considerably better modeling of texture transitions (c.f. Figure 5). Second, using the mass center locations for seeding points, prevents the occurrence of a misaligned clustering done on finer mosaic scales for the consecutive iterations. The congruency of SP boundaries is of vital importance for simplifying the encoding approach and improving the speed and quality performance of the originally proposed compression methods [44], [45].

### C. DEPTH RECONSTRUCTION

The operator $\mathcal{B}$ (7) is expected to exploit the relation between color and depth through a cross-modality guided reconstruction filter [33], [32]. In practice, we adopt the cross-

bilateral filter as modified in [37]. Two weight laterals are applied per pixel $\mathbf{V}_j$ in pixel neighborhood (e.g. square block) $\psi_j$:

$$\varpi_m = \lambda_s \left\|\mathbf{x}_j - \mathbf{x}_m\right\| \lambda_c \left|\mathbf{V}_j - \mathbf{V}_m\right|, \quad m \in \psi_j, \qquad (11)$$

where $\{\lambda_s, \lambda_c\}$ are parametrized Gaussian smoothing kernels [32] for the spatial proximity and intensity similarity correspondingly. Then a bilateral weighted average is applied to each depth pixel

$$\hat{Z}_j = \sum_m \varpi_m Z_m \Big/ \sum_m \varpi, \qquad (12)$$

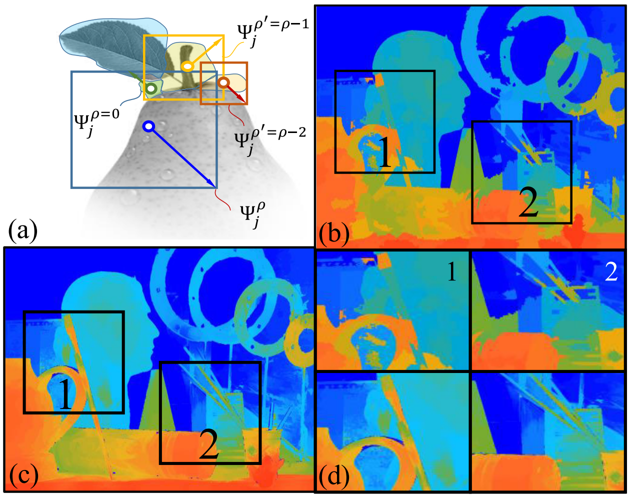


**FIGURE 6.** **Proposed reconstruction filter applied on *"Art"* data set: (a) an adapting principle, (b) super-pixel masking operator output *Z* for 64 segments and several refinement updates, (c) filtered output$\hat{Z}$, and (d) zoomed regions (labeled by white rectangles "1" and "2"); colors are exaggerated for better perception of details.**

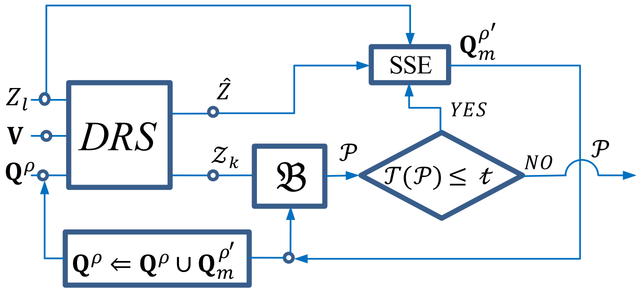


**FIGURE 7.** **Block diagram of depth map encoding employing DRS.**

to form the reconstructed map $\hat{Z}$ (7). The neighborhood $\psi_j$ (c.f. Figure 6 (a)) is selected to be proportional to the segment size of the current refinement level $\rho'$

$$\#\left(\psi_j\right) \propto M^0 2^{\rho'}. \qquad (13)$$

The spatial proximity kernel $\lambda_s$ must be related to the size of the neighborhood $\psi_j$. An example of the filter performance is demonstrated by visual outputs given in Figure 6 (b-d).

## IV. APPLICATION CASES

### A. DEPTH MAP ENCODING APPLICATION

First application is encoding of depth map in the V+Z representation, where the color and depth modalities are already aligned. With reference to Figure 4, this means that the input pixel and depth maps are with the same spatial resolution.

Figure 7 illustrates the proposed technique. The decimated depth map $\mathcal{Z}$ being output of DRS undergoes arithmetic encoding exemplified by the operator $\mathfrak{B}$. It outputs an encoded binary sequence $\mathcal{P}$. The reconstructed depth map $\hat{Z}$ is compared against the original one by means of Sum of Squared Errors (SSE) on super-pixel level, and regions of high reconstruction error are split into finer and embedded

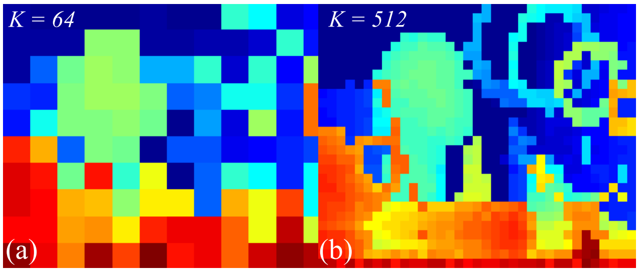


**FIGURE 8.** Isotropic map generation applied on "*Art*" data set for different number of super-pixels, $K$ = (a) 64, (b) 512 elements (c.f. Figure 5 (b)).

super-pixels. Their centroids are returned to the DRS module which updates the outputs for next-iteration reconstructed depth map $\hat{Z}$ and its decimated version $\mathcal{Z}$. The latter one along with the localization information for partitioned SP segments is stored in a predictive sequence unified with $\mathcal{P}$. The refinement process is applied iteratively subject to an encoding bit-budget $\mathscr{t}$ compared with the bit length $\mathcal{T}$ of the sequence $\mathcal{P}$, i.e. $\mathcal{T}(\mathcal{P}) \leq \mathscr{t}$.

### 1) DEPTH REFINEMENT BY SUPER-PIXEL PARTITIONING

The SSE is calculated for each super-pixel

$$\varepsilon_k^\rho = \left\| Z_j - \hat{Z}_j \right\|^2, \quad s.t. \, j \in C_k^\rho, \tag{14}$$

SPs with highest errors $\varepsilon_k^\rho$ are marked for further refinement by going to the finer scale $\rho' = \rho - 1$ being kept after the multi-layer clustering. The seeding points $\mathbf{G}_k^{\rho'}$ and the associated $C_k^{\rho'}$ segments are fed back to DRS.

### 2) ENCODING SCHEMES

We encode three components: (A) The uniformly-decimated depth map $Z$ produced at iteration $\rho$ is encoded in predictive sequence $P^Z$; (B) the depth values corresponding to partitioned SPs are encoded in predictive sequence $P^{pt}$; and (C) the partitioned SP structure is encoded in the a binary sequence $B$.

(A) The decimated depth map $\mathcal{Z}^\rho$ at stage $\rho$ has an isotropic structure with dimensions $s^\rho_{\{\mathcal{H},\mathcal{W}\}}$ and values $\mathcal{Z}_k^\rho$, corresponding to each segment $C_k^\rho$, as illustrated in Figure 8. The segmentation structure comes from the color modality and can be reproduced, thus it does not need to be encoded. The map itself is encoded in a predictive sequence $\mathcal{P}^Z$ similarly to "JPEG-LS" standard [21] and described in detail in our previous work [45].

(B) Consider $M$ partitioned SPs with corresponding depth values $\mathcal{Z}_m^{\rho'}$, $m = (1, .., M)$. These are predicted in a tree structure $\mathcal{P}^{pt}$ by the difference with their parent sub-pixel $\mathcal{Z}_k^\rho$

$$\mathcal{P}_m^{pt} = \mathcal{Z}_k^\rho - \mathcal{Z}_m^{\rho'}. \tag{15}$$

The entire sequence $\mathcal{P}^{pt}$ is subsequently encoded by an adaptive multi-alphabet range coder [52], [45].

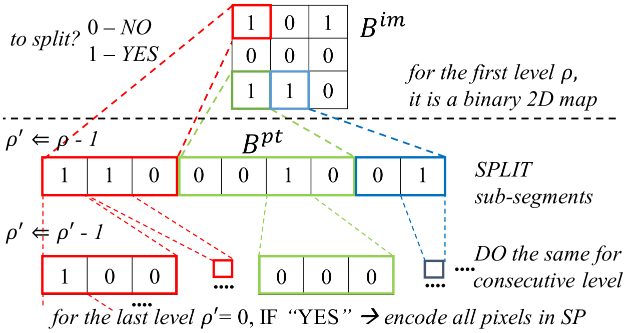


**FIGURE 9.** Binary tree structure for encoding super-pixel refinement partitioning.

(C) The partitioning is encoded in a binary sequence $B$, formed by two sub-sequences $\{B^{im}, B^{pt}\}$, which encode the partitioning of the isotropic map $Z^\rho$ and the partitioning tree structure $\mathcal{P}^{pt}$ respectively, as shown in Figure 9. Partitioned SPs are indexed by 1 (*split*) and non-refined SPs are indexed by 0 (*no split*). The map $B^{im}$ encodes the partitioned SPs for the initial stage $\rho$ and the shape and indexing follow those of the isotropic map $\mathcal{Z}^\rho$. The binary map is scanned column-wise to initialize the first index tree level in $B^{pt}$. Next level is for partitioned SPs belonging to consecutive refinement stages $\rho' = \rho - 1$. Those are encoded in a concatenated sequence in $B^{pt}$. Note that it does not need to store information about the number of children of refined SPs, as this is automatically found when the SP clustering for a refinement level is run. For the last refinement level $\rho' = 0$, there is an exception: SPs which are marked for partitioning are not indexed further, since the segment is entirely encoded by from the original depth values. The sequence $B^{im}$ is encoded separately from the rest of the tree by a "*Context-Adaptive Binary Range Coder*" (CABRC) [53]. For the context modeling, it is assumed that "*split/no-split*" of current SP depends on "*split/no-split*" of its neighbors. Using this assumption, the value of a binary element $B_k^{im}$, is assigned to four possible binary sub-contexts indexed by the sum of neighboring pixels.

### 3) ENCODING EDGE OUTLIERS BY "YIELD-FLOW" PROTOCOL

The efficiency of the proposed depth encoding approach relies on the ideal consistency between color and depth modalities. However, in real case of V+Z capture, depth maps can come with various artifacts caused by stereo-correspondence errors, low-resolution non-confocal depth sensors along with projection misalignment and resampling ambiguities introduced by measurement errors [43], [44]. Examples of regions with such artifacts are given for a frame of "*Ballet*" data set in Figure 10 (c, d). While artifacts of the above-mentioned types are affecting a relatively small number of pixels, the encoding residual error will be concentrated precisely around them (c.f. Figure 10 (e)). We denote such problematic areas as

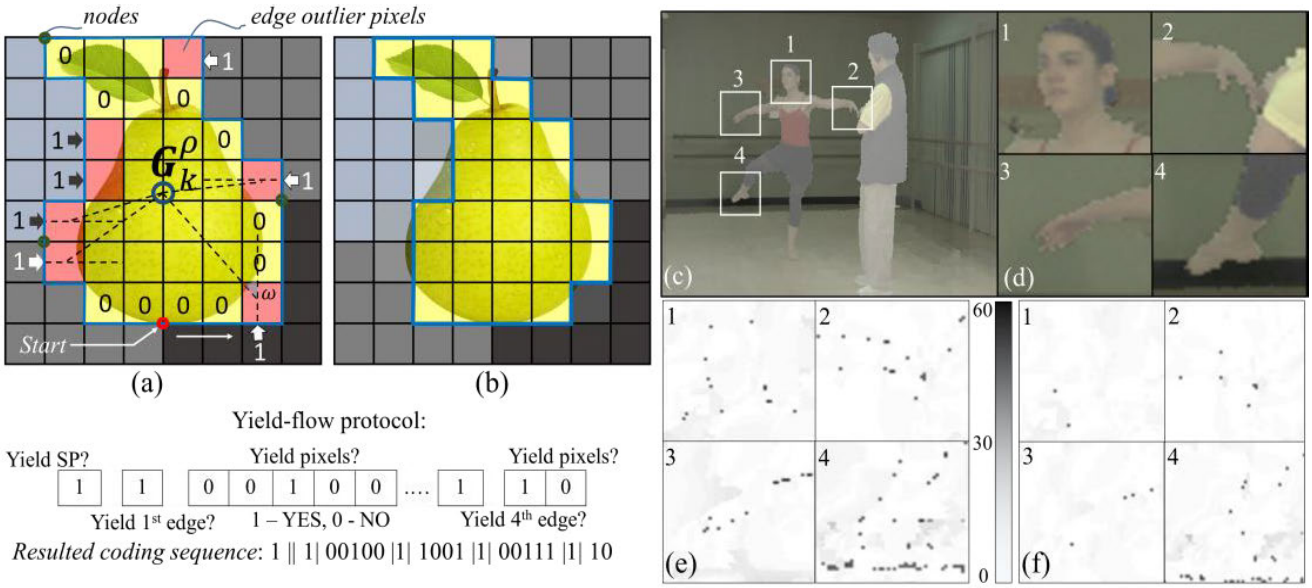


**FIGURE 10.** Proposed edge outlier coding approach and resulting coding sequence: (a) input segment, (b) modified output; Example of an edge outlier encounter for a frame from "*Ballet*" data set: (c) depth and color fusion (blended with transparency), (d) zoomed regions, absolute residual error for encoded depth for the same bitrate of t~0.016 bpp: (e) non-applied (~36 dB) and (f) edge outlier encoding applied (~38.92 dB).

*edge-consistency outliers* (ECO). The SPs which contain ECO, will indicate high SSE values (14), then the refinement partitioning will concentrate on those SPs attempting better quality which might go until the last refinement stage is reached and pixels are encoded individually. Apparently, the refinement scheme applied in such manner will be inefficient and could fail producing an optimal encoding output for the given bit-budget $t$. To tackle the problem, we propose an optional ECO binary encoding scheme called "*yield-flow*" protocol (YF). It indicates an encounter of possible ECO, if the partitioned SP children have at least two members with the same depth value as of the parent SP. In such case, the encoding system activates YF process that consists of sequence of {"YES" - 1, "NO" –0} flags (c.f. Figure 10 (a, b). The first bit of YF "*tells*" the encoder whether the SP is to be encoded for ECO. If YES, then the YF follows the internal pixel boundary of the SP counter clock-wisely (c.f. Figure 10 (a, b)) starting from lowest-left boundary node of neighboring SPs. The positive bit value indicates whether the boundary segment has to be processed. The positive bits in YF will indicate a "*yield*" procedure: The depth value of processed pixel - $Z_j$ is replaced with the value of neighboring pixels in the horizontal and vertical nearest direction that belongs to other SP clusters of the same refinement stage. In case of many choices, the decision is done for the neighboring pixel that forms the smallest angle $\omega$ between the neighborhood direction and the direction to SP centroid $\mathbf{G}_k^\rho$. In our realization, the yield process meets the following requirements. First, the SP should belong to a refinement stage higher than a certain threshold $\rho > t_\rho$, (e.g. $t_\rho = 5$). Second, pixels considered for the yield process are those that have no error comparing to GT for the new assigned depth value. Since the yield-processed pixels belong to GT, then those are excluded from the SP clusters of all higher refinement stages and should be skipped also by the regularization filtering step. The performance of the proposed edge outlier encoding is exemplified in Figure 10 (f), where it is shown that most of ECO are suppressed for a significant quality metric gain (c.f. Figure 10 (c-f)).

### B. FUSION FOR ASYMMETRIC V+Z CAMERA SETUP

The general DRS can be applied for 3D fusion of asymmetric V+Z data, provided some pre-processing is performed before feeding the DRS module as shown in Figure 11. In the above-mentioned setting, the two modalities are not aligned as they come from two non-confocal sensors and the dedicated depth sensor is usually of lower resolution. The depth pixels $Z_k$, $k = (1, .., K)$ , $K < J$ should undergo a re-projection step $\mathbf{\Pi}$ to locate them onto the image grid of the color sensor

$$Z_k^p = \mathbf{\Pi}\left(Z_k, \boldsymbol{f}\right), \tag{16}$$

where $Z_k^p$ are re-projected samples and $\boldsymbol{f}$ is a set of camera parameters related to some multi-view geometry model [48], [54]. At the initial SP clustering stage, there are strictly $K$ seeding points $\mathbf{Q}_k^p$ coinciding with the projected locations $\mathbf{x}_k^p$ of $Z_k$. The same association is done for the output samples $\mathcal{Z}_k$. The projected locations $\mathbf{x}_k^p$ appear non-uniformly located with respect to the color map grip. Therefore, $\mathbf{Q}_k^p$ are found by a standard interpolation $L$ (e.g. by bi-cubic splines [55])

$$\mathbf{Q}_k^p = \mathcal{L}\left\{\mathbf{V}, \mathbf{x}_k^p\right\}. \tag{17}$$

The size of the seeding support region $S_j$ for the SP clustering operator $\Xi$ (5) is fixed by the scale difference between the

**TABLE 1. Metric results for V+Z Fusion Resampling Techniques.**

| | "Book" | | "Shark" | | "Devils" | |
|---:|---|---|---|---|---|---|
| Metrics | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| Voronoi(NN) | 21.21 | *4.05* | 24.745 | *5.01* | 15.15 | *3.18* |
| Bilinear | 20.56 | *4.93* | 28.21 | *5.02* | 16.49 | *3.22* |
| Bicubic | 27.77 | *17.60* | 31.68 | *19.48* | 26.25 | *18.03* |
| BF [36] | 21.05 | *6.19* | 24.76 | *7.79* | 15.30 | *5.27* |
| AD [28] | 26.10 | *15.30* | 30.20 | *17.08* | 24.55 | *16.14* |
| GF [41] | 17.02 | *5.32* | 19.87 | *6.51* | 12.38 | *3.85* |
| Hyp [37] | 18.41 | ***3.46*** | 23.06 | *4.47* | 14.83 | *3.25* |
| Yang [38] | 24.87 | *13.28* | 29.85 | *15.86* | 23.50 | *14.72* |
| JGF [29] | 25.67 | *15.06* | 28.93 | *16.54* | 23.86 | *15.75* |
| IMLS [30] | 26.14 | *15.47* | 35.38 | *17.53* | 24.45 | *15.95* |
| CLMF [31] | 25.67 | *15.06* | 28.93 | *16.54* | 23.86 | *15.75* |
| TGV [42] | 19.70 | *3.86* | 23.92 | ***4.15*** | 14.31 | *3.24* |
| *Proposed* (SP) | 19.84 | *4.69* | 22.85 | *4.56* | 14.84 | *3.12* |
| *Proposed* (1 iter.) | 17.94 | *4.22* | 21.13 | *5.33* | 12.77 | *3.20* |
| *Proposed* (3 iter.) | **16.16** | *4.21* | **19.33** | *4.48* | **11.60** | ***2.81*** |

* - Bold text indicates best performing result in each data set

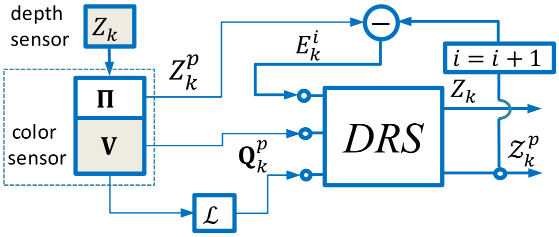


**FIGURE 11. Block diagram of proposed 3D fusion employing DRS.**

dimensions of the two sensors: $\{\mathcal{W}, \mathcal{H}\}_V$ / $\{\mathcal{W}, \mathcal{H}\}_Z$. Furthermore, a *Richardson-Lucy iterative scheme* [56] is applied iteratively

$$E_k^i = Z_k^{i-1} - \mathcal{Z}_k^{i-1}, \quad E_k^0 = Z_k^0 \tag{18}$$

$$\hat{Z}^{i+1} = \hat{Z}^0 + \lambda_L \mathcal{B}\left\{\mathcal{M}\left(C, E_k^i\right)\right\}, \tag{19}$$

where $\lambda_L$ is a regularization constant. For each iteration $i$, the error residual $E_k^i$ is used as a feedback input to DRS, and further, the reconstructed result from $\mathcal{B}$ is accumulated for initial reconstruction $\hat{Z}^0$. Usually, very few iterations $i$ (e.g. $i \sim 3$) are enough to converge to optimal output of $\hat{Z}$.

## V. EXPERIMENTAL RESULTS

We present experiments demonstrating the utilization of the proposed framework in two cases: depth encoding and fusion of asymmetric V+Z data. To quantify the performance, we use the standard mean absolute error (MAE), root mean squared error (RMSE) and the related Peak-Signal-to-Noise Ratio (PSNR) in [*dB*] between the processed and ground true depth maps. For datasets, where geometry is represented by disparity maps, we use also the percentage of bad pixels (BAD) which shows the percentage of disparities which differ from the ground true disparity map by more than one pixel [8]. The following datasets are used in the experiments: *Microsoft's* "*Breakdancer*" and "*Ballet*" [57]; *Middlebury*'s "*Aloe*", "*Art*", "*Baby*", "*Dolls*", "*Teddy*", "*Cones*", and "*Bowling*" [8]; and ToF data. The latter contain scenes captured by asymmetric non-confocal V+Z stereo-camera setup, where the depth sensor is a noisy Time-of-Flight (ToF) camera with 120 × 160 pixels spatial resolution, while the color camera is of resolution 610 × 810 pixels.

### A. DEPTH COMPRESSION FOR VIEW-PLUS-DEPTH DATA

The quality metrics are calculated versus the encoding (compression) rate in bits-per-pixel (*bpp*) measured on the encoded isotropic map $P^{im}$. The first experiment characterizes the gain obtained by applying the reconstruction filter $B$ (7). The results are shown in Figure 12 (a). The quality is varied by varying the SP segmentation point on the plot indicates the PSNR between either $\bar{Z}$ or reconstructed (regularized) $\hat{Z}$ and the non-compressed depth. By increasing the number of SP elements $K$, one gets higher quality for the price of high bit rate. No further optimization is applied. Still, the proposed technique reaches PSNR of about 40 *dB* for $\mathcal{t} \leq 0.1$ *bpp* with additional improvement of at least 2 *dB* when the reconstruction filter is applied.

For optimized encoding, we employ the classical "*Rate-distortion Optimization Scheme*" [58]. The only varying parameter in the system is the choice of the refinement stage $\rho$ for the initial segmentation, which in all experiments was fixed to 256 elements, $\rho s.t.K \approx 256$. For the depth encoding scheme that utilizes also predictive refinement and the proposed YF encoding, the output results are compared

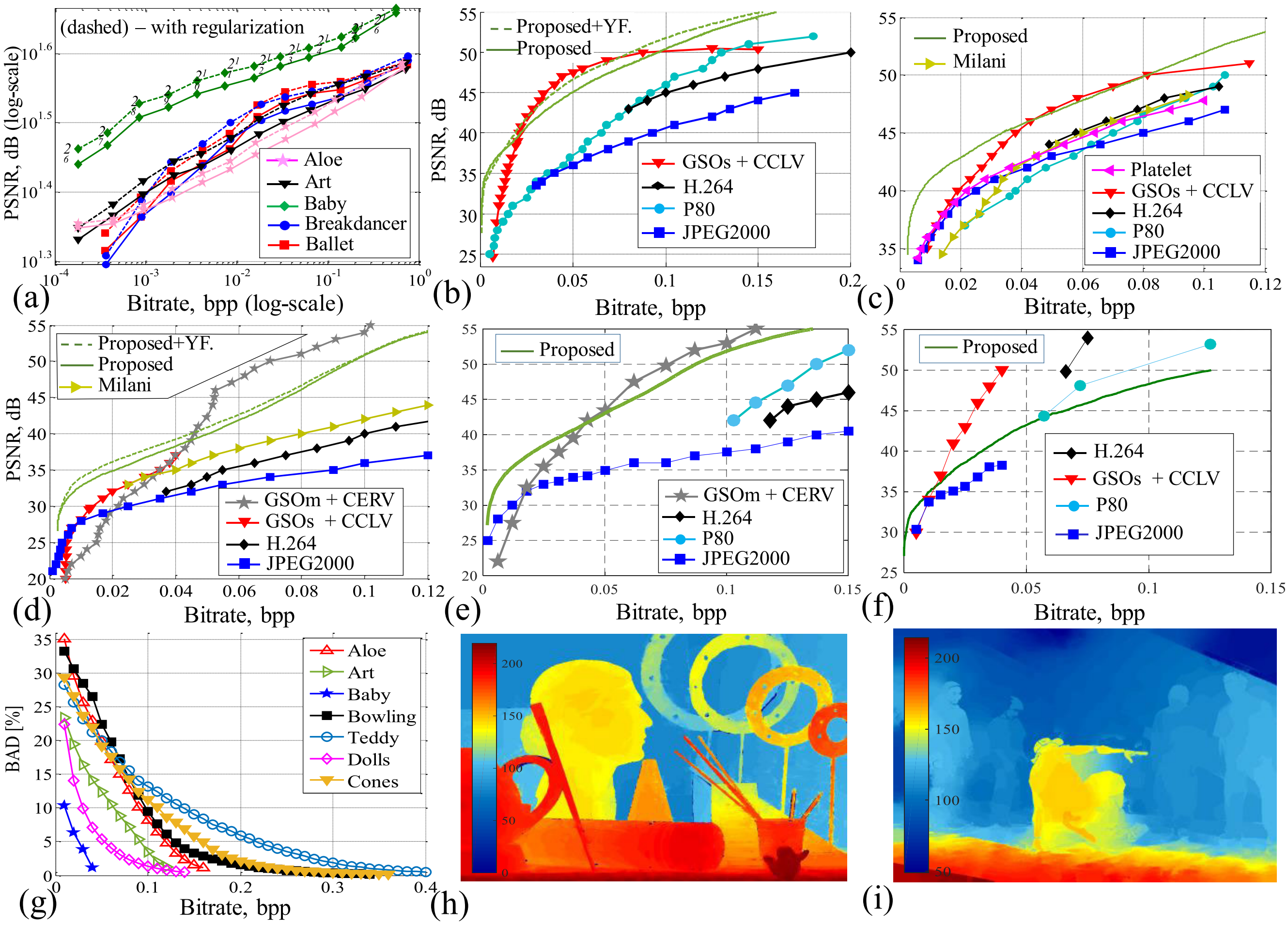


**FIGURE 12. Evaluation results for proposed depth encoding application for V+Z data: (a) performance of non-refined SP segmentation of depth maps (number of elements are denoted as orders of two for each point); comparison to state-of-art methods; in terms of PSNR for some datasets: (b) "*Ballet*", (c) "*Breakdancer*", (d) "*Art*", and (e) "*Baby*", (f) "*Bowling*"; (g) in terms of BAD metric; example of encoded depth maps: (h) "*Art*" (t∼0.008 *bpp*), (i) "*Breakdancer*" (t∼0.0014 *bpp*); method notation reference: "*Platelet*" [12], "*P80*" [15], "*GSOs+CCLV*", "*GSOm+CERV*" [16], "*Milani*" [18], "*JPEG2000*" [22], "*H.264*" [23].**

against the works denoted as: "*Platelet*" [12], "*Milani*" [18], "P80" [15], "*GSOs+CCLV*", "*GSOm+CERV*" [16], "*H.264*" [23], "*JPEG2000*" [22]. Since "*GSOs+CCLV*" and "*GSOm +CERV*" perform optimally for different bitrate zones, for those, a single plot is given that holds the better metric value. The results are given in the plots of Figure 12 (b-f) for different test data. The proposed method is clearly highly competitive and performs best for very low bitrate regions (e.g. $t \leq 0.05$), where the quality of the decompressed output is above 45 *dB*. This is considered near- lossless for most of the rendering applications utilizing depth maps [59]. A depiction of decoded and reconstructed depth map for "*Art*" data set for bit budget $t \cong 0.008$ *bpp* is shown in Figure 12 (h) and "Breakdancer" data set for bit-budget $t \cong 0.0014$ *bpp* is shown in Figure 12 (i). When YF is applied for test sets with problematic zones (e.g. "*Ballet*"), the results are highly competitive for the entire range. In another test, we calculate the BAD metric as plotted in Figure 12 (g). The curves show that for a wide range of tested stereo-matching datasets of *Middlebury* [8], the proposed technique robustly fades the BAD percentage to about $3-5\%$ for bitrates below $t < 0.2$ *bpp* [60]. Such performance is in par with the performance of the highest-ranked stereo-matching estimation algorithms [8]. The performance of our method is slightly inferior for datasets of low-depth contrast and low-resolution (e.g. "*Bowling*" (c.f. Figure 12 (f)).

### B. 3D RESAMPLING AND FUSION OF ASYMMETRIC VIEW-PLUS-DEPTH DATA

For this experiment we use the dataset from [42] which are commonly adopted benchmarking datasets. The datasets provide projected irregular data samples $Z_k^p$ ready to be applied for 3D fusion and resampling. The GT depth maps have been captured by another high-end high-definition depth sensor. The scenes are referred to as "*Shark*", "*Books*" (c.f. Figures 13, 14), and "*Devil*".

Along with basic methods of "*Voronoi* (*NN*)", "*Bilinear*", and "*Bicubic*" resampling, the proposed fusion technique has

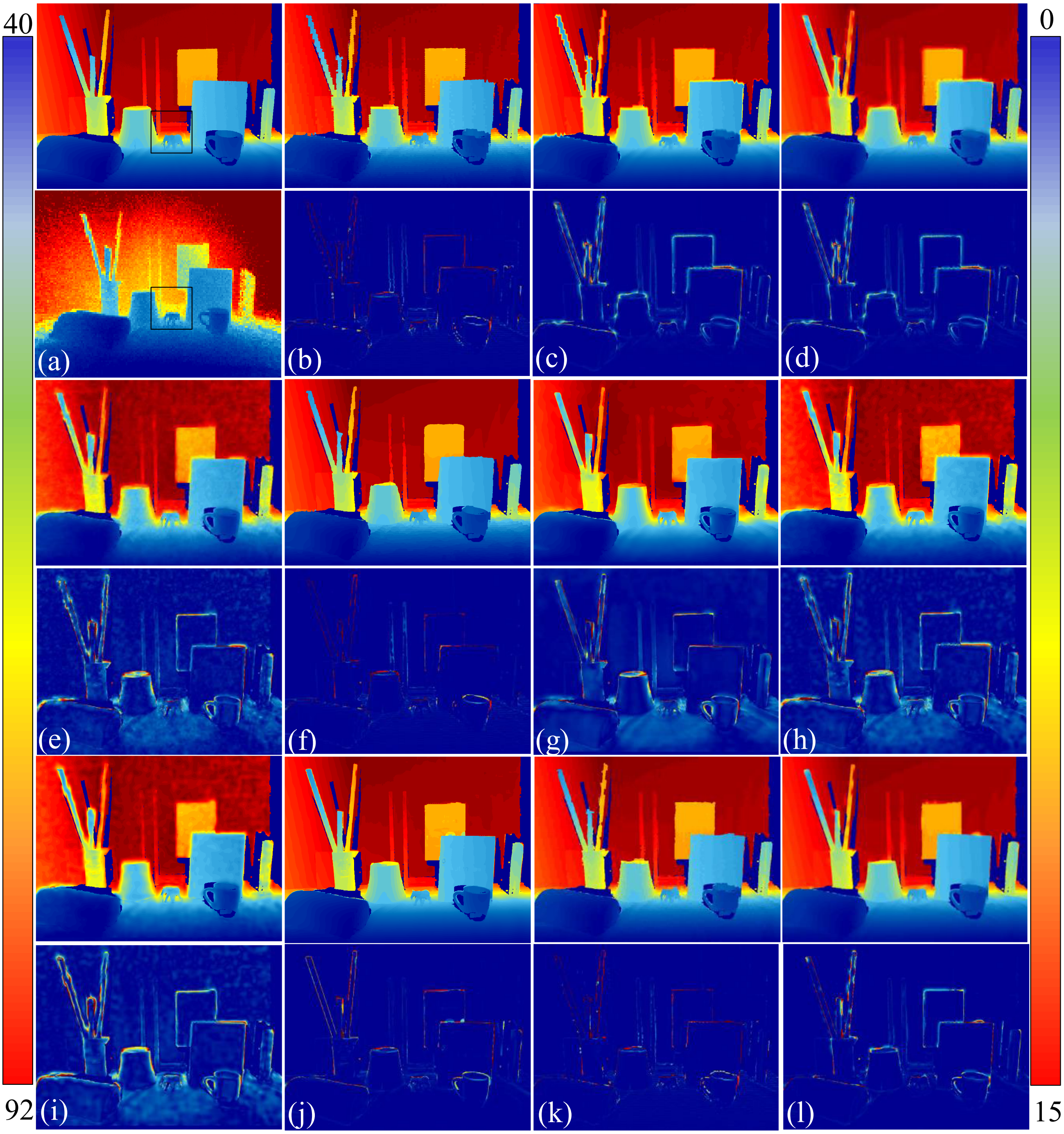


**FIGURE 13.** Results for various View-plus-depth fusion techniques for *"Book"* data set (visuals and zoomed details for the black window are given in Figure 14): (a) Ground-truth with the low-resolution noisy input (bottom, scaled to fit); fusion results for (up – resulted depth maps, bottom - map of residuals): (b) *"Voronoi* (*NN*)*"*, (c) *"Bilinear"*, (d) *"GF"* [41], (e) *"AD"* [28], (f) *"Hyp"* [37], (g) *"Yang"* [38], (h) *"CLMF"* [31], (i) *"IMLS"* [30], (j) *"TGV"* [42], (k) *"Proposed* (*SP*)*"*, (l) *"Proposed* (3 *iter.*)*"*; color map indices are shown for depth (left) and residuals(right), in centimeters.

been compared to the performance of à number of state-of-art 3D fusion methods "*BF*" [36], "*AD*" [28], "*GF*" [41], "*Hyp*" [37], "*Yang*" [38], "*JGF*" [29], "*IMLS*" [30], "*CLMF*" [31], "*TGV*" [42], and "*Yang*" [38]. The code scripts for all the referenced methods have been obtained online and run for the tuned or default settings, when the authors of the particular approach provide code scripts for the evaluation of same benchmarking test. The calculated MAE and RMSE are given in Table 1; visual outputs of some of the methods and scenes are given in Figure 13; along with depiction of the absolute difference maps (maps of residuals) with respect to GT data. The visual outputs for zoomed region

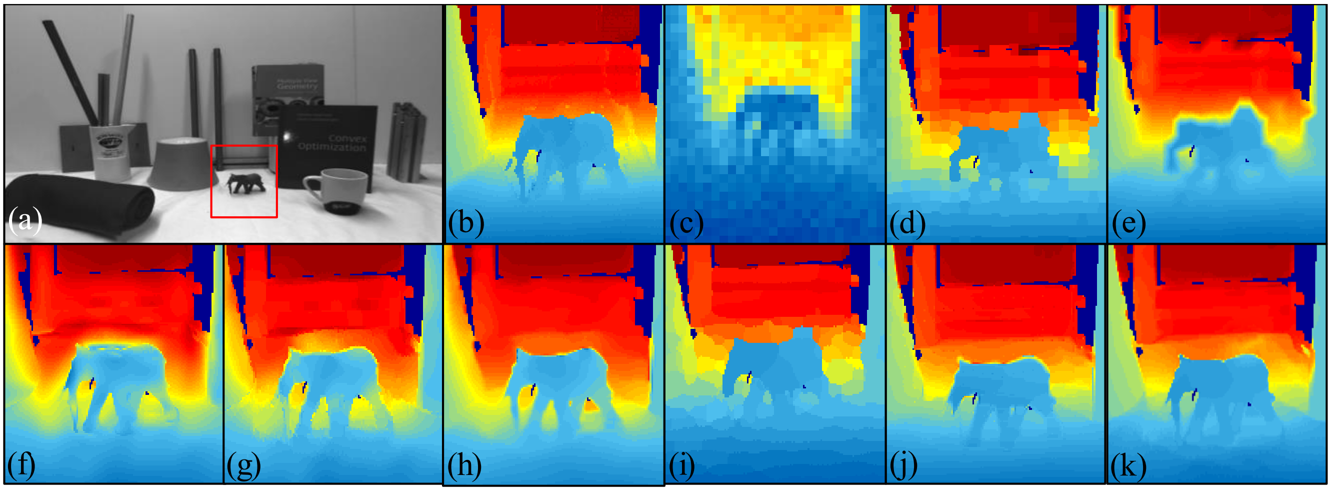


**FIGURE 14. Results for the zoomed detail in "*Books*" data set (c.f. Figure 13): (a) visible modality, (b) ground-truth reference, (c) noisy low-resolution input, (d) "*Voronoi* (*NN*)", (e) "*Bilinear*", (f) "*IMLS*" [30], (g) "*CLMF*" [31], (h) "*Yang*" [38], (i) "*Hyp*" [37], (j) "*TGV*" [42], (k) "*Proposed* (*3 iter.*)".**

(shown with black edge in Figures 13 (a) and 14 (a)) of a miniature elephant sculpture is provided in Figure 14). The proposed framework has been tested for three cases: "*Proposed* (*SP*)" with no iterative refinement applied, and when iterative refinement has been applied for $i = 3$ iterations ("*Proposed* 3 *iter.*"). The results can be analyzed as follows: the proposed framework in its basic form provides a balanced output in terms of error metrics, when compared to similarly performing methods e.g. "*Hyp*", "*TGV*" and "*GF*", where "*TGV*" has the most competitive results. However, "*TGV*" is slow and took about 10 minutes on our computing platform, while our proposed technique offers real-time performance. Basic interpolation methods involving no cross-modality filtering e.g. "*Voronoi* (*NN*)" and "*Bilinear*" perform surprisingly well in some cases (c.f. Table 1), which can be explained by the imperfectly aligned data modalities for GT data. Cross-modality filtering methods aim at finding edge congruency between V and Z modalities, and any initial misalignment leads to high error (c.f. Figure 13 (f-l)), which is not manifested in the direct resampling methods (c.f. Figure 13 (b, c)). However, visual appearance of the latter is not good in overall (c.f. Figure 14 (d, e)).

## VI. CONCLUSIONS

The presented work improved and streamlined our previous depth compression method [45] to a more general aspect of treating View-plus-depth data. Specifically, we relate the depth representation with the underlined color modality in terms of super-pixels. To this end, we have proposed a novel hierarchical super-pixel segmentation which keeps the boundary congruency of successive layers. In this way, the segmentation structure is very suitable for depth modelling in terms of constant depth segments, and its subsequent down-sampling for effective encoding or for its color-adaptive reconstruction. More specifically, the SPs allow for embedding also the down-sampled depth isotropic maps and thus achieving better performance of the encoding scheme. The reconstruction filter, which leads to smoothed and well-aligned depth maps, has been made adaptive to the size of the refining SPs. We have added a boundary correction in terms of the proposed edge outlier encoding protocol. Apart from effectively avoiding code redundancies related to misaligned V+Z data, such boundary correction provides a suitable alignment of the two modalities, which is important for rendering virtual views.

The proposed encoding technique is highly competitive in the very low bit rate region. The general framework is also suitable for fusing non-confocal sensor data with asymmetric spatial resolution. It is easily tunable for other image processing tasks such as segmentation and multi-sensor data sparsification.

## REFERENCES

[1] K. Müller, P. Merkle, and T. Wiegand, "3-D video representation using depth maps," *Proc. IEEE*, vol. 99, no. 4, pp. 643–656, Apr. 2011.

[2] P. Ndjiki-Nya, M. Koppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand, "Depth image-based rendering with advanced texture synthesis for 3-D video," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 453–465, Jun. 2011.

[3] X. Yang, J. Liu, J. Sun, X. Li, W. Liu, and Y. Gao, "DIBR based view synthesis for free-viewpoint television," in *Proc. 3DTV Conf., True Vis.-Capture, Transmiss. Display 3D Video (3DTV-CON)*, Antalya, Turkey, May 2011, pp. 1–4.

[4] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," in *Proc. ACM SIGGRAPH Papers (SIGGRAPH)*, Los Angeles, CA, USA, 2005, pp. 765–776.

[5] J. Hol, T. Schon, F. Gustafsson, and P. Slycke, "Sensor fusion for augmented reality," in *Proc. 9th Int. Conf. Inf. Fusion*, Florence, Italy, Jul. 2006, pp. 1–6.

[6] A. Kolb, E. Barth, R. Koch, and R. Larsen, "Time-of-Flight cameras in computer graphics," *Comput. Graph. Forum*, vol. 29, no. 1, pp. 141–159, Mar. 2010.

[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. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.

[9] PMD [Vision] CamCube 2.0, Time-of-flight camera. *PMDTechchnologies Gmbh*. Accessed: 2014. [Online]. Available: www.pmdtec.com/news_media/video/camcube.php

[10] A. Smolic, "3D video and free viewpoint video—From capture to display," *Pattern Recognit.*, vol. 44, no. 9, pp. 1958–1968, Sep. 2011.

[11] A. Chuchvara, M. Georgiev, and A. Gotchev, "A speed-optimized RGB-Z capture system with improved denoising capabilities," *Proc. SPIE*, vol. 9019, pp. 1533–1537, Feb. 2014.

[12] Y. Morvan, P. With, and D. Farin, "Platelet-based coding of depth maps for the transmission of multiview images," *Proc. SPIE*, vol. 6055, Jan. 2006, Art. no. 60550K.

[13] N. Ponomarenko, V. Lukin, A. Gotchev, and K. Egiazarian, "Intra-frame depth image compression based on anisotropic partition scheme and plane approximation," in *Proc. 2nd Int. ICST Conf. Immersive Telecommun.*, Berkeley, CA, USA, 2009, pp. 1–6.

[14] R. Mathew, P. Zanuttigh, and D. Taubman, "Highly scalable coding of depth maps with arc breakpoints," in *Proc. Data Compress. Conf.*, Snowbird, UT, USA, Apr. 2012, pp. 42–51.

[15] R. Mathew, D. Taubman, and P. Zanuttigh, "Scalable coding of depth maps with R-D optimized embedding," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1982–1995, May 2013.

[16] I. Schiopu and I. Tabus, "MDL segmentation and lossless compression of depth images," in *Proc. Workshop Inf. Theoretic Methods Sci. Eng. (WITMSE)*, Helsinki, Finland, Aug. 2011.

[17] I. Schiopu and I. Tabus, "Lossy depth image compression using greedy rate-distortion slope optimization," *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1066–1069, Nov. 2013.

[18] S. Milani, P. Zanuttigh, M. Zamarin, and S. Forchhammer, "Efficient depth map compression exploiting segmented color data," in *Proc. IEEE Int. Conf. Multimedia Expo*, Barcelona, Spain, Jul. 2011, pp. 1–6.

[19] S. Hoffmann, M. Mainberger, J. Weickert, and M. Puhl, "Compression of depth maps with segment-based homogeneous diffusion," in *Scale Space and Variational Methods in Computer Vision (SSVM)* (Lecture Notes in Computer Science), vol. 7893, A. Kuijper, K. Bredies, T. Pock, and H. Bischof, Eds. Berlin, Germany: Springer, May 2013, pp. 319–330.

[20] I. Tosic and S. Drewes, "Learning joint intensity-depth sparse representations," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2122–2132, May 2014.

[21] *Lossless and Near-Lossless Coding of Continuous Tone Still Images*, ITU-T Recommendations, document FCD-14495, (JPEG-LS) ISO/IEC JTC1/SC29 WG1 (JPEG/JBIG), 1997.

[22] *JPEG 2000 Image Coding System: Core Coding System*, ITU-T Recommendations, document 15444-1:2004, ISO/IEC, (JPEG2000) JTC 1/SC, vol. 29, 2004.

[23] *Advanced Video Coding for Generic Audiovisual Services*, ITU-T Recommendations, document 14496-10, ISO/IEC, (H.26L), 2003.

[24] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2005, pp. 291–298.

[25] W. Hannemann, A. Linarth, B. Liu, G. Kokai, and O. Jesorsky, "Increasing depth lateral resolution based on sensor fusion," *Int. J. Intell. Syst. Technol. Appl.*, vol. 5, nos. 3–4, pp. 393–401, 2008.

[26] B. Huhle, S. Fleck, and A. Schilling, "Integrating 3D time-of-flight camera data and high resolution images for 3DTV applications," in *Proc. 3DTV Conf.*, Kos Island, Greece, May 2007, pp. 1–4.

[27] J. Lu, D. Min, R. S. Pahwa, and M. N. Do, "A revisit to MRF-based depth map super-resolution and enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Prague, CR, USA, May 2011, pp. 985–988.

[28] J. Liu and X. Gong, "Guided depth enhancement via anisotropic diffusion," in *Advances in Multimedia Information Processing (PCM)* (Lecture Notes in Computer Science), vol. 8294, B. Huet, C. W. Ngo, J. Tang, Z. H. Zhou, A. G. Hauptmann, and S. Yan, Eds. Cham, Switzerland: Springer, 2013, pp. 408–417.

[29] M.-Y. Liu, O. Tuzel, and Y. Taguchi, "Joint geodesic upsampling of depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 169–176.

[30] N. K. Bose and N. A. Ahuja, "Superresolution and noise filtering using moving least squares," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2239–2248, Aug. 2006.

[31] J. Lu, K. Shi, D. Min, L. Lin, and M. Do, "Cross-based local multipoint filtering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2012, pp. 430–437.

[32] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vis.*, Bombay, India, Jan. 1998, pp. 839–847.

[33] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 60–65.

[34] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[35] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[36] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noise-aware filter for real-time depth upsampling," in *Proc. ECCV Workshop Multi-Camera Multi-Modal Sensor Fusion Algorithms Appl.*, Marseille, France, Oct. 2008, pp. 1–13.

[37] S. Smirnov, A. Gotchev, and K. Egiazarian, "Methods for depth-map filtering in view-plus-depth 3D video representation," *EURASIP J. Adv. Signal Process.*, vol. 25, no. 2, . 25, Feb. 2012.

[38] J. Yang, X. Ye, K. Li, C. Hou, and Y. Wang, "Color-guided depth recovery from RGB-D data using an adaptive autoregressive model," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3443–3458, Aug. 2014.

[39] B. Huhle, T. Schairer, P. Jenke, and W. Straßer, "Fusion of range and color images for denoising and resolution enhancement with a non-local filter," *Comput. Vis. Image Understand.*, vol. 114, no. 12, pp. 1336–1345, Dec. 2010.

[40] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 1623–1630.

[41] 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.

[42] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sidney, NSW, Australia, Dec. 2013, pp. 993–1000.

[43] M. Georgiev, A. Gotchev, and M. Hannuksela, "Joint de-noising and fusion of 2D video and depth map sequences sensed by low-powered tof range sensor," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, San Jose, CA, USA, Jul. 2013, pp. 1–4.

[44] M. Georgiev and A. Gotchev, "On the asymmetric view+depth 3D scene representation," in *Proc. Int. Workshop Video Process. Qual. Metrics Consumer Electron. (VPQM)*, Phoenix, AZ, USA, 2015, pp. 1–7.

[45] M. Georgiev, E. Belyaev, and A. Gotchev, "Depth map compression using color-driven isotropic segmentation and regularised reconstruction," in *Proc. Data Compress. Conf.*, Snowbird, UT, USA, Apr. 2015, pp. 153–162.

[46] M. Georgiev and A. Gotchev, "Improved depth compression by depth downsampling guided by color super-pixel refinement segmentation," in *Proc. Data Compress. Conf.*, Snowbird, UT, USA, Mar. 2018, p. 409.

[47] R. Hunter, "Photoelectric color difference meter," *J. Opt. Soc. Amer.*, vol. 38, no. 7, pp. 985–995, 1948.

[48] R. Hartley and A. Zissermann, *Multiple View Geometry in Computer Vision*. 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[49] Ren and Malik, "Learning a classification model for segmentation," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Nice, France, Apr. 2003, pp. 10–17.

[50] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to State-of-the-Art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[51] M. Van den Bergh, X. Boix, G. Roig, and L. Van Gool, "SEEDS: Superpixels extracted via energy-driven sampling," *Int. J. Comput. Vis.*, vol. 111, no. 3, pp. 298–314, Feb. 2015.

[52] E. Belyaev, K. Liu, M. Gabbouj, and Y. Li, "An efficient adaptive binary range coder and its VLSI architecture," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1435–1446, Aug. 2015.

[53] E. Belyaev, A. Veselov, A. Turlikov, and K. Liu, "Complexity analysis of adaptive binary arithmetic coding software implementations," in *Smart Spaces and Next Generation Wired/Wireless Networking. ruSMART (NEW2AN 2011)* (Lecture Notes in Computer Science), vol. 6869, S. Balandin, Y. Koucheryavy, and H. Hu, Eds. Berlin, Germany: Springer, 2011.

[54] J. Kannala and S. S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1335–1340, Aug. 2006.

[55] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing. I. theory," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 821–833, Feb. 1993.

[56] G. Zech, "Iterative unfolding with the Richardson–Lucy algorithm," *Nucl. Instrum. Methods Phys. Res. A, Accel., Spectrometers, Detectors Associated Equip.*, vol. 716, no. 1, pp. 1–9, 2013.

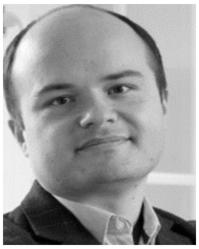
[57] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Madison, WI, USA, Jun. 2003, pp. 195–203.

[58] G. Schuster and A. Katsagellos, *Rate-Distortion Based Video Compression*. New York, NY, USA: Springer, 1997.

[59] T. Senoh, J. Kenji, T. Nobuji, Y. Hiroshi, and K. Wegner, *View Synthesis Reference Software (VSRS) 4.2 With Improved Inpainting and Hole Filling*, document ISO/IEC JTC1/SC29/WG11, n. MPEG2017/M40657, 2017.

[60] E.-C. Forster, T. Lowe, S. Wenger, and M. Magnor, "RGB-guided depth map compression via compressed sensing and sparse coding," in *Proc. Picture Coding Symp. (PCS)*, Cairns, QLD, Australia, May 2015.

**MIHAIL GEORGIEV** (Member, IEEE) received the Dipl.Ing. and M.Sc. degrees in information technologies from the Faculty of Electronics, Technical University of Varna, Varna, Bulgaria, in 2002 and 2003, respectively, and the Dr.Sc. (Tech.) degree from the Tampere University of Technology, Finland, in 2018. His research interests include everything related to 3D capturing: computational geometry, active sensing devices, non-confocal multi-sensor setups data acquisition, denoising, calibration, fusion, rendering, occlusion in-painting, and depth compression.



**EVGENY BELYAEV** received the M.S. (Engineer) degree in automated systems of information processing and control, in 2005, the Ph.D. degree in technical sciences from the State University of Aerospace Instrumentation (SUAI), Saint-Petersburg, Russia, in 2009, and the Dr.Sc. (Tech.) degree from the Tampere University of Technology, Finland, in 2015. He is currently working as a Research Fellow with ITMO University, Saint-Petersburg. His research interests include low-complexity joint source-channel video coding, arithmetic coding, and compressive sensing.

**ATANAS GOTCHEV** (Member, IEEE) received the M.Sc. degrees in radio and television engineering, in 1990, and in applied mathematics, in 1992, the Ph.D. degree in telecommunications from the Technical University of Sofia, in 1996, and the D.Sc. (Tech.) degree in information technologies from the Tampere University of Technology, in 2003. He is currently a Professor with the Laboratory of Signal Processing and the Director of the Centre for Immersive Visual Technologies at the Tampere University of Technology. His research interests consist of sampling and interpolation theory as well as spline and spectral methods with applications for multidimensional signal analysis. His recent work concentrates on the algorithms for multi-sensor 3-D scene capture, transform-domain light-field reconstruction, and Fourier analysis of 3-D displays.