Joint-Feature Guided Depth Map Super-Resolution With Face Priors

Shuai Yang, Jiaying Liu, *Member, IEEE*, Yuming Fang, *Member, IEEE*, and Zongming Guo, *Member, IEEE*

***Abstract*—In this paper, we present a novel method to super-resolve and recover the facial depth map nicely. The key idea is to exploit the exemplar-based method to obtain the reliable face priors from high-quality facial depth map to improve the depth image. Specifically, a new neighbor embedding (NE) framework is designed for face prior learning and depth map reconstruction. First, face components are decomposed to form specialized dictionaries and then reconstructed, respectively. Joint features, i.e., low-level depth, intensity cues and high-level position cues, are put forward for robust patch similarity measurement. The NE results are used to obtain the face priors of facial structures and smooth maps, which are then combined in an uniform optimization framework to recover high-quality facial depth maps. Finally, an edge enhancement process is implemented to estimate the final high resolution depth map. Experimental results demonstrate the superiority of our method compared to state-of-the-art depth map super-resolution techniques on both synthetic data and real-world data from Kinect.**

***Index Terms*—Depth enhancement, depth map, face priors, Kinect, neighbor embedding (NE), super-resolution (SR).**

I. Introduction

IN RECENT years, with the development of consumer-level depth cameras such as time-of-flight and Microsoft Kinect, depth images gain increasing popularity and have been extensively studied. Since the depth map is more robust to the environment and can provide spatial information, it has been widely used in cybernetics applications. Many works concerning depth images have been carried out, including scene classification [1], 3-D modeling [2], object detection [3], human activity detection [4] and posture reconstruction [5], as well as human face recognition [6]. More depth-sensor-based computer vision applications are comprehensively reviewed in [7]. Unfortunately, raw depth maps obtained by the sensor are often far from the satisfactory. The low resolution (LR) and the degradation of sampled depth maps become a critical issue and limit the applications of depth information significantly.

Manuscript received June 16, 2016; revised October 18, 2016; accepted December 4, 2016. Date of publication December 22, 2016; date of current version November 15, 2017. This work was supported by the National Natural Science Foundation of China under Contract 61472011 and Contract 61571212. This paper was recommended by Associate Editor L. Shao. *(Corresponding author: Jiaying Liu.)*

The authors are with the Institute of Computer Science and Technology, Peking University, Beijing 100871, China (e-mail: williamyang@pku.edu.cn; liujiaying@pku.edu.cn; fa0001ng@e.ntu.edu.sg; guozongming@pku.edu.cn).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. The total size of the file is 45.5 MB.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2638856

To facilitate the use of depth data, many researchers have focused on the research of depth super-resolution (SR), by which a high resolution (HR) depth map is reconstructed from an LR input. The methods for general depth map SR can be divided into three categories: 1) multiple depth map fusion; 2) image-guided depth map SR; and 3) single depth map SR. Multiple depth map fusion techniques [8]–[10] merge several unaligned low-quality depth maps of one same scene to reconstruct a high-quality depth map. However, multiple depth maps may not be available, while a corresponding high-quality color image is usually available and can help the SR process. In this kind of image-guided depth map SR [11]–[13], high frequency components in color images are used as the guidance to improve the depth map recovery. Recently, depth map SR using stereo-vision-assisted model has also been developed. Given a pair of color images and a low-resolution depth map as inputs, this model uses the stereo matching theory to further improve the depth map SR process [14]–[16] and even to accomplish a tougher depth estimation task [17]. Single depth map SR [18]–[20] recovers depth information with a single low-quality input. This method often takes exemplar-based strategy to learn from extra data, which makes up for the lack of multiple frames or color images. Most existing studies above focus on the general depth map SR. Much less has been done to use face prior information to improve facial depth map SR.

In this paper, we propose an exemplar-based approach to deal with image-guided facial depth map SR. For facial depth reconstruction, examplar-based strategy is superior because human faces share regular patterns that can be learned as priors to help reconstruction. A new neighbor embedding (NE) SR framework is designed, in which external examples can be directly used as dictionaries to form high-quality facial priors. We present adaptations of NE for general depth map SR, which focuses on the depth boundary recovery. Moreover, our method exploits face priors accurately by considering both low-level and high-level cues of depth, intensity and position of facial components. By integrating learned facial structure and smoothness priors into an uniform optimization framework, our method achieves high-quality results from a severely degraded LR depth map and its corresponding HR color image. To the best of our knowledge, this is the first work that incorporates these two important facial priors into a unified NE optimization framework.

The performance of our method is evaluated with state-of-the-art depth map SR methods. Experimental results demonstrate the superior of our method in both synthetic facial depth

maps and degraded Kinect-captured face data. In summary, the main contributions of this paper are as follows.

1) *Edge-Aware NE for Depth Map SR:* We propose an edge-aware NE method in which high-quality edge map is learned based on both depth cues and color cues, and a constrained optimization function that matches the characteristics of depth maps is designed to obtain the sharp and clean edges.
2) *Joint Scale-Independent DIP (Depth, Intensity, and Position) Feature:* We propose scale-robust joint feature that combines high-level and low-level cues to effectively handle the ambiguity problem between HR/LR patches and between intensity/depth patches, which helps learn accurate facial priors.
3) *Face Prior Analysis and Utilization in NE Framework:* We propose a dual gradient regularization (DGR) optimization to impose learned facial structure and smoothness priors onto the raw depth map to recover distinct face structures and smooth out noise simultaneously. We show that face structures can be well recovered without over-smoothing by taking full use of face priors in this unified NE framework.

The rest of this paper is organized as follows. In Section II, we review related works in image-guided depth map SR, edge-aware SR, and NE. In Section III, the details of the proposed algorithm is presented. We validate our method by comparing it with state-of-the-art image-guided depth map SR algorithms on both synthetic data and real-world data from Kinect in Section IV. Finally, we conclude this paper in Section V.

II. Related Work

In this section, we discuss related approaches associated with image-guided depth map SR, edge-aware SR, and NE.

A. Image-Guided Depth Map Super-Resolution

For image-guided depth map SR, filter-based methods [21]–[23] are widely adopted by early works. They usually filter the depth map adaptively according to the depth structures and color intensities. Yang *et al.* [24] iteratively applied a bilateral filter to the cost volume in the stereo vision literature. Liu *et al.* [25] proposed to use geodesic distance to calculate the filter weight and recover sharper edges. Local-linear-model-based guided filter [11] shows superiority in computational efficiency and gradient preservation. But it tends to over-smooth regions where values in the guidance image are close to each other. Tan *et al.* [26] improved guided filter using a spatial adaptive scheme to prevent blurring edges. However, since the color of human faces lacks changes, the filter weights can be misled. Thus, it is not suitable to apply these methods on human faces.

In recent years, constrained optimization methods have been developed for image-guided depth map SR. In [27], an Markov random filed (MRF) formulation for depth map SR is introduced. Yang *et al.* [28] formulated the depth refinement as a minimization of auto-regressive prediction errors. In [29], guided by the structures in the color image, the depth map is upsampled using a weighted least squares optimization. Ferstl *et al.* [12] regarded the depth image recovery problem as a global energy optimization problem using total generalized variation (TGV) regularization. For these approaches, the performance depends on good image priors as the regularization term to constrain the optimization. When applying these methods to facial depth maps, defining a universal face prior artificially is a tough task. Intuitively, we can obtain face priors through learning the external dataset.

Exemplar-based methods [13], [30], or learning-based methods for depth map enhancement attempt to model statistical dependencies between intensity and depth through proper dictionaries. Li *et al.* [31] trained a joint dictionary consisting of both the gradient of the depth map and the edges of the color image. However, this method does not consider the discontinuities between color textures and depth edges. To tackle this problem, Kwon *et al.* [32] proposed the normalized absolute gradient dot product to predict the coincidence between depth edges and intensity edges and Kiechle *et al.* [13] proposed to directly learn the dependencies between depth and intensity using co-sparse analysis model, resulting in promising performance. Different from these related studies, our method goes further by considering high-level face position prior for accurate nearest neighbor searching.

B. Edge-Aware Super-Resolution

Edge preserving is an important topic in SR. Tai *et al.* [33] proposed to use point spread function guided by curveness map to reconstruct HR color images. Meanwhile, the NE method in [34] preserves edges by classifying image patches into two types (edge/non-edge), and searching similar patches in the training patches that are of the same type, but with no additional edge enhancement process. Li *et al.* [35] proposed to preserve edges by combining low-frequent images and the high-frequent gradient images learned using neighborhood regression. By comparison, our method learns edge priors based on both depth cues and color cues, and uses TV regularization weighted by learned edge priors to preserve edges, which matches the characteristics of depth maps to obtain sharper and cleaner edges.

C. Neighbor Embedding

NE assumes that image patches in LR and HR images form manifolds with the similar local geometry and neighborhood relationships. Chang *et al.* [36] introduced locally linear embedding for image SR. Local geometry features were characterized by linearly representing a feature vector with its similar patches in the feature space. The HR patches were reconstructed as weighted averages of neighbors in the HR space, using the same coefficients estimated in the LR space. The recent anchored neighborhood regression (ANR) [37] proposed a joint NE with sparse learned dictionaries to anchor the neighborhood embedding of a given LR patch to the nearest atoms in the dictionary. In addition, the ANR approach adopted the ridge regression and precomputed the corresponding embedding projection matrices, which reduced algorithm complexity.

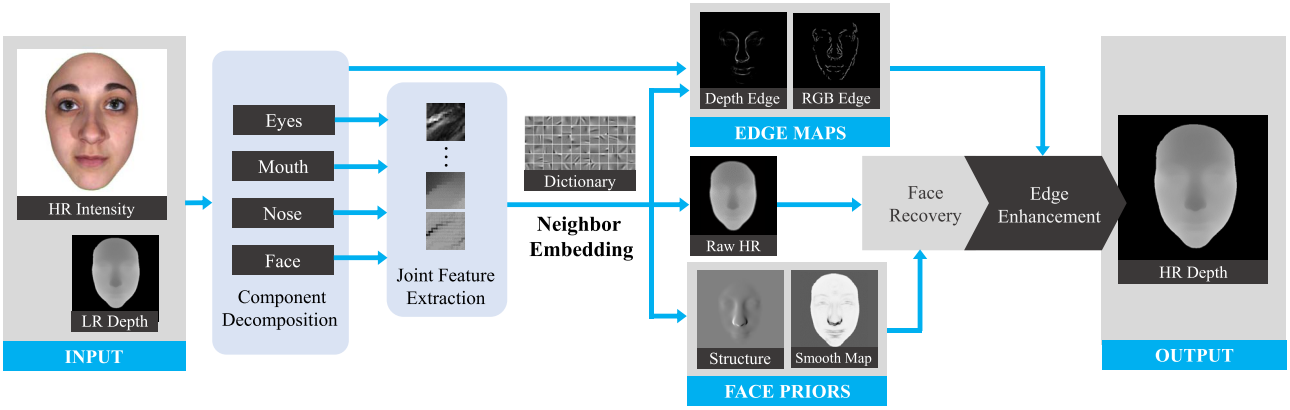


Fig. 1. Framework of the proposed method. We first decompose a whole face into facial components based on the HR color image and reconstruct them, respectively. Joint features, including low-level depth, intensity cues, and high-level position cues are extracted to represent each patch for robust nearest neighbor searching. The face priors of facial structures and smooth map estimated by these nearest neighbors are used to recover the facial depth map. In addition, our method further enhances depth boundaries and makes them clean and sharp using learned edge maps.

Many researches [38]–[43] have adopted NE framework to learn high-quality face information for face hallucination. The pioneering work in [38] assumes that similar textures are shared by the patches located at the same position of face images and proposed a position-based NE framework. The main problem is that it requires face images in the training set and testing set to be well aligned, or the inaccurate position prior will degrade the reconstruction results. The subsequent works incorporate face position prior based on [38] and mostly improve it by designing new sparsity priors [39]–[41] or locality priors [41]–[43]. Our method surpasses these methods in that: 1) we do not require the training set and testing set to be well or roughly aligned and 2) we consider facial structure prior and facial smoothness prior that are crucial for facial depth map reconstruction but ignored by these existing methods.

III. Proposed Method

Fig. 1 shows the framework of the proposed method. The main idea is to obtain face priors (structures and smooth maps) using the proposed NE framework to improve the reconstruction of facial depth map. First of all, we introduce the proposed edge-aware NE method for general depth map SR. The method uses learned edge maps to enhance depth boundaries and make them clean and sharp. We use this method as a baseline and further propose a joint-feature-based depth map recovery method with learned face prior to address the SR problem for faces. Specifically, we first decompose a whole face into facial components based on the HR color image and reconstruct them, respectively. Then, we propose joint scale-independent DIP features to measure the patch similarity for robust nearest neighbor searching. The face priors estimated by these nearest neighbors are finally used to recover the facial depth map.

A. Edge-Aware Neighbor Embedding for Depth Map Super-Resolution

This section presents the proposed edge-aware NE method for general depth map SR. The conventional NE [36] restores images using coupled dictionaries. We use X and Y to represent LR and HR depth maps, and x and y to represent LR and HR patches. In addition, we use I to denote the HR color image and each LR depth patch x is accompanied with an HR color patch z. Given a target LR depth map X_t as input, we estimate the target HR depth map Y_t with the help of the coupled dictionaries $\mathcal{D} = \{\mathcal{D}_x, \mathcal{D}_y\} = \{x_s^i, y_s^i\}_{i=1}^N$, where x_s^i/y_s^i are paired LR/HR patches from external source depth maps and N represents the dictionary size.

Our framework follows standard procedures in NE SR. The main procedures can be summarized as follows.

1) For each patch x_t^i in X_t, find the set $\mathcal{N}_x^i$ of K nearest neighbors in $\mathcal{D}_x$ and use the corresponding HR neighbors $\mathcal{N}_y^i$ in $\mathcal{D}_y$ to reconstruct y_t^i.
2) Construct Y_t using overlapped patches y_t^i.

To find accurate nearest neighbors, we combine depth cues and intensity cues. The patch similarity is measured by

$$\text{dist}(x_t, x_s) = \|x_t - x_s\|_2^2 + w_c \left\| z_t^{\text{edge}} - z_s^{\text{edge}} \right\|_2^2 \tag{1}$$

where w_c is the weight to combine depth and intensity cues. z^{edge} depicts the edge information of I. It is obtained as the maximum gradient magnitude among the RGB channels. Based on (1), we use nearest neighbor searching (K-NN) to find the K neighbors $\mathcal{N}_x^i = [x_s^{i_1}, x_s^{i_2}, \ldots, x_s^{i_K}]$ of x_t^i, and the optimal reconstruction coefficients can be solved by

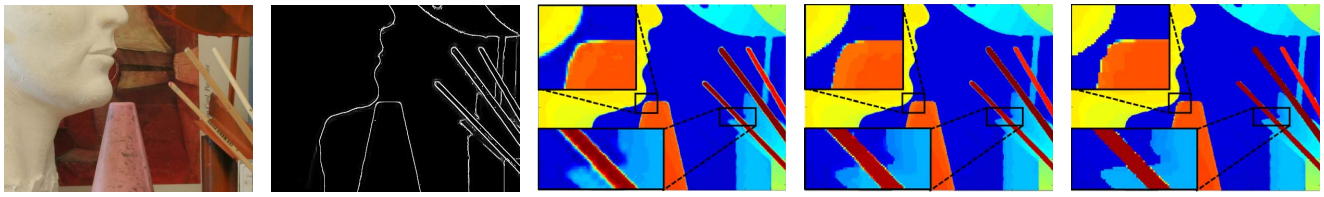
$$\arg\min_{\alpha^i} \left\| x_t^i - \mathcal{N}_x^i \alpha^i \right\|_2^2 + \mu \left\| \alpha^i \right\|_2^2 \tag{2}$$

where μ is the regularization term coefficient. The counterpart y_t^i is then obtained by applying the same coefficient to the corresponding HR neighbors $\mathcal{N}_y^i = [y_s^{i_1}, y_s^{i_2}, \ldots, y_s^{i_K}]$

$$y_t^i = \mathcal{N}_y^i \alpha^i. \tag{3}$$

Finally, we average the overlapped patches y_t^i to construct the raw target HR depth map Y_t^{raw}.

Edges are of particular importance in textureless depth map [20]. However, traditional SR methods often suffer edge



(a) (b) (c) (d) (e)

Fig. 2. Proposed edge-aware NE method produces clean and sharp edges. The RMSE between the results in (c) and (d) against (e) are 3.09 and 2.74, respectively. In this example, the raw depth map is upsampled by 4×. (a) Color. (b) E_t. (c) Y_t^{raw}. (d) $Y_t(\nu = 0.8)$. (e) Ground truth.

blurring problem. To solve this problem, we propose a TV-based edge enhancement method. The main idea is to use total variation (TV) regularization to smooth two sides of the edge while keeping large gradients on the edge. To accomplish this task, we first estimate HR edge map E_t of Y_t. Let $e(\cdot)$ denotes the edge map extraction operation and $e(\mathcal{N}_y^i) = [e(y_s^{i_1}), e(y_s^{i_2}), \ldots, e(y_s^{i_K})]$. The edge map of y_t^i can be estimated by $e_t^i = e(\mathcal{N}_y^i)\alpha^i$. Then, we average e_t^i to form the estimated edge map E_t^d. E_t is the combination of depth cues and intensity cues

$$E_t = \max\left(E_t^d, E_t^c\right) \tag{4}$$

where $E_t^c = e(I_t)$. Fig. 2(b) shows an example of E_t. Next, a TV map is derived to indicate where to smooth

$$M_{\text{TV}} = R\left(Y_t^{\text{raw}}\right) \otimes (\mathbf{1} - E_t). \tag{5}$$

In the above equation, $R(Y_t^{\text{raw}})$ is the local variance of Y_t^{raw}. Specifically, the local variance at pixel (i, j) is the variance of depth values in the patch $y_{i,j}^{\text{raw}}$ centered at pixel (i, j) in Y_t^{raw}

$$R\left(Y_t^{\text{raw}}\right)_{i,j} = \min\left(1, \text{var}\left(y_{i,j}^{\text{raw}}\right)\right) \tag{6}$$

which has high values near the edge. Meanwhile $(\mathbf{1} - E_t)$ has low values on the edge. $\otimes$ is the element-wise multiplication to combine $R(Y_t^{\text{raw}})$ and $(\mathbf{1} - E_t)$. Fig. 3(b) shows an example of M_{TV}. Finally, the TV regularization is performed on Y_t^{raw} (for simplicity, we denote Y_t^{raw} as d_0)

$$Y_t = \arg\min_d U_1(d, d_0) + \lambda_e V(d) \tag{7}$$

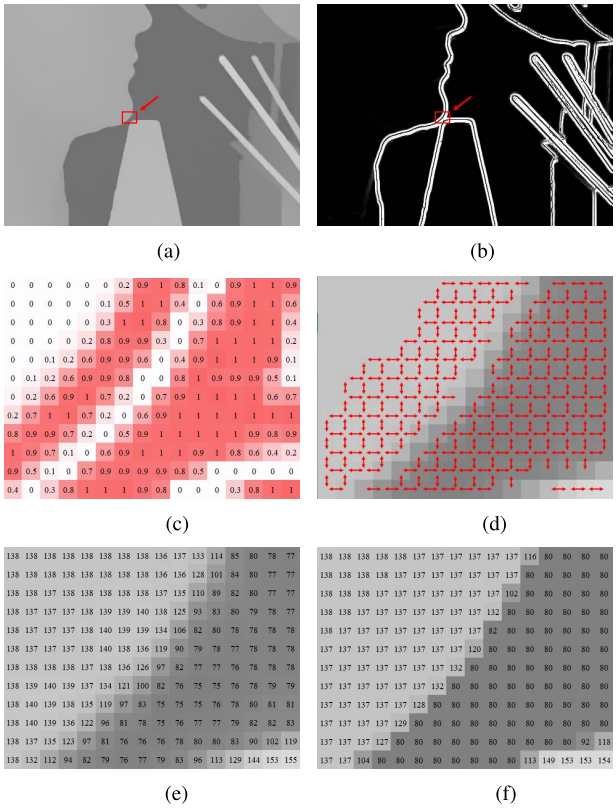
where

$$U_1(d, d_0) = \|M_c \otimes (d - d_0)\|_2^2 \tag{8}$$

$$V(d) = \|M_{\text{TV}} \otimes \nabla_x d\|_1 + \left\|M_{\text{TV}} \otimes \nabla_y d\right\|_1. \tag{9}$$

U_1 is the data term. The confidence map $M_c = \max(\nu, \mathbf{1} - E_t)$ is added based on the fact that pixel values on the edge are less reliable and ν controls the edge sharpness of the optimization results. V is the TV regularization term weighted by M_{TV}. λ_e is the weight to combine two terms. As shown in Fig. 3(d)–(f), under the guidance of M_{TV}, depth values at one side of the edge get uniformed and depth values at different sides of the edge are separated. The performance of our edge enhancement is illustrated by the comparison in Fig. 2(c) and (d).

In the following, we describe our main contributions on how we integrate face priors with this baseline method to improve facial depth map reconstruction quality.



(a) (b) (c) (d) (e) (f)

Fig. 3. Illustrations for the TV term to enhance the edges. The edges in raw super-resolved depth map. (a) Y_t^{raw} is refined under the guidance of the TV map. (b) M_{TV}. (c) Cropped zoomed region of (b). Pixels are colorized for better visualization. (d) Pixel with high M_{TV} weight is under strong TV constraints (represented as red arrows) to have similar depth with its adjacent pixels. After edge enhancement, (f) Y_t has shaper edges than (e) Y_t^{raw}.

B. Joint Scale-Independent Feature Representation for Patches

For image-guided depth map recovery problem, two issues must be addressed. The first is the scale problem for depth map. Color is an inherent changeless features for human face but the depth of a face will change along with its distance to the depth sensor. The second is the low correlation between color and depth. For example, one may find two similar skin-colored patches have very different depth values. To tackle these issues, in this section, we introduce the joint DIP feature to improve the nearest neighbor searching for facial depth map. For robustness, we extract scale-independent depth, intensity, and position features from the depth and color patches to represent the raw x_t.

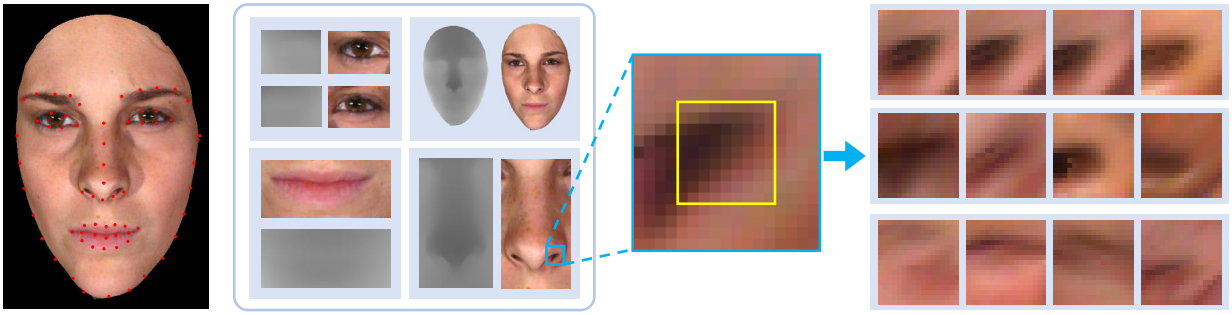


Fig. 4. Facial component decomposition and joint DIP feature to improve the nearest neighbor searching. (a) Face image with detected landmarks. (b) Facial component regions. (c) Target patch is shown in the yellow rectangle. To better recognize the content of an patch, expanded boundaries are added as shown in the blue rectangle. For space saving, the corresponding depth map is not shown. (d) Four most similar patches of (c) obtained using different methods. First row: component decomposition + DIP feature. Second row: DIP feature without component decomposition. Last row: component decomposition with only depth feature.

1) Facial Component Decomposition: We start with a face detection and landmark localization [44]. Each face is annotated by landmark points that locate facial components of interest. As shown in Fig. 4(b), we concentrate on the eye, nose, and mouth regions. These component regions together with the whole face region form four specialized dictionaries $\mathcal{D}^i, i \in \{1, 2, 3, 4\}$. And each region of the target image is reconstructed using the corresponding dictionaries. For simplicity, we use $\mathcal{D}$ to refer to the four dictionaries $\mathcal{D}^i$ in the following sections.

The benefits of the facial component decomposition are obvious.

1) After decomposition, each patch is implicitly classified. With the help of this high-level classification cue, the search spaces for similar patches are restricted and the ambiguity problem between LR/HR pairs is relieved, thus obtaining more accurate neighbors, as shown in the first and second rows of Fig. 4(d).
2) Each component is aligned implicitly after facial component decomposition. Therefore, the localization of a patch in its corresponding region is meaningful and can be used as an important cue for nearest neighbor searching. Thus, we propose the joint DIP features.

2) Joint Scale-Independent DIP Features: We represent patch x^i as its joint DIP features: $F(x^i) = \{x^i, c^i, p^i\}$, where x^i is the depth feature. c^i describes the intensity feature of the color image. Furthermore, p^i depicts the position feature.

1) *Depth Feature:* $x = [\nabla x; \nabla^2 x]$ consists of the first-order and second-order gradients of the depth map. The gradients consider relative depth changes independent of geometric scales. Physically, the first-order gradients depict relative elevations of face sense organs and the second-order gradients evaluate the smoothness of faces. Both are well-defined features for facial depth maps.
2) *Intensity Feature:* $c = [z^{\text{edge}}; z^{\text{norm}}]$ contains edge and intensity information. z^{norm} is the normalized intensity with zero means to cope with the scale problem. In addition to the edge information, we also consider to use the intensity information because human faces share certain color patterns that are correlated with facial depth. Moreover, the intensity information is less degraded and thus more reliable than the depth information.
3) *Position Feature:* $p = [u/W; v/H]$, where (u, v) are the coordinates of the patch center and $[W, H]$ are the width and height of the facial component region. As described in the last section, the structures (and therefore, the positions) of the eyes, nose, and mouth in their corresponding regions are relatively stationary, which means nearby positions share reliable similar patches.

Here, we briefly analyze the complementarity of the proposed features that jointly alleviate the ambiguity problem in nearest neighbor search. The two low-level cues of depth and intensity are complementary in that the former ensures fidelity of face structures while the latter provides high-quality cues to make up for the former's degradation. Meanwhile, position feature considers high-level face priors to help alleviate the ambiguity problem in depth and intensity features. And yet the abstract position feature itself has no information on depth and requires depth/intensity cues to locate similar depth patches.

Given the joint scale-independent DIP features, we formulate a measure for two patches

$$\text{dist}(x_t, x_s) = \|x_t - x_s\|_2^2 + w_c\|c_t - c_s\|_2^2 + w_p\|p_t - p_s\|_2^2 \quad (10)$$

where w_c and w_p are the weights to combine the depth, intensity, and position cues. The effect of the DIP feature for nearest neighbor searching is illustrated by the comparison in the first and last rows in Fig. 4(d). From this figure, we can see that by jointly considering high-level and low-level features, the ambiguity problem is effective resolved and we are capable to find reliable similar patches for learning.

C. Examplar-Based Face Prior Learning

To overcome the severe noise and quantization artifacts in the LR depth map, we learn two key face priors from the clean HR dictionary $\mathcal{D}_y$ to recovery the degraded facial depth map. The first prior is the facial structures, which depict relative elevations of face sense organs. In this paper, these relative elevations are measured by the first-order gradients. The second prior is the facial smooth map M_s, which

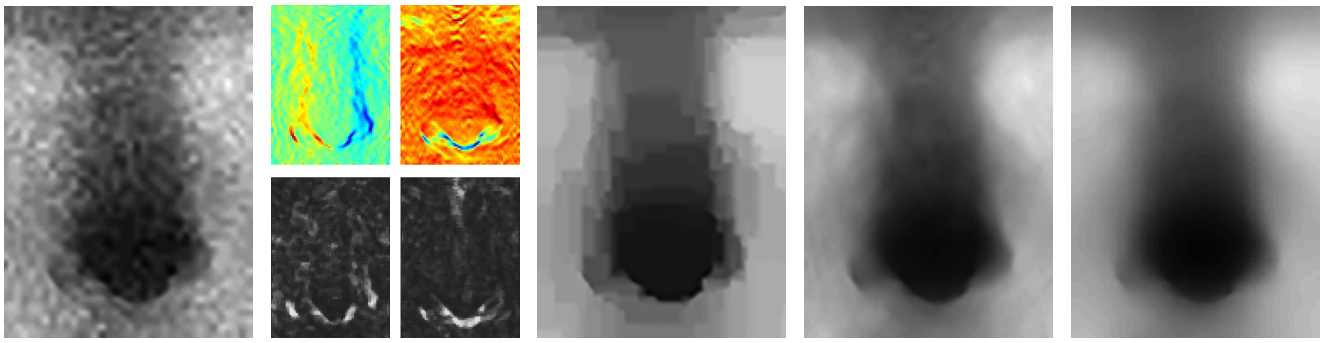


Fig. 5. Performance of the DGR. (a) Raw upsampled result Y_t^{raw}. (b) Learned face priors. First row is the facial structures ∇Y_t^{raw}. Second row is the facial smooth map M_s. Both are shown as the x component and y component. (c), (d), and (e) are reconstructed results Y_t using TV optimization, single gradient optimization and the proposed DGR optimization, respectively. The depth contrast is enhanced for better visualization.

indicates the smooth region of the face, for instance, the cheeks. The smoothness can be naturally evaluated by the second-order gradients. Given a target patch x_t^i, we therefore find its K nearest neighbors $\mathcal{N}_x^i$ based on (10) and obtain the corresponding HR nearest neighbors $\mathcal{N}_y^i$. Their first-order gradients $\nabla\mathcal{N}_y^i = [\nabla y_s^{i_1}, \nabla y_s^{i_2}, \ldots, \nabla y_s^{i_K}]$ and second-order gradients $\nabla^2\mathcal{N}_y^i = [\nabla^2 y_s^{i_1}, \nabla^2 y_s^{i_2}, \ldots, \nabla^2 y_s^{i_K}]$ are calculated. Then the optimal reconstruction coefficient α^i is calculated following (2) and is applied to $\mathcal{N}_y^i$, $e(\mathcal{N}_y^i)$, $\nabla\mathcal{N}_y^i$, and $\nabla^2\mathcal{N}_y^i$:

$$\left[y_t^i; e_t^i; \nabla y_t^i; \nabla^2 y_t^i\right] = \left[\mathcal{N}_y^i; e\left(\mathcal{N}_y^i\right); \nabla\mathcal{N}_y^i; \nabla^2\mathcal{N}_y^i\right]\alpha^i. \quad (11)$$

Next, these learned patch features are merged to the image space, resulting in the estimated raw HR depth map Y_t^{raw}, edge map E_t^d, facial structure prior ∇Y_t^{raw}, and facial smoothness prior $\nabla^2 Y_t^{\text{raw}}$, respectively. After that, $(\mathbf{1} - \nabla^2 Y_t^{\text{raw}})$ is applied to the median filter to remove noise artifacts and is normalized to finally form the facial smooth map M_s. Fig. 5(b) gives an example of learned facial structures and smooth map.

D. Advanced Depth Map Recovery With Face Priors

To recover distinct face structures while smoothing out noises, we propose a novel DGR term, which impose structure and smoothness constraints by two gradients. Then we reconstruct the facial depth map by solving the following optimization function (for simplicity, we denote Y_t^{raw} as d_0 and ∇Y_t^{raw} as g_0):

$$\arg\min_{d,g} U_2(d, g, d_0, g_0) + R(d, g_0) \quad (12)$$

where the auxiliary variable g is introduced as the refined gradient of Y_t and

$$U_2(d, g, d_0, g_0) = \lambda_d\|d - d_0\|_2^2 + \|g - g_0\|_2^2 \quad (13)$$

is the data term, with λ_d controlling the fidelity to respect the characteristics of the target face, and

$$R(d, g) = \|\nabla d - g\|_2^2 + \lambda_s\|M_s \otimes \nabla g\|_2^2 \quad (14)$$

is the proposed DGR term, with λ_s to control the smoothness. The first gradient term of DGR performs structural restoration. The second gradient term, weighted by M_s, enforces second-order smoothness over the learned facial smooth region.

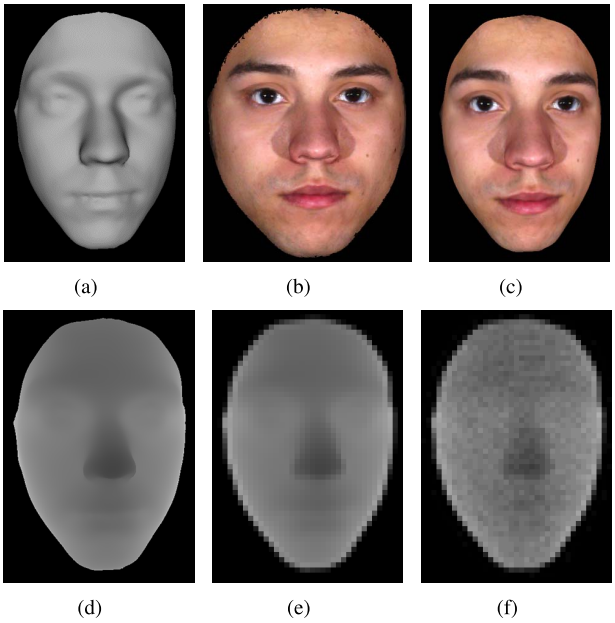


Fig. 6. Facial depth data construction. (a) and (b) 3-D face mesh and the corresponding texture from BU-3DFE dataset. (c) and (d) Color image and the synthetic facial depth map to form dictionaries. (e) Clean LR depth map for SR. (f) LR depth map with simulated degradations for SR.

Here, we discuss the superiority of the proposed DGR optimization. First, compared to the widely used TV term, which favors constant solutions and results in stepping artifacts, DGR term favors piecewise smooth gradients (second order smoothness), which accords with the normal physical structures of human faces. Second, the usage of the second gradient term brings another benefit of strong denoising ability. It is illustrated by the comparison in Fig. 5(d) and (e), which exhibit optimization results with and without $\|M_s \otimes \nabla g\|_2^2$. The proposed DGR term outperforms single gradient regularization (SGR) term in smoothing out noises.

E. Multi-Scale Solution

Since the ambiguity between depth/intensity pairs and LR/HR depth pairs will get severer when the scale difference gets greater, we take the multiscale strategy to address

TABLE I
UPSAMPLING OF CLEAN BU-3DFE DATA

Methods	Woman1			Woman2			Man1			Man2		
	2×	4×	8×	2×	4×	8×	2×	4×	8×	2×	4×	8×
Diebel *et al.* [27]	0.774	0.825	0.864	0.758	0.813	0.841	0.761	0.828	0.936	0.769	0.825	0.854
Yang *et al.* [24]	0.861	0.832	0.877	0.750	0.759	0.798	0.864	0.859	0.919	0.747	0.747	0.784
He *et al.* [11]	0.659	0.904	1.637	0.607	0.807	1.363	0.639	0.926	1.632	0.640	0.795	1.284
Kiechle *et al.* [13]	0.604	0.624	0.801	0.575	0.596	0.694	0.574	0.596	0.863	0.617	0.639	0.746
Ma *et al.* [45]	1.037	1.009	1.050	0.903	0.905	0.922	1.041	1.018	1.075	0.891	0.877	0.903
Ferstl *et al.* [12]	0.677	0.796	1.052	0.630	0.730	1.059	0.689	0.841	1.294	0.656	0.720	0.850
Ours	**0.541**	**0.596**	**0.627**	**0.508**	**0.567**	**0.588**	**0.511**	**0.569**	**0.618**	**0.544**	**0.608**	**0.632**

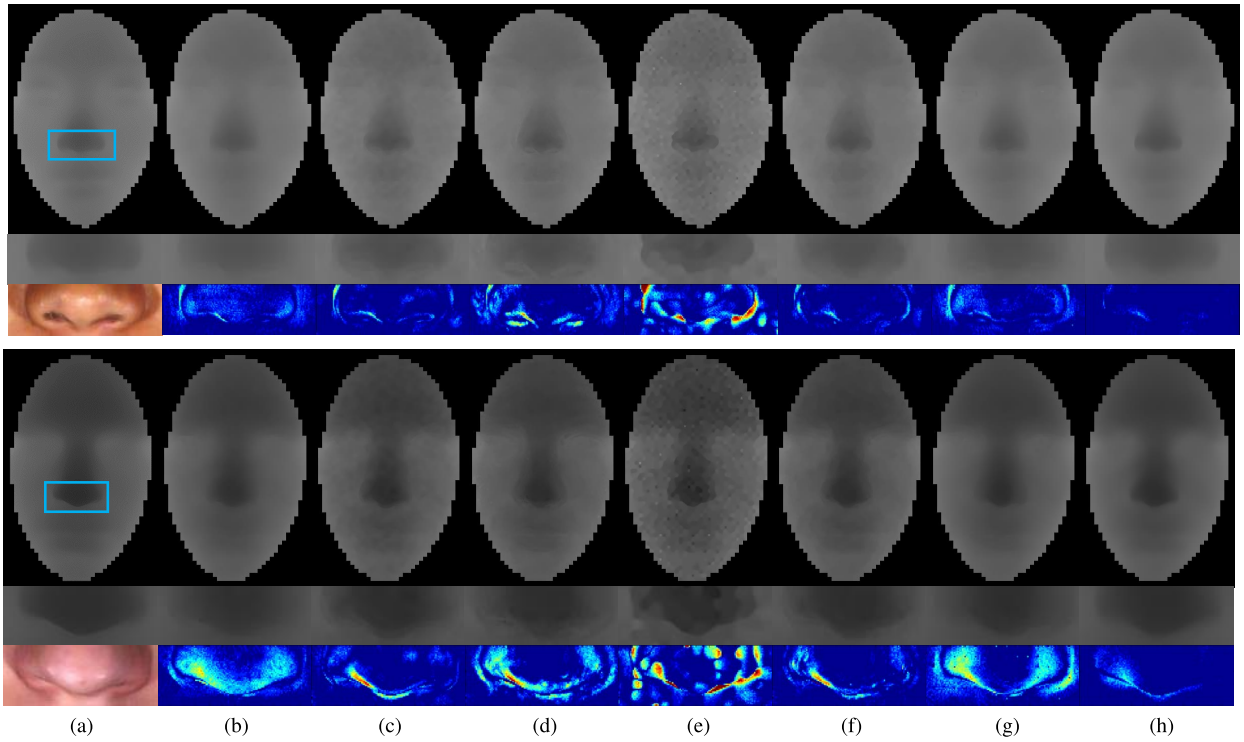


Fig. 7. Visual comparison with state-of-the-art methods for 8× upsampling on the degraded BU-3DFE dataset. Our method eliminates the noise while effectively restoring facial structures. And the error maps demonstrate that our reconstructed depth map is highly consistent with the ground truth. (a) Ground truth. (b) Diebel and Thrun [27]. (c) Yang *et al.* [24]. (d) He *et al.* [11]. (e) Kiechle *et al.* [13]. (f) Ma *et al.* [45]. (g) Ferstl *et al.* [12]. (h) Proposed method. For visual inspection, regions highlighted by blue rectangles are enlarged, and the error maps between the recovered depth map and ground truth are shown below the results. More examples can be found in the supplementary material.

this issue. Specifically, for the scale l, the dictionaries are obtained using the downsampled source HR color images and HR depth maps with factor $1/2^{l-1}$, and the reconstruction result at scale l forms the LR target depth map at scale $l-1$.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present a comparative performance evaluation of the proposed method. Specifically, we detail the experiment settings and illustrate the performance of our method compared to state-of-the-art techniques in both synthetic and Kinect-captured facial depth maps. In addition, the effect of parameters is discussed. Finally, an experiment on general depth map SR is also conducted to validate the effectiveness of the proposed edge-aware NE method.

A. Methodology

1) Parameter Settings: The proposed method is implemented on MATLAB R2014a platform. In the experiments, we use $n \times n$ HR patches with an overlap of one pixel between adjacent patches where $n = 7$. The dictionary size N is 100 000, and for each patch $K = 7$ nearest neighbors are searched. For the proposed joint scale-independent DIP features, the weights between different terms are set to $w_c = 9$ and $w_p = 2n^2 = 98$. For the neighborhood regression, the regularization term coefficient μ in (2) is 0.15. In depth map recovery using the DGR term, the smoothness factor λ_s is set to 2. Meanwhile, another factor λ_d is set to 4 for clean depth maps and 1 for degraded depth maps. Finally, in the edge enhancement process, the canny edge detector is used. The parameter ν to control the edge sharpness is set to 0.8 and the factor λ_e of the TV term is empirically chosen as 0.015, 0.005, and 0.001 for the SR of the target LR depth map at scale 1–3, respectively.

2) Facial Depth Data Construction for Learning and Testing: The BU-3DFE dateset [46] is used to construct dictionaries. Fig. 6(c) and (d) shows an example of the depth map and the corresponding color image of the face data.

TABLE II
UPSAMPLING OF DEGRADED BU-3DFE DATA

Methods	Woman1			Woman2			Man1			Man2		
	2×	4×	8×	2×	4×	8×	2×	4×	8×	2×	4×	8×
Diebel *et al.* [27]	1.152	1.239	2.002	1.062	1.220	2.070	1.068	1.229	1.936	1.031	1.181	1.640
Yang *et al.* [24]	1.056	1.310	1.916	0.922	1.216	1.713	1.017	1.278	1.746	0.887	1.185	1.645
He *et al.* [11]	1.180	1.373	1.981	1.081	1.264	1.718	1.108	1.348	1.941	1.057	1.237	1.588
Kiechle *et al.* [13]	1.633	3.056	4.120	1.319	2.806	3.884	1.454	2.785	3.808	1.339	2.766	3.718
Ma *et al.* [45]	1.245	1.385	1.771	1.062	1.232	1.564	1.185	1.323	1.657	1.037	1.198	1.518
Ferstl *et al.* [12]	1.234	1.324	1.899	1.030	1.204	1.770	1.136	1.302	1.843	1.107	1.173	1.512
baseline	2.552	2.590	2.675	2.549	2.552	2.459	2.397	2.460	2.420	2.436	2.485	2.385
baseline+DGR	0.914	1.112	1.568	0.810	1.016	1.338	0.876	1.116	1.396	0.836	1.043	1.374
baseline+DGR+DI	0.890	1.094	1.504	0.799	0.999	1.284	0.864	1.102	1.357	0.819	1.024	1.317
baseline+DGR+DP	0.882	1.063	1.473	0.790	0.973	1.229	0.840	1.048	1.325	0.820	1.011	1.259
baseline+TV+DIP	1.194	1.447	2.022	1.020	1.314	1.813	1.046	1.368	1.781	1.042	1.302	1.775
baseline+SGR+DIP	1.432	1.211	1.472	1.354	1.139	1.259	1.304	1.152	1.342	1.307	1.132	1.290
baseline+DGR+DIP	**0.869**	**1.039**	**1.421**	**0.782**	**0.949**	**1.187**	**0.827**	**1.023**	**1.295**	**0.812**	**0.988**	**1.211**

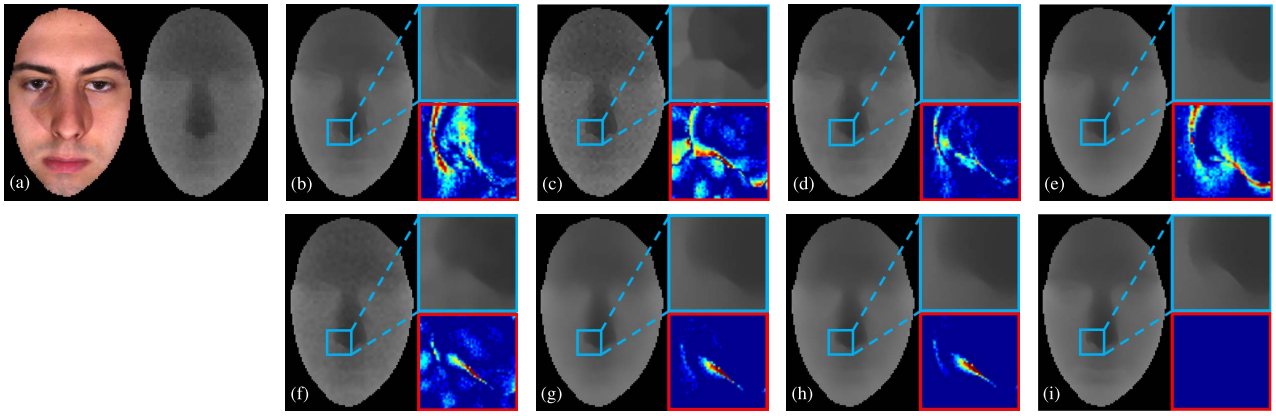


Fig. 8. Visual comparison of the proposed method with and without the key components of the DIP features and the DGR term for 8× upsampling on the degraded BU-3DFE dataset. These two key components help to obtain high-quality face priors and produce superior performance over other state-of-the-art methods. (a) Input color image and LR depth map. (b) He *et al.* [11]. (c) Kiechle *et al.* [13]. (d) Ma *et al.* [45]. (e) Ferstl *et al.* [12]. (f) Our baseline. (g) Our baseline + DGR optimization. (h) Our baseline + DGR optimization + DIP feature. (i) Ground truth. More examples can be found in the supplementary material.

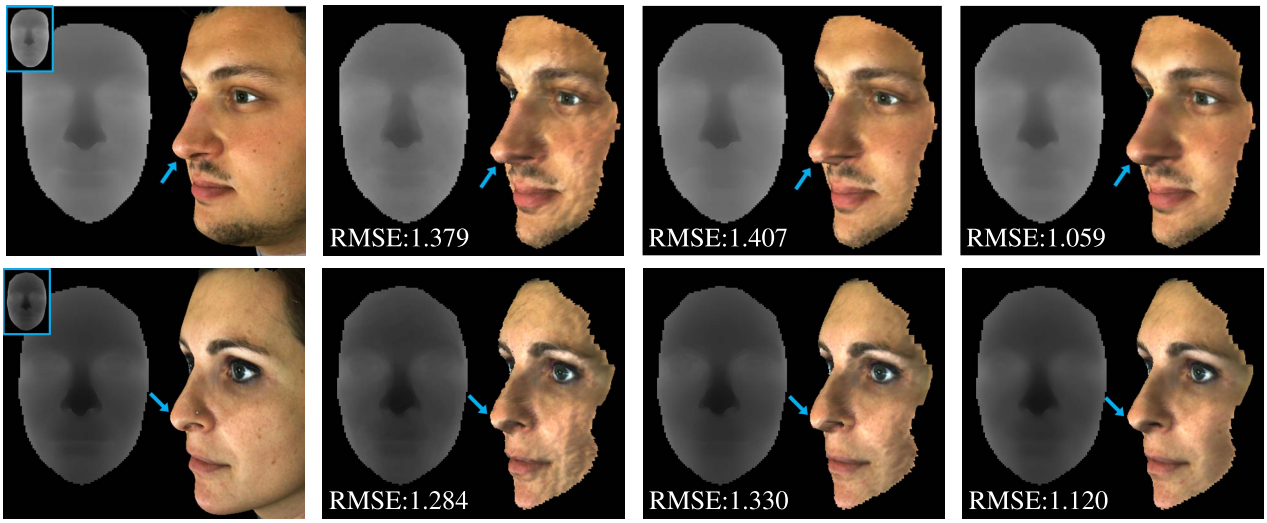


Fig. 9. Comparison of the 3-D surface reconstruction for 4× upsampling on the degraded Florence Superface dataset. The RMSE is shown at the bottom of the reconstructed depth maps. We show that 3-D reconstruction accuracy can be well improved by the proposed method. Our method obtains the lowest RMSE, indicating the effectiveness of our method quantitatively. As shown by the blue arrow (the tip of nose in the images), our result is most similar to the ground truth among the compared results. (a) Ground truth. (b) Ma *et al.* [45]. (c) Ferstl *et al.* [12]. (d) Ours.

We take 70 3-D face models and synthesize depth maps using their z-axis data. Color images are obtained from the textured models.

In the testing phase, we use the rest of the face models in the BU-3DFE dateset [46] and the Florence Superface dataset [47] to form our testing data for synthetic facial depth map SR.

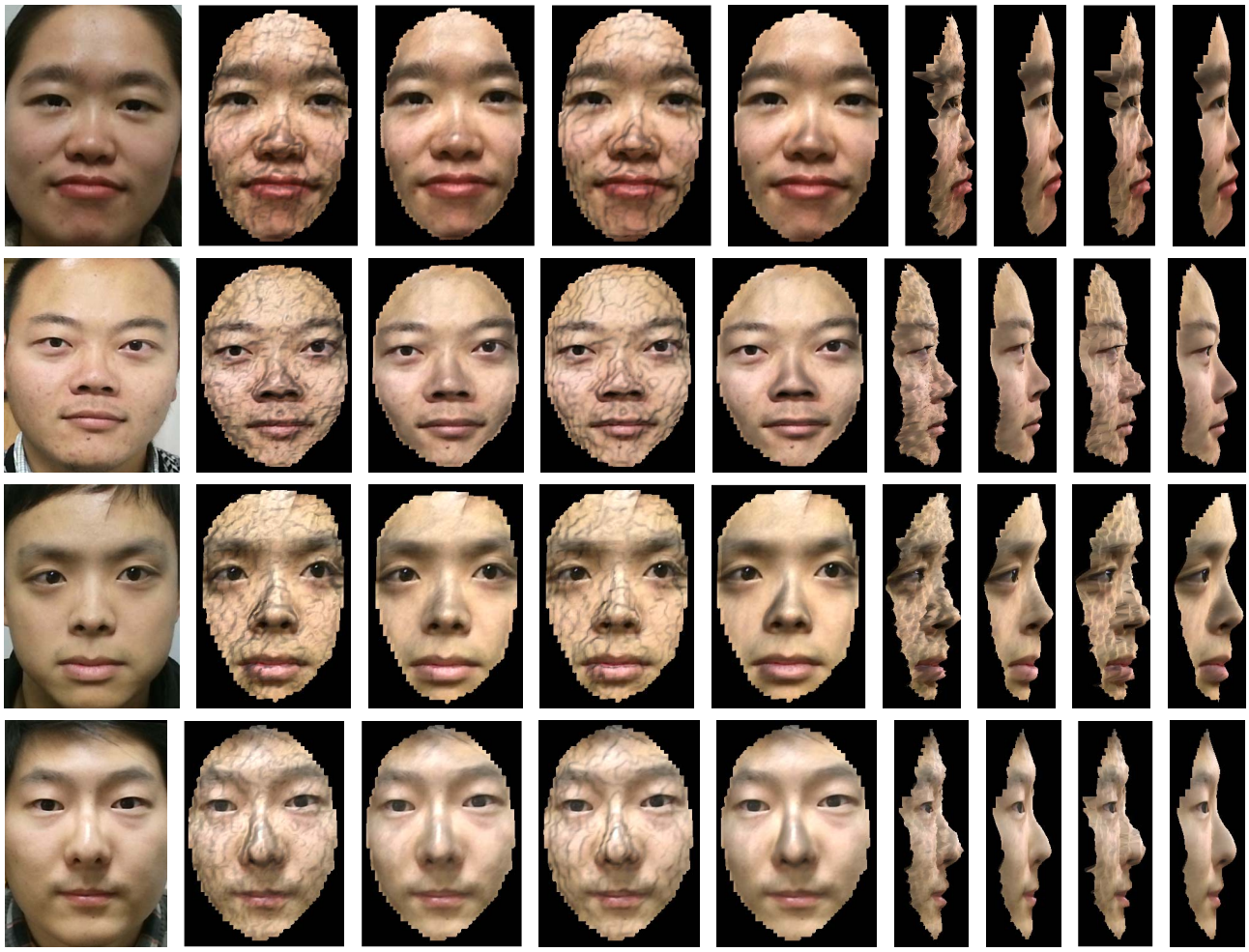


Fig. 10. Comparison on Kinect face data. For real-world face data, our method recovers 3-D surfaces that best match the normal facial physical structures. (a) Color. (b) Raw. (c) Ferstl *et al.* [12]. (d) Kiechle *et al.* [13]. (e) Ours.

As reported in [48], distance-dependent noise and quantization error are two main degradations for Kinect depth data. To simulate real-world consumer-level depth cameras, these two artifacts are added. In particular, distance-dependent noise that satisfies a Gaussian distribution $N(0, \kappa d^2)$ is added, where $\kappa = 1.43 \times 10^{-5}$ is Kinect-oriented constant [48] and d is the original depth. In our case, we set the distance (depth) between the tip of nose and the camera as 1.2 m. After that, the noisy depth map is quantized using quantization steps of 3 mm. An example of the degradation simulation is shown in Fig. 6(e) and (f). We refer to [48] and [49] for details of the Kinect degradation.

Besides, we have collected several face data from Kinect 2.0 for real-world applications. In the experiment, we use a three-scale strategy. Specifically, the raw Kinect depth map is first aligned to the corresponding HR color image using nearest neighbor interpolation and then downsampled by a factor of 8 to obtain the target LR depth map for recovery.

B. Performance Evaluation With Synthetic Facial Depth Map

We compare our method with state-of-the-art image-guided SR methods on synthetic depth maps derived from 3-D face models in the BU-3DFE dateset [46] and the Florence Superface dataset [47]. Both clean and degraded cases are considered.

1) Clean BU-3DFE Dateset: We apply the proposed method to LR noise-free facial depth maps. We compare with Diebel's and Thrun's MRF-based method [27], Yang *et al.*'s 3-D joint bilateral filter method [24], He *et al.*'s guided filter method [11], Ma *et al.*'s weighted median filter method [45], Kiechle *et al.*'s bimodal co-sparse analysis method [13], and Ferstl *et al.*'s TGV method [12]. The last three methods can be representative for state-of-the-art filter-based, optimization-based, and exemplar-based techniques, respectively. The MATLAB reimplementation of Diebel's and Yang's methods can be found in [24] and [27]. The softwares of He's, Ma's, Kiechle's, and Ferstl's methods are available on their project websites [11]–[13], [45]. The training data used for [13] is identical to that used in our method. Table I reports the comparison of $2\times$, $4\times$, and $8\times$ upsampling in terms of root-mean-square error (RMSE). Kiechle *et al.*'s method [13] and the proposed method obtains lowest RMSE, indicating that the depth reconstruction can be well improved through learning from high-quality depth maps.

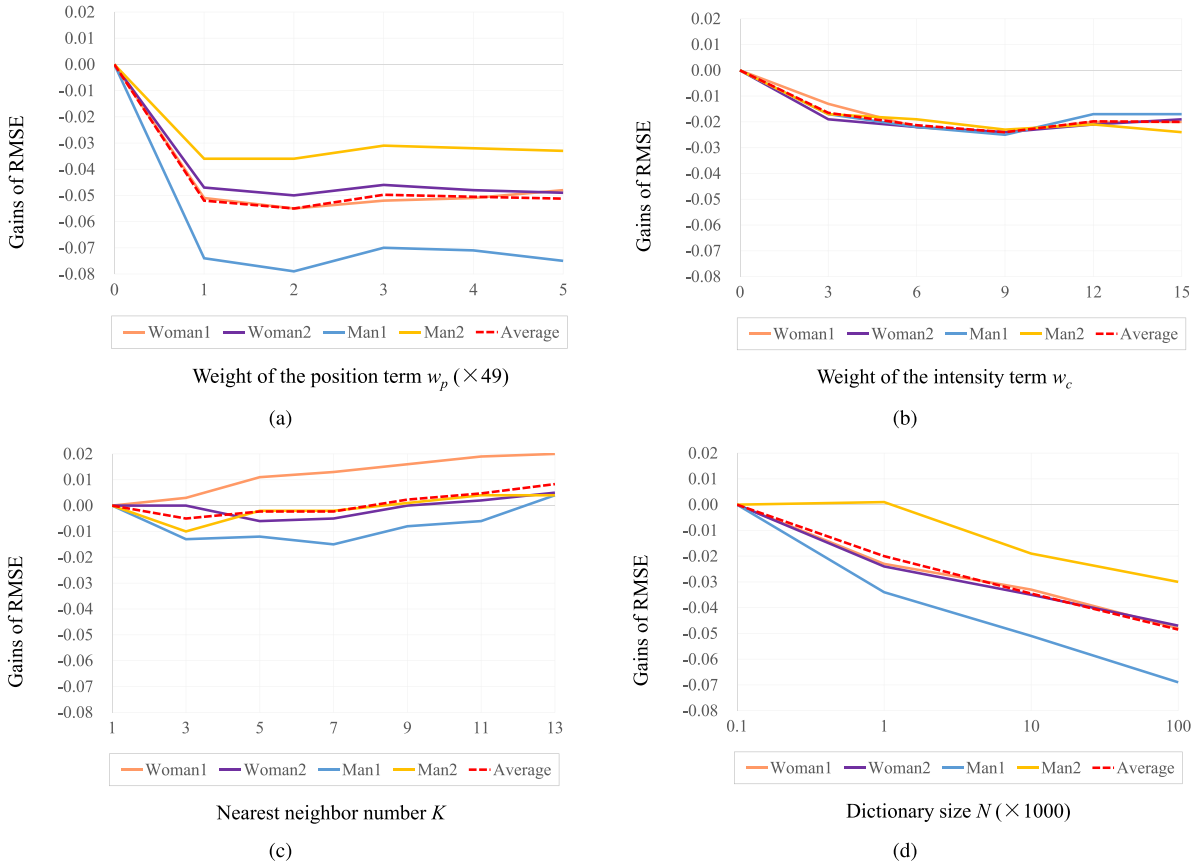


Fig. 11. Effect of the parameters. (a) Weight w_p of the position term. (b) Weight w_c of the intensity term. (c) Nearest neighbor number K. (d) Dictionary size N. Here, we show the gains of RMSE instead of the original RMSE for better comparisons. It is computed by subtracting the RMSE obtained using the smallest parameter (e.g., $w_p = 0, w_c = 0, K = 1$) from the original RMSE. Lower values indicate more accurate results.

2) Degraded BU-3DFE Dateset: To show how our algorithm stands out from other state-of-the-art methods, we conduct an experiment on the SR of facial depth maps that have severe noise and quantization problems. For visual comparison, 8× upsampled depth maps for Woman3 and Man2 are shown in Fig. 7. Please enlarge and view these figures on the screen for better comparison. Without sufficient reliable depth information, the filter-based methods [11], [24], [45] produce distinct texture copying artifacts. Diebel and Thrun's method [27] and Ferstl *et al.*'s method [12] does well in denoising but tends to produce over-smooth results, in which the nose details are mostly lost. The Kiechle *et al.*'s learning-based method [13] is severely affected by the noises and even produces some impulsive noises. In comparison, our method is capable of utilizing the low-level intensity and high-level position information to obtain high-quality face priors to improve the reconstruction. From the zoomed regions of the wing of nose, the depth is recovered to match the normal facial physical structures.

For quantitative comparison, recovery results in terms of RMSE are reported in Table II, and our method obtains the lowest RMSE for all cases. Comparisons are also presented for our approach without and with the key components of the DIP features and the DGR term. We use the proposed edge-aware NE as baseline and the high RMSE indicates that unreliable input can severely degrade the performance of exemplar-based methods. It can be also seen in Fig. 8(f) that conventional NE fails to undo noises. The errors are dramatically reduced in the last row of Table II, which demonstrates that our adaptations of the baseline method nicely address the degradation problem in the depth refinement process. To study the structure of the proposed DIP feature, we remove the position cues (by setting $w_p = 0$ and canceling facial component decomposition) and intensity cues (by setting $w_c = 0$), and see the changes of RMSE. As shown in Table II, removing any one of the cues will increase the RMSE but removing position cues affects more. This demonstrates that the proposed DIP feature is well designed for robust nearest neighbor search under the degradation condition, among which the high-level position cues contribute quite a lot. In addition, we further give the quantitative comparison between the proposed DGR term and TV/SGR terms in the last three rows of Table II. The RMSE results verify the superiority of the DGR optimization. From the error maps of Fig. 8(g) and (h), we can also see that the DGR optimization effectively eliminates noise, and the DIP features further help refine the facial structures, contributing to high-quality depth maps.

3) Degraded Florence Superface Dataset: We additionally tested our method on the Florence Superface dataset [47]. Two examples of our results are shown in Fig. 9. 3-D face models are reconstructed using the recovered depth maps. Ma *et al.*'s [45] results suffer noises and have uneven surfaces,

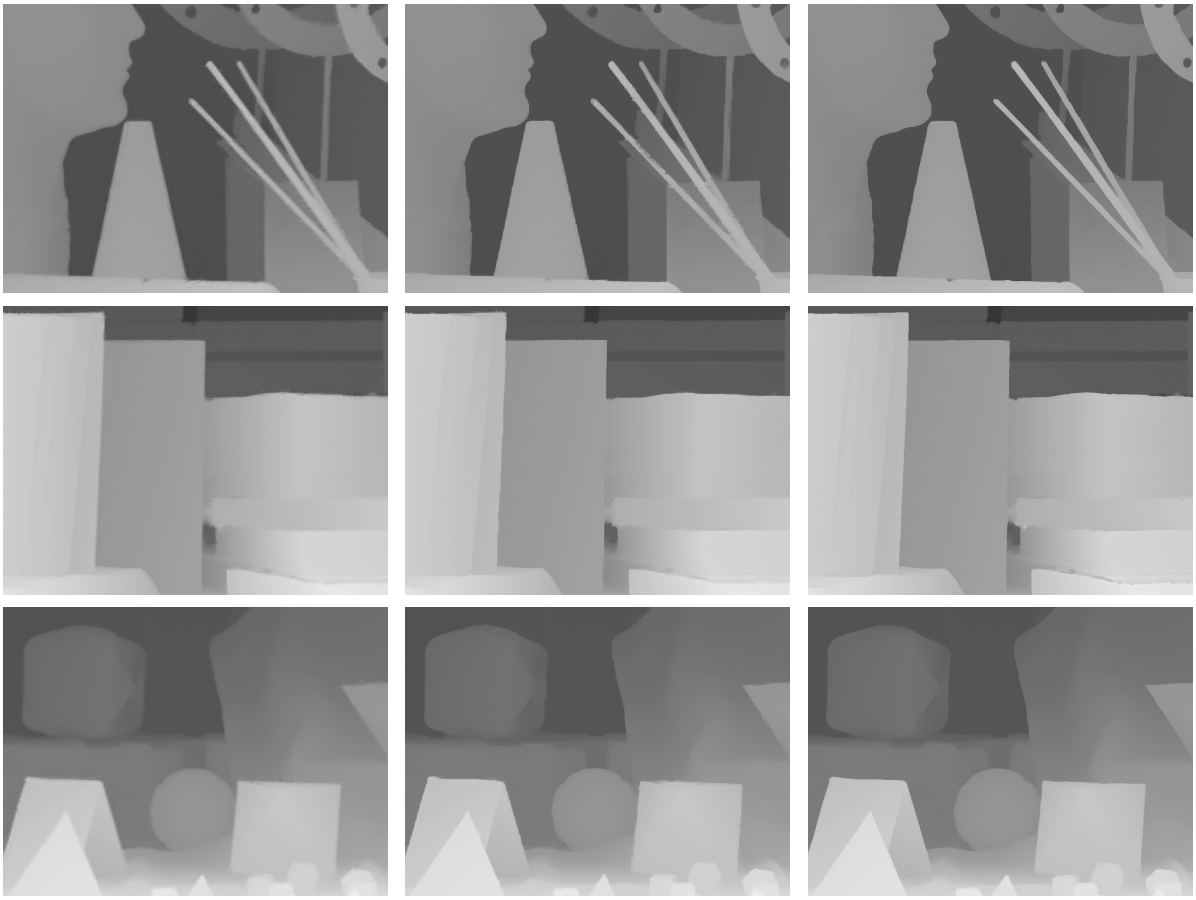


Fig. 12. Depth map 4× SR for noise-free Middlebury dataset. Our method produces sharp and clean edges. (a) He *et al.*'s method [11]. (b) Ferstl *et al.*'s method [12]. (c) Our results. Please enlarge and view the edges in these figures on the screen for better comparison. The results are cropped for the visualization. Full resolution comparisons with more methods are provided in the supplementary materials.

which reduces the visual quality. Meanwhile, Ferstl *et al.*'s method [12] does well in denoising, but creates some excessively flat region which is at variance with physical structures of human faces. See the nose of the woman in the second row. By comparison, our method achieves best results in both visual quality and quantitative evaluation. The 3-D face model synthesized from our enhanced depth map is highly consistent with the ground truth. More examples can be found in the supplementary material.

C. Performance Evaluation With Real-World Kinect Face Data

We also apply the proposed method to real-world Kinect facial depth data. Fig. 10 illustrates that the proposed method preserves most of the facial components, and the boundaries in the side views are quite smooth. The sudden change in depth map reconstructed by Kiechle *et al.*'s method [13] leads to the stepping artifacts on 3-D surfaces. Ferstl *et al.*'s method [12] generates rough boundaries, and it fails to recover plausible facial components, such as yielding pointy noses and concave upper lips.

D. Effect of the Parameters

In this section, we discuss the effects of parameters. Fig. 11 illustrates the relationship between the RMSE results and the weight w_p of the position term, the weight w_c of the intensity term, nearest neighbor number K and the dictionary size N for 4× upsampling on the degraded BU-3DFE dataset.

By incorporating position cues, our method achieves more accurate results, as shown in Fig. 11(a). It seems that our method is robust to the choice of w_p when $w_p \in [49, 245]$.

The similar conclusion is found for w_c in Fig. 11(b) that the better results are obtained by introducing intensity cue which provides robust nearest neighbor search to depth noises. The acceptable value for w_c is within $[6, 12]$ and the performance will slightly drop if w_c is larger than 12. The reason is that higher weight on intensity cues will undervalue the depth cues for the fidelity of face structures.

We observe that our method is robust to K. In a wide range of $[1, 11]$, the change in terms of RMSE is subtle which is within 0.01 on average. As for the dictionary size N, increasing N will improve our SR results. But the gain has a marginal decreasing effect. As shown in Fig. 11(d), the exponential growth of N brings only a linear decrease of RMSE. Considering the computing complexity and memory capacity, we set N to 100 000.

In the end, the effect of the patch size n on RMSE is illustrated in Table III. Large patch size (e.g., $n = 11, 13$) makes it hard to fit the depth structures using NE regression in (2), thus decreasing the performance. We also observe that for 2×

TABLE III
EFFECT OF PATCH SIZE

Patch Size	Average RMSE		
	2×	4×	8×
3×3	0.898	**0.991**	1.280
5×5	0.826	0.994	**1.273**
7×7	**0.823**	1.000	1.279
9×9	0.825	1.011	1.301
11×11	0.830	1.016	1.338
13×13	0.837	1.017	1.365

TABLE IV
RMSE COMPARISON FOR NOISE-FREE MIDDLEBURY DATASET

Methods	art		books		moebius	
	2×	4×	2×	4×	2×	4×
Diebel *et al.* [27]	3.124	3.810	1.201	1.537	1.188	1.442
Yang *et al.* [24]	4.073	4.071	1.608	1.686	1.069	1.386
He *et al.* [11]	3.804	3.804	1.562	1.562	1.435	1.435
Ferstl *et al.* [12]	3.038	3.800	1.286	1.592	1.130	1.460
Ours (without TV)	1.214	1.993	0.421	0.752	0.471	0.774
Ours	**1.125**	**1.909**	**0.416**	**0.749**	**0.462**	**0.771**

upsampling, the results improve with the increase of patch size as a small patch size (e.g., $n = 3, 5$) tends to overfit the noise. For 4× and 8× upsamplings, since decreasing the resolution of an image is equivalent to increasing its patch size, the overfitting problem is relieved.

E. Performance Evaluation With Middlebury Dataset

In the end, to validate the advantages of edge enhancement, we tested the proposed edge-aware NE method on the noise-free Middlebury 2005 dataset [50] provided by [12], which contains much more distinct edges than facial depth maps. In this experiment, our dictionaries are obtained from the Middlebury 2006 dataset [50]. Comparisons are presented in Table IV for our approach without and with edge enhancement process as well as other related methods in terms of RMSE. Examples of our results are also illustrated in Fig. 12. Compared to He *et al.*'s [11] and Ferstl *et al.*'s [12] methods, our method yields sharper and more clean edges.

V. CONCLUSION

In this paper, we present a novel joint-feature guided depth map SR method with face priors. A joint DIP feature is designed to consider both low-level and high-level facial characteristics for better similarity measurement. Through integrating face priors with the modified NE framework, significant improvements on facial depth reconstruction are achieved comparing to state-of-the-art technologies. In future work, we will investigate the general structure and smoothness priors and expand the applicability to the depth map reconstruction for general scenes.

REFERENCES

[1] D. Tao, L. Jin, Z. Yang, and X. Li, "Rank preserving sparse learning for Kinect based scene classification," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1406–1417, Oct. 2013.

[2] M. Camplani, T. Mantecón, and L. Salgado, "Depth-color fusion strategy for 3-D scene modeling with Kinect," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1560–1571, Dec. 2013.

[3] A. González, D. Vázquez, A. M. Lóopez, and J. Amores, "On-board object detection: Multicue, multimodal, and multiview random forest of local experts," *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–11, Aug. 2016.

[4] B. Ni, Y. Pei, P. Moulin, and S. Yan, "Multilevel depth and image fusion for human activity detection," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1383–1394, Oct. 2013.

[5] H. P. H. Shum, E. S. L. Ho, Y. Jiang, and S. Takagi, "Real-time posture reconstruction for Microsoft Kinect," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1357–1369, Oct. 2013.

[6] 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.

[7] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013.

[8] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "LidarBoost: Depth superresolution for ToF 3D shape scanning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 343–350.

[9] S. Izadi *et al.*, "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol.*, Santa Barbara, CA, USA, 2011, pp. 559–568.

[10] S. A. Gudmundsson, H. Aanaes, and R. Larsen, "Fusion of stereo vision and Time-of-Flight imaging for improved 3D estimation," *Int. J. Intell. Syst. Technol. Appl.*, vol. 5, nos. 3–4, pp. 425–433, Nov. 2008.

[11] 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. [Online]. Available: http://research.microsoft.com/en-us/um/people/kahe/eccv10/

[12] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 993–1000. [Online]. Available: http://rvlab.icg.tugraz.at/project_page/project_tofusion/project_tofsuperresolution.html

[13] M. Kiechle, S. Hawe, and M. Kleinsteuber, "A joint intensity and depth co-sparse analysis model for depth map super-resolution," in *Proc. Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 1545–1552. [Online]. Available: http://www.gol.ei.tum.de/index.php?id=6&L=1

[14] D. C. Garcia, C. Dorea, and R. L. D. Queiroz, "Depth-map super-resolution for asymmetric stereo images," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Vancouver, BC, Canada, 2013, pp. 1548–1552.

[15] Y. Yang, J. Cai, Z. Zha, M. Gao, and Q. Tian, "A stereo-vision-assisted model for depth map super-resolution," in *Proc. IEEE Int. Conf. Multimedia Expo*, Chengdu, China, 2014, pp. 1–6.

[16] Y. Yang, M. Gao, J. Zhang, Z. Zha, and Z. Wang, "Depth map super-resolution using stereo-vision-assisted model," *Neurocomputing*, vol. 149, pp. 1396–1406, Feb. 2015.

[17] J. Zhang *et al.*, "A unified scheme for super-resolution and depth estimation from asymmetric stereoscopic video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 479–493, Mar. 2016.

[18] D. Ferstl, M. Rüther, and H. Bischof, "Variational depth superresolution using example-based edge representations," in *Proc. Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 513–521.

[19] J. Xie, R. S. Feris, S.-S. Yu, and M.-T. Sun, "Joint super resolution and denoising from a single depth image," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1525–1537, Sep. 2015.

[20] 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.

[21] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, 2007.

[22] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noise-aware filter for real-time depth upsampling," in *Proc. Workshop Multi-Camera Multi-Modal Sensor Fusion Algorithms Appl.*, 2008, pp. 1–12.

[23] D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1176–1190, Mar. 2012.

[24] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8. [Online]. Available: https://bitbucket.org/shshzaa/depth-enhancement

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

[26] X. Tan, C. Sun, and T. D. Pham, "Edge-aware filtering with local polynomial approximation and rectangle-based weighting," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2693–2705, Dec. 2016.

[27] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 291–298. [Online]. Available: https://bitbucket.org/shshzaa/depth-enhancement

[28] 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.

[29] 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.

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

[31] Y. Li, T. Xue, L. Sun, and J. Liu, "Joint example-based depth map super-resolution," in *Proc. IEEE Int. Conf. Multimedia Expo*, Melbourne, VIC, Australia, Jul. 2012, pp. 152–157.

[32] H. H. Kwon, Y.-W. Tai, and S. Lin, "Data-driven depth map refinement via multi-scale sparse representation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 159–167.

[33] Y.-W. Tai, W.-S. Tong, and C.-K. Tang, "Perceptually-inspired and edge-directed color image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, 2006, pp. 1948–1955.

[34] T.-M. Chan, J. Zhang, J. Pu, and H. Huang, "Neighbor embedding based super-resolution algorithm through edge detection and feature selection," *Pattern Recognit. Lett.*, vol. 30, no. 5, pp. 494–502, 2009.

[35] Y. Li, J. Liu, W. Yang, and Z. Guo, "Neighborhood regression for edge-preserving image super-resolution," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 1201–1205.

[36] H. Chang, D. Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, USA, Jun./Jul. 2004, p. 1.

[37] R. Timofte, V. De, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 1920–1927.

[38] X. Ma, J. Zhang, and C. Qi, "Hallucinating face by position-patch," *Pattern Recognit.*, vol. 43, no. 6, pp. 2224–2236, 2010.

[39] C. Jung, L. Jiao, B. Liu, and M. Gong, "Position-patch based face hallucination using convex optimization," *IEEE Signal Process. Lett.*, vol. 18, no. 6, pp. 367–370, Jun. 2011.

[40] Z. Wang, R. Hu, S. Wang, and J. Jiang, "Face hallucination via weighted adaptive sparse regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 802–813, May 2014.

[41] J. Jiang, J. Ma, C. Chen, X. Jiang, and Z. Wang, "Noise robust face image super-resolution through smooth sparse representation," *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–12, Aug. 2012.

[42] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4220–4231, Oct. 2014.

[43] J. Jiang, C. Chen, K. Huang, Z. Cai, and R. Hu, "Noise robust position-patch based face super-resolution via Tikhonov regularized neighbor representation," *Inf. Sci.*, vols. 367–368, pp. 354–372, Nov. 2016.

[44] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

[45] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu, "Constant time weighted median filtering for stereo matching and beyond," in *Proc. Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 49–56. [Online]. Available: http://research.microsoft.com/en-us/um/people/kahe/

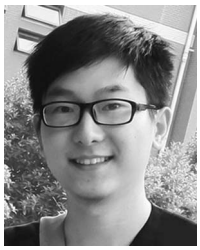
[46] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 2006, pp. 211–216. [Online]. Available: http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html

[47] S. Berretti, A. Del Bimbo, and P. Pala, "Superfaces: A super-resolution model for 3D faces," in *Proc. Int. Conf. Comput. Vis.*, Florence, Italy, 2012, pp. 73–82. [Online]. Available: http://www.micc.unifi.it/vim/datasets/4d-faces/

[48] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[49] M. J. Landau, B. Y. Choo, and P. A. Beling, "Simulating Kinect infrared and depth images," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3018–3031, Dec. 2016.

[50] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Madison, WI, USA, Jun. 2003, pp. I-195–I-202.



Shuai Yang received the B.S. degree in computer science from Peking University, Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree with the Institute of Computer Science and Technology.

His current research interests include image inpainting, depth map enhancement, and image stylization.

Jiaying Liu (S'09–M'10) received the B.E. degree in computer science from Northwestern Polytechnic University, Xi'an, China, in 2005, and the Ph.D. (Hons.) degree in computer science from Peking University, Beijing, China, in 2010.

She is currently an Associate Professor with the Institute of Computer Science and Technology, Peking University. She was a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA, from 2007 to 2008. She was a Visiting Researcher with Microsoft Research Asia, Beijing, in 2015, supported by "Star Track for Young Faculties." She has authored over 80 technical articles in refereed journals and proceedings, and holds 13 granted patents. Her current research interests include image/video processing, compression, and computer vision.

Dr. Liu has been also serving as a TC member in APSIPA IVM since 2015, and APSIPA Distinguished Lecture from 2016 to 2017.

Yuming Fang (M'13) received the B.E. degree from Sichuan University, Chengdu, China, the M.S. degree from the Beijing University of Technology, Beijing, China, and the Ph.D. degree from Nanyang Technological University, Singapore.

He is currently an Associate Professor with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. He was a (Visiting) Researcher with IRCCyN Laboratory, PolyTech' Nantes, University of Nantes, Nantes, France, National Tsinghua University, Hsinchu, Taiwan, and the University of Waterloo, Waterloo, ON, Canada. His current research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, and 3-D image/video processing.

Dr. Fang currently serves as an Associate Editor for the IEEE ACCESS and is on the editorial board of *Signal Processing: Image Communication*.

Zongming Guo (M'09) received the B.S. degree in mathematics, and the M.S. and Ph.D. degrees in computer science from Peking University, Beijing, China, in 1987, 1990, and 1994, respectively.

He is currently a Professor with the Institute of Computer Science and Technology, Peking University. His current research interests include video coding, processing, and communication.

Dr. Guo was a recipient of the First Prize of the State Administration of Radio Film and Television Award in 2004, the First Prize of the Ministry of Education Science and Technology Progress Award in 2006, the Second Prize of the National Science and Technology Award in 2007, the Wang Xuan News Technology Award, the Chia Tai Teaching Award in 2008, the Government Allowance granted by the State Council in 2009, and the Distinguished Doctoral Dissertation Advisor Award of Peking University in 2012 and 2013. He is the Executive Member of the China-Society of Motion Picture and Television Engineers.