EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Facial image super-resolution guided by adaptive geometric features

Zhenfeng Fan[1,2], Xiyuan Hu[1,2*], Chen Chen[1], Xiaolian Wang[1] and Silong Peng[1]

*Correspondence:
xiyuan.hu@ia.ac.cn
[1] Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing, China
[2] University of Chinese Academy of Sciences, 80 Zhongguancun East Road, Beijing, China

Abstract

This paper addresses the traditional issue of restoring a high-resolution (HR) facial image from a low-resolution (LR) counterpart. Current state-of-the-art super-resolution (SR) methods commonly adopt the convolutional neural networks to learn a non-linear complex mapping between paired LR and HR images. They discriminate local patterns expressed by the neighboring pixels along the planar directions but ignore the intrinsic 3D proximity including the depth map. As a special case of general images, the face has limited geometric variations, which we believe that the relevant depth map can be learned and used to guide the face SR task. Motivated by it, we design a network including two branches: one for auxiliary depth map estimation and the other for the main SR task. Adaptive geometric features are further learned from the depth map and used to modulate the mid-level features of the SR branch. The whole network is implemented in an end-to-end trainable manner under the extra supervision of depth map. The supervisory depth map is either a paired one from RGB-D scans or a reconstructed one by a 3D prior model of faces. The experiments demonstrate the effectiveness of the proposed method and achieve improved performance over the state of the arts.

Keywords: Convolutional neural networks (CNNs), Depth map, Face super-resolution

1 Introduction

Human-centered image and video analysis have gained ever-increasing attention in both academic and industrial areas worldwide in recent years. Since face is a key character of human, machine-aided facial analysis becomes a popular issue in various applications. For example, the manufactories of mobile devices are highly interested in developing both hardware and software systems for the collection of facial images. This paper addresses a traditional issue, i.e., face super-resolution (SR) in the field of facial image analysis.

Face super-resolution, also known as face hallucination for heavily blurred images, aims at recovering a high-resolution (HR) facial image from a low-resolution (LR) one. It is fundamental in a number of applications for facial analysis, such as face alignment [1, 2], face recognition [3, 4], and E-commerce platforms [5]. When the acquired facial images are of very low resolutions especially in the surveillance videos, it becomes extremely difficult for machines as well as humans to identify useful information therein. Face SR alleviates this problem to some extent by restoring the missing pixels.

Face SR is a sub-problem of the general image SR. It is an ill-posed traditional problem that attracts a lot of academic studies and has straightforward applications in the photographic industry. Recently, the advancement of convolutional neural networks (CNNs) has activated a lot of studies on both general image and face SR. The state-of-the-art methods [6–11] commonly adopt a well-customized CNN structure to learn a complex non-linear mapping between many pairs of LR and HR images. The structure of the CNN tends to be deeper and wider, with increasing data to feed. The prior knowledge is recently used in the form of facial landmarks [1], parsing maps [12], or facial component heatmaps [13] for the special task of face SR. In this paper, we explore another form of information, i.e., the depth map of face, which can be incorporated to assist the face SR task.

The motivation originates from understanding the basic convolutional operation for the SR task. A deep CNN structure in fact learns to interpolate the missing pixels based on the local patterns of the input image. The convolution is a translation invariant operation that deals with the neighboring pixels in the 2D image coordinates. For example, the cascade of several convolution layers' receptive field for a given pixel p is shown in Fig. 1a. However, the intrinsic 3D proximity of the neighboring pixels within the receptive field is not equal to the 2D proximity of them, as shown in Fig. 1b. This is generally ignored in the literature and may lead to blurry effect on the edges of the faces. We argue that a network for face SR should also consider the 3D proximity of neighboring pixels, which can be inferred from a facial depth map (Fig. 1c).

In this paper, we propose an SR network architecture guided by the depth map to enhance the face SR performance. While it is not easy to learn the depth map for a general image, the face has limited geometric variations to infer a relatively accurate shape. The proposed network includes two branches: one for auxiliary depth map estimation and the other for the main SR task. Adaptive geometric features are further learned from the depth map and used to modulate the mid-level features of the SR branch. The whole network is end-to-end trainable, with a common SR reconstruction loss and an extra loss of depth supervision. During training, the supervision of paired depth map is involved. During testing, only raw LR input image is required. The experiments are carried out on two publicly available datasets FRGC v2.0 [14] and FFHQ [15] and demonstrate the effectiveness of the proposed method.

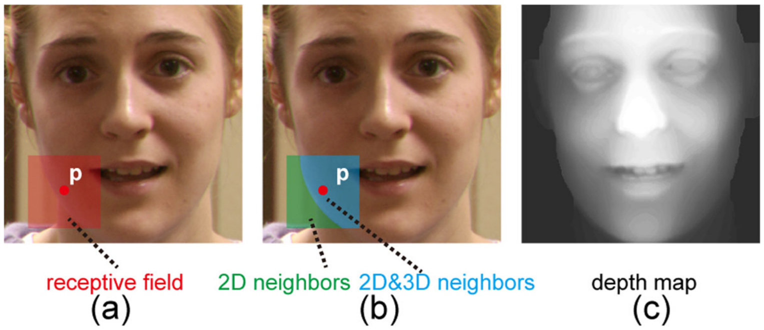

Fig. 1 Receptive field of 2D and 3D cases for input image. Illustrations of **a** the receptive field and **b** the difference of 2D and 3D neighbors for a given pixel *p*, and **c** the corresponding depth map

The main contributions of this paper are as follows: (1) we first propose to incorporate depth map for the face SR task. The depth map is either a matched one from RGB-D scans or a reconstructed one by a 3D prior model of faces. (2) We build a tailored network to estimate the depth map and to assist the main SR task. We use adaptive geometric features to modulate the mid-level features for face SR. The modulation operation is new and complementary to the general convolutional operations for SR tasks. (3) The proposed method leads to sharper edges for super-resolved images, which is a desirable goal for face SR.

This paper is organized as follows. Section 2 reviews the related works. Section 3 elaborates the proposed network architecture, the loss for training the network, and the way to prepare training data. Section 4 presents the experiments for the validation of the proposed method. Section 5 discusses limitations and future directions for this work. We conclude this paper in Section 6.

2 Related work

Since face SR belongs to the larger class of general image SR, this section reviews the related works in both image SR and face SR. We also focus on the state-of-the-art deep learning-based methods.

2.1 Image super-resolution

The advancement of CNNs has activated a lot of studies on image SR. Dong et al. [16] first introduce a three-layer CNN to learn a non-linear mapping between many pairs of LR and HR images. Then, Kim et al. [17] propose a deep CNN with 20 layers to learn the residuals between the paired LR and HR images. Early works focus on improving the performance of SR in terms of quantitative metrics such as the peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM).

It is later found that minimizing the mean square error (MSE) prefers high PSNR/SSIM but lacks high-frequency details. Ledig et al. [6] introduce a perceptual loss which is a combination of an adversarial loss and a pixelwise MSE loss in a high-level feature layer of the VGGnet [18] pretrained by the imageNet classification task. Parts of their work, also known as SRResNet, use an advanced CNN structure [19] at that time to achieve better PSNR/SSIM. Since then, many works on image SR focus not only on PSNR/SSIM but also on better visualized results that contain high-frequency details. However, there is a trade-off between them since the loss function is usually a combination of several terms. While the earlier works evaluate the results by visualizing the super-resolved images, the current works propose quantitative metrics such as the perceptual index (PI) [20]. In this work, we focus on enhancing the performance of face SR in terms of PSNR and SSIM, since this will also set a higher baseline for results of good perceptual qualities.

2.2 Face super-resolution

Exploiting facial prior in face SR, such as the spatial locations of landmarks is the key difference from general image SR. There are many existing works dedicated to face SR using the prior knowledge of face. Before the coming of deep learning-based methods, early works [21, 22] attempt to learn the super-resolved faces in some low-dimensional representations. This reduces the dimensions of the original ill-posed problem, thus leads to more realistic restored facial images. Although these methods only deal with the frontal pose, they provide clues to learn the face SR from the intrinsic facial structures.

The recent CNN-based face SR methods [1, 12, 13, 23, 24] have progressed a lot, in terms of both quantitative metrics as PSNR/SSIM and perceptive visual qualities. The CNN-based methods do not replace the classical methods totally, but rather seem to combine and re-implement the old ideas, with the powerful new tool now. Meanwhile, the CNN-based face alignments have also progressed a lot, reducing the difficulties of face SR under pose variations. Song et al. [24] propose a facial structure generation network to restore a coarse HR face and use exemplar HR faces for detail enhancement. Zhu et al. [25] super-resolve unaligned faces with very low resolutions in a task-alternating cascade framework. Since more accurate face alignment promotes better SR results and vice versa, the task-alternating framework leads to improvements on both SR and face alignment. More recent works [1, 12, 13] implement some multitask networks and train in an end-to-end manner. They use either facial landmarks, parsing maps, or facial component heatmaps which are in fact different forms of face alignment. The aforementioned methods belong to single image-based SR. Contrary to that some other works [26, 27] super-resolve a facial image with the help of a high-quality guided image. In this work, we propose to use the facial depth map, which is an intrinsic property of face to assist the face SR task.

3 Proposed method

In this section, we introduce the detailed network architecture, loss for training the network, and ways to prepare supervisory depth maps.

3.1 Network architecture

The architecture of the proposed network is illustrated in Fig. 2, and the detailed structures of the reused units are specified in Fig. 3 in different colors. The output channel size (n), the kernel size (k), and the stride (s) are indicated for each convolutional layer. We do not mark out the input channel sizes since they can be deduced by comparing the output channel sizes of two neighboring layers. The input and output layers should also be compatible with the input and output images. In summary, the network consists of three sub-blocks: the main SR block, the depth estimation block, and the modulation block, with the detailed architectures concluded in Table 1.

The main SR block is designed based on the basic network architecture of SRResNet [6]. We adopt the residual unit as the basic element of the network. The residual unit has two main advantages: (1) it makes the network deeper with minor computational cost, (2) the skip connection (residual) helps to solve problems of vanishing and exploding gradients and enables the network easier to be trainable. We also take the two improvements from [28] and [29]: removing the batch normalization layers and adding rescaling for the residual unit (see Fig. 3), which can enhance the performance of the network. In addition, the designed network learns the global residual in two stages rather than one as in [6]. The purpose is to maintain the full information for the mid-level features from the input image for modulation, while still keeping the local and global residual structures as in [6].

The depth estimation block adopts an hourglass architecture inherited from the well-known Unet [30]. The feature concatenations in different feature expansions enable the network to learn the depth map in different level of features and in a more robust manner.

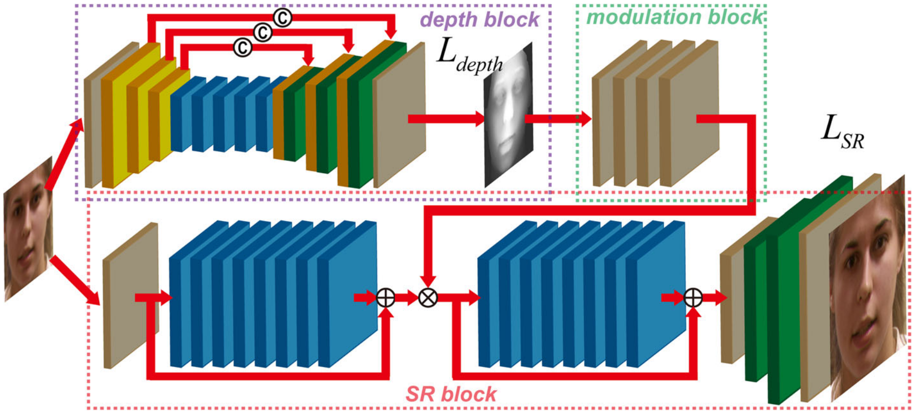

Fig. 2 Network architecture. The basic network structure with the depth block, the modulation block, and the SR block. The input and output channel sizes are commonly annotated in Fig. 3 for each convolutional layers specified in different colors. In the special cases for the input and output layers that take the 3-channel images as inputs or outputs, the input and output channel sizes should be modified to be compatible with the input and output images

We design this architecture to estimate the depth map from an RGB image. This task is commonly difficult for general images in an uncontrolled environment. However, it is a popular issue worthy of study for indoor sceneries [31] because of the limited geometric variations. The face has also limited geometric structure which we believe that the depth can be learned from the RGB image.

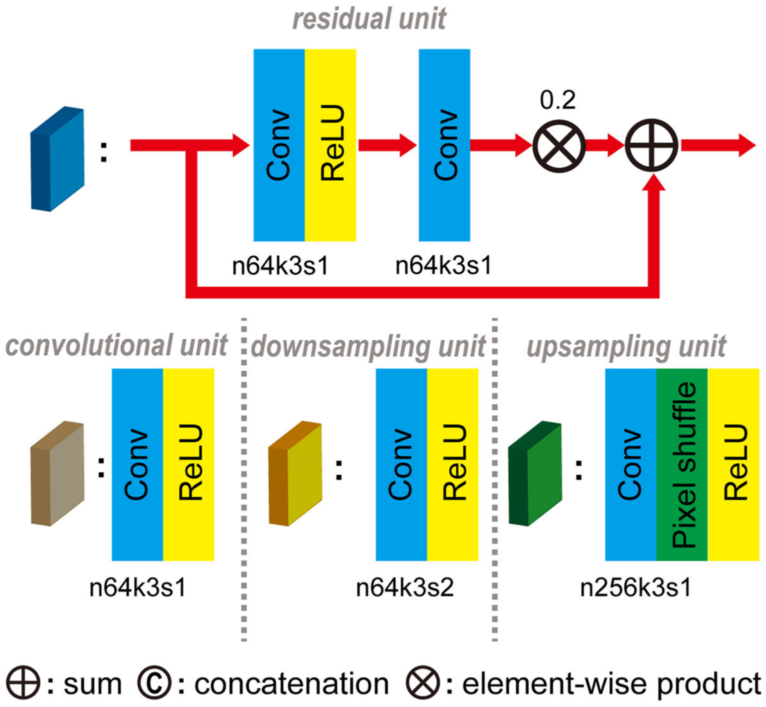

Fig. 3 Network units. The basic units of the network in Fig. 2. The output channel size (*n*), the kernel size (*k*), and the stride (*s*) are indicated for each convolutional layer. The input channel size can be deduced by comparing the output channel sizes of two neighboring layers

Table 1 Architectures of the proposed network including three sub-blocks

Network architecture		
The main SR block	The depth estimation block	The modulation block
Conv: $[3 \times 3, 3, 64]$	Conv: $[3 \times 3, 3, 64]$	
Residual: $[3 \times 3, 64, 64] \times 8$	Downsample: $[3 \times 3, 3, 64] \times 3$	
Feature addition	Residual: $[3 \times 3, 64, 64] \times 5$	Conv: $[3 \times 3, 1, 64]$
Feature multiplication	Upsample: $[3 \times 3, 64, 64]$	
Residual: $[3 \times 3, 64, 64] \times 8$	Feature concatenation	
Feature addition	Upsample: $[3 \times 3, 64, 64]$	
Conv: $[3 \times 3, 64, 64]$	Feature concatenation	
Upsample: $[3 \times 3, 64, 64] \times 2$	Upsample: $[3 \times 3, 64, 64]$	Conv: $[3 \times 3, 64, 64] \times 3$
Conv: $[3 \times 3, 64, 3]$	Feature concatenation	
Output image	Conv: $[3 \times 3, 64, 1]$	

Building blocks are shown in brackets (see also Figs. 2 and 3), with the numbers of blocks stacked. Downsampling is performed by convolution with a stride of 2. Upsampling is performed by convolution followed by pixel shuffle. The convolutional layer parameters are denoted as "<convolutional kernel size>, <input channel size>, and <output channel size>" within square brackets. The ReLU activation function is not shown for brevity

The modulation block uses a cascade of several convolutional layers (followed by ReLU [32] activation) to learn adaptive geometric features from the depth map. The learned features are then fed into a mid-level layer of the main SR block. The mid-level feature layer is used for modulation because it has a medium size of receptive field to include the local image patterns. The modulation is implemented as *element-wise products* between the two feature maps. Since each individual convolutional feature can be seen as (non-linear) combinations of neighboring pixels within the receptive field, the modulation in fact weights the individual convolutional features differently. The purpose is to adaptively adjust the convolutional features according to the geometric features learned from the depth map. This gives access to blend the useful 3D geometric information into the main SR task.

3.2 Loss for training the network

Recently many state-of-the-art works [1, 6, 28, 33, 34] on SR focus not only on reducing the *mean square error* (MSE) between the reconstructed and HR images, but also on better visualized results containing high-frequency details. As a result, they use different kinds of loss functions, e.g., the perceptual loss [6] and the adversarial loss [35]. However, there is a trade-off [20] between these pursued goals. In this work, we aim to enhance the performance of face SR in the sense of MSE, since this will also set a higher baseline for balancing results of good perceptual qualities.

Depth guided loss. The depth guided loss is used for the reconstruction of the guiding depth map. We employ the pixelwise MSE loss as:

$$L_{depth} = \left\| I_{depth} - \hat{I}_{depth} \right\|_F^2 = \left\| I_{depth} - D(I_{LR}) \right\|_F^2, \tag{1}$$

where I_{depth} and $\hat{I}_{depth}$ are the supervisory and estimated depth map, respectively, $D(\cdot)$ is the network of the depth block, and I_{LR} is the input LR image.

SR loss. We also adopt the common pixel-wise MSE loss as the SR loss:

$$L_{SR} = \| I_{HR} - I_{SR} \|_F^2, \tag{2}$$

where I_{HR} and I_{SR} are the HR image and super-resolved image by the whole network, respectively.

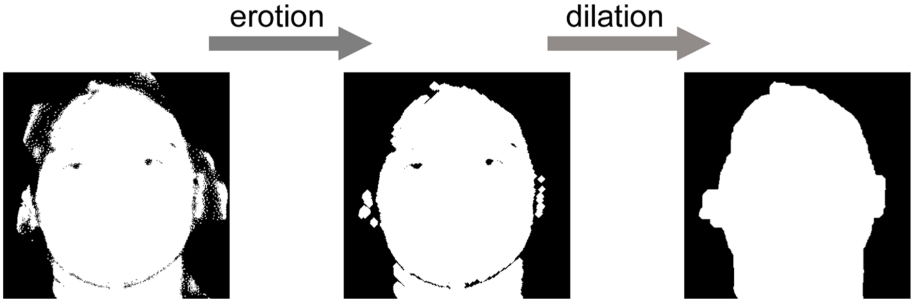

Fig. 4 Image morphological operations. An example of image dilation and erosion operation to remove the outliers and holes for a depth mask map

Total loss. We set the total loss function as the combination of the depth guided loss and the SR loss:

$$L_T = \alpha L_{depth} + L_{SR}, \tag{3}$$

where α is an adjustable parameter. In our experiments, the quantitative results on the validation set are almost stable for a wide range of α only if the two losses are numerically comparable, and we set $\alpha = 5$. The whole network is implemented in an end-to-end trainable manner with the total loss in Eq. 3 (also refer to Fig. 2).

3.3 Preparing the depth maps

In this section, we provide two ways for the preparation of matched depth maps for training the face SR network.

3.3.1 Processing matched depth scans

One way is to directly get the depth data from raw RGB-D cameras. Since the raw depth scans may contain many outliers and holes, we use image dilation and erosion operations to remove the outliers and fill the holes. Image dilation and erosion are basic morphological operations used to handle raw depth scans in this study. Figure 4 shows an example. We first use erosion operations to remove the outliers for the valid region (mask) for the depth scans and then use dilation operations to fill the holes. The structuring array used in this work is as follows.

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{4}$$

After the dilation and erosion are operated on the facial mask, we use bicubic interpolation to recover the missing pixels within the mask, as shown in Fig. 5.

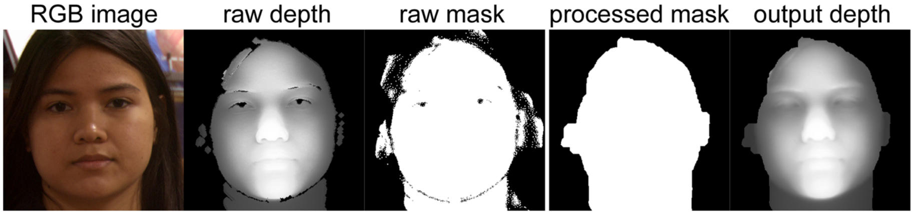

Fig. 5 Depth map processing. An example of depth map interpolation for RGB-D scans based on the processed mask

This generally provides high-quality depth maps but is restricted to matched RGB-D datasets.

3.3.2 3D face depth reconstruction

The other way is to use a prior 3D face model [36, 37] to reconstruct the shape of a facial image.

We assume that the 3D facial shape has a Gaussian distribution which can be expressed by a linear principal component analysis (PCA) model. Specifically, let the shape vector of a face be $S = (x_1, y_1, z_1, ..., x_l, y_l, z_l)^T \in \mathbb{R}^{3l}$, where l is the number of 3D vertices. After Procrustes alignment, we conduct standard PCA analysis to the concatenated vectors of shapes for all exemplar 3D facial shapes. A new facial shape therefore can be represented as eigenvectors s_i (in descending order according to their eigenvalues) of the covariance matrices, as

$$S = S_\beta = \bar{S} + \sum_{i=1}^{3l-1} \beta_i s_i, \quad \beta_i \sim N(0, \sigma_i), \tag{5}$$

where $\bar{S}$ is the average shape, β_i is the coefficient of each orthogonal base, and σ_i is the variance of Gaussian distribution. This is actually a linear model for the representation of a facial shape.

Then, the depth map can be calculated from the resulted pose and shape parameters. The reconstruction process is to minimize the following function:

$$E(\beta, s, R, T) = \sum_i \left\| x_i - SOP(R, T, s, S_\beta) \right\|^2, \tag{6}$$

where x_i is the value for each pixel and $SOP(\cdot)$ denotes scaled orthographic projection with respect to the rotation R, translation T, scaling factor s, and 3D shape S_β as

$$SOP(R, t, s, S_\beta) = s \left[\begin{array}{cccccc} 1 & 0 & 0 & 0 & 1 & 0 \end{array} \right] RS_\beta + T. \tag{7}$$

Finally, we convert the 3D shapes to depth maps with Z-buffer technology [38] and feed them into the training of the proposed network. Figure 6 shows an example of the reconstructed depth map. This way generates reasonable depth maps but lacks high-frequency details restricted by the specific 3D face prior models [36, 39–43].

4 Experiments

Datasets. We carry out our experiments on two publicly available datasets. (1) *FRGCv2.0* [14] is one of the largest RGB-D datasets with matched depth maps. We take out some problematic faces from the 4007 samples of this dataset and retain the other 3683 faces

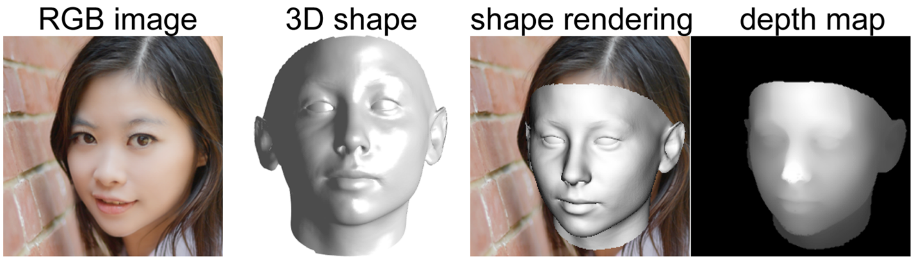

Fig. 6 Depth map by 3D reconstruction. An example of reconstructed depth map from a facial image by 3D reconstruction methods

of good quality. We select 3500 images as the training set, 100 images as the validation set, and the remaining 83 images as the test set. We then crop the images to obtain the facial regions according to the facial landmarks provided by [44]. The cropped images of faces are then resized to the resolution 256 × 256 with matched depth maps. (2) *FFHQ* [15] is a recently released high-quality dataset. We select 5000 images as the training set, 100 images as the validation set, and another 100 images as the test set. We then resize the original images to the resolution 256 × 256. The matched depth maps for training are synthesized by a 3D prior face model as described in Section 3. For both datasets, the HR images are downsampled to obtain the LR images of resolution 64 × 64 with the *bicubic* kernel.

Training details. We implement the networks with the Pytorch platform. The proposed method is trained with the Adam optimizer [45] with an initial learning rate of 2×10^{-4}. The mini-batch size is set to 32. Other parameters of the optimizer follow the default settings in Pytorch. The learning rate of the proposed network is decayed by a factor 10 every 100 epoches at a total number of 300 epoches. For the networks of other methods, we fine tune the learning rates to achieve the best performances. It takes $\sim$ 3 h on a GPU of GTX2080Ti to train the proposed model on the image resolution of 256×256.

Evaluation metrics. In the experiments, we adopt three quantitative metrics for the evaluation of the SR reconstructed results: (1) the peak signal to noise ratio (*PSNR*) measures the pixel-wise similarity between the reconstructed image and the HR reference image, (2) the structural similarity index measure (*SSIM*) [46] considers the local structural similarity between the two images, and (3) the perceptual index (**PI**) [20] prefers the high-frequency details as a non-reference metric that is different to the reference metrics such as PSNR and SSIM.

4.1 Learned depth maps

We expect that the depth block of the network can learn the depth map from the original LR image. The output of the trained depth block is extracted to validate it. Figure 7 shows some examples of the learned depth maps from the validation set, together with the ground-truth ones. We can see that the depth map is learned reasonably compared to the ground-truth. The learned depth maps show similar structures with the ground-truth ones in their geometric information. This demonstrates that the 3D shape of face can be well learned and further integrated into the main SR task.

4.2 Quantitative evaluations

We compare the proposed method with the state-of-the-art SR networks, in terms of PSNR, SSIM, and PI. We train the networks of VDSR [17], SRResNet [6], and RDN [47] with the same RGB images from the FRGCv2.0 and FFHQ dataset. We do not include some recently proposed networks for specific face SR [23, 24] because they generally focus on visual qualities instead of quantitative metrics. Table 2 summarizes the quantitative results. We also include a *trimmed* version of the proposed network with only the main SR block as the baseline network for comparison. The results show that the modulation with adaptive geometric features leads to superior results over the baseline methods in quantitative metrics. All the evaluation metrics have achieved significant improvements by the proposed method for the FRGC v2.0 dataset with matched depth scans. For example, the PSNR shows 0.40 dB gains over the baseline network. Although the gain in PSNR

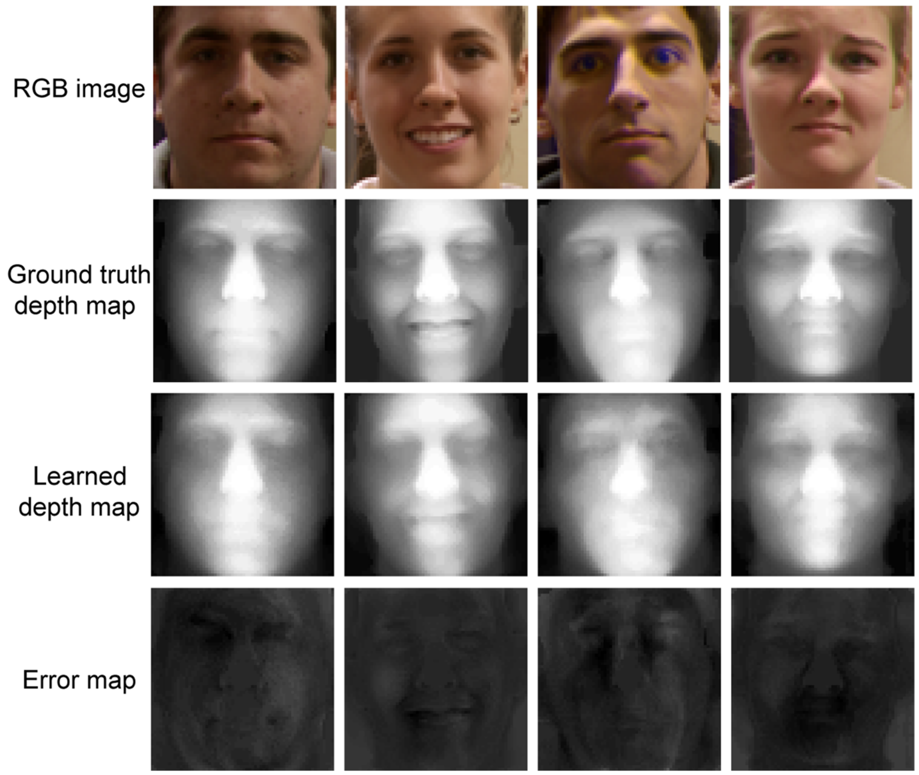

Fig. 7 Results of learned depth maps. Some examples of the learned depth maps in comparison with the ground truths. From top to bottom: the original RGB images, the supervisory depth maps (as the ground-truths), the learned depth maps, and the error maps between the learned depth maps and the ground-truths

for the FFHQ dataset with synthetic depth is only 0.12 dB, the resulted PI index achieves considerable improvement, which shows great advantages for *sharper edges* as in Fig. 8.

4.3 Qualitative evaluations

We show two examples from the two datasets in Fig. 8, marked with individual evaluation metrics. While the state-of-the-art methods such as SRResNet and RDN, and our trimmed SR network are competitive with each other for individual test samples, we find that the intervention with the depth map leads to almost uniformly superior results for the FRGC v2.0 dataset. The regions in the boundary of the face are remarkably sharper than

Table 2 Quantitative result

Method\dataset	FRGC v2.0			FFHQ		
	PSNR	SSIM	PI	PSNR	SSIM	PI
Bicubic	34.96	0.9119	9.5696	28.04	0.8224	7.6602
VDSR [17]	36.66	0.9237	8.3482	29.49	0.8584	6.4264
SRResNet [6]	37.20	0.9315	8.3070	30.07	0.8717	5.9043
RDN [47]	37.37	0.9328	8.2789	30.23	0.8722	6.0352
Ours (trimmed)	37.32	0.9324	8.2755	30.19	0.8720	5.9586
Ours (full)	37.72	0.9413	8.0159	30.31	0.8734	5.5071
HR	$+\infty$	1.0000	7.0542	$+\infty$	1.0000	4.5806

Comparisons on the test set for the scaling factor ×4. Red/blue color indicates the best/second best performance

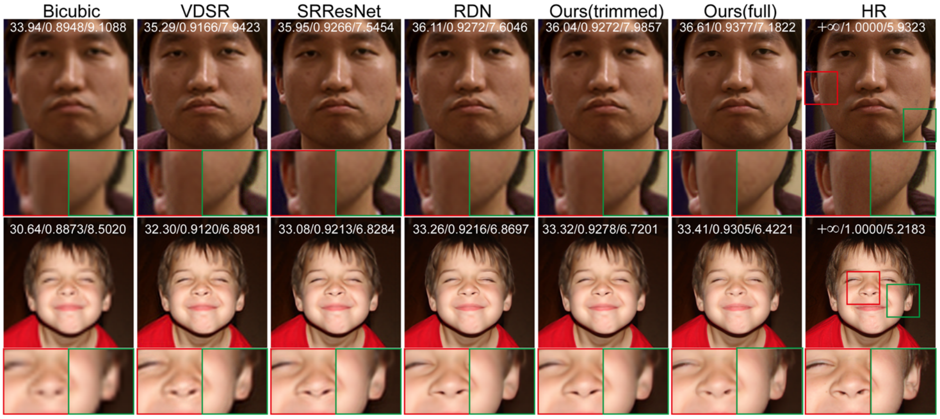

Fig. 8 Qualitative results. Two examples with the individual quantitative metrics from the FRGCv2.0 dataset (upper) and FFHQ dataset (lower), respectively. Please zoom in and pay attention to the *sharp edges* achieved by the proposed network

that obtained by the other methods for both of the two datasets. We owe these improvements to the distinction of 3D coordinates by virtue of the learned geometric features. This proves that the proposed method has positive effect on the face SR task, especially for sharper details of edges.

4.4 Ablation study

In this work, we propose the auxiliary depth and modulation blocks for the main SR task. We conduct two additional comparative experiments on the FRGC v2.0 validation set to demonstrate the effect of the auxiliary blocks. We also provide two ways for the preparation of matched depth maps, and an additional experiment with different supervisory depth maps is carried out for the ablation study.

First, we exclude the auxiliary depth and modulation blocks to get a trimmed version of the proposed network. The purpose is to learn the effect of the added network structure. We train the trimmed network with the same settings and report the quantitative results on the validation set in Table 3. It shows that removing the depth and modulation blocks leads to 0.44 dB drop of PSNR, which demonstrates the added network structure has positive effect on the final results.

Then, we retain the whole network but exclude the guided depth loss for training. This is equal to the setting $\alpha = 0$. This experiment aims to demonstrate the effectiveness of the guided geometric features without regard to the network architecture. The result in Table 3 shows 0.56 dB drop of PSNR without the guided loss even with the auxiliary blocks, which demonstrates that the supervisory depth actually contributes to the improvement of the face SR task.

Table 3 Ablation studies for the network architectures

Component	Comparison		
Auxiliary blocks	✓	×	✓
Depth-guided loss	✓	×	×
PSNR	37.97	37.53	37.41

Comparisons of results with the added blocks and depth-guided loss, respectively

Finally, we conduct an additional experiment on the FRGC v2.0 dataset. The supervisory depth maps are the matched raw scans, the reconstructed ones, and the preprocessed ones from the raw scans, respectively. Table 4 shows the results in terms of PSNR using different depth maps. It shows that the raw depth map after preprocessing leads to superior results than both the raw one and the reconstructed one. This indicates that removing the outliers and holes for raw depth scans is beneficial for robust training of the network. Although the reconstructed depth maps show superior results over the baseline SR network, it is inferior to that by the matched depth scans. This may result from reconstruction errors to the ground-truth. Thus, the proposed method prefers raw depth maps from matched RGB-D scans.

5 Discussion

There are some remaining problems for the depth maps used to train the proposed network. First, the raw depth scans of the current hardware devices usually contain a lot of noise and errors. In this paper, we use morphological image operations to suppress the noise, and we also use bicubic interpolation to fill the “holes”. Although these operations largely relieve the depth data from noise and errors, the depth scans are still far from the ground-truth ones. This should be a main factor hindering the performance for the proposed method. It is useful to collect a dataset with more advanced devices to train the proposed network for better performance. Secondly, the depth maps obtained by 3D face reconstruction methods are limited by the prior model, which lacks high-frequency details and novel structures that cannot be expressed by the model. Combining the prior model and raw depth scans is a possible way to construct closer depth map to the ground-truth, which promotes better solutions for the face SR problem guided by the depth maps. Finally, the improved performance with the depth map is at the cost of extra network architectures for the inference of depth map. Better network architectures can be explored to incorporate the depth information into the SR task.

It is worth mentioning that the success of the proposed method is not achievable without the use of large amount of available data, like most deep learning-based methods. The developments of effective data mining [48–50] and computing technologies [51–54] will push the face enhancement methods to real-life applications. Also, the prevalence of portable RGB-D scanning devices (e.g., iPhoneX) will provide more data and platforms for these methods. In addition, this model may incorporate with the state-of-the-art GAN model for a better visual performance based on the improved quantitative performance. In the future, we will develop more effective ways to incorporate depth information into the SR task, and specific data processing methods for various applications of face SR.

6 Conclusion

In this paper, we propose to use adaptive geometric features for the modulation of the face SR task. The face image is a special case of the general images that has limited geo-

Table 4 Ablation studies for the supervisory depth maps

Depth maps	Reconstructed ones	Raw scans	Preprocessed scans
PSNR	37.66	37.78	37.97

Comparisons of results with the reconstructed depth maps, those from the raw scans, and the preprocessed ones, respectively metric variations. We design a specific network structure to estimate the depth map from a facial image and then use it to produce adaptive geometric features to modulate the mid-level features for the main SR task. The supervisory depth map is either a matched one from RGB-D scans or a reconstructed one by a 3D prior model of faces. The experiments demonstrate that the acquired SR results are superior to the state-of-the-art works without depth guidance, especially with the help of real matched depth maps. We hope that the fast development and widely promotion of RGB-D cameras will lead to better solutions and applications for the face SR problem.

Abbreviations

HR:High-resolution; LR:Low-resolution; SR; Super-resolution; CNNs: Convolutional neural networks; MSE: Mean square error; PSNR: Peak signal to noise ratio; SSIM: Structural similarity index measure; PI: Perceptual index; PCA: Principal component analysis

Acknowledgements

This work is supported in part by the National Key R&D Program of China (2017YFC0803505) and the Natural Science Foundation of China (61571438).

Authors' contributions

Zhenfeng Fan and Xiyuan Hu are the principal contributors in terms of the key ideas and the experimental results. Chen Chen and Xiaolian Wang contribute to the writing of this manuscript together with the main authors. In a supervising role, Silong Peng contributes to the check of the final manuscript and is in charge of the funding for this research. The authors read and approve the final manuscript.

Funding

This work is supported in part by the National Key R&D Program of China (2017YFC0803505) and the Natural Science Foundation of China (61571438).

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Competing interests

The authors declare that they have no competing interests.

Received: 27 March 2020 Accepted: 6 July 2020
Published online: 17 July 2020

References

1. A. Bulat, G. Tzimiropoulos, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans, (2018), pp. 109–117. https://doi.org/10.1109/cvpr.2018.00019
2. F. Liu, D. Zeng, Q. Zhao, X. Liu, in *European Conference on Computer Vision*, Joint face alignment and 3D face reconstruction, (2016), pp. 545–560. https://doi.org/10.1007/978-3-319-46454-1_33
3. P. Li, L. Prieto, D. Mery, P. J. Flynn, On low-resolution face recognition in the wild: comparisons and new techniques. IEEE Trans. Inf. Forensic Secur. **14**(8), 2000–2012 (2019)
4. J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, Y. Xu, Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. IEEE Trans. Pattern Anal. Mach. Intell. **39**(1), 156–171 (2016)
5. Y. Huang, Y. Chai, Y. Liu, J. Shen, Architecture of next-generation e-commerce platform. Tsinghua Sci. Technol. **24**(1), 18–29 (2018)
6. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Photo-realistic single image super-resolution using a generative adversarial network (IEEE Computer Society, 2017), pp. 4681–4690. https://doi.org/10.1109/CVPR.2017.19
7. Y. Wang, L. Wang, H. Wang, P. Li, Information-compensated downsampling for image super-resolution. IEEE Sig. Process. Lett. **25**(5), 685–689 (2018)
8. D. L. Cosmo, E. O. T. Salles, Multiple sequential regularized extreme learning machines for single image super resolution. IEEE Sig. Process. Lett. **26**(3), 440–444 (2019)
9. W. Yang, W. Wang, X. Zhang, S. Sun, Q. Liao, Lightweight feature fusion network for single image super-resolution. IEEE Sig. Process. Lett. **26**(4), 538–542 (2019)
10. C. Ren, X. He, Y. Pu, Nonlocal similarity modeling and deep CNN gradient prior for super resolution. IEEE Sig. Process. Lett. **25**(7), 916–920 (2018)
11. D. Fan, S. Fang, G. Wang, S. Gao, X. Liu, The visual human face super-resolution reconstruction algorithm based on improved deep residual network. EURASIP J. Adv. Sig. Process. **2019**, 32 (2019)
12. Y. Chen, Y. Tai, X. Liu, C. Shen, J. Yang, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, FSRNet: end-to-end learning face super-resolution with facial priors, (2018), pp. 2492–2501. https://doi.org/10.1109/cvpr.2018.00264

13. X. Yu, B. Fernando, B. Ghanem, F. Porikli, R. Hartley, in *European Conference on Computer Vision*, Face super-resolution guided by facial component heatmaps, (2018), pp. 217–233. https://doi.org/10.1007/978-3-030-01240-3_14
14. P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1*, Overview of the face recognition grand challenge, (2005), pp. 947–954. https://doi.org/10.1109/cvpr.2005.268
15. T. Karras, S. Laine, T. Aila, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, A style-based generator architecture for generative adversarial networks, (2019), pp. 4401–4410. https://doi.org/10.1109/cvpr.2019.00453
16. C. Dong, C. C. Loy, K. He, X. Tang, in *European Conference on Computer Vision*, Learning a deep convolutional network for image super-resolution (Springer, 2014), pp. 184–199. https://doi.org/10.1007/978-3-319-10593-2_13
17. K. Jiwon, J. Kwon Lee, K. Mu Lee, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Accurate image super-resolution using very deep convolutional networks, (2016), pp. 1646–1654. https://doi.org/10.1109/cvpr.2016.182
18. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
19. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Deep residual learning for image recognition, (2016), pp. 770–778. https://doi.org/10.1109/cvpr.2016.90
20. Y. Blau, T. Michaeli, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, The perception-distortion tradeoff, (2018), pp. 6228–6237. https://doi.org/10.1109/cvpr.2018.00652
21. B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, R. M. Mersereau, Eigenface-domain super-resolution for face recognition. IEEE Trans. Image Process. **12**(5), 597–606 (2003)
22. W. W. Zou, P. C. Yuen, Very low resolution face recognition problem. IEEE Trans. Image Process. **21**(1), 327–340 (2011)
23. E. Zhou, H. Fan, Z. Cao, Y. Jiang, Q. Yin, in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, Learning face hallucination in the wild (AAAI Press, 2015). http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9752
24. Y. Song, J. Zhang, L. Gong, S. He, L. Bao, J. Pan, Q. Yang, M.-H. Yang, Joint face hallucination and deblurring via structure generation and detail enhancement. Int. J. Comput. Vis. **127**(6-7), 785–800 (2019)
25. S. Zhu, S. Liu, C. C. Loy, X. Tang, in *European Conference on Computer Vision*, Deep cascaded bi-network for face hallucination (Springer, 2016), pp. 614–630. https://doi.org/10.1007/978-3-319-46454-1_37
26. X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, R. Yang, in *European Conference on Computer Vision*, Learning warped guidance for blind face restoration, (2018), pp. 272–289. https://doi.org/10.1007/978-3-030-01261-8_17
27. B. Dogan, S. Gu, R. Timofte, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Exemplar guided face image super-resolution without facial landmarks, (2019), pp. 0–0. https://doi.org/10.1109/cvprw.2019.00232
28. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, in *European Conference on Computer Vision Workshops*, ESRGAN: enhanced super-resolution generative adversarial networks (Springer, 2018), pp. 0–0. https://doi.org/10.1007/978-3-030-11021-5_5
29. B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Enhanced deep residual networks for single image super-resolution, (2017), pp. 1132–1140. https://doi.org/10.1109/cvprw.2017.151
30. O. Ronneberger, P. Fischer, T. Brox, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, U-net: convolutional networks for biomedical image segmentation, (2015), pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
31. Z. Chen, V. Badrinarayanan, G. Drozdov, A. Rabinovich, in *European Conference on Computer Vision*, Estimating depth from RGB and sparse sensing, (2018), pp. 167–182. https://doi.org/10.1007/978-3-030-01225-0_11
32. V. Nair, G. E. Hinton, in *Proceedings of the International Conference on Machine Learning*, Rectified linear units improve restricted Boltzmann machines (Omnipress, 2010), pp. 807–814. https://icml.cc/Conferences/2010/papers/432.pdf
33. X. Deng, Enhancing image quality via style transfer for single image super-resolution. IEEE Sig. Process. Lett. **25**(4), 571–575 (2018)
34. J. W. Soh, G. Y. Park, J. Jo, N. I. Cho, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Natural and realistic single image super-resolution with explicit natural manifold discrimination, (2019), pp. 8122–8131. https://doi.org/10.1109/cvpr.2019.00831
35. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, in *Advances in Neural Information Processing Systems*, Generative adversarial nets (Curran Associates., 2014), pp. 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarialnets
36. V. Blanz, T. Vetter, in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, A morphable model for the synthesis of 3D faces, (1999), pp. 187–194. https://doi.org/10.1145/311535.311556
37. S. Romdhani, Face image analysis using a multiple features fitting strategy. PhD thesis, University of Basel (2005)
38. W. Straßer, Schnelle kurven-und flächendarstellung auf grafischen sichtgeräten. PhD thesis (1974)
39. J. Booth, A. Roussos, A. Ponniah, D. Dunaway, S. Zafeiriou, Large scale 3D morphable models. Int. J. Comput. Vis. **126**(2-4), 233–254 (2018)
40. Z. Fan, X. Hu, C. Chen, S. Peng, in *Proceedings of the European Conference on Computer Vision*, Dense semantic and topological correspondence of 3D faces without landmarks, (2018), pp. 523–539. https://doi.org/10.1007/978-3-030-01270-0_32
41. T. Bolkart, S. Wuhrer, in *Proceedings of the IEEE International Conference on Computer Vision*, A groupwise multilinear correspondence optimization for 3D faces, (2015), pp. 3604–3612. https://doi.org/10.1109/iccv.2015.411
42. A. Patel, W. A. Smith, in *Proceedings of the IEEE International Conference on Computer Vision*, 3D morphable face models revisited, (2009), pp. 1327–1334. https://doi.org/10.1109/cvpr.2009.5206522
43. Z. Fan, X. Hu, C. Chen, S. Peng, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boosting local shape matching for dense 3D face correspondence, (2019), pp. 10944–10954. https://doi.org/10.1109/cvpr.2019.01120

44. C. Creusot, N. Pears, J. Austin, A machine-learning approach to keypoint detection and landmarking on 3D meshes. Int. J. Comput. Vis. **102**(1-3), 146–179 (2013)
45. D. P. Kingma, J. Ba, in *International Conference on Learning Representations*, Adam: a method for stochastic optimization (Elsevier, 2015). http://arxiv.org/abs/1412.6980
46. Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al., Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
47. Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Residual dense network for image super-resolution (IEEE Computer Society, 2018), pp. 2472–2481. https://doi.org/10.1109/CVPR.2018.00262
48. H. Zhu, W. Hu, Y. Zeng, in *CCF International Conference on Natural Language Processing and Chinese Computing*, Flexner: a flexible LSTM-CNN stack framework for named entity recognition (Springer, 2019), pp. 168–178. https://doi.org/10.1007/978-3-030-32236-6_14
49. X. Xu, Y. Li, T. Huang, Y. Xue, K. Peng, L. Qi, W. Dou, An energy-aware computation offloading method for smart edge computing in wireless metropolitan area networks. J. Netw. Comput. Appl. **133**, 75–85 (2019)
50. X. Xu, Q. Liu, Y. Luo, K. Peng, X. Zhang, S. Meng, L. Qi, A computation offloading method over big data for IOT-enabled cloud-edge computing. Future Gener. Comput. Syst. **95**, 522–533 (2019)
51. L. Qi, W. Dou, W. Wang, G. Li, H. Yu, S. Wan, Dynamic mobile crowdsourcing selection for electricity load forecasting. IEEE Access. **6**, 46926–46937 (2018)
52. L. Qi, X. Zhang, W. Dou, Q. Ni, A distributed locality-sensitive hashing-based approach for cloud service recommendation from multi-source data. IEEE J. Sel. Areas Commun. **35**(11), 2616–2624 (2017)
53. L. Qi, W. Dou, Y. Zhou, J. Yu, C. Hu, A context-aware service evaluation approach over big data for cloud applications. IEEE Trans. Cloud Comput. (2015). https://doi.org/10.1109/tcc.2015.2511764
54. G. Li, S. Peng, C. Wang, J. Niu, Y. Yuan, An energy-efficient data collection scheme using denoising autoencoder in wireless sensor networks. Tsinghua Sci. Technol. **24**(1), 86–96 (2018)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.