Reference-Guided Deep Super-Resolution via Manifold Localized External Compensation

Wenhan Yang, *Student Member, IEEE*, Sifeng Xia, Jiaying Liu, *Senior Member, IEEE*, and Zongming Guo, *Member, IEEE*

***Abstract*— The rapid development of social network and online multimedia technology makes it possible to address traditional image and video enhancement problems, with the aid of online similar reference data. In this paper, we tackle the problem of super-resolution (SR) in this way, specifically aiming to handle the "one-to-many" problem between the image patches of low resolution (LR) and high resolution (HR). We propose a manifold localized deep external compensation (MALDEC) network to additionally utilize reference images, *i. e.*, retrieved similar images in cloud database and reference HR frame in a video, to provide an accurate localization and mapping to the HR manifold, and compensate the lost high-frequency details. The proposed network employs a three-step recovery: 1) internal structure inference, which uses the LR image itself and the internally inferred high frequency information to preserve main structure of the HR image; 2) manifold localization, which localizes the HR manifold and constructs the correspondence between the internal inferred image and the external images; and 3) external compensation, which introduces the external references of retrieved similar patches based on manifold localization information to reconstruct the high-frequency details. The learnable components of MALDEC, internal structure inference, and external compensation, are trained jointly to make a good tradeoff between these two terms for an optimal SR result. Finally, the proposed method is examined under three tasks: cloud-based image SR, multi-pose face reconstruction, and reference frame-guided video SR. Extensive experiments demonstrate the superiority of our method than the state-of-the-art SR methods in both objective and subjective evaluations, and our method offers new state-of-the-art performance.**

***Index Terms*— Super-resolution, manifold localization, external compensation, internal structure inference.**

I. Introduction

IMAGE super-resolution (SR) aims to estimate a high-resolution (HR) image from low-resolution (LR) observations. Due to the quality degradation in acquisition, saving and storage processes, LR images lose parts of high-frequency information, which makes single image SR an ill-posed problem. Thus, the key issue is to impose a *priori* knowledge to regularize the SR recovery and to localize the manifold of HR images.

The earliest upscaling methods model the missing HR pixels as the polynomial function of known LR pixels and their relative locations. They have relatively low complexity, and however generate blurred results with ring artifacts. Later works adopt a maximum posterior probability (MAP) framework to constrain the estimation of the HR image. They describe various desirable properties of natural images with priors in the form of regularization terms. Typical regularizations, including gradient [1], nonlocal [2] and total variation [3], make the SR reconstructed results closer to the HR image manifold in visual appearances, *i.e.* sharp edges and alleviation of ring artifacts. However, these regularization terms are designed heuristically, and thus are hard to represent diversified patterns or features of natural images.

Learning-based methods obtain prior knowledge and build the mapping from the LR and HR space by training on a large dataset. Compared with previous methods, they present visually promising results with rather high computational efficiency. Several machine learning models and their variants are used to build the mapping, including sparse representation [4]–[7], neighbor embedding, anchor regression [8], random forest [9], and deep networks [10]–[13] *et al*. Especially, combining the effectiveness of data-driven priors extracted from large scale training sets and the powerful modeling capacity of deep network, the recently proposed deep-based methods [10]–[13] offer rather impressive SR performance. SRCNN [10] is a seminal work. It is a three-layer CNN, which performs the SR process following an equivalent process to sparse coding reconstruction. The sparse prior [12] is explicitly incorporated into the network with learned iterative shrinkage and thresholding unit. In [14], very deep convolutional network is constructed by stacking many 3 × 3 convolutions to obtain a very large receptive field, which brings in significant performance gain.

Despite the impressive results obtained by the learning-based methods with the adaptive learned priors, especially deep learning-based approaches, they still suffer from the ill-posed nature of image SR. Constrained by mean square error, learning-based methods cannot accurately localize the natural image manifold and usually drop into the problem of "regression to mean" [15] that the mean of several similar HR

Manuscript received September 28, 2017; revised March 3, 2018 and April 24, 2018; accepted May 10, 2018. Date of publication May 18, 2018; date of current version May 3, 2019. This work was supported in part by the National Natural Science Foundation of China under Contract 61772043 and in part by the CCF-Tencent Open Research Fund. This paper was recommended by Associate Editor Z. Wang. *(Corresponding author: Jiaying Liu.)*

The authors are with the Institute of Computer Science and Technology, Peking University, Beijing 100080, China (e-mail: yangwenhan@pku.edu.cn; sfxia18@163.com; liujiaying@pku.edu.cn; guozongming@pku.edu.cn).

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/TCSVT.2018.2838453

patches other than an HR patch is regressed. To address this, two branches of methods were proposed. One is perception-loss based image hallucination [16], [17]. These works aim to use deep networks to construct the distance metric simulating human perception to constrain the SR reconstruction. They generate visually pleasing but inauthentic high-frequency details, which may breakdown the usability of practical applications, such as criminal detection or action recognition in surveillance videos.

The other is cloud-based image enhancement. Nowadays, we have been in a new era of computational imaging. The popular online image sharing communities and pervasive cameras generate extensive images and videos everyday. Thus, it has become feasible to enhance a low quality image or video with similar high-quality data retrieved online, *i. e.* the images from different views or the portrait images with different poses. The side information from the cloud database provides effective clues to infer the HR image manifold and to recover the missing details in the degradation. For image SR, the retrieved data provides context of a local patch, *i. e.*, semantic attributes and texture features, to localize the HR manifold, and enables a customized high-frequency detail enhancement based on online retrieved samples. Yue *et al.* [18] firstly explored the cloud-based image SR in a patch matching and fusion framework. In [19] and [20], online retrieved data is utilized to facilitate a more accurate parameter estimation in a sparse representation model. Timofte *et al.* [15] proposed to employ semantic annotations to benefit image super-resolution.

However, these methods neglect several important issues. First, the retrieved similar patches are fused by shallow models, which are insufficient to handle the complex dependency between internal structures of LR images and high-frequency details extracted from external references. Second, in these hand-crafted approaches, the detail extraction and fusion are designed separately, which sets a barrier to accurate manifold localization. Thus, these methods may generate inaccurate introduced high-frequency details including noises and artifacts. Third, these approaches contain lots of parameters in each step, they need to be carefully tuned by hand to achieve promising performance, which increases their difficulties to handle real applications.

In this paper, we follow the cloud-based SR methods and propose a **MA**nifold **L**ocalized **D**eep **E**xternal **C**ompensation (**MALDEC**) to utilize reference images, *i. e.* online retrieved similar images or those from different views to facilitate the image SR. **MALDEC** jointly models and optimizes the internal structure inference and external compensation.

Contributions of this paper lie in three aspects:

- This is the first work that uses deep networks to jointly model the structure inference and external details compensation for image SR. The high-frequency details are transferred based on the manifold localization considering both inferred structural information and external data. Thus, the SR results not only accord with the internal structures of images, but as well are close to the HR image manifold.
- The internal structure inference and external compensation networks are jointly trained with various kinds of degradations and wrong matchings, to make a good trade-off between these two components for an accurate SR recovery.
- The extensive evaluations on three tasks: cloud-based image SR, multi-pose face image reconstruction and reference frame-guided video SR, show significant superiority over previous deep learning-based and cloud-based methods objectively and subjectively.

The rest of this paper is organized as follows. Section II briefly reviews the related work. Section III illustrates our MALDEC. Section IV presents its three applications, including cloud-based image SR, multi-pose face image reconstruction and reference frame-guided video SR. Experimental results and concluding remarks are presented in Sections V and VI, respectively.

II. Related Work

A. Deep Learning Single Image SR

Dong *et al.* [10] made the first attempt of deep-based image SR by constructing a three-layer CNN. Instead of using a generic CNN model, Wang *et al.* [12] incorporated the sparse prior into CNN with learned iterative shrinkage and thresholding algorithm. In [14] and [21], very deep convolutional networks are constructed to obtain a very large receptive field, which brings in significant performance gain. In [13], the sub-band recovery is built with edge guidance, leading to better high-frequency detail recovery. To accelerate this SR process, in [22] and [23], the feature extraction and transformation are performed in LR space instead of HR space. To regularize the deep network to generate visually pleasing results, perceptive loss [16] and adversarial generative network [17] are proposed to constrain the SR process. In this paper, we propose to generate both visually pleasing and objectively authentic HR estimation under the guidance of manifold localization based on both the internally inferred image and reference images.

B. Data-Driven Based Image Processing

Several methods exploit high-level information, such as semantic attributes and geometric features, to facilitate image SR. Yue *et al.* [18], [24] first explored the problem of cloud-based image SR and denoising, and addressed them by utilizing online retrieved similar data to compensate lost details. Li *et al.* [19] used the retrieved HR image patches to learn a more accurate distribution for sparse representation coefficients. Liu *et al.* [20] utilized the semantic information to build an adaptive group-structured sparsity model for image SR. There also are several works generating the prior knowledge online from a cloud database to regularize a series of image reconstruction applications, including JPEG image restoration [25], stylization [26], [27], color adjustment [28], joint filter [29], rain removal [30] *etc.* In this paper, we follow this path and differently we construct a deep network to jointly optimize the internal structure inference and external compensation for image SR.

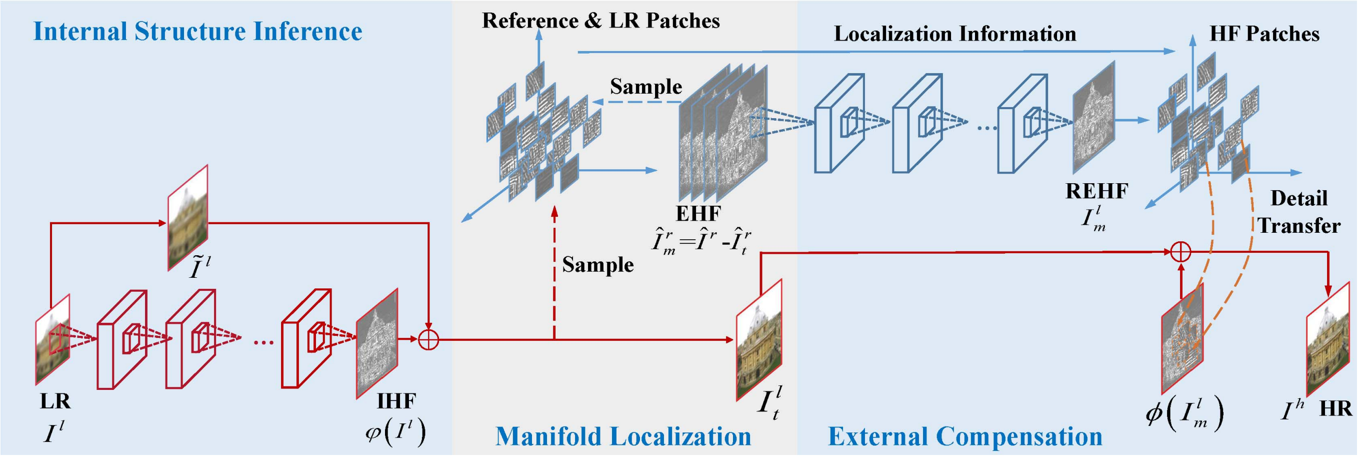


Fig. 1. Framework of the proposed three-step MALDEC to perform joint *internal structure inference* and *external compensation* guided by *manifold localization*. The internal structure inference network (ISIN) learns to predict the HR image as accurately as possible solely based on an LR image. In manifold localization, we measure the similarity between the reference patches and internally inferred patches, and construct the correspondences between them. The external compensation network (ECN) makes up residual lost high-frequency details with the retrieved HR references and manifold localization guidance. IHF, EHF, and REHF signify internal, external, and refined external maps, respectively.

C. Example-Based Image SR

Comparing with learning-based methods obtaining an HR estimation by inference, example-based approaches [31] generate an HR result by sampling. There are a large collection of coupled LR-HR patches. Based on the patch similarity in LR space, several corresponding HR patches are retrieved from the database. Then, these retrieved HR patches are fused to generate an HR estimation. Many models are used for patch fusion, including random Markov field [32], neighbor embedding [31], [33], factor graph [34] and in-place regression [35] *et al.* These models blend the sampled patches while keeping local regional consistency. In this paper, we are interested in the complementary capacity of example-based methods to learning-based approaches, and design a deep network to jointly infer high-frequency information and fuse the sampled high-frequency details to obtain a better HR estimation. In our MALDEC, the high-frequency details are transferred based on manifold localization measured by hierarchical features.

D. Video Super-Resolution

According to the way of exploiting motion information, the video SR methods can be divided into two categories: the *explicit motion-based methods* that align LR frames according to either optical flow [36], [37] or motion compensation [38] and the *implicit motion-based methods* that embed motion as a weighting term [39]–[42] or a regularization term [36], [43]–[45] for tuning the HR estimation. Recently, several deep learning methods [46]–[50] have been proposed to address the video SR problems in both explicit and implicit ways. Compared with conventional methods, in these works, CNNs and RNNs are used to model some parts of the video SR pipeline, *i.e.* feature extraction, motion compensation, and multi-frame fusions, achieving superior video SR performance. In this paper, we apply our MALDEC for reference-guide video super-resolution, where adjacent high-resolution reference frames are given. With additional reference information, our MALDEC can effectively transfer high-frequency details from references to results, leading to superior SR reconstruction performance.

III. MANIFOLD LOCALIZED DEEP EXTERNAL COMPENSATION

In this section, we first give an intuitive explanation for our MALDEC. Then, we present its overall framework as shown in Fig. 1 and illustrate the details of its each component. For readers' convenience, we provide a summarization of all symbols and denotations in Table I.

A. Joint Internal Structure Inference and External Compensation

For image SR, the main aim is to recover the missing high-frequency details. However, as shown in Fig. 2 (a), due to ambiguity of single image SR problem, solely based on LR information, only parts of lost high-frequency details can be restored. The mean of several HR patches corresponding to the LR patch is inferred (denoted in dark blue), and the difference between each HR patch (denoted in red and by black arrows) and this mean value is lost. SRGAN [17] maps the LR signal to one of these HR signals, without guarantee for fidelity, and generates the SR results presenting pixel-level detail artifacts, as shown in Fig. 2 (c).

To compensate the high frequency details, we exploit the high-level context information, *i. e.*, geometric correspondences and semantic attributes, to localize the corresponding HR signals (denoted in yellow and by the yellow arrow). We construct a three-step recovery network to infer the "structure (mean)" HR signal, and "detail (distinguished)" HR signal, and then compensate the "detail (distinguished)" HR signal, respectively. The internal structure inference predicts the HR image I^h as accurately as possible solely based on an LR image I^l. In manifold localization, we measure

TABLE I
SYMBOLS AND DENOTATIONS SUMMARIZATION

Symbol	Implication	Symbol	Implication
I^l	LR image	$\tilde{I}^l$	The image simply up-sampled from I^l
I_t^l	The intermediate result image	$\varphi(\cdot)$	The process that ISIN infers the IHF map
$\hat{I}_m^r$	Extracted external high-frequency map	I^r	HR reference
$E_s(\cdot)$	Energy function for high frequency detail transfer	I_m^l	Refined external high-frequency map
$\phi(\cdot)$	The joint process of manifold localization and detail fusion	$\psi(\cdot)$	The process that ECN extracts EHF map
$\Psi(\cdot)$	A specified set of hierarchical features	$\Phi(\cdot)$	The list of all extracted local patches
$\Phi_i(\cdot)$	Patch index	$\Phi_{NN(i)}(\cdot)$	The indexes of the best matched patches
$\hat{I}^r$	Aligned reference images	$\hat{I}_t^r$	Intermediate SR reference image
$\hat{I}_t^{r'}$	Transformed intermediate SR reference image	σ	The standard deviation of all pixels in the image
τ	The mean values of all pixels in the image	P_i	Query patch in I_t^l
Q_j^i	Candidate patch in $\hat{I}_t^{r'}$	∇	The operation to calculate the gradient of the patches
I^g	The ground truth HR image	I^{sr}	The synthesized reference image
I_l^{sr}	The down-sampled synthesized reference image	I_t^{sr}	The intermediate SR synthesized reference image
$\hat{I}_l^r$	The down-sampled reference image	x_i	HR training sample
z_i	Synthetic reference training sample	y_i	LR training sample
$\mathrm{G}(\cdot)$	The process of learned external compensation network	$\mathrm{F}(\cdot)$	The process of learned internal structure inference network
Θ_2	The parameter of external compensation network	Θ_1	The parameter of internal structure inference network

the similarity between the reference patches and internally inferred patches, and construct the correspondences between them. The external detail compensation makes up residual lost high-frequency details based on the external retrieved HR references. MALDEC presents both perceptually convincing and structurally promising results, as shown in Fig. 2 (d).

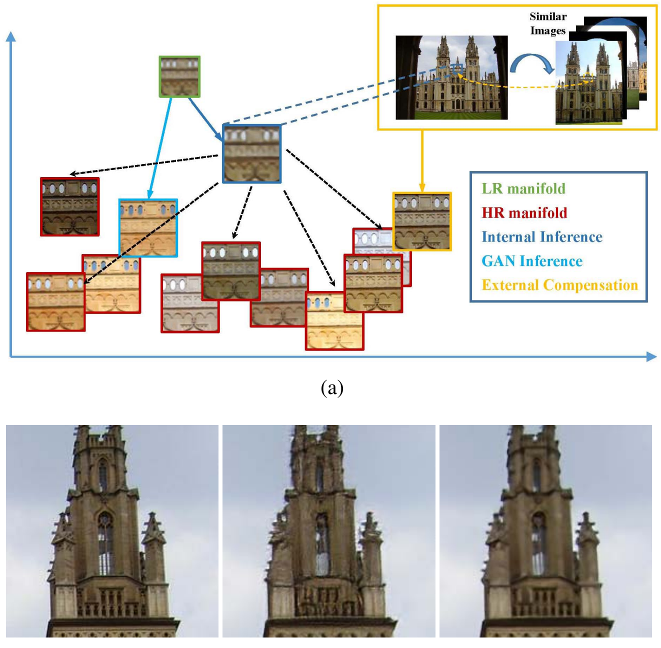
B. Internal Structure Inference Network

The first component Internal Structure Inference Network (ISIN) is utilized to initially reconstruct the HR image solely based on the LR image I^l itself. As shown in Fig. 1, ISIN takes I^l as its input and outputs an Internal High-Frequency map (IHF).

With the inferred IHF map, the intermediate result image I_t^l is then reconstructed as follows:

$$I_t^l = \tilde{I}^l \oplus \varphi(I^l), \tag{1}$$

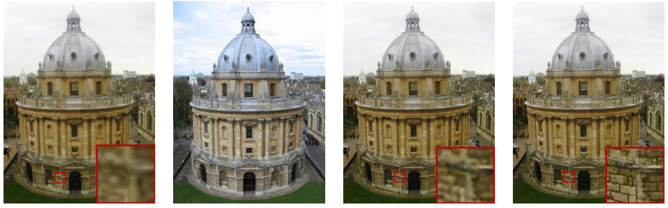
where $\oplus$ is the summation between $\tilde{I}^l$ and $\varphi(I^l)$. $\varphi(I^l)$ represents the process that ISIN infers the IHF map from the LR image I^l. $\tilde{I}^l$ is the image that simply up-sampled from I^l. In general, ISIN can take any SR network as its structure. In our paper, due to the outstanding performance, we select VDSR [14] and DEGREE [13] as our ISINs in two experimental settings, respectively. For these two network structures, we follow the original settings in [13] and [14] to use Bicubic and Nearest interpolations as their corresponding up-sampling methods to generate $\tilde{I}^l$, respectively.



(a)

(b) (c) (d)

Fig. 2. Internal methods "regress to mean" [15] and generate generally smooth estimation. GAN drives the reconstruction towards natural image manifold but presents pixel-level detail artifacts. MALDEC generates both perceptually convincing and structurally promising results. (a) Intuitive explanation for the image SR in the image manifold. (b) Ground truth. (c) SRGAN [17]. (d) MALDEC.



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

Fig. 3. An example of the reconstructed result from ISIN upscaled by 3 times with a reference image by retrieval. (a) Input LR I^l. (b) Reference I^r. (c) Result I_t^l by ISIN. (d) Ground truth HR image I^g.

C. External Compensation Network

ISIN works well in predicting the common part of the high-frequency map from an LR image. However, due to ambiguity mentioned-above, not all high-frequency details can be well recovered, as shown in Fig. 3 (c). To address the issue, we construct an External Compensation Network (ECN), as shown in Fig. 1, to further extract External High-Frequency (EHF) map $\hat{I}_m^r$, which could not be covered by ISIN, from each HR reference $\hat{I}^r$.

As shown in Fig. 3, there are usually common illumination and color variations between the input LR image (Fig. 3 (a)) and its reference image (Fig. 3 (b)). Moreover, the existing low-frequency information in the references maybe misleading to construct correspondences of HF information. Therefore, some measures are taken to improve the robustness of extracting $\hat{I}_m^r$. Firstly, in the training phase, contrast of the ground-truth reference images is additionally adjusted to simulate common illumination and color differences between the input LR image and the HR reference image. We alternatively utilize the differential image between $\hat{I}^r$ and its intermediate SR image $\hat{I}_t^r$ as the input of ECN, rather than directly input $\hat{I}^r$. $\hat{I}_t^r$ is obtained through up-sampling the down-sampled image of $\hat{I}^r$ by ISIN. The differential image is chosen because of its high efficiency in reducing illumination and color differences and removing redundant low-frequency information.

Then, ECN extracts EHF map from the input by a fully convolutional network as shown in Fig. 1. Afterwards, based on manifold localization, the EHF map is adapted to Refined External High-Frequency map (REHF) I_m^l. Finally, the reconstructed result I^h is derived by:

$$\hat{I}_m^r = \hat{I}^r - \hat{I}_t^r, \tag{2}$$

$$I_m^l = \psi(\hat{I}_m^r), \tag{3}$$

$$I^h = I_t^l \oplus \phi(I_m^l), \tag{4}$$

where ψ signifies the process that ECN extracts the high-frequency map, and ϕ signifies the joint process of manifold localization and detail fusion. The details of $\phi(\cdot)$ will be illustrated in Section III-D and Section IV.

D. Manifold Localized High-Frequency Detail Transfer

Similar to example-based approaches, in ECN, we need to transfer the high-frequency details based on the structure of LR images. To achieve this, the detail is transferred between the input patch and similar reference patches measured by hierarchical features, including many factors, *i. e.* location, intensity, and gradient. These features jointly describe the context of an patch, help localize the HR manifold accurately and facilitate the single image SR.

Let $\Phi\left(\Psi\left(I_t^l\right)\right)$ denote the list of all local patches extracted from $\Psi\left(I_t^l\right)$ – a specified set of hierarchical features of I_t^l. Each reference patch is indexed as $\Phi_i(\Psi(I_t^l))$. The high-frequency detail is transferred by minimizing the following energy function,

$$E_s\left(\Psi\left(I_t^l\right),\Psi\left(\hat{I}_t^r\right)\right)=\sum_{i=1}^{m}\left\|\Phi_i\left(\Psi\left(I_t^l\right)\right)-\Phi_{NN(i)}\left(\Psi\left(\hat{I}_t^r\right)\right)\right\|_2^2, \tag{5}$$

where $\Phi_{NN(i)}(\Psi(\hat{I}_t^r))$ is the best matched patch to $\Phi_i(\Psi(I_t^l))$ in the feature space of $\hat{I}_t^r$. m denotes the number of all local patches extracted from $\Psi\left(I_t^l\right)$. We first search the best matched location based on feature similarity measured by (5), then compensate the details with ECN. The features selected for manifold localization, and the high-frequency detail fusion process are application-dependent. Thus, the relevant details will be illustrated in Section IV.

IV. Applications:Reference Guided Image SR

In this section, we apply our MALDEC to three applications:

- Cloud-based image SR. Besides the input LR image, several HR images in the same scene from different views are available by image retrieval.
- Multi-pose face reconstruction. Besides the input LR face image, several HR face images of the same person with different poses are available.
- Reference frame-guided video SR. When constructing an LR frame, several reference HR frames at the same scene in the same video are available.

TABLE II

The Processing Flow of Cloud-Based Image SR, Multi-Pose Face Reconstruction and Reference Frame-Guided Video SR. The Phases Retrieval, Alignment and Feature Calculation Form the Manifold Localization

Steps	Hierarchical Feature	Cloud-based Image	Multi-Pose Face	Reference Video
Retrieval	\	√		
Alignment	\	√		√
Feature Calculation	Aligned Loc.	√		√
	Flow Loc.		√	
	Patch Sim.	√	√	√
	Gradient Sim.	√	√	√
Transfer	\	√	√	√

Applying MALDEC to these three scenarios, we specify the features to calculate manifold localization. These features, providing context information, are used to create correspondences for input and reference images. Then, the high-frequency details of these reference images are transferred to the super-resolved result constrained by the hierarchical feature matching. The processing flow of each task is summarized in Table II. The manifold localization decomposes into several specific steps: reference retrieval, alignment, hierarchical feature and distance calculation.

A. Reference Retrieval

For multi-pose face reconstruction and reference-frame guided video SR, the reference images with the same content and abundant high-frequency details are given. For cloud-based image SR, we first search the HR reference images following the approach in [20]. The initially reconstructed immediate SR image I_t^l is used for retrieving reference images in a dataset. A BOW [51] model that effectively encodes the patch statistic around SURF key points is built for indexing and retrieving reference images.

B. Alignment

Because the retrieved reference images are still different with I_t^l in scales and viewpoints, we first align each I^r to I_t^l. SIFT feature [52] of I_t^l and each $\tilde{I}^r$ are detected and matched. Then, RANSAC algorithm is performed over the matched points to find the homography transformation matrix, which is finally used to transform I^r to $\hat{I}^r$. For multi-pose face reconstruction, the homography matrix is set as identity matrix.

C. Hierarchical Feature and Distance Calculation

After alignment, the reference image $\hat{I}^r$ and the internal reconstruction result I_t^l are matched in a global view but still have location shifts at pixel level. Thus, the direct fusion cannot be employed, and a pixel level location matching constrained by hierarchical features is needed.

There are usually significant differences in illumination, color and resolution between the intermediate SR image I_t^l and the aligned HR references. As a result, for the purpose of better matching results we first utilize the intermediate SR reference image $\hat{I}_t^r$ of each I^r for matching, which shares similar resolution-level with I_t^l . Then, we transform each $\hat{I}_t^r$ to reduce illumination difference:

$$\hat{I}_t^{r'} = (\hat{I}_t^r - \tau(\hat{I}_t^r))\frac{\sigma(I_t^l)}{\sigma(\hat{I}_t^r)} + \tau(I_t^l), \tag{6}$$

where $\hat{I}_t^{r'}$ is the transformed result, $\tau(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation values of all pixels in the image, respectively.

Then, based on $\hat{I}_t^{r'}$, we calculate hierarchical features, which cover many factors to facilitate the location matchings:

a) Aligned Location: The pixel index j in $\hat{I}_t^{r'}$ and i in I_t^l are selected to represent the spatial locations of two patches, which can be further used to measure the spatial proximity.

b) Flow Location: SIFT flow [53] is capable to create local dense correspondences. Thus, we use flow compensated locations, $\pi(i)$ in $\hat{I}_t^{r'}$, as another clue to calculate the spatial proximity.

c) Patch Intensity: Let $\mathbf{P}_i$ denote the query patch of size $\sqrt{n} \times \sqrt{n}$ in I_t^l centered at position i and $\mathbf{Q}_j^i$ denote the $\sqrt{n} \times \sqrt{n}$ candidate patch in $\hat{I}_t^{r'}$ centered at j. Small patches contain little structural information of raw images, and meanwhile it may be difficult for large patches to have an exact matching. In our MALDEC, we select 17 as the patch size, and other choices are also compared in our experiments in Section V-E.

d) Patch Gradient: The sole intensity-guided fidelity criteria cannot fully capture the intrinsic image structures, especially the high-frequency information. When the reference patch is roughly estimated, the problem can be more severe. To describe the structural similarity, we extract patch gradient $\nabla(\mathbf{P}_i)$ and $\nabla(\mathbf{Q}_j^i)$ as parts of hierarchical features. ∇ is the operation to calculate the gradient of the patches.

After obtaining the hierarchical features, we then calculate the distance measured by these features and construct the final location correspondences for the further high-frequency detail fusion. The distance is defined as:

$$\begin{aligned} &d(\mathbf{P}_i, \mathbf{Q}_j^i) \\ &\quad = \left(||\mathbf{P}_i - \mathbf{Q}_j^i||_2^2 + \rho||\nabla(\mathbf{P}_i) - \nabla(\mathbf{Q}_j^i)||_2^2\right) \\ &\qquad \cdot \mathcal{H}_{(\kappa\sqrt{n}\times\kappa\sqrt{n})}(\|i - j\|) \cdot \mathcal{H}_{(\kappa\sqrt{n}\times\kappa\sqrt{n})}(\|\pi(i) - j\|), \end{aligned} \tag{7}$$

where $\sqrt{n}\times\sqrt{n}$ is the size of the query patch, and $\kappa\sqrt{n}\times\kappa\sqrt{n}$ denotes a larger window for location matchings. ρ is a weighting parameter that controls the relative importance of pixel value differences and their gradient differences. In particular, DC components of the patches are removed before the distance computation. $\mathcal{H}_s(\cdot)$ is the inverse hard thresholding operator,

$$\mathcal{H}_s(x) = \begin{cases} 1 & x < s \\ +\infty & x \geq s \end{cases} \tag{8}$$

Intuitively, the location information plays a role in screening. When the adjusted locations in $\hat{I}_t^{r'}$, including the aligned location and flow compensated locations, are close enough to the corresponding location of the query patch in $\tilde{I}^l$, they are considered as the candidates of further detail transfer. When a patch in $\hat{I}_t^{r'}$ passes this screening, its distance to the query patch is measured jointly by patch intensity and gradient.

D. High-Frequency Detail Transfer

After the hierarchical feature and distance calculation, pixels at the same position in the matched patches between I_t^l and each $I_t^{r'}$ are corresponded. Then, the REHF map is fused with I_t^l based on pixel-wise matching correlation. The REHF I_m^l defined in Section III-C is calculated as follows:

$$I_{m,\mathbf{p}}^l = \begin{cases} \dfrac{\sum\limits_{\mathbf{q}\in\Omega_\mathbf{p}} \hat{I}_{m,\mathbf{q}}^r \cdot e^{\frac{-d(\mathbf{p},\mathbf{q})}{\zeta}}}{\sum\limits_{\mathbf{q}\in\Omega_\mathbf{p}} e^{\frac{-d(\mathbf{p},\mathbf{q})}{\zeta}}}, & |\Omega_\mathbf{p}| \neq 0, \\ 0, & |\Omega_\mathbf{p}| = 0, \end{cases} \tag{9}$$

where ζ is the parameter to control the shape of the exponential function for the high-frequency detail extraction. $I_{m,\mathbf{p}}^l$ is value of the pixel $\mathbf{p}$ in map I_m^l and similarly $\hat{I}_{m,\mathbf{q}}^r$ is the value of pixel $\mathbf{q}$. The set $\mathbf{\Omega_p}$ contains the matched pixels of $\mathbf{p}$ from the HF maps extracted from all of the references. $d(\mathbf{p}, \mathbf{q})$ calculates GMSE distance between the patches that $\mathbf{p}$ and $\mathbf{q}$ belong to. Note that pixel-wise correlations between I_m^l and each $\hat{I}_m^r$ here are the same as the builded pixel-wise correlations between I_t^l and each $\hat{I}_t^{r'}$. $|\mathbf{\Omega_p}|$ represents the number of elements in set $\mathbf{\Omega_p}$.

Finally, the SR result is obtained by directly adding the REHF map I_m^l to the intermediate reconstructed SR image I_t^l as $I^h = I_t^l \oplus I_m^l$.

V. Experiments

A. Training Methodology

Directly training our MALDEC shown in Fig. 1 faces two difficulties: 1) a large number of reference images need to be collected for training; 2) the forward prediction and backward gradient propagation of our network are time-consuming because the manifold localization includes several steps that need to be performed off-line, *i. e.* external high-frequency detail fusion, image retrieval, alignment and location matching. Therefore, to make training more efficient, we use synthetic data for network training. The process of training phase is shown in the top panel of Fig. 5. We use the ground-truth HR image I^g to generate the corresponding LR image I^l and the synthesized reference image I^{sr}. The ground-truth HR images I^g are applied with illumination transformations, and contrast transformations to generate the synthesized reference images I^{sr}. Then, the intermediate SR synthesized reference image I_t^{sr} is generated by ISIN based on the down-sampled synthesized reference input I_l^{sr}. Then, I_t^{sr} and I^{sr} are input to ECN to generate the HR prediction I^h. Given n groups of HR, LR and synthetic reference samples $\{(x_i, y_i, z_i)\}_{i=1}^n$ for training. Let $F(\cdot)$ and $G(\cdot)$ represent the learned ISIN and

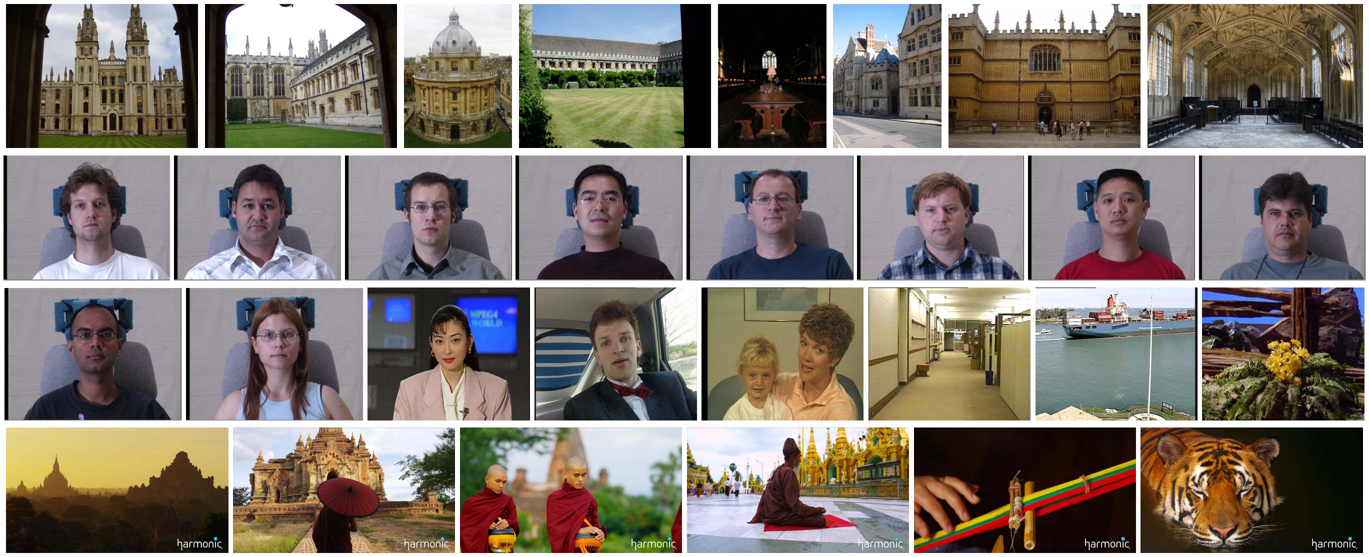


Fig. 4. The testing images selected from *Oxford Building dataset* [54] (the first row), *CAS-PEAL-R1* [55] (the second row), *Myanmar* HDTV sequence (the third row) and standard CIF testing sequences.

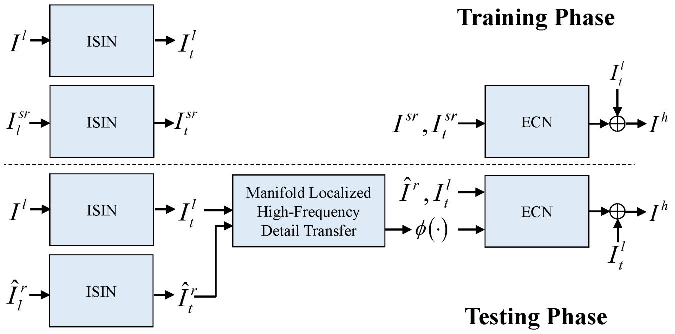


Fig. 5. The training and testing processes of MALDEC.

ECN processes for recovering I_t^l and I^h based on I^l and I^{sr}, respectively. We assume that ISIN is parameterized by Θ_1 and ECN is parameterized by Θ_2, then these two networks are trained by the following joint mean squared error (MSE):

$$L(\Theta_1, \Theta_2) = \frac{1}{n}\sum_{i=1}^{n}\Big(||\mathbf{F}(\mathbf{y}_i; \Theta_1) - \mathbf{x}_i||^2 + ||\mathbf{G}(\mathbf{z}_i; \Theta_2) - \mathbf{x}_i||^2\Big). \tag{10}$$

During the testing phase, the manifold localized high-frequency detail transfer, including the image retrieval, alignment and position matching, is performed off-line as shown in the bottom panel of Fig. 5. The experimental results show that, training the network with our synthesized data and testing additionally with the corresponding off-line operations, achieves promising SR results.

B. Experimental Settings

We train our MALDEC based on 91 images [4] and 200 images in *BSD500* training set [56]. Besides, as mentioned in Section III-C, during the training process of ECN, contrast of the ground truth images is first adjusted with random perturbation for more robustly extracting high-frequency information from the references.

With these 291 images, we first transform the images to YCbCr color space and only utilize the Y channel, which plays a major role in human perception for visual quality. The chrominance channels are later simply up-sampled by Bicubic interpolation in the testing phrase. Then, we generate sub-images at the size of 32 × 32 from images in the dataset with the stride step of 16. To make a solid and extensive comparison, we use two down-sampling configurations in our experiments: 1) *Setting I*, down-sampling method in [2], where images are first blurred with a kernel whose spatial size is 7 × 7 and blur level is 1.2, and then directly down-sampled by scaling factors of 2, 3 and 4; 2) *Setting II*, directly down-sampling by Bicubic interpolation. Around 10 thousand sub-images are obtained for training. The learning rate for ECN is initially set as 10^{-4} and drops to 10^{-5} after 50,000 iterations. The training ends when 100,000 iterations are reached.

We compare our MALDEC with different methods in three tasks, respectively:

- Cloud-based image SR. We compare MALDEC with Bicubic, neighbor regression super-resolution (NRSR) [31], landmark image SR (Landmark) [18], online compensated group structured sparse representation (GSSR) [20], deep edge guided recurrent residual learning (DEGREE) [13], very deep super-resolution (VDSR) [14], super-resolution generative adversarial network (SRGAN) [17], deep laplacian pyramid network (LapSRN) [57]. NRSR is a conventional neighbor embedding-based method. Landmark and GSSR are two cloud-based image SR methods. DEGREE, VDSR and LapSRN are state-of-the-art deep learning-based image SR methods.
- Multi-pose face reconstruction. We compare MALDEC with Bicubic, VDSR and neighbor embedding face components (NEFC) [58]. NEFC is one of the state-of-the-art

TABLE III

PSNR and SSIM Results of Different Methods on Cloud-Based Image SR. (·) Denotes Performance Gain of MALDEC Compared With Other Methods. SF Denotes the Scaling Factor. DEGREE Is Chosen as the Backbone Model of MALDEC. (Configuration I)

Methods	SF	Bicubic	Landmark	GSSR	DEGREE	MALDEC
PSNR	2	24.16 (9.50)	30.41 (3.25)	31.39 (2.27)	32.56 (1.10)	**33.66** -
SSIM		0.711 (.226)	0.860 (.077)	0.894 (.043)	0.922 (.015)	**0.937** -
PSNR	3	23.77 (7.16)	29.31 (1.63)	29.20 (1.74)	29.48 (1.45)	**30.93** -
SSIM		0.686 (.198)	0.826 (.058)	0.840 (.044)	0.849 (.035)	**0.884** -
PSNR	4	23.23 (6.13)	27.71 (1.64)	27.69 (1.67)	27.85 (1.50)	**29.35** -
SSIM		0.652 (.183)	0.786 (.049)	0.785 (.051)	0.791 (.044)	**0.835** -

TABLE IV

PSNR and SSIM Results of Different Methods on Cloud-Based Image SR. (·) Denotes Performance Gain of MALDEC Compared With Other Methods. SF Denotes the Scaling Factor. Note That, LapSRN Does Not Support 3× Enlargement VDSR Is Chosen as the Backbone Model of MALDEC (Configuration II)

Methods	SF	Bicubic	NE	VDSR	LapSRN	MALDEC
PSNR	2	28.13 (5.81)	32.19 (1.75)	33.12 (0.81)	32.92 (1.02)	**33.94** -
SSIM		0.842 (.100)	0.922 (.020)	0.931 (.011)	0.930 (.012)	**0.942** -
PSNR	3	26.07 (5.21)	29.47 (1.81)	29.90 (1.38)	\ \	**31.28** -
SSIM		0.749 (.143)	0.850 (.042)	0.860 (.033)	\ \	**0.892** -
PSNR	4	24.97 (5.07)	28.08 (1.96)	28.34 (1.70)	28.37 (1.67)	**30.04** -
SSIM		0.685 (.171)	0.796 (.060)	0.803 (.053)	0.807 (.049)	**0.856** -

TABLE V

PSNR and SSIM Results of Different Methods on Multi-Pose Face Reconstruction. (·) Denotes Performance Gain of MALDEC Compared With Other Methods. SF Denotes the Scaling Factor. VDSR Is Chosen as the Backbone Model of MALDEC. (Configuration II)

Methods	Bicubic		NEFC		VDSR		MALDEC	
SF	2	3	2	3	2	3	2	3
PSNR	37.74 (2.57)	34.68 (4.14)	39.13 (1.18)	38.31 (.50)	39.63 (.67)	38.00 (.81)	**40.31** -	**38.82** -
SSIM	0.945 (.012)	0.917 (.021)	0.943 (.014)	0.934 (.004)	0.955 (.002)	0.935 (.003)	**0.957** -	**0.939** -

TABLE VI

PSNR and SSIM Results of Different Methods on Reference Frame-Guided Video SR. (·) Denotes Performance Gain of MALDEC Compared With Other Methods. SF Denotes the Scaling Factor. VDSR Is Chosen as the Backbone Model of MALDEC. (Configuration II)

Methods	Bicubic			VSRNet			VDSR			MALDEC		
SF	2	3	4	2	3	4	2	3	4	2	3	4
PSNR	33.08 (5.58)	30.07 (4.81)	28.23 (4.37)	36.84 (1.81)	32.83 (2.05)	30.52 (2.07)	37.56 (1.09)	33.38 (1.50)	30.87 (1.73)	**38.65** -	**34.88** -	**32.60** -
SSIM	.916 (.053)	.842 (.092)	.793 (.113)	.951 (.018)	.882 (.052)	.831 (.075)	.959 (.010)	.905 (.029)	.860 (.045)	**.969** -	**.934** -	**.906** -

face hallucination methods. For a fair comparison, we add multi-pose reference images to the training set of NEFC.

- Reference frame-based video SR. We compare MALDEC with Bicubic, VDSR, video super-resolution network (VSRNet) [58] and draft learning video super-resolution (DraftLearn) [46]. VSRNet and DraftLearn are state-of-the-art deep video SR methods. Note that, the degradation setting of DraftLearn is different with ours, and its SR results based on our input LR images have pixel shifts to the HR images. Thus, to provide a fair comparison, we only compare with the results of DraftLearn subjectively as shown in Fig. 8.

For SRGAN, we use an online public available implementation.[1] GSSR and DEGREE are implemented by ourselves, and other codes of competing methods are also kindly provided by the authors. The detailed links are provided in Table VII.

[1] https://github.com/tensorlayer/srgan

TABLE VII
CODE LINKS OF COMPETING METHODS

Methods	Code Link
NRSR	https://github.com/lyttonhao/NRSR/tree/release
VDSR	https://github.com/huangzehao/caffe-vdsr
SRGAN	https://github.com/tensorlayer/srgan
LapSRN	https://github.com/phoenix104104/LapSRN
NEFC	https://github.com/lyttonhao/Face-Hallucination-of-Facial-Components/tree/release
VSRNet	http://ivpl.eecs.northwestern.edu/content/research-projects/17151
DraftLearn	http://www.cse.cuhk.edu.hk/leojia/projects/DeepSR

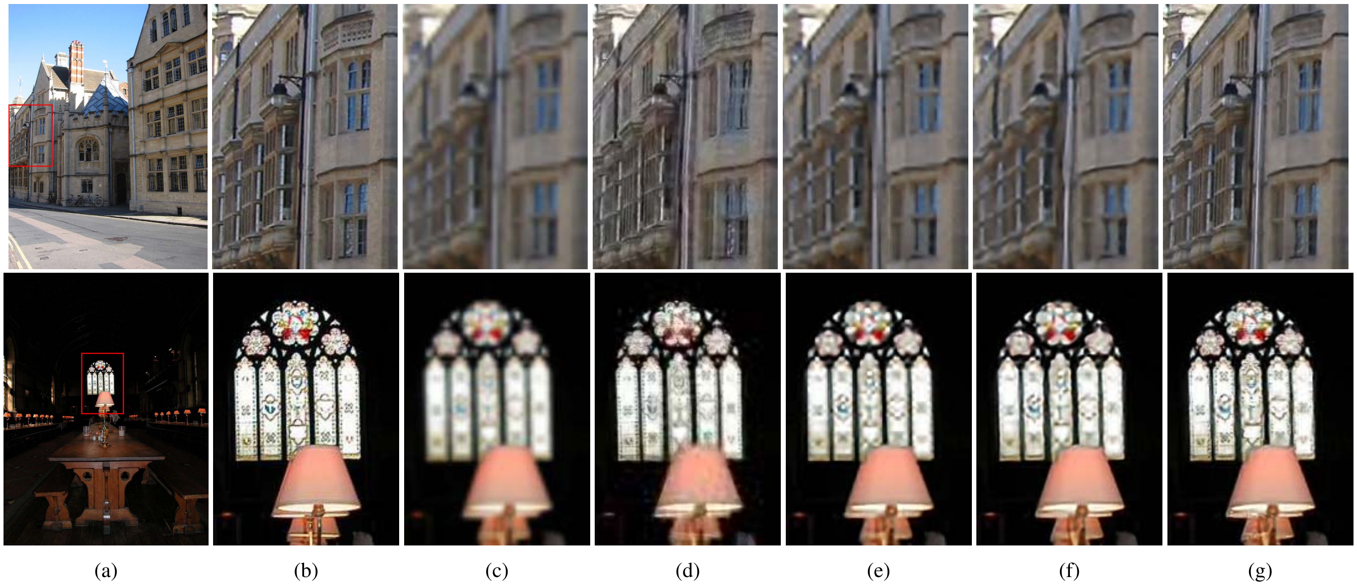


Fig. 6. Visual comparisons of different methods on cloud-based image SR (Configuration I). Top panel: in 4× enlargement, MALDEC successfully reconstructs textures of the windows with fewest artifacts. Bottom panel: in 3× enlargement, MALDEC generates the most clear and visually pleasing decorations on the window. (a) HR. (b) HR region. (c) Bicubic. (d) Landmark. (e) GSSR. (f) DEGREE. (g) MALDEC.

The testing images used for evaluating the performance of MALDEC in three tasks are respectively as follows:

- Eight images from *Oxford Building* dataset [54], used for testing cloud-based image SR as shown in the first row of Fig. 4. For each testing image, we retrieve four reference images for external high-frequency details compensation.
- Ten sets of images selected from *Multi-PIE* [59] as shown in the second row and the first two images of the third row as shown in Fig. 4 used for testing multi-pose face reconstruction. For each set, four reference images with certain poses are utilized to facilitate the front face reconstruction.
- Six video sequences cropped from *Myanmar* HDTV sequence[2] as shown in the last row of Fig. 4 and six standard CIF testing sequences[3] as shown in the last six images of the third row of Fig. 4, used for evaluating reference frame-guided video SR. For each testing frame, we use four former frames at the certain interval as reference frames.

C. Objective Evaluation

Tables III-V show the objective results of the compared methods in three tasks. Tables III and IV present the results in the cloud-based image SR in configuration I and II, respectively. Our MALDEC achieves more than 1dB and 0.015 performance gain in average PSNR and SSIM over other methods for all scaling factors. Even compared with VDSR and DEGREE, MALDEC still achieves 0.81 dB, 1.38 dB, 1.70 dB and 1.10 dB, 1.45 dB, 1.50 dB average PSNR gain in 2,3 and 4× enlargement, respectively. For multi-pose face reconstruction as shown in Table V, even without using facial landmark points as constraints, our MALDEC still gains over NEFC more than 0.5dB in average PSNR. Tables VI presents the results in the reference frame-guided video SR. Compared with the state-of-the-art video SR method – VSRNet, our MALDEC obtains 1.81, 2.05 and 2.07 dB higher PSNR result in 2,3 and 4× enlargement, respectively.

D. Subjective Results

Subjective results are shown in Figs. 6-9. Figs. 6 and 7 present the visual results of cloud-based image SR. Bicubic interpolation generates blurred results. NRSR successfully obtains sharper edges as shown in Fig. 7(d). However, it fails to reconstruct many high-frequency details. Landmark successfully introduces high-frequency details from the reference image. However, artifacts sometimes are brought by incorrect patch matchings or inappropriate patch blending, as shown in Fig. 6(d). Sparse representation based GSSR

[2] https://www.harmonicinc.com/4k-demo-footage-download/
[3] https://media.xiph.org/video/derf/

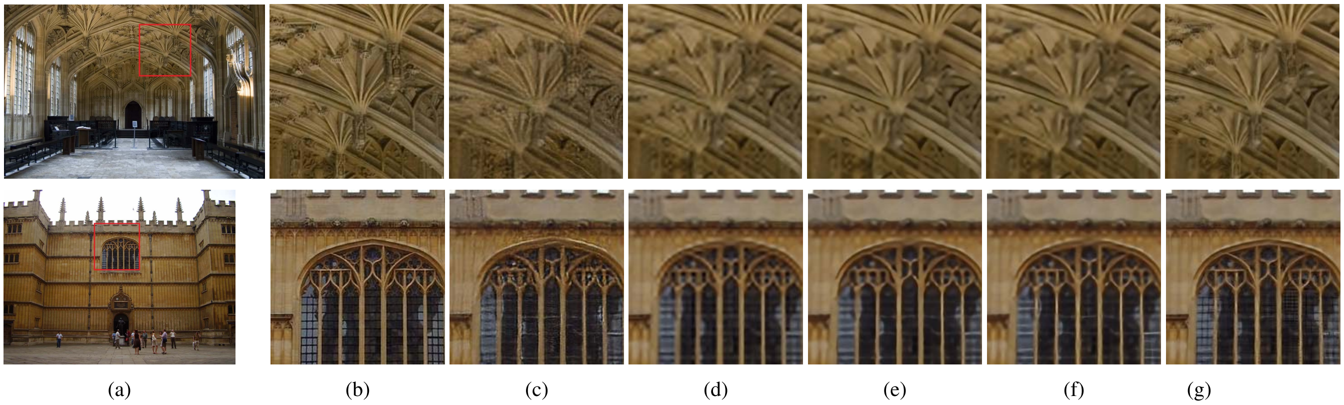


Fig. 7. Visual comparisons of different methods on cloud-based image SR (Configuration II). Top panel: in 2× enlargement, MALDEC reconstructs the clearest decorations of the ceiling. Bottom panel: in 4× enlargement, MALDEC generates the most clear and visually pleasing details of the window. (a) HR. (b) HR region. (c) SRGAN. (d) NRSR. (e) LapSRN. (f) VDSR. (g) MALDEC.

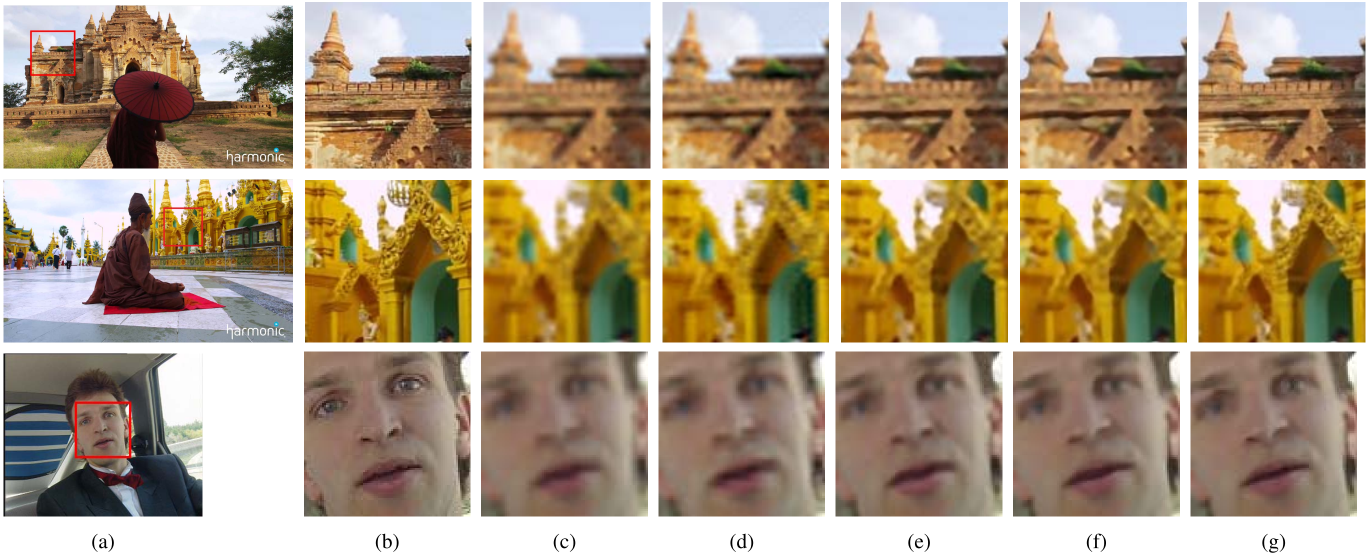


Fig. 8. Visual comparisons of different methods on reference frame-guided video SR (Configuration II). Top two panels: in 4× enlargement, MALDEC reconstructs the most detailed decorations of the roofs and walls. Bottom panel: in 4× enlargement, MALDEC generates the clearest face details. (a) HR. (b) HR region. (c) Bicubic. (d) DraftLearn. (e) VSRNet. (f) VDSR. (g) MALDEC.

does not utilize the position information of the reference patches. While there are many similar reference patches, noise and inaccurate details are brought into the results of GSSR, as shown in Fig. 6(e). Due to the excellent modeling capacity of DEGREE and VDSR, they accurately reconstruct the main structures of nature images as shown in Figs. 6(f) and 7(f), respectively. LapSRN shows superior capacity to reconstruct salient features of natural images, *i. e.* long edges, to DEGREE and VDSR as shown in Fig. 7(e). Nevertheless, due to ambiguity, some high-frequency details are inevitably missing in texture regions. Comparatively, our MALDEC obtains the clearest high-frequency detail reconstruction. Owing to the robustness of high-frequency details extraction and patch matching, the result of MALDEC contains the least visual artifacts and provides the best visual quality. Fig. 9 presents the visual results of multi-pose face reconstruction. MALDEC presents the most visually promising reconstructed details. The subjective results of the reference frame-guided video SR are presented in Fig. 8. It is clearly shown that, MALDEC reconstructs more clear details than both VSRnet and VDSR.

E. More Evaluation Metrics

To demonstrate the superiority of our method to other methods, we further compare the proposed method to other state-of-the-art methods on different metrics, including PSNR, SSIM, PSNR-HVS [60], MS-SSIM [61], perceptual loss (dentoed as Perceptual-2 and Perceptual-3) [16], ConPatch [62], and subjective rank product [63]. The testing task used for the evaluation is cloud-based image SR. As mentioned in Section V-B, eight images from *Oxford Building* dataset [54], shown in the first row of Fig. 4, are used as testing images. For each testing image, four reference images are retrieved for external high-frequency details compensation. PSNR and SSIM are conventional signal processing metrics to measure whether the recovered signals are similar to the original ones. PSNR-HVS and MS-SSIM are the improved PSNR and SSIM,

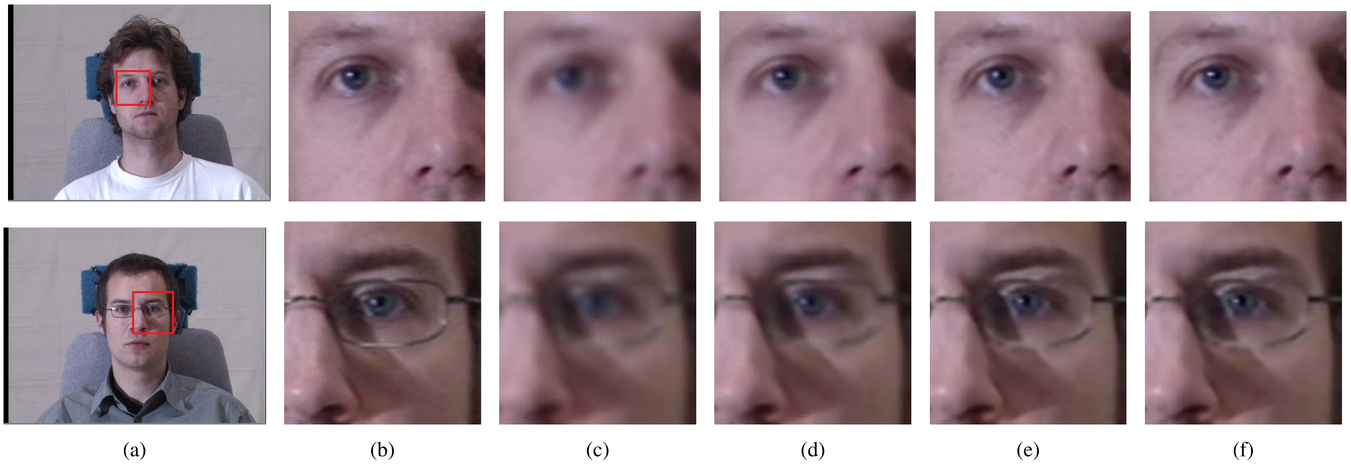


Fig. 9. Visual comparisons of different methods on multi-pose face hallucination (Configuration II). Top panel: 2× enlargement. Bottom panel: 3× enlargement. MALDEC reconstructs the clearest facial details. (a) HR. (b) HR region. (c) Bicubic. (d) NEFC. (e) VDSR. (f) MALDEC.

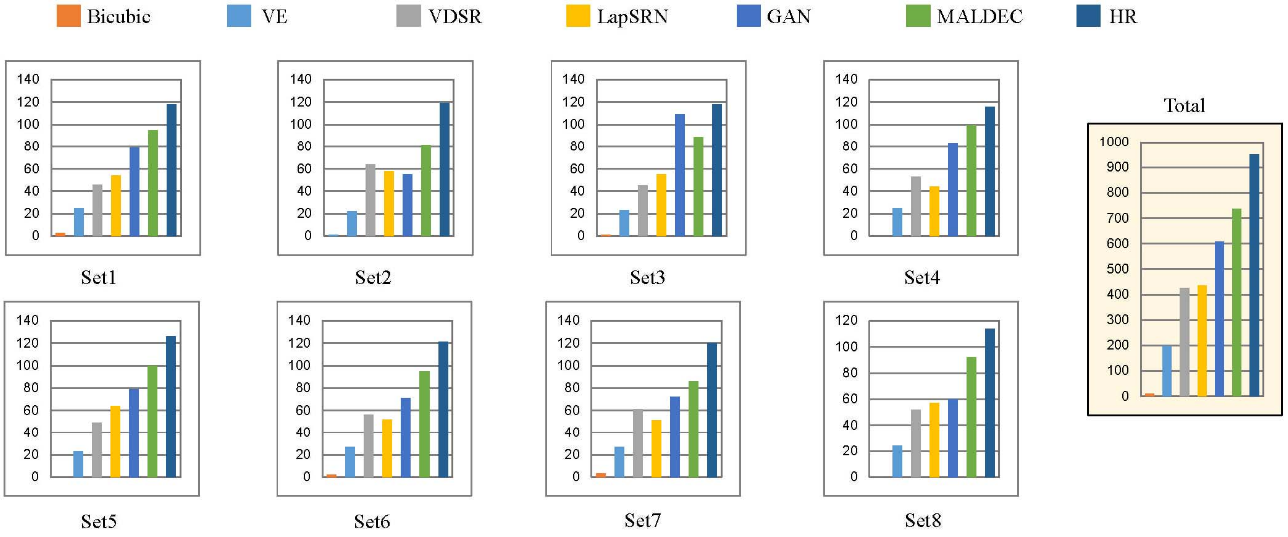


Fig. 10. The number of votes per testing image and the total ranking of seven methods.

respectively, considering more on human visual perception. Perceptual loss measures the similarity between two images based on their features extracted from the pretrained deep networks, aiming to simulate human visual perception. Perceptual-2 and Perceptual-3 are calculated based on the second and third layers of a pretrained VGG19 network [64]. ConPatch (denoted as CP) [62] is a powerful feature to measure the similarity between two patches along with their surrounding contexts. In our experiment, we follow the configuration of depth image SR in [62] to sample 11 × 11 patches with a 21 × 21 context from SR results and HR images, calculate the mean square error of the conPatch features by locations, and then obtain the per-pixel average value as the metric. For PSNR, SSIM, PSNR-HVS, MS-SSIM and CP, a larger number signifies better reconstruction quality. For Perceptual-2, Perceptual-3 and RP, a smaller number signifies better reconstruction quality. It is observed from Table VIII that, MALDEC obtains superior objective results on all metrics.

We also conduct subjective evaluations. To compare different SR results from the perspective of observers, we employ the paired comparison approach, where the participants are shown two SR images at a time, side by side, and are asked to simply choose the preferred one by visual quality. We have a total of 32 participants, including both domain experts and generally knowledgeable individuals, each given 105 pairwise comparisons in average over a set of eight images with seven different SR methods (including the HR image). Fig. 10 illustrates the seven methods, ranked by the number of votes received. It can be seen that the proposed MALDEC outperforms other methods in most cases, and achieves an overall superior performance. The RP column in Table VIII shows the results of the rank product $\psi(O) = \left(\prod_i r_{O,i}\right)^{1/b}$, where $r_{O,i}$ is the specific ranking for method O and image i $(i = 1 \ldots b)$. Compared with others, MALDEC produces the best consistency among different test cases to achieve the best visual quality.

F. Evaluation in Running Time

We also provide the running time comparison. The time-consuming patch matching in MALDEC can be accelerated by fast approximation search such as KD-tree [65].

TABLE VIII

THE RESULTS OF USING DIFFERENT METRICS TO MEASURE THE SIMILARITY BETWEEN THE SR RESULTS AND THE HR IMAGES. FOR PSNR, SSIM, PSNR-HVS, MS-SSIM AND CP, A LARGER NUMBER SIGNIFIES BETTER RECONSTRUCTION QUALITY. FOR PERCEPTUAL-2, PERCEPTUAL-3 AND RP, A SMALLER NUMBER SIGNIFIES BETTER RECONSTRUCTION QUALITY

Metrics	PSNR	SSIM	PSNR-HVS	MS-SSIM
BC	27.11	0.7596	24.43	0.9438
NE	28.07	0.7960	26.31	0.9600
VDSR	28.33	0.8032	26.83	0.9618
LapSRN	28.35	0.8039	26.88	0.9619
GAN	26.83	0.7498	25.18	0.9455
MALDEC	30.02	0.8549	29.44	0.9736
Metrics	Peceptual-2	Peceptual-3	CP	RP
BC	66122.47	86124.54	3.36	7.00
NE	53959.49	69487.99	2.60	6.00
VDSR	50033.66	64909.00	1.78	4.31
LapSRN	49808.27	65505.41	2.72	4.40
GAN	52352.68	58026.43	2.59	3.04
MALDEC	31429.25	40502.50	3.62	2.104

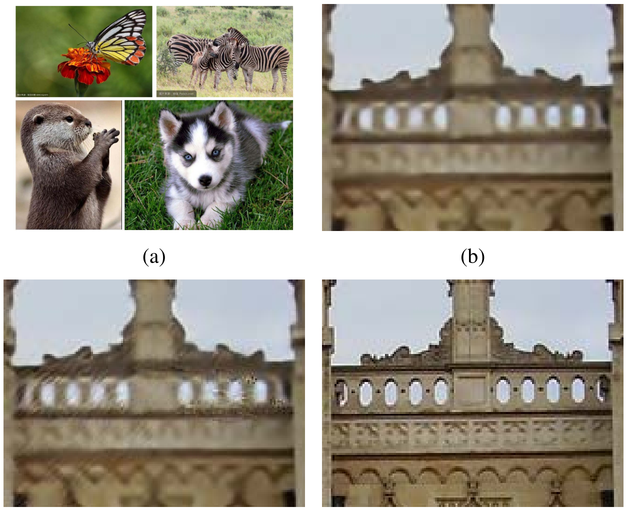


Fig. 11. A failure case of MALDEC. When all retrieved references are not similar to the input LR image, the reconstructed result of MALDEC may present noises and artifacts. (a) Reference. (b) VDSR. (c) MALDEC. (d) Ground truth.

After the acceleration, our method only need 30 seconds to super-resolve an image from 256*192 to 1024*768 on CPU i7-5930K. It can be further accelerated by multi-thread technique and thus is potential to be a real-time application. The running times of two newest deep based methods CSCN [12] and VDSR [14] at CPU mode are respectively 14s and 15s.

VI. CONCLUSION AND DISCUSSIONS

In this paper, a manifold localized deep external compensation network is developed to additionally utilize reference images, *i. e.* retrieved similar images in cloud database and reference HR frame in a video, to localize the HR image manifold and compensate the lost high-frequency details. Compared with traditional methods, the structure inference and external compensation of MALDEC are jointly trained to make a good trade-off between these two terms for an optimal SR result. Extensive experiments on cloud-based image SR, multi-pose face reconstruction and reference frame-guided video SR demonstrate the superiority of our method in both objective and subjective evaluations. Our work also provides a flexible framework to utilize online retrieved data. It has the potential of being applied for other image processing tasks, *i. e.* image denoising, stylization and dehazing. The texture details are extracted from the online retrieved reference images, and then compensated to the processed results of internal methods. Our MALDEC has two drawbacks: 1) when all retrieved images are not similar to the LR image, the reconstruction results may present some noises or artifacts, as shown in Fig. 11; 2) the current time efficiency of MALDEC cannot meet the need of practical real-time industrial applications. In the future, we will carry out attempts to overcome these drawbacks, extend our approach to other applications, and fasten our MALDEC to pursuit an industrial application.

ACKNOWLEDGEMENT

The authors thank Dr. Huanjing Yue from Tianjin University for her help with the experimental aspects of this study.

REFERENCES

[1] J. Sun, J. Sun, Z. Xu, and H.-Y. Shum, "Gradient profile prior and its applications in image super-resolution and enhancement," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1529–1542, Jun. 2011.

[2] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1620–1630, Apr. 2013.

[3] A. Marquina and S. J. Osher, "Image super-resolution by TV-regularization and Bregman iteration," *J. Sci. Comput.*, vol. 37, no. 3, pp. 367–382, Dec. 2008.

[4] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[5] L. He, H. Qi, and R. Zaretzki, "Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 345–352.

[6] Y. Zhang, J. Liu, W. Yang, and Z. Guo, "Image super-resolution based on structure-modulated sparse representation," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2797–2810, Sep. 2015.

[7] J. Ren, J. Liu, and Z. Guo, "Context-aware sparse decomposition for image denoising and super-resolution," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1456–1469, Apr. 2013.

[8] R. Timofte, V. D. Smet, and L. V. Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. Asian Conf. Comput. Vis.*, Springer, 2014, pp. 111–126.

[9] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3791–3799.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[11] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3194–3207, Jul. 2016.

[12] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 370–378.

[13] W. Yang *et al.*, "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5895–5907, Dec. 2017.

[14] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[15] R. Timofte, V. De Smet, and L. Van Gool, "Semantic super-resolution: When and where is it useful?" *Comput. Vis. Image Understand.*, vol. 142, pp. 1–12, Jan. 2016.

[16] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[17] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 105–114.

[18] H. Yue, X. Sun, J. Yang, and F. Wu, "Landmark image super-resolution by retrieving Web images," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4865–4878, Dec. 2013.

[19] Y. Li, W. Dong, G. Shi, and X. Xie, "Learning parametric distributions for image super-resolution: Where patch matching meets sparse coding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 450–458.

[20] J. Liu, W. Yang, X. Zhang, and Z. Guo, "Retrieval compensated group structured sparsity for image super-resolution," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 302–316, Feb. 2017.

[21] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1637–1645.

[22] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 391–407.

[23] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1874–1883.

[24] H. Yue, X. Sun, J. Yang, and F. Wu, "CID: Combined image denoising in spatial and frequency domains using Web images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2933–2940.

[25] X. Liu, X. Wu, J. Zhou, and D. Zhao, "Data-driven sparsity-based restoration of JPEG-compressed images in dual transform-pixel domain," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5171–5178.

[26] Y. Zhang, W. Dong, O. Deussen, F. Huang, K. Li, and B.-G. Hu, "Data-driven face cartoon stylization," in *Proc. ACM Int. Conf. Exhib. Comput. Graph. Interact. Techn. Asia*, vol. 14, 2014, pp. 1–4.

[27] J. Liu, W. Yang, X. Sun, and W. Zeng, "Photo stylistic brush: Robust style transfer via superpixel-based bipartite graph," *IEEE Trans. Multimedia*, Dec. 2017.

[28] B. Wang, Y. Yu, T.-T. Wong, C. L. P. Chen, and Y.-Q. Xu, "Data-driven image color theme enhancement," in *Proc. ACM Int. Conf. Exhib. Comput. Graph. Interact. Techn. Asia*, vol. 29, no. 6, Dec. 2010, pp. 146:1–146:10.

[29] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.

[30] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1685–1694.

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

[32] W. T. Freeman and C. Liu, "Markov random fields for super-resolution and texture synthesis," in *Markov Random Fields for Vision and Image Processing*. Cambridge, MA, USA: MIT Press, 2011.

[33] S. Yang, J. Liu, Y. Fang, and Z. Guo, "Joint-feature guided depth map super-resolution with face priors," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 399–411, Jan. 2018.

[34] Z. Xiong, D. Xu, X. Sun, and F. Wu, "Example-based super-resolution with soft information and decision," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1458–1465, Oct. 2013.

[35] J. Yang, Z. Lin, and S. Cohen, "Fast image super-resolution based on in-place example regression," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1059–1066.

[36] C. Liu and D. Sun, "On Bayesian adaptive video super resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 346–360, Feb. 2014.

[37] R. Fransens, C. Strecha, and L. V. Gool, "Optical flow based super-resolution: A probabilistic approach," *Comput. Vis. Image Understand.*, vol. 106, no. 1, pp. 106–115, 2007.

[38] S. Baker and T. Kanade, *Super-Resolution Optical Flow*, document CMU-RI-TR-99-36, 1999.

[39] H. He and L. P. Kondi, "An image super-resolution algorithm for different error levels per frame," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 592–603, Mar. 2006.

[40] A. V. Kanaev and C. W. Miller, "Multi-frame super-resolution algorithm for complex motion patterns," *Opt. Express*, vol. 21, no. 17, pp. 19850–19866, Aug. 2013.

[41] O. A. Omer and T. Tanaka, "Region-based weighted-norm approach to video super-resolution with adaptive regularization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2009, pp. 833–836.

[42] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[43] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, 1992.

[44] Q. Yuan, L. Zhang, and H. Shen, "Regional spatially adaptive total variation super-resolution with spatial information filtering and clustering," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2327–2342, Jun. 2013.

[45] X. Zhang, R. Xiong, S. Ma, G. Li, and W. Gao, "Video super-resolution with registration-reliability regulation and adaptive total variation," *J. Vis. Commun. Image Represent.*, vol. 30, pp. 181–190, Jul. 2015.

[46] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 531–539.

[47] Y. Huang, W. Wang, and L. Wang, "Bidirectional recurrent convolutional networks for multi-frame super-resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 235–243.

[48] J. Caballero *et al.*, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2848–2857.

[49] D. Liu *et al.*, "Robust video super-resolution with learned temporal dynamics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2526–2534.

[50] W. Yang, J. Feng, G. Xie, J. Liu, Z. Guo, and S. Yan, "Video super-resolution based on spatial-temporal recurrent residual networks," *Comput. Vis. Image Understand.*, vol. 168, pp. 79–92, Mar. 2017.

[51] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 524–531.

[52] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[53] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[54] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[55] W. Gao *et al.*, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.

[56] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[57] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. Int. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5835–5843.

[58] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 109–122, Jun. 2016.

[59] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–8.

[60] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," in *Proc. 2nd Int. Workshop Video Process. Qual. Metrics*, vol. 4, 2006, pp. 1–4.

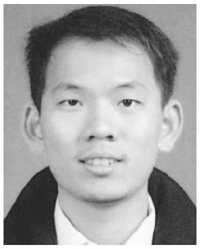
[61] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[62] Y. Romano and M. Elad, "Con-patch: When a patch meets its context," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 3967–3978, Sep. 2016.

[63] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," in *Proc. ACM SIGGRAPH Asia Papers*, 2010, pp. 160:1–160:10.

[64] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[65] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2227–2240, Nov. 2014.



Wenhan Yang (S'17) received the B.S. degree in computer science, from Peking University, Beijing, China, in 2012, where he is currently pursuing the Ph.D. degree with the Institute of Computer Science and Technology. He was a Visiting Scholar with National University of Singapore from 2015 to 2016. His current research interests include image processing, sparse representation, image restoration, and deep learning-based image processing.

Sifeng Xia received the B.S. degree in computer science, from Peking University, Beijing, China, in 2017, where he is currently pursuing the master's degree with the Institute of Computer Science and Technology, Peking University. His current research interests include deep learning-based image processing and video coding.

Jiaying Liu (S'08–M'10–SM'17) received the B.E. degree in computer science, from Northwestern Polytechnic University, Xi'an, China, in 2005 and the Ph.D. degree (Hons.) 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 has authored over 100 technical articles in refereed journals and proceedings and holds 24 granted patents. Her current research interests include image/video processing, compression, and computer vision.

She was a Visiting Scholar with University of Southern California, Los Angeles, CA, USA, from 2007 to 2008. In 2015, she was a Visiting Researcher with Microsoft Research Asia, supported by Star Track for Young Faculties. She served as a TC Member for the IEEE CAS MSA and APSIA IVM. She is a Senior Member of CCF. She served as a APSIA Distinguished Lecturer from 2016 to 2017.

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 is an Executive Member of the China Society of Motion Picture and Television Engineers. He 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, and the Wang Xuan News Technology Award and the Chia Tai Teaching Award in 2008. He received the Government Allowance granted by the State Council in 2009. He received the Distinguished Doctoral Dissertation Advisor Award from Peking University in 2012 and 2013.