Learning Super-Resolution Jointly from External and Internal Examples

Zhangyang Wang†, Yingzhen Yang‡, Zhaowen Wang†, Shiyu Chang†, Jianchao Yang‡, Thomas S. Huang†
†Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
‡Adobe Systems Inc, San Jose, CA 95110, USA
{zwang119, yyang58}@illinois.edu, zhaowen@adobe.com, chang87@illinois.edu, jiayang@adobe.com,
t-huang1@illinois.edu

Abstract—Single image super-resolution (SR) aims to estimate a high-resolution (HR) image from a low-resolution (LR) input. Image priors are commonly learned to regularize the otherwise seriously ill-posed SR problem, either using external LR-HR pairs or internal similar patterns. We propose joint SR to adaptively combine the advantages of both external and internal SR methods. We define two loss functions using sparse coding based external examples, and epitomic matching based on internal examples, as well as a corresponding adaptive weight to automatically balance their contributions according to their reconstruction errors. Extensive SR results demonstrate the effectiveness of the proposed method over the existing state-of-the-art methods, and is also verified by our subjective evaluation study.

Index Terms—Super-resolution, example-based methods, sparse coding, epitome

EDICS Category: TEC-ISR Interpolation, Super-Resolution, and Mosaicing

I. INTRODUCTION

Super-resolution (SR) algorithms aim to constructing a high-resolution (HR) image from one or multiple low-resolution (LR) input frames [1]. This problem is essentially ill-posed because much information is lost in the HR to LR degradation process. Thus SR has to refer to strong image priors, that range from the simplest analytical smoothness assumptions, to more sophisticated statistical and structural priors learned from natural images [2], [3], [4], [5]. The most popular single image SR methods rely on example-based learning techniques. Classical example-based methods learn the mapping between LR and HR image patches, from a large and representative external set of image pairs, and is thus denoted as external SR. Meanwhile, images generally possess a great amount of self-similarities; such a self-similarity property motivates a series of internal SR methods. With much progress being made, it is recognized that external and internal SR methods each suffer from their certain drawbacks. However, their complementary properties inspire us to propose the joint super-resolution (joint SR), that adaptively utilizes both external and internal examples for the SR task. The contributions of this paper are multi-fold:

- We propose joint SR exploiting both external and internal examples, by defining an adaptive combination of different loss functions.
- We apply epitomic matching [6] to enforcing self-similarity in SR. Compared the the local nearest neighbor (NN) matching adopted in [5], epitomic matching features more robustness to outlier features, as well as the ability to perform efficient non-local searching.
- We carry out a human subjective evaluation survey to evaluate SR result quality based on visual perception, among several state-of-the-art methods.

II. A MOTIVATION STUDY OF JOINT SR

A. Related Work

The joint utilization of both external and internal examples has been most studied for image denoising [17]. Mosseri et. al. [18] first proposed that some image patches inherently prefer internal examples for denoising, whereas other patches inherently prefer external denoising. Such a preference is in essence the tradeoff between noise-fitting versus signal-fitting. Burger et. al. [16] proposed a learning-based approach that automatically combines denoising results from an internal and an external method. The learned combining strategy outperforms both internal and external approaches across a wide range of images, being closer to theoretical bounds.

In SR literature, while the most popular methods are based on either external or internal similarities, there have been limited efforts to utilize one to regularize the other. The authors in [19] incorporated both a local autoregressive (AR) model and a nonlocal self-similarity regularization term, into the sparse representation framework, weighted by constant coefficients. Yang et. al. [20] learned the (approximated) nonlinear SR mapping function from a collection of external images with the help of in-place self-similarity. More recently, an explicitly joint model is put forward in [21], including two loss functions by sparse coding and local scale invariance, bound by an indicator function to decide which loss function will work for each patch of the input image. Despite the existing efforts, there is little understanding on how the external and internal examples interact with each other in SR, how to judge the external versus internal preference for each patch, and how to make them collaborate towards an overall optimized performance.

External SR methods use a universal set of example patches to predict the missing (high-frequency) information for the HR image. In [17], during the training phase, LR-HR patch pairs
B. Comparing External and Internal SR Methods

Both external and internal SR methods have different advantages and drawbacks. See Fig. 1 for a few specific examples.

These comparisons display the generally different, even complementary behaviors of external and internal SR. Based on the observations, we expect that the external examples contribute to visually pleasant SR results for smooth regions as well as some irregular structures that barely recur in the input. Meanwhile, internal examples serve as a powerful source to reproduce unique and singular features that rarely appear externally but repeat in the input image (or its different scales). Note that similar arguments have been validated statistically in the image denoising literature [16].

III. A Joint SR Model

Let \( X \) denote the HR image to be estimated from the LR input \( Y \). \( X_{ij} \) and \( Y_{ij} \) stand for the \((i,j)\)-th \((i,j = 1,2,...)\) patch from \( X \) and \( Y \), respectively. Considering almost all SR methods work on patches, we define two loss functions \( \ell_G(\cdot) \) and \( \ell_I(\cdot) \) in a patch-wise manner, which enforce the external and internal similarities, respectively. While one intuitive idea
is to minimize a weighted combination of the two loss functions, a patch-wise (adaptive) weight $\omega(\cdot)$ is needed to balance them. We hereby write our proposed joint SR in the general form:

$$\min_{x_{ij}, \Theta_G, \Theta_I} \ell_G(x_{ij}, \Theta_G|y_{ij}) + \omega(\Theta_G, \Theta_I) \ell_I(x_{ij}, \Theta_I|y_{ij}).$$  \hspace{1cm} (1)

$\Theta_G$ and $\Theta_I$ are the latent representations of $x_{ij}$ over the spaces of external and self examples, respectively. The form $f(x_{ij}, \Theta|y_{ij})$, $f$ being $\ell_I$, $\ell_G$ or $\omega$, represents the function dependent on variables $x_{ij}$ and $\Theta$ ($\Theta_G$ or $\Theta_I$), with $y_{ij}$ known (we omit $y_{ij}$ in all formulations hereinafter). We will discuss each component in (1) next.

One specific form of joint SR will be discussed in this paper. However, note that with different choices of $\ell_G(\cdot)$, $\ell_I(\cdot)$, and $\omega(\cdot)$, a variety of methods can be accommodated in the framework. For example, if we set $\ell_G(\cdot)$ as the (adaptively reweighted) sparse coding term, while choosing $\ell_I(\cdot)$ equivalent to the two local and non-local similarity based terms, then (1) becomes the model proposed in [19], with $\omega(\cdot)$ being some empirically chosen constants.

A. Sparse Coding for External Examples

The HR and LR patch spaces $\{x_{ij}\}$ and $\{y_{ij}\}$ are assumed to be tied by some mapping function. With a well-trained NN and some empirically chosen constants.

2) EPI: Epitomic Matching for Internal SR: The matching of $x_{ij}^E$ over the smoothed input image $y'$ makes the core step of the high frequency transfer scheme. However, the performance of NN matching $\ell_3$ is degraded with the presence of noise and outliers. Moreover, the NN matching in $\ell_3$ is restricted to a local window for efficiency, which potentially accounts for some rigid artifacts.

Instead, we propose epitomic matching to replace NN matching in the above frequency transfer scheme. As a generative model, epitome [23, 24] summarizes a large set of raw image patches into a condensed representation in a way similar to Gaussian Mixture Models. We first learn an epitome $e_{y'}$ from $y'$, and then match each $x_{ij}^E$ over $e_{y'}$ rather than $y'$ directly. Assume $(m, n) = f_{ep}(x_{ij}^E, e_{y'})$, where $f_{ep}$ denotes the procedure of epitomic matching by $e_{y'}$. It then follows the same way as in $\ell_3$:

$$x_{ij}^E = y_{mn} - y'_{mn};$$  \hspace{1cm} (6)

Since $e_{y'}$ summarizes the patches of the entire $y'$, the proposed epitomic matching benefits from non-local patch matching. In the absence of self-similar patches in the local neighborhood, epitomic patch matching can thus be applied to SR by itself as well, named EPI for short, which will be included in our experiments in Section 4 and compared to the method using NN matching in $\ell_3$.

Since $e_{y'}$ summarizes the patches of the entire $y'$, the proposed epitomic matching benefits from non-local patch matching. In the absence of self-similar patches in the local neighborhood, epitomic patch matching can thus be applied to SR by itself as well, named EPI for short, which will be included in our experiments in Section 4 and compared to the method using NN matching in $\ell_3$.

C. Learning the Adaptive Weights

In [13], Mosseri et.al. showed that the internal versus external preference is tightly related to the Signal-to-Noise-Ratio (SNR) estimate of each patch. Inspired by that finding, we could seek similar definitions of "noise" in SR based on the latent representation errors. The external noise is defined by the residual of sparse coding

$$N_y(a_{ij}) = ||D_ya_{ij} - y_{ij}||^2.$$  \hspace{1cm} (5)

Meanwhile, the internal noise finds its counterpart definition by the epitomic matching error within $f_{ep}$:

$$N_x(x_{ij}^E) = ||y_{mn} - x_{ij}^E||^2,$$  \hspace{1cm} (6)

where $y_{mn}$ is the matching patch in $y'$ for $x_{ij}^E$.

Usually, the two "noises" are on the same magnitude level, which aligns with the fact that external- and internal-examples will have similar performances on many (such as homogenous regions). However, there do exist patches where the two
have significant differences in performances, as shown in Fig. 1 which means the patch has a strong preference toward one of them. In such cases, the “preferred” term needs to be sufficiently emphasized. We thus construct the following patch-wise adaptive weight ($p$ is the hyperparameter):

$$\omega(\alpha_{ij}, X^E_{ij}) = \exp(p \cdot |N_g(\alpha_{ij}) - N_i(X^E_{ij})|).$$ (7)

When the internal noise becomes larger, the weight decays quickly to ensure that external similarity dominates, and vice versa.

D. Optimization

Directly solving (1) is very complex due to the intrinsic nonlinearity and entanglement among all variables. Instead, we follow the coordinate descent fashion [25] and solve the following three sub-problems iteratively.

1) $a_{ij}$-subproblem: Fixing $X_{ij}$ and $X^E_{ij}$, we have the following minimization w.r.t $a_{ij}$

$$\min_{a_{ij}} \lambda|a_{ij}|_1 + ||D_1a_{ij} - Y_{ij}||_F^2 + ||D_ha_{ij} - X_{ij}||_F^2 + ||\ell_2(X_{ij}, X^E_{ij}) \cdot \exp(-p \cdot N_i(X^E_{ij})) \cdot \exp(p \cdot N_g(\alpha_{ij}))||_F $$ (8)

The major bottleneck of exactly solving (8) lies in the last exponential term. We let $a_{ij}^0$ denote the $a_{ij}$ value solved in the last iteration. We then apply first-order Taylor expansion to the last term of the objective in (8), with regard to $N_g(\alpha_{ij})$ at $\alpha_{ij} = a_{ij}^0$, and solve the approximated problem as follows:

$$\min_{a_{ij}} \lambda|a_{ij}|_1 + (1 + C)||D_1a_{ij} - Y_{ij}||_F^2$$

$$+ ||D_ha_{ij} - X_{ij}||_F^2,$$ (9)

where $C$ is the constant coefficient:

$$C = ||\ell_2(X_{ij}, X^E_{ij}) \cdot \exp(-p \cdot N_i(X^E_{ij})) \cdot \exp(p \cdot N_g(\alpha_{ij}))||_F$$

$$= p \ell_2(X_{ij}, X^E_{ij}) \cdot \omega(\alpha_{ij}^0, X^E_{ij}).$$ (10)

(9) can be conveniently solved by the feature sign algorithm [9]. Note (9) is a valid approximation of (8) since $a_{ij}$ and $a_{ij}^0$ become quite close after a few iterations, so that the higher-order Taylor expansions can be reasonably ignored.

Another noticeable fact is that since $C > 0$, the second term is always emphasized more than the third term, which makes sense as $Y_{ij}$ is the “accurate” LR image, while $X_{ij}$ is just an estimate of the HR image and is thus less weighted. Further considering the formulation (10), $C$ grows up as $\omega(\alpha_{ij}^0, X^E_{ij})$ turns larger. That implies when external SR becomes the major source of “SR noise” on a patch in the last iteration, (9) will correspondingly rely less on the last solved $X_{ij}$.

2) $X^E_{ij}$-subproblem: Fixing $a_{ij}$ and $X_{ij}$, the $X^E_{ij}$-subproblem becomes

$$\min_{X^E_{ij}} \exp(-p \cdot ||Y^m_{mn} - X^E_{ij}||_F^2) \ell_2(X_{ij}, X^E_{ij}).$$ (11)

Note that (11) is in essence a reweighed version of epitomic matching described in Section 3.2, and could be solved by similar algorithms.

3) $X_{ij}$-subproblem: With both $a_{ij}$ and $X^E_{ij}$ fixed, the solution of $X_{ij}$ simply follows a weight least square (WLS) problem:

$$\min_{X_{ij}} ||D_ha_{ij} - X_{ij}||_F^2 + \omega(a_{ij}, X^E_{ij})||X - X^E_{ij}||_F^2,$$ (12)

with an explicit solution:

$$X_{ij} = \frac{D_ha_{ij} + \omega(a_{ij}, X^E_{ij})}{1 + \omega(a_{ij}, X^E_{ij})} X^E_{ij}.$$ (13)

IV. Experiments

A. Implementation Details

We itemize the parameter and implementation settings for the following group of experiments:

- We use $5 \times 5$ patches with one pixel overlapping for all experiments except those on SHD images in Section 4.4, where the patch size is $25 \times 25$ with five pixel overlapping.
- In (2), we adopt the $D_1$ and $D_h$ trained in the same way as in (4), due to the similar roles played by the dictionaries in their formulation and our $\ell_2$ function. However, we are aware that such $D_1$ and $D_h$ are not optimized for the proposed method, and will integrate a specifically designed dictionary learning part in future work. $\lambda$ is empirically set as 1.
- In (4), the size of the epitome is $\frac{1}{4}$ of the image size.
- In (10), we set $p = 1$ for all experiments. We also observed in experiments that a larger $p$ will usually lead to a faster decrease in objective value, but the SR result quality may degrade a bit.
- We initialize $a_{ij}$ by solving coupled sparse coding in (4). $X_{ij}$ is initialized by bicubic interpolation.
- We set the maximum iteration number to be 10 for the coordinate descent algorithm. For SHD cases, the maximum iteration number is adjusted to be 5.
- For color images, we apply SR algorithms to the illuminance channel only, as humans are more sensitive to illuminance changes. We then interpolate the color layers (Cb, Cr) using plain bi-cubic interpolation.

B. Comparison with State-of-the-Art Results

We compare the proposed method with the following selection of competitive methods as follows.

- **Bi-Cubic Interpolation (“BCI” for short and similarly hereinafter),** as a comparison baseline.
- **Coupled Sparse Coding (CSC) [4],** as the classical external-example-based SR.
- **Local Self-Example based SR (LSE) [5],** as the classical internal-example-based SR.
- **Epitome-based SR (EPI).** We compare EPI to LSE to demonstrate the advantage of epitomic matching over the local NN matching.
- **SR based on In-place Example Regression (IER) [20],** as the previous SR utilizing both external and internal information.
- **The proposed joint SR (JSR).**
Fig. 2. 3× SR results of the Temple image.
We list the SR results (best viewed on a high-resolution display) for two test images: Temple and Train, by an amplifying factor of 3. PSNR measurements and zoomed local regions (using nearing neighbor interpolation) are available for different methods as well.

In Fig. 2 although greatly outperforming the naive BCI, the external-example based CSC tends to lose many fine details. In contrast, LSE brings out an overly sharp SR result with observable blockiness. EPI produces a more visually pleasing result, through searching for the matches over the entire input efficiently by the pre-trained epotope rather than a local neighborhood. Therefore, EPI substantially reduces the artifacts compared to LSE. But without any external information available, it is still incapable of inferring enough high-frequency details from the input solely, especially under a large amplifying factor. The result of IER greatly improves but is still accompanied with occasional small artifacts. Finally, JSR provides a clear recovery of the steps, and it reconstructs the most pillar textures. In Fig. 3 JSR is the only algorithm which clearly recovers the number on the carrier and the bricks on the bridge simultaneously. The performance superiorities of JSR are also verified by the PSNR comparisons, where larger margins are obtained by JSR over others in both cases.

Next, we move on to the more challenging 4× SR case, using the Chip image which is quite abundant in edges and textures. Since we have no ground truth for the Chip image of 4× size, only visual comparisons are presented. Given such a large SR factor, the CSC result is a bit blurry around the characters on the surface of chip. Both LSE and EPI create jaggy artifacts along the long edge of the chip, as well as small structure distortions. The IER result causes less artifacts but sacrifices detail sharpness. The JSR result presents the best SR with few artifacts.

The key idea of JSR is utilizing the complementary behavior of both external and internal SR methods. Note when one inverse problem is better solved, it also makes a better parameter estimate for solving the other. JSR is not a simple static weighted average of external SR (CSC) and internal SR (EPI). When optimized jointly, the external and internal subproblems can “boost” each other (through auxiliary variables), and each performs better than being applied independently. That is why JSR gets details that exist in neither internal or external SR result.

C. Effect of Adaptive Weight

To demonstrate how the proposed joint SR will benefit from the learned adaptive weight \( \omega \), we compare 4× SR results of Kid image, between joint SR solving (1), and its counterpart with fixed global weights, i.e. set the weight \( \omega \) as constant for all patches. Table 1 shows that the joint SR with an adaptive weight gains a consistent PSNR advantage over the SR with a large range of fixed weights.

More interestingly, we visualize the patch-wise weight maps of joint SR results in Fig. 2 - 4 using heat maps, as in Fig. 5. The \((i,j)\)-th pixel in the weight map denote the final weight of \( x_{ij} \) when the joint SR reaches a stable solution. All weights are normalized between \([0,1]\), by the form of sigmoid function: \( \frac{1}{1+\omega(\alpha_{ij},X_{ij}^E)} \), for visualization purpose. A larger pixel value in the weight maps denote a smaller weight and thus a higher emphasis on external examples, and vice versa. For Temple image, Fig. 5(a) clearly manifests that self examples dominate the SR of the temple building that is full of texture patterns. The most regions of Fig. 5(b) are close to 0.5, which means that \( \omega(\alpha_{ij},X_{ij}^E) \) is close to 1 and external and internal examples have similar performances on most patches.

However, internal similar makes more significant contributions in reconstructing the brick regions, while external examples works remarkably better on the irregular contours of forests. Finally, the Chip image is an example where external examples have advantages on the majority of patches. Considering self examples prove to create artifacts here (see Fig. 4 (c) (d)), they are avoided in joint SR by the adaptive weights.

D. SR Beyond Standard Definition: From HD Image to UHD Image

In almost all SR literature, experiments are conducted with Standard-Definition (SD) images (720 × 480 or 720 × 576 pixels) or smaller. The High-Definition (HD) formats: 720p (1280 × 720 pixels) and 1080p (1920 × 1080 pixels) have become popular today. Moreover, Ultra High-Definition (UHD) TVs are hitting the consumer markets right now with the 3840 × 2160 resolution. It is thus quite interesting to explore whether SR algorithms established on SD images can be applied or adjusted for HD or UHD cases. In this section, we upscale HD images of 1280 × 720 pixels to UHD results of 3840 × 2160 pixels, using competitor methods and our joint SR algorithm.

Since most HD and UHD images typically contain much more diverse textures and a richer collection of fine structures than SD images, we enlarge the patch size from 5 × 5 to 25 × 25 (the dictionary pair is therefore re-trained as well) to capture more variations, meanwhile increasing the overlapping from one pixel to five pixels to ensure enough spatial consistency. Hereby JSR is compared with its two “component” algorithms, i.e., CSC and EPI. We choose several challenging SHD images (3840 × 2160 pixels) with very cluttered texture regions, downsampling them to HD size (1280 × 720 pixel) on which we apply the SR algorithm with a factor of 3. In all cases, our results are consistently sharper and clearer. The SR results (zoomed local regions) of the Leopard image are displayed in Fig. 8 for examples, with the PSNR measurements of full-size results.

E. Subjective Evaluation

We conduct an online subjective evaluation survey on the quality of SR results produced by all different methods
Fig. 3. 3× SR results of the *Train* image.
Fig. 4. 4× SR results of the Chip image.
in Section 4.2. Ground truth HR images are also included when they are available as references. Each participant of the survey is shown a set of HR image pairs obtained using two different methods for the same LR image. For each pair, the participant needs to decide which one is better than the other in terms of perceptual quality. The image pairs are drawn from all the competitive methods randomly, and the images winning the pairwise comparison will be compared again in the next round, until the best one is selected. We have a total of 101 participants giving 1,047 pairwise comparisons, over six images which are commonly used as benchmark images in SR, with different scaling factors (Kid×4, Chip×4, Statue×4, Leopard×3, Temple×3 and Train×3). We fit a Bradley-Terry [26] model to estimate the subjective scores for each method so that they can be ranked. More experiment details are included in our Appendix. Figure 7 shows the estimated scores for the six SR methods in our evaluation. As expected, all SR methods receive much lower scores compared to ground truth (set as score 1), showing the huge challenge of the SR problem itself. Also, the bicubic interpolation is significantly worse than others. The proposed JSR method outperforms all other state-of-the-art methods by a large margin, which proves that JSR can produce more visually favorable HR images by human perception.

V. CONCLUSION

This paper presents a joint single image SR model, by learning from both external and internal examples. We define the two loss functions by sparse coding and epitomic matching, respectively, and construct an adaptive weight to balance the two terms. Experimental results demonstrate that joint SR outperforms existing state-of-the-art methods for various test images of different definitions and scaling factors, and is also significantly more favored by user perception. We will further integrate dictionary learning into the proposed scheme, as well as reducing its complexity.

APPENDIX

a) Epitomic Matching Algorithm: We assume an epitome e of size $M_e \times N_e$, for an input image of size $M \times N$, where $M_e < M$ and $N_e < N$. Similarly to GMMs, e contains three parameters [6], [22], [23]: $\mu$, the Gaussian mean of size $M_e \times N_e$; $\phi$, the Gaussian variance of size $M_e \times N_e$; and $\pi$, the mixture coefficients. Suppose $Q$ densely sampled, overlapped patches from the input image, i.e. $\{Z_k\}_{k=1}^Q$. Each $Z_k$ contains pixels with image coordinates $S_k$, and is associated with a hidden mapping $T_k$ from $S_k$ to the epitome coordinates. All the $Q$ patches are generated independently from the epitome and the corresponding hidden mappings as below:

$$\prod_{k=1}^Q p(Z_k | T_k, e) = \prod_{k=1}^Q p(Z_k | T_k, e)$$  \hspace{1cm} (14)

The probability $p(Z_k | T_k, e)$ in (14) is computed by the Gaussian distribution where the Gaussian component is specified by the hidden mapping $T_k$. $T_k$ behaves similar to the hidden variable in the traditional GMMs.

Figure 8 illustrates the role that the hidden mapping plays in the epitome as well as the graphical model illustration for epitome. With all the above notations, our goal is to find the epitome $e$ that maximizes the log likelihood function $e = \arg \max_e \log p(\{Z_k\}_{k=1}^Q | e)$, which can be solved by the Expectation-Maximization (EM) algorithm [6], [24].

The Expectation step in the EM algorithm which computes the posterior of all the hidden mappings accounts for the most time consuming part of the learning process. Since the posterior of the hidden mappings for all the patches are
Fig. 6. 3× SR results of the *Leopard* image (local region displayed).
corresponding patch of the same size in e and Z_k can be mapped to any possible epitome patch in accordance with T_{k}. (b) The epitome graphical model

independent of each other, they can be computed in parallel. Therefore, the learning process can be significantly accelerated by parallel computing.

With the epitome \( e_Y \) learned from the smoothed input image \( Y' \), the location of the matching patch in the epitome \( e_Y \) for each patch \( X_{ij}^{E} \) is specified by the most probable hidden mapping for \( X_{ij}^{E} \):

\[
T_{ij}^* = \arg \max_{T_{ij}} p\left(T_{ij} | X_{ij}^{E}, e\right)
\]

The top \( K \) patches in \( Y' \) with large posterior probabilities \( p\left(T_{ij}^* | e\right) \) are regarded as the candidate matches for the patch \( X_{ij}^{E} \), and the match \( Y_m' \) is the one in these \( K \) candidate patches which has minimum Sum of Squared Distance (SSD) to \( X_{ij}^{E} \). Note that the indices of the \( K \) candidate patches in \( Y' \) for each epitome patch are pre-computed and stored when training the epitome \( e_Y \), from the smoothed input image \( Y' \), which makes epitomic matching efficient.

Moreover, we can incorporate Nearest Neighbor (NN) matching to our epitomic matching, leading to a enhanced patch matching scheme that features both non-local (by epitome) and local (by NN) matching. Suppose the high frequency components obtained by epitomic matching and NN matching for patch \( X_{ij}^{E} \) are \( H_{ij,e} \) and \( H_{ij,NN} \) respectively, we use a smart weighted average of the two as the final high frequency component \( H_{ij} \):

\[
H_{ij} = w H_{ij,e} + (1 - w) H_{ij,NN}
\]

where the weight \( w = p\left(T_{ij}^* | X_{ij}^{E}, e\right) \) which is probability of the most probable hidden mapping given the patch \( X_{ij}^{E} \). A higher \( w \) indicates that the patch \( X_{ij}^{E} \) is more likely to have a match by epitomic matching with the probability measured through the corresponding most probable hidden mapping, thereby more weight is associated with the epitomic matching, and vice versa. This is the practical implementation of epitome-based SR (EPI) we used in the experiment section of the submitted paper.

EPI significantly reduces the artifacts and produces more visually pleasing SR results by the dynamic weighting (16), compared to the local NN matching method [5].

b) Subjective Review Experiment: The methods under comparison include BIC, CSC, LSE, IER, EPI, JSR. Ground truth HR images are also included when they are available as references. Each of the human subject participating in the evaluation is shown a set of HR image pairs obtained using two different methods for the same LR image. For each pair, the subject needs to decide which one is better than the other in terms of perceptual quality. The image pairs are drawn from all the competitive methods randomly, and the images winning the pairwise comparison will be compared again in the next round until the best one is selected.

We have a total of 101 participants giving 1,047 pairwise comparisons over 6 images with different scaling factors (“Kid”×4, “Chip”×4, “Statue”×4, “Leopard”×3, “Temple”×3 and “Train”×3). Not every participant completed all the comparisons but their partial responses are still useful. All the evaluation results can be summarized into a \( 7 \times 7 \) winning matrix \( W \) for 7 methods (including ground truth), based on which we fit a Bradley-Terry [25] model to estimate the subjective score for each method so that they can be ranked. In the Bradley-Terry model, the probability that an object \( X \) is favored over \( Y \) is assumed to be

\[
p(X > Y) = \frac{e^{s_X}}{e^{s_X} + e^{s_Y}} = \frac{1}{1 + e^{s_Y - s_X}},
\]

where \( s_X \) and \( s_Y \) are the subjective scores for \( X \) and \( Y \). The scores \( s \) for all the objects can be jointly estimated by maximizing the log likelihood of the pairwise comparison observations:

\[
\max_s \sum_{i,j} w_{ij} \log\left(\frac{1}{1 + e^{s_j - s_i}}\right)
\]

where \( w_{ij} \) is the \((i,j)\)-th element in the winning matrix \( W \), representing the number of times when method \( i \) is favored over method \( j \). We use the Newton-Raphson method to solve Eq. (15) and set the score for ground truth method as 1 to avoid the scale issue.

Fig. 7 shows the estimated scores for six SR methods in our evaluation. As expected, all the SR methods have much lower scores than ground truth, showing the great challenge in SR problem. Also, the bicubic interpolation is significantly worse than other SR methods. The proposed JSR method outperforms other previous state-of-the-art methods by a large margin, which verifies that JSR can produce visually more pleasant HR images than other approaches.

REFERENCES

[1] S. C. Park, M. K. Park, and K. G. Kang, “Super-resolution image reconstruction: A technical overview,” IEEE Signal Processing Magazine, vol. 20, no. 3, pp. 21-36, 2003.
[2] R. Fattal, “Image upsampling via imposed edge statistics,” ACM Transactions on Graphics, vol. 26, no. 3, pp. 95-102, 2007.
[3] Z. Lin, and H. Y. Shum, “Fundamental limits of reconstruction-based superresolution algorithms under local translation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 1, pp. 83-97, 2004.
[4] J. Yang, Z. Wang, Z. Lin, S. Chen, and T. Huang, “Coupled Dictionary Training for Image Super-resolution,” IEEE Transactions on Image Processing, vol. 21, no. 8, pp. 3467-3478, 2012.
[5] G. Freedman and R. Fattal, “Image and video upscaling from local self-examples,” ACM Transactions on Graphics, vol. 28, no. 3, 2010.
[6] N. Jojic, B. J. Frey, and A. Kannan, “Epitomic analysis of appearance and actions on Graphics,” vol. 26, no. 3, pp. 95-102, 2007.
[7] W. T. Freeman, T. R. Jones, and E. C. Pasztor, “Example-based super-resolution,” IEEE Computer Graphics and Applications, vol. 22, no. 2, pp. 56-65, 2002.
[8] K. I. Kim, and Y. Kwon. “Single-image super-resolution using sparse regression and natural image prior,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 6, pp. 1127-1133, 2010.

[9] H. Lee, A. Battle, R. Raina and A. Y. Ng. “Efficient sparse coding algorithms,” in Proceedings of Neural Information Processing Systems (NIPS), pp. 801-808, 2007.

[10] J. Yang, J. Wright, T. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Transactions on Image Processing, vol. 19, no. 11, pp. 2861-2873, 2010.

[11] C. Yang, J. Huang, and M. Yang, “Exploiting self-similarities for single frame super-resolution,” in Proceedings of Asian Conference on Computer Vision (ACCV), pp. 1807-1818, 2010.

[12] W. Dong, L. Zhang, G. Shi and X. Li. “Nonlocally centralized sparse representation for image restoration,” IEEE Transactions on Image Processing, vol. 22, no. 4, pp. 1620-1630, 2012.

[13] D. Glasner, S. Bagon and M. Irani, “Super-resolution from a single image,” in Proceeding of IEEE International Conference on Computer Vision (ICCV), pp. 349-356, 2009.

[14] J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman, “Non-local sparse models for image restoration,” in Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 2272-2279, 2009.

[15] P. Chatterjee, and P. Milanfar, “Is denoising dead?” IEEE Transactions on Image Processing, vol. 19, no. 4, pp. 895-911, 2010.

[16] H. Burger, C. Schuler, and S. Harmeling, “Learning how to combine internal and external denoising methods,” Lecture Notes in Computer Science: Pattern Recognition, vol. 8142, pp. 121-130, 2013.

[17] M. Zontak, “Internal statistics of a single natural image,” in Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 977-984, 2011.

[18] I. Mosseri, M. Zontak, and M. Irani, “Combining the power of internal and external denoising”, IEEE International Conference on Computational Photography (ICCP), pp. 1-9, 2013.

[19] W. Dong, D. Zhang, G. Shi, and X. Wu, “Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization”, IEEE Transactions on Image Processing, vol. 20, no. 7, pp. 1838-1857, 2011.

[20] J. Yang, Z. Lin, and S. Cohen, “Fast image super-resolution based on in-place example regression,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1059-1066, 2013.

[21] Z. Wang, Z. Wang, S. Chang, J. Yang and T. S. Huang, “A Joint Perspective Towards Image Super-Resolution: Unifying External- and Self-Examples,” accepted by IEEE Winter conference on Applications of Computer Vision(WACV), 2014.

[22] K. Ni, A. Kannan, A. Criminisi, and J. Winn, “Epitomic location recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 12, pp. 2158-2167, 2009.

[23] X. Chu, S. Yan, L. Li, K. Chan, and T. Huang, “Spatialized epitome and its applications”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 311-318, 2010.

[24] Y. Yang, X.Chu, T. Ng, A. Chia, J. Yang, H. Jin and T. Huang, “Epitomic image colorization”, in Proceedings of IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014.

[25] D. Bertsekas, Nonlinear Programming, 2nd ed. Nashua, NH: Athena Scientific, 1999.

[26] R. A. Bradley, and M. E. Terry, “Rank analysis of incomplete block designs: I. The method of paired comparisons,” Biometrika, pp. 324-345, 1952.