Super-resolution from Internet-scale Scene Matching

Libin Sun
Brown University
lbsun@cs.brown.edu

James Hays
Brown University
hays@cs.brown.edu

Abstract

In this paper, we present a highly data-driven approach to the task of single image super-resolution. Super-resolution is a challenging problem due to its massively under-constrained nature - for any low-resolution input there are numerous high-resolution possibilities. Our key observation is that, even with extremely low-res input images, we can use global scene descriptors and Internet-scale image databases to find similar scenes which provide ideal example textures to constrain the image upsampling problem. We quantitatively show that the statistics of scene matches are more predictive than internal image statistics for the super-resolution task. Finally, we build on recent patch-based texture transfer techniques to hallucinate texture detail and compare our super-resolution with other recent methods.

1. Introduction

Single image super-resolution is a well-studied problem where one tries to estimate a high-resolution image from a single low-resolution input. Unlike multi-frame super-resolution where a sequence of low-resolution images of the *same scene* aligned to subpixel shifts reveal some high frequency detail signals, it is impossible to unambiguously restore high frequencies in a single image super-resolution framework. Single image super-resolution is an extremely under-constrained problem: there are many plausible natural images that would downsample exactly to a given low-resolution input. As a result, existing works in the field present intelligent ways to *hallucinate* plausible image content instead of recovering the ground truth.

Over the past decades there has been impressive work on single image super-resolution, but no method is able to address the fundamental problem of synthesizing novel object and texture detail. Many methods can sharpen edges, but the long term challenge in this field is to produce realistic, context-appropriate texture, material, and object detail. This limitation becomes more apparent as the desired magnification factor increases. State-of-the-art methods [11, 31] can sometimes synthesize convincing detail for blurs equivalent to a 2x or 3x loss of resolution, but we compare algorithms with a more challenging task – 8x super-resolution. While 8x magnification might seem extreme, the equivalent amount of detail loss is commonly caused by imaging artifacts such as defocus or motion blur.

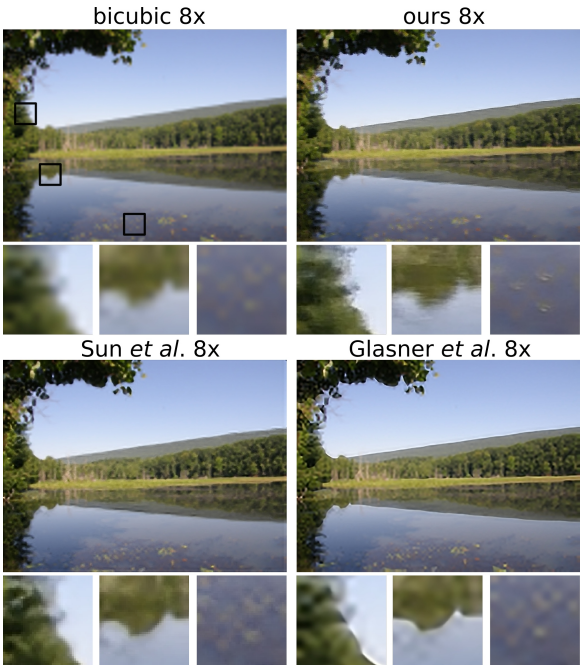

Figure 1. Super-resolution results for 8x upsampling. The input image is 128 pixels wide. We compare our results to those of Sun and Tappen [31] and Glasner et al. [11].

How can an algorithm synthesize appropriate detail for an arbitrary, low-resolution scene? Our visual experience is extraordinarily varied and complex – photos depict a huge variety of scene types with different viewpoints, scales, illuminations, and materials. How can an algorithm have image appearance models specific to each possible scene configuration? Recent works [6, 13, 14, 16] show that with diverse, "Internet scale" photo collections containing millions of scenes, for most query photos there exist numerous

examples of very similar scenes. A key insight of this paper is that research in scene representation and matching has advanced such that one can find similar enough scenes *even when a query is very low-resolution* and then use these matching scenes as a context-specific high-resolution appearance model to enhance the blurry scene. This lets us convincingly enhance image detail at magnification factors beyond previous super-resolution methods.

Our primary contributions are that: (1) We examine scene matching in a low-resolution regime that has rarely been studied. The notable exception is "Tiny Images" [33] which limited experiments to an intentionally impoverished representation. (2) We quantify the expressiveness and predictive power of matched scenes and show that they are competitive with single-image priors. This contrasts with and expands upon the findings of Zontak and Irani [38]. (3) We produce super-resolution results with plausible image detail beyond the capabilities of existing super-resolution method for diverse photographic scenes. Compared to previous work, our results are especially convincing for *texture transitions* which challenge previous region-matching super-resolution methods.

1.1. Repairing Image Blur

There is an enormous body of research aimed at alleviating the effects of blur-inducing imaging phenomena – defocus, motion, and scattering to name a few. Photographic blur can not be unambiguously inverted in realistic imaging conditions [3], therefore "...the central challenge ... is to develop methods to disambiguate solutions and bias the processes toward more likely results given some prior information"[17]. Deblurring algorithms tend to use relatively compact, parametric image priors, often learned from natural image statistics, that encode principles such as "edges should be sharp", "gradients should be rare", "colors should be locally smooth" [37, 26, 27, 4, 35, 17, 5]. These parametric models are helpful but limited. Their assumptions, such as a heavy-tailed gradient distribution, are not universally true [5]. In general, these models can sharpen edges but will not enhance texture, material, or object detail because these phenomena are too complex for the models.

1.2. Super-resolution

Unlike the previous causes of blur in which inverting a point spread function can sometimes yield useful detail, with single image super-resolution it is clearer which detail can be "recovered" (none of it) and what detail must be "hallucinated" or "synthesized" (all of it). This ambiguity makes super-resolution a demanding application for statistical image models or priors.

While some recent super-resolution methods use parametric image priors similar to those used in deblurring applications (e.g. [8, 22]), many super-resolution methods in the last decade utilize *data-driven* image priors, starting with the seminal work of Freeman et al. [10, 9]. Such data-driven methods implicitly or explicitly "learn" the mapping between low and high-resolution image patches [19, 12, 32, 31]. A data-driven prior does not make the super-resolution problem any less ambiguous – it is simply a more expressive model for proposing high-frequency versions of low-resolution image content.

Consider a hypothetical, ideal super-resolution algorithm. When presented with a low-resolution mountain, it would insert details only appropriate to mountains. When presented with a face, it would insert details specific to faces. This idea led to the development of very effective *domain specific* face super-resolution algorithms [3, 21]. But for real scenes, to insert the most plausible detail, one must first *recognize* the context[1]. The seminal work of Baker and Kanade [3] refers to this process as "recogstruction", a portmanteau of "recognition" and "reconstruction".

To achieve "recogstruction" one needs to go beyond compact, parametric models or data-driven models trained from tiny image patches that are not expressive enough for recognition or reconstruction. Recent works [12, 31] add explicit or implicit material/texture recognition to help alleviate the limits of these local, compact representations. In both methods, low-resolution input images are segmented and each segment is constrained to synthesize details by drawing patches from matched material or texture regions which are hopefully semantically and visually similar. These methods are very promising, but in both cases the material matching is not reliable – material recognition is very hard [20] and it is even harder at low-resolution. [12] alleviates this difficulty with manual intervention. Another difficulty with these approaches is handling boundaries between texture regions. [12] resorts to self-similarity for edge refinement because they do not have training examples of texture transitions. In [31], the segments do not capture the diverse texture transition scenarios either, and their algorithm relies on an edge smoothness prior to produce sharp edges. Our algorithm requires no such special case because our matched scenes typically contain the same texture transitions as our query scene.

One complementary and surprisingly effective way to synthesize scene-appropriate detail is to build a statistical image model from the low-resolution input image itself [11]. This only works when a scene exhibits self-similarity across scales, but this is common because perspective projection causes surfaces to span many scales. More recently, Zontak and Irani [38] argue that these "internal" image statistics are often a *better* prior than "external" image databases for image restoration tasks. One of our key

[1] The "recognition" does not need to be explicit – an algorithm needs to establish correspondence among visually and semantically similar image content, whether that involves explicit classification or not.

results is to use the evaluation protocol of [38] to show that it is possible to compete with single-image internal statistics by intelligently leveraging a large image database (Section 3.1).

The methods by Sun and Tappen [31] and Glasner et al. [11] are representative of the state-of-the-art in automatic super-resolution, but they still do not reliably insert texture or object detail into photographs. More often than not the results show sharper edges but are not convincing beyond magnification factors of 2 or 3. We compare our results to these algorithms in Section 5.

Beyond single image super-resolution, there is ongoing research for which the input is multiple photographs of the same physical scene. For instance, "Photozoom" [7] relates photographs with a hierarchy of homographies and then transfers details. Lastly, image enhancement methods such as "CG2Real" [16] can be modified to perform super-resolution by inputting blurry scenes. However, CG2Real assumes that the input is corrupted in some way and thus is not faithful to the input image, as is desirable in super-resolution.

1.3. Super-resolution Goals and Evaluation

In typical image restoration and super-resolution literature (e.g. [17, 5]) the formal goal is to recover "clean" scene x given blurred scene y, a known PSF (or blur kernel) k[2], and a known downsampling function D. These variables have the following relationship: $y = D(x \otimes k) + n$ where $\otimes$ is the convolution operator and n is a noise term. One can then evaluate a result by comparing the estimated x to the known, "ground truth" x which generated y.

This evaluation makes sense when (1) k is small or invertible and (2) you are interested in forensically accurate reconstructions. But when either k or the downsampling factor becomes large, the possible values for x grow enormously. For 8x super-resolution, the output space is 64 times higher-dimension than the observed low-resolution input. There is an enormous space of detailed and plausible output images that are faithful to the low-resolution input. Why should one penalize a convincing result just because it doesn't resemble the observed "ground truth"? Recognizing this problem, recent work has adopted a more forgiving comparison between the estimated x and "ground truth" – SSIM [34] – which rewards local structural similarity rather than exact pixel to pixel correspondence. However, SSIM and other existing measurements of reconstruction error penalize texture hallucination (See Figure 2 for an example).

Rather than evaluating reconstruction error, an alternative is to perform human perceptual studies [22]. Such experiments are difficult, though, because of the subjective biases of individual, non-expert observers. In Section 5 we perform such a study. However, we think the most diagnostic results are qualitative in nature – in Section 5 we show that our approach is able to insert edge and texture detail in diverse scenes where previous methods could not.

[2]For super-resolution a Gaussian blur of appropriate width can be used as the PSF [11].

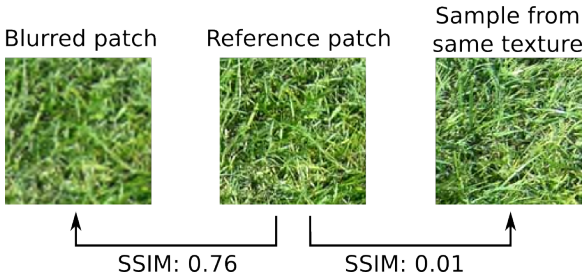

Figure 2. SSIM scores calculated with respect to the reference patch on the left. The middle patch, cropped from the same texture, scores poorly while the patch on the right, a blurred version of the reference, scores very highly. Because SSIM and other reconstruction measures favor blur over texture misalignment, they favor conservative algorithms which do not insert texture details.

2. Algorithm Overview

Our algorithm (Figure 3) first finds matching scenes from a large Internet database (Section 3). The input image and each matching scene are segmented and a correspondence is found between each segment in the input and several best matching segments from the similar scenes (Section 4.1). Finally, each input segment is upsampled by matching low-resolution patches and transferring in high-resolution details from its corresponding segments (Section 4.2).

The local patch matching at the heart of most data-driven super-resolution algorithms is fundamentally ambiguous – it is hard to match to semantically similar texture based on local image evidence regardless of the size of the training database. Our pipeline follows a coarse-to-fine structure not just to reduce computational complexity, but also to alleviate this ambiguity by constraining matching to segments from scenes which are hopefully semantically similar. Instead of making decisions entirely locally, we make easier decisions at the scene and segment level first. Constraining the synthesis process to a small number of regions from similar scenes also increases perceived texture coherence.

3. Scene Matching

Our proposed detail synthesis pipeline can be thought of as taking the data-driven super-resolution trend to its extreme by using a massive, "Internet-scale" photo collection as an extremely detailed statistical image model. While the state-of-the-art method of [31] uses a training set of four thousand images, the largest to date, our algorithm uses

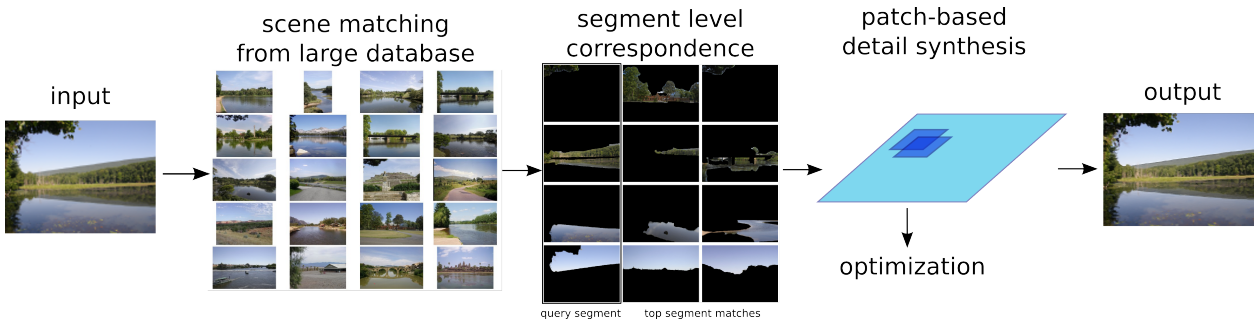

Figure 3. Our proposed pipeline. From left to right, for a low-resolution input we find most similar scenes from a large database. Each input segment is corresponded with best matching segments in these similar scenes. Then a patch-based super-resolution algorithm is used to insert detail from the matched scene segments.

more than six million images. We follow in the footsteps of several successful massively data-driven scene matching algorithms, e.g. [13, 29, 6, 28, 16], which sample the space of scenes so densely that for most query scenes one can find semantically and structurally similar scenes.

A key insight for this paper is that while other super-resolution representations and models can not understand the context presented by low-resolution scenes, scene matching can succeed even in the presence of extreme blur. If we can find very similar scenes for a low-resolution query then those scenes provide an *ideal* set of context-appropriate textures of similar scale, illumination, and viewpoint to use for detail synthesis.

However, our application of scene matching is especially difficult because the input images are low-resolution and thus have degraded textures, which are the most discriminative scene features [36]. To make the most of what scene statistics remain we use a combination of scene descriptors – color histograms, tiny images [33], gist descriptors [24], dense texton histograms [23], sparse bags-of-visual-words [30] built with "soft assignment" [25], geometric layout [15], and surface-specific color and texton histograms [36]. The distances in each feature space are weighted such that each feature contributes roughly equally to the ranking of top scene matches.

For accurate scene matching, the scene features in our photo collection need to be computed at the same resolution as a query scene. However, the query scene can be of arbitrarily low resolution and recomputing features for an entire database is computationally expensive. Therefore we use a hierarchical scene matching process where initial matches are found at a low, fixed resolution, then for each initial match the scene descriptors are recomputed at the query resolution and the matches are re-ranked. Figure 4 shows examples of scene matches for several queries where each input image is only 128 pixels wide.

To find similar scenes we need a diverse photo collection with millions of example scenes. We use the Flickr-derived database of [14] which contains over 6 million high resolution photographs. Because we use this photo database to learn the relationship between low-resolution scenes and high-frequency details, it is important that all scenes *actually contain* high-frequency details. Therefore we filter out all blurry photographs using the "blur" classifier of [18]. This disqualifies about 1% of photographs. We use the top 20 matches for each input image as a scene-specific training database for detail enhancement.

3.1. Understanding the Quality of Scene Matches

Data-driven super-resolution methods estimate a high-resolution image by matching to a database of low and high resolution pairs of patches. In our case, the database is a set of query-specific scene matches. Recently, Zontak and Irani [38] proposed criteria to assess the value of training databases for image restoration tasks. First, **expressiveness** quantifies how well patch matches from a database *could possibly* reconstruct the ground truth. Second, **predictive power** quantifies how effective patch matches from a database are at constraining the solution toward the ground truth. Expressiveness is similar to the "reconstruction error" examined in [2] for image databases with trillions of patches.

In the following subsections, we analyze two "external" databases: (1) our query-specific scene matches and (2) random scenes from the Berkeley Segmentation Dataset (BSD) [23], and two "internal" databases: (1) a database of all scales of the full resolution *ground truth*, except for a 21x21 window around the current patch under consideration and (2) a limited internal database of all scales of the *input* image. Of the two internal databases, only the "limited" variant is applicable to the task of super-resolution because one does not have access to the full-resolution ground truth during super-resolution. Internal databases include patches at scales of $0.8^i, i = \{0, 1, 2, 3, 4, 5, 6\}$ while external databases are not multi-scale.

We use a test set of 80 diverse scenes and evaluate ex-

Figure 4. For four low-resolution query scenes, we show six of the top twenty scene matches that our algorithm will use to insert high-frequency detail. The last row shows an example of scene match failure. For a small portion of test cases the scene matching finds some instance-level matches, as in the Venice image, but generally this is not the case. We will explicitly indicate when a result was generated using instance-level matches.

pressiveness and predictive power for the task of 2x super-resolution. We analyze 2x super-resolution to be consistent with [38] even though we show results for 8x super-resolution in Section 5. At higher levels of magnification the internal image statistics are increasingly unhelpful for super-resolution. Even though the task is 2x super-resolution, scene matches are found from input images at 1/8 resolution. We resize all images to a maximum dimension of 512 pixels and convert to grayscale. Query patches are sampled uniformly across all gradient magnitudes from input images.

3.1.1 Expressiveness

Expressiveness provides an upper-bound for image restoration tasks if there were an oracle guiding selection of high-resolution patches out of a database. An infinite database of random patches would have perfect expressiveness (but poor predictive power). Expressiveness is defined by the average L_2 distance between each ground truth patch and its nearest neighbor in a database. Patch comparisons are made with 5×5 patches with DC removed. Figure 5 compares the expressiveness of the ground truth high resolution image, the limited internal scales derived from the input image itself, 20 random images from BSD [23], and the 20 best scene matches for each query.

Zontak and Irani [38] show that it is favorable to exploit the stability of single image statistics for tasks such as denoising and super-resolution because the same level of *expressiveness* can only be achieved by external databases with hundreds of random images. Indeed, the "internal (all scales)" database outperforms 20 random images and 20 scene matches. But the "internal (limited)" scenario which simulates the super-resolution task is less expressive than both external databases. The 20 scene matches are only slightly more expressive than 20 random images. We believe this is because expressiveness favors variety. However, this variety causes the random BSD images to have less predictive power. Overall, this analysis shows that, compared to other approaches, our scene matches contain slightly more relevant appearance statistics to drive a super-resolution algorithm.

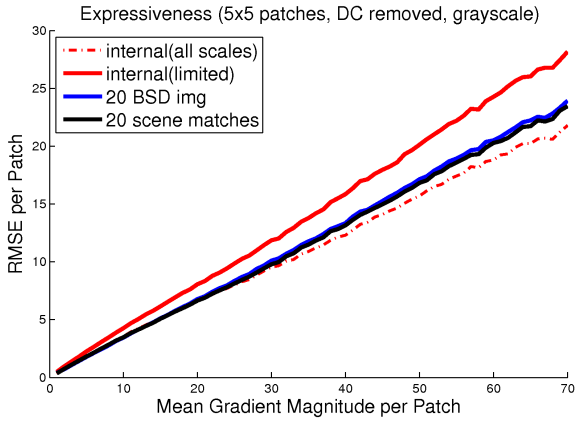

Figure 5. Comparison of expressiveness of internal vs external databases. Using up to 20 scene matches, the expressiveness of external database can be significantly better than internal. The "limited" internal database is the low frequencies of the input image that would be usable for a super-resolution algorithm. 150,000 query patches from 80 query images were sampled to generate the plots.

3.1.2 Predictive Power

The predictive power involves two measurements: (i) *prediction error* and (ii) *prediction uncertainty*. For each 5×5 low-resolution query patch l (DC removed), we find the 9 most similar low-res patches $\{l_i\}_1^9$ and set the predicted high-res patch to $\hat{h} = \frac{\sum_i w_i \cdot h_i}{\sum_i w_i}$, where h_i is the high-res

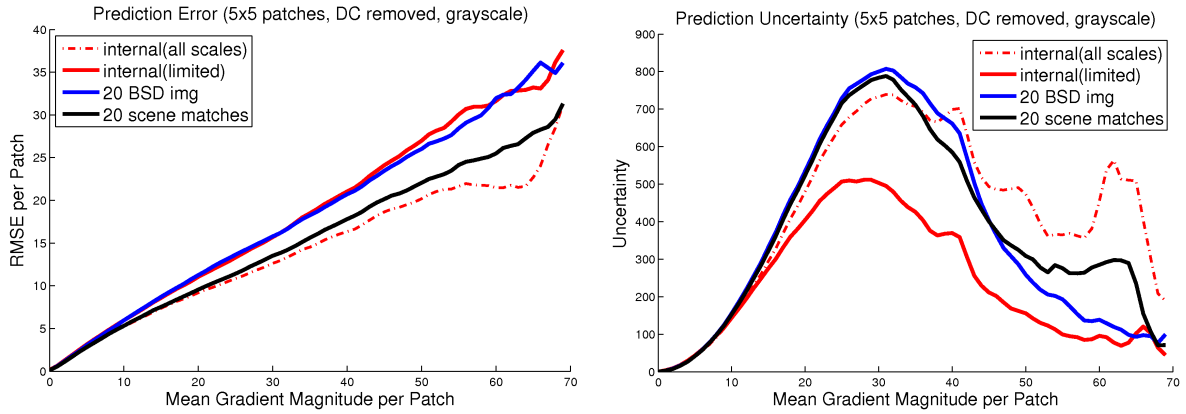

Figure 6. Comparison of prediction error and uncertainty of internal vs external databases. A total of 180,000 query patches sampled uniformly from our 80 test cases are used for this experiment.

patch corresponding to l_i, and w_i is a similarity score defined by $w_i = \exp\{-\frac{||l-l_i||_2^2}{2\sigma^2}\}$. Then, prediction error is simply the SSD between the ground truth and estimated high-res patch: $||h_{GT} - \hat{h}||_2^2$; and prediction uncertainty is approximated by $trace(cov_w(h_i, h_j))$, using the same weighting scheme. In our experiments, we set $\sigma^2 = 50$.

Figure 6 plots the prediction error (left) and prediction uncertainty (right) against the mean gradient magnitude per patch. Prediction error is arguably the most important metric for a super-resolution database, and here our scene matches outperform the internal and random external super-resolution databases. In fact, the "internal (all scales)" condition which is something of an upper-bound for this task is only slightly more predictive than the scene matches.

In the prediction uncertainty evaluation, an unexpected observation is that toward high gradient magnitude the curve starts to drop. We speculate the reason is that (1) high gradient patches contain sufficient information (even at low-res) to make the matching unambiguous, and (2) high gradient patches are rare, thus there are fewer patches to possibly match to.

Overall, our external database of scene matches is more expressive, has lower prediction error, and comparable prediction uncertainty compared with single image statistics (the "limited" scenario which corresponds to super-resolution).

However, the relative expressiveness and prediction power of these strategies can change depending on which transformations are considered and which representation is used for the matching. For instance, expressiveness can be improved significantly by considering transformations of each database such as rotations, scalings, mirroring, and contrast scaling. However, enriching a database in this manner tends to decrease predictive power. Therefore we did not apply these transformations to our external databases. In [38] the internal database includes rotated versions of the input, but adding rotations did not significantly impact our evaluations. Also note that while these plots are a valuable quantitative evaluation of the training database, they are not a direct predictor of synthesis quality. For instance, a good patch-based synthesis algorithm will overcome prediction uncertainty by considering spatial overlap between patches and this analysis intentionally ignores that.

4. Super-resolution Method

Our detail synthesis algorithm is similar to the method proposed in [31]. The significant difference is that our synthesis method is constrained to sample from a small set of scene matches while [31] uses a *universal* database of image segments. We also differ from [31] in that we use a greedy texture transfer method which considers high frequency coherence instead of picking candidate patches independently.

4.1. Segmentation and Texture Correspondence

While our scene matches provide expressive, context-specific image content for hallucination, we want to constrain the local patch matching further. An exhaustive search over all scene matches while synthesizing textures is inefficient, but more importantly it leads to texture incoherence as each local patch could potentially draw textures from very different sources. Constraining the local texture search by first matching at the region level significantly reduces output incoherence and helps push back against the prediction uncertainty observed in Figure 6.

We use a recent hierarchical segmentation algorithm [1] to segment our input and matched scenes. Extremely small segments are merged to nearby ones to provide more stable segment matching results. Each segment is represented by color histograms and texton histograms, and the top 5 most similar scene match segments for each input segment are

found using chi-square distance. These segments provide a relevant yet highly constrained search space for detail insertion. An example segment-level correspondence is shown in Figure 7.

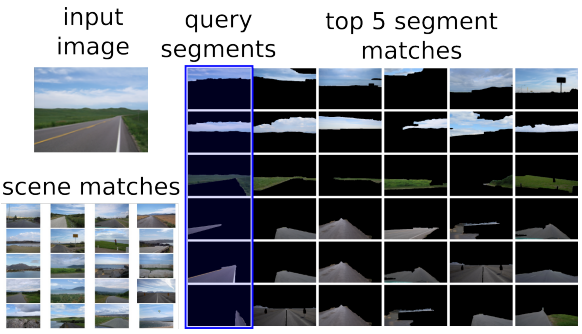

Figure 7. Counter-clockwise from upper left: Input image, top 20 scene matches, and the top 5 matching segments for the largest input segments. Each input segment is restricted to draw texture from slightly expanded versions of these matched segments.

Using non-overlapping segments presents a problem at segment transitions. By definition, segments tend not to contain these transition zones. Such transitions are also hard to find in a universal database of image segments [12, 31]. E.g., even if each region is correctly matched to brick, vegetation, sky, etc., there may be no examples of the transitions between those regions. For this reason, previous methods rely on single-image self-similarity [12] or parametric priors [31] to handle segment boundaries. Alternatively, scene matches allow us to elegantly handle texture transitions and boundaries because our scene matches often contain the same transitions (e.g. building to grass, tree to sky) as a query scene. We simply expand each segmented region to include the transition region of textures and boundaries. Thus our segmentations are actually overlapping and not a strict partitioning of the images.

4.2. Segment-level Synthesis of Coherent Textures

As shown in [38] and Figure 6, the under-constrained nature of super-resolution causes large uncertainty in the missing high frequencies in the image. When the upsampling factor is large, i.e. 8x, finding appropriate patches based on local evidence alone is fundamentally ambiguous [3].

We use a greedy tiling procedure similar to the "single-pass" algorithm of [9], allowing each subsequent patch choice to be conditioned on existing high frequencies and thus providing a well-constrained environment for synthesizing details. We do not expect this step to generate perfect textures, but allow for *opportunistic* insertion of details while remaining faithful to the low frequencies. Let P^l be a low-resolution input patch with DC removed and let I^h_x, I^h_y be the existing image gradient of the output image in the x and y direction respectively. Initially I^h_x, I^h_y are set to 0. Let S_P be the segment containing patch P, and $\mathbf{S}(S_P)$ be the top 5 most similar example segments to S_P. We seek to find among $\mathbf{S}(S_P)$ a patch Q (with DC removed) that is both *faithful* and *coherent*:

$$Q = \arg \min_{Q \in \mathbf{S}(S_P)} D_f(P^l, Q^l) + \beta D_c(I^h_x, I^h_y, Q^h) \quad (1)$$

where

$$D_f(P^l, Q^l) = \sum_i |P^l(i) - Q^l(i)| \quad (2)$$

$$D_c(I^h_x, I^h_y, Q^h) = \sum_{j \in overlap} |I^h_x(j) - \nabla Q^h_x(j)| + |I^h_y(j) - \nabla Q^h_y(j)| \quad (3)$$

Then we update the existing high frequencies by a weighted average of the gradients copied from Q^h, with the weights for each Q defined by $w = (D_f + \beta D_c)^{-0.8}$. Query patches are sampled over a half-overlapping grid while database patches are densely sampled.

After we have our set of overlapped patches, we carry out the super-resolution optimization described in [31], using the set of patches $\{Q\}$ to generate the pixel candidates for the hallucination term, so that neighboring pixels in the output image will be collectively constrained by a group of coherent pixel values. Similar to other super-resolution methods, we find it advantageous to incrementally upsample the image, so we upsample the input image by a factor 2 three times to achieve 8x magnification. We make no effort to optimize the running time of our algorithm so it is quite slow – roughly four hours per image, most of which is spent on the last 2x upsampling.

5. Results

We compare our algorithm against two recent methods which we consider exemplary of the state-of-the-art in super-resolution – [31] uses segment-level matching to a database of thousands of images, and [11] uses internal image statistics. We also show bicubic interpolation as a baseline. Figure 8 and figure 9 show results on man-made scenes and natural scenes, respectively. Figure 10 shows results where some of our scene matches are instance-level matches. Finally, figure 11 shows cases in which our algorithm produces undesirable artifacts. To help visualize the level of detail achieved by each method we zoom in on three crops from each result. In general, out results exhibit sharp edges, natural transition of textures and distinctive details.

Figure 12 compares results from our algorithm, which draws texture from matched scenes, against a baseline which instead uses random scenes from our database. This "random scene" baseline is similar to the early data-driven super-resolution method of [9] in which patches were

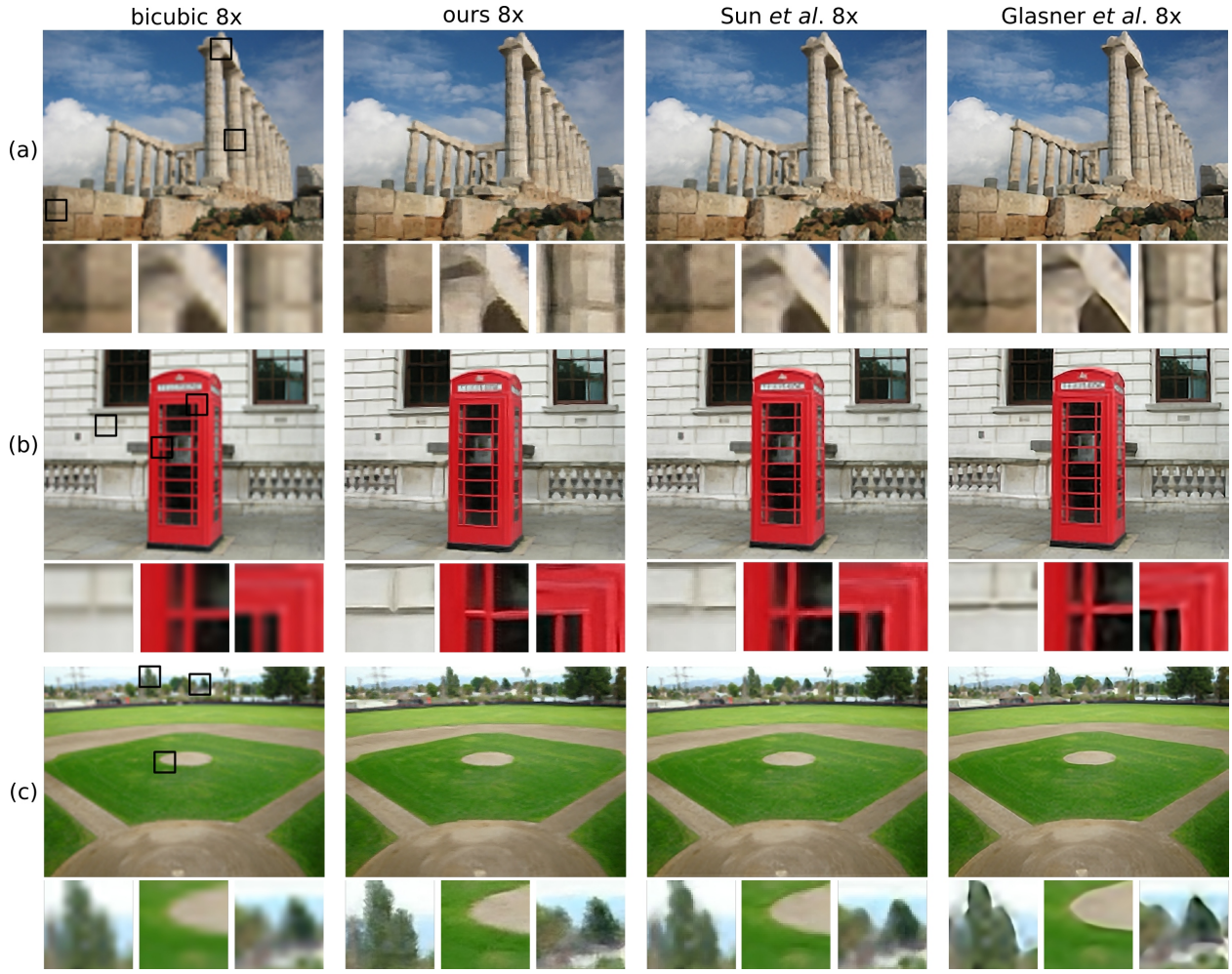

Figure 8. Results on man-made scenes. Appropriate textures/materials can be observed among the trees in (c) and surfaces in (a). Edges appear realistic and detailed in (b).

matched in a small, universal image database. The diverse random scenes still guide the algorithm to produce sharper edges than bicubic interpolation, but there is no further detail added.

To help evaluate the quality of our results and to further evaluate the contribution of scene matching, we perform a comparative perceptual study with 22 participants. We use 20 test cases for the study – 10 which have good scene matches and 10 which have bad scene matches, as evaluated by the authors. As in [22], we show participants pairs of super-resolution outputs from the same input image but different algorithms and ask them to select "the image they think has better quality". We also allow a participant to indicate that the images are equally good. The left/right placement of outputs is randomized. In a pilot study we found that that our results and those of Sun and Tappen [31] were almost universally favored over [11] and bicubic, so we exclude them from the main study. Figure 13 shows the preference of participants towards each algorithm for the test

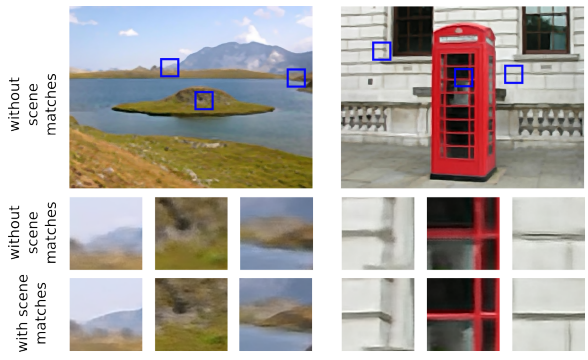

Figure 12. From top to bottom: super-resolution results using random scenes rather than matching scenes, zoomed in crops, and the corresponding crops from our algorithm using matched scenes.

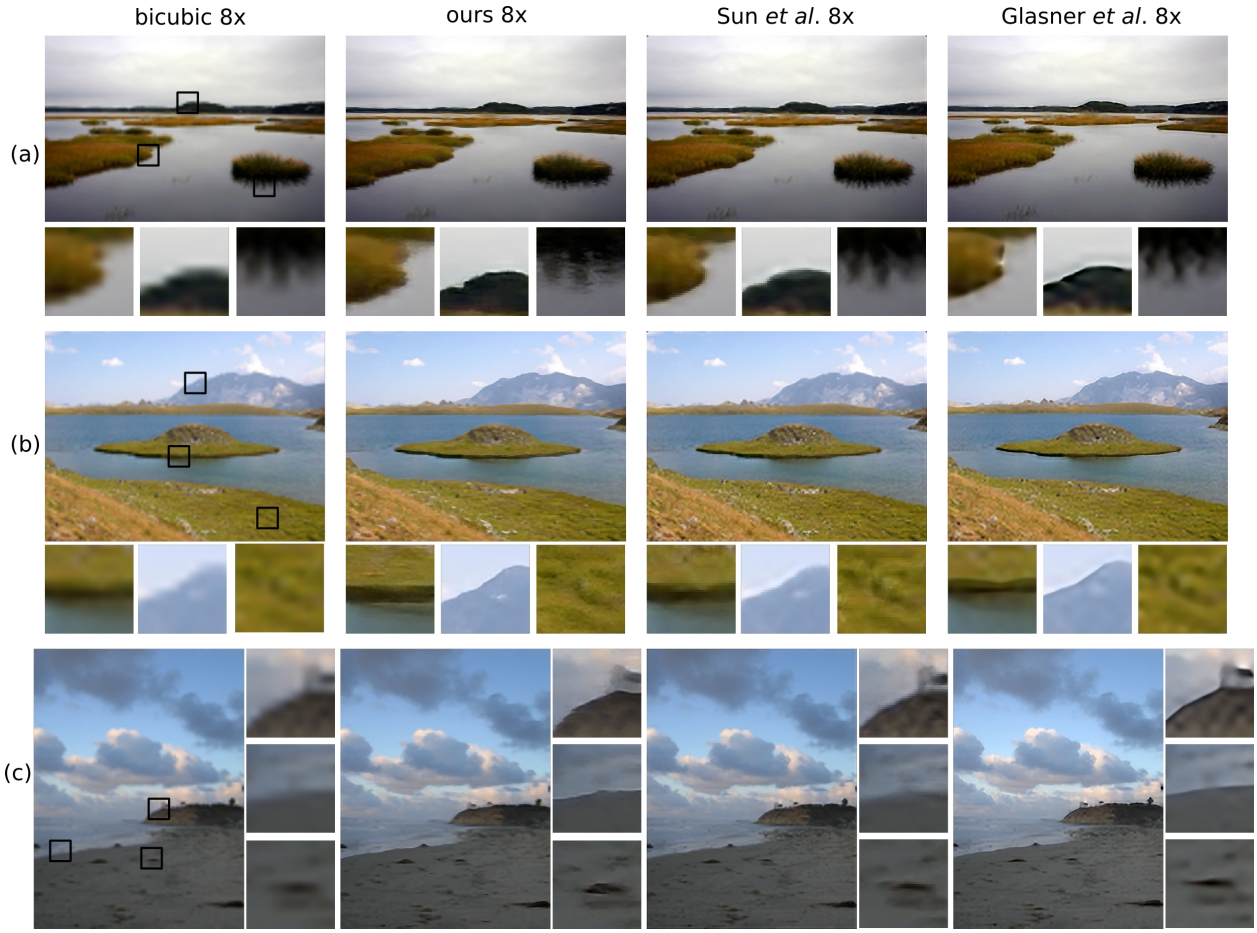

Figure 9. Results on natural scenes. Our results show successful hallucination of details in water, grass and sand. Some of the details might actually violate the downsampling reconstruction to some extent, but they certainly appear reasonable and appropriate.

cases with "good" and "bad" scene matches. While participants seem to favor our algorithm when the scene matching is successful, the task is quite subjective – a few users preferred our algorithm on almost all outputs while some users exclusively preferred [31]. We believe this discrepancy arises because our results tend to have more detail but also more artifacts and individual participants weigh these factors differently. We experimented with study designs which ask users about "detail" and "realism" separately, but we find that observers have trouble disentangling these factors.

6. Discussion

Our algorithm is somewhat more likely to introduce artifacts than other state-of-the-art algorithms because it is more aggressive about inserting texture detail. Most algorithms err on the side of caution and avoid committing to texture details because a single bad patch-level correspondence can produce a glaring artifact which ruins a result.

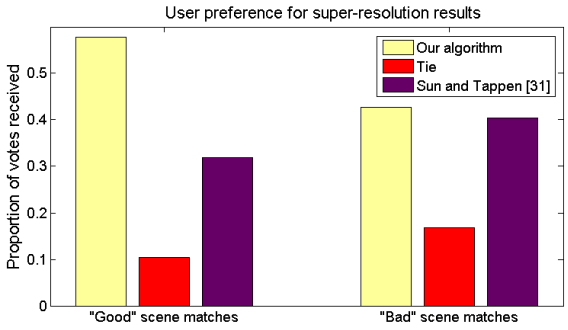

Figure 13. The breakdown of votes when participants compared our results to those of Sun and Tappen [31]. For scenes where the scene matches offered very similar textures to the input (left), participants favor our results. For scenes where the scene matches are spurious or mismatched in scale neither algorithm is favored.

Only with a large database and our scene matching pipeline can we safely insert textures for many low-resolution im-

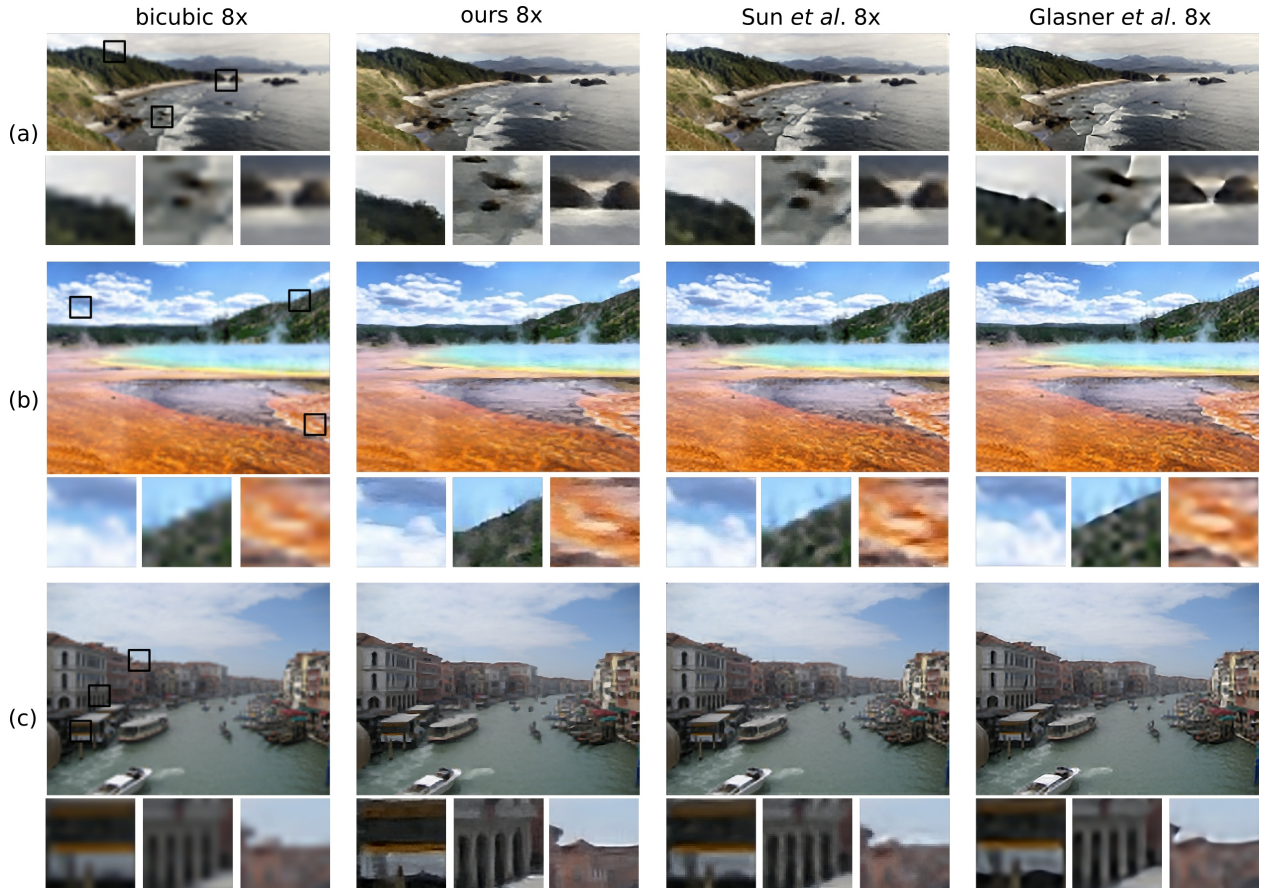

Figure 10. Results where we have at least one instance level scene match. Our algorithm is able to hallucinate salient image structures. For example, the ferry and arches in (c) are successfully hallucinated. In this case, they also approximate the ground truth.

ages. We do not claim that our algorithm represents the unambiguous state-of-the-art in super-resolution. The algorithms we compare against perform well, especially at lower levels of magnification. Existing algorithms are less likely to make mistakes for inputs where our scene matching algorithm may have trouble, such as indoor or rarely photographed scenes. However, we expect scene matching to perform more reliably as better image descriptors are developed and larger image databases become commonplace. We also think that our approach is complementary to prior methods which make use of internal image statistics, but we think that the quality of "external" databases is likely to increase faster than the quality of "internal" databases which are fundamentally quite limited.

While scene matching quality is important, we believe that the quality of our results is strongly bottlenecked by the well-studied texture transfer problem. Even for scenes with excellent scene matches our algorithm can produce surprisingly poor results. For example, when the scene consists of highly intricate textures without dominant structure it is hard to synthesize coherent textures, as shown in Figure 14. Such difficulties persist with alternative texture transfer schemes such as those based on Markov Random fields [10] or texture optimization [12]. While the super-resolution task is certainly "vision hard", it seems as if there is much progress to be made by improving relatively low-level texture transfer optimizations.

7. Acknowledgement

We are grateful to the authors of [31, 11] for running their algorithms on our test set images. We would also like to thank Jarrell Travis Webb, John F. Hughes, Erik B. Sudderth, Pedro Felzenszwalb, and our anonymous reviewers for their insightful suggestions. This work is funded by NSF CAREER Award 1149853 to James Hays.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. volume 33, pages 898–916, 2011.

Figure 11. Results which contain noticeable artifacts.

Figure 14. Failure example with excellent scene matches. Top row: input image (left) and scene matches (right). Bottom row: close-up view of output result at locations indicated by the blue squares.

[2] S. Arietta, J. Lawrence, and J. Lawrence. Building and using a database of one trillion natural-image patches. pages 9–19, 2011.

[3] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.

[4] E. P. Bennett, M. Uyttendaele, C. L. Zitnick, R. Szeliski, and S. B. Kang. Video and image bayesian demosaicing with a two color image prior. In *ECCV*, 2006.

[5] T. S. Cho, N. Joshi, C. L. Zitnick, S. B. Kang, R. Szeliski, and W. T. Freeman. A content-aware image prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[6] K. Dale, M. K. Johnson, K. Sunkavalli, W. Matusik, and H. Pfister. Image restoration using online photo collections. In *International Conference on Computer Vision*, 2009.

[7] M. Eisemann, E. Eisemann, H.-P. Seidel, and M. Magnor. Photo zoom: High resolution from unordered image collections. In *GI '10: Proceedings of Graphics Interface 2010*, pages 71–78. Canadian Information Processing Society, 2010.

[8] R. Fattal. Image upsampling via imposed edge statistics. *ACM Trans. Graphics (Proc. SIGGRAPH 2007)*, 26(3), 2007.

[9] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. In *IEEE Computer Graphics and Applications*, 2002.

[10] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000.

[11] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *ICCV*, 2009.

[12] Y. HaCohen, R. Fattal, and D. Lischinski. Image upsampling via texture hallucination. In *International Conference on Computational Photography*, 2010.

[13] J. Hays and A. A. Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (SIGGRAPH 2007)*, 26(3), 2007.

[14] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *CVPR*, 2008.

[15] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *Int. J. Comput. Vision.*, 75(1), 2007.

[16] M. K. Johnson, K. Dale, S. Avidan, H. Pfister, W. T. Freeman, and W. Matusik. Cg2real: Improving the realism of computer-generated images using a large collection of photographs. *IEEE Transactions on Visualization and Computer Graphics*, 2010.

[17] N. Joshi, C. L. Zitnick, R. Szeliski, , and D. Kriegman. Image deblurring and denoising using color priors. In *CVPR*, 2009.

[18] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, pages 419–426, 2006.

[19] K. I. Kim and Y. Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(6), 2010.

[20] C. Liu, L. Sharan, R. Rosenholtz, and E. H. Adelson. Exploring features in a bayesian framework for material recognition. In *CVPR*, 2010.

[21] C. Liu, H. Y. Shum, and W. T. Freeman. Face hallucination: theory and practice. *International Journal of Computer Vision (IJCV)*, 75(1):115–134, 2007.

[22] F. Liu, J. Wang, S. Zhu, M. Gleicher, and Y. Gong. Visual-quality optimizing super resolution. *Computer Graphics Forum*, 28(1):127–140, 2009.

[23] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. ICCV*, July 2001.

[24] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. In *Visual Perception, Progress in Brain Research*, volume 155, 2006.

[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[26] S. Roth and M. J. Black. Fields of experts: a framework for learning image priors. In *CVPR*, 2005.

[27] S. Roth and M. J. Black. Steerable random fields. In *ICCV*, 2007.

[28] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, , and A. Zisserman. Segmenting scenes by matching image composites. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[29] J. Sivic, B. Kaneva, A. Torralba, S. Avidan, and B. Freeman. Creating and exploring a large photorealistic virtual space. In *First IEEE Workshop on Internet Vision at CVPR*, 2008.

[30] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003.

[31] J. Sun and M. F. Tappen. Context-constrained hallucination for image super-resolution. In *CVPR*, 2010.

[32] Y.-W. Tai, S. Liu, M. S. Brown, and S. Lin. Super resolution using edge prior and single image detail synthesis. In *CVPR*, 2010.

[33] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE PAMI*, 30(11):1958–1970, 2008.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[35] Y. Weiss and W. T. Freeman. What makes a good model of natural images? In *CVPR*, 2007.

[36] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[37] S. C. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *IJCV)*, 1998.

[38] M. Zontak and M. Irani. Internal statistics of a single natural image. In *CVPR*, pages 977–984, 2011.