Image Super-Resolution with Cross-Scale Non-Local Attention and Exhaustive Self-Exemplars Mining

Yiqun Mei[1], Yuchen Fan[1], Yuqian Zhou[1], Lichao Huang[2],
Thomas S. Huang[1], Humphrey Shi[3,1]

[1]IFP Group, UIUC, [2]Horizon Robotics, [3]University of Oregon

Abstract

Deep convolution-based single image super-resolution (SISR) networks embrace the benefits of learning from large-scale external image resources for local recovery, yet most existing works have ignored the long-range feature-wise similarities in natural images. Some recent works have successfully leveraged this intrinsic feature correlation by exploring non-local attention modules. However, none of the current deep models have studied another inherent property of images: ***cross-scale feature correlation****. In this paper, we propose the first Cross-Scale Non-Local (CS-NL) attention module with integration into a recurrent neural network. By combining the new CS-NL prior with local and in-scale non-local priors in a powerful recurrent fusion cell, we can find more cross-scale feature correlations within a single low-resolution (LR) image. The performance of SISR is significantly improved by exhaustively integrating all possible priors. Extensive experiments demonstrate the effectiveness of the proposed CS-NL module by setting new state-of-the-arts on multiple SISR benchmarks. Our code will be available at: https://github.com/SHI-Labs/Cross-Scale-Non-Local-Attention*

1. Introduction

Single image super resolution (SISR) aims at recovering a high-resolution (HR) image from its low-resolution (LR) counterpart. SISR has numerous applications in the areas of satellite imaging, medical imaging, surveillance monitoring and high-definition display and imaging *etc* [3, 32, 40, 41, 43]. The mapping between LR and HR image is not bijective, which yields more possibilities for a faithful and high-quality HR recovery. Due to this ill-posed nature, SISR remains challenging in the past decades.

Early efforts in traditional methods provide good practices for resolving SISR. By fully using the intrinsic property of the LR images, they mostly focus on local prior and non-local prior for patch matching and reconstruction. Specifically, local prior based methods, like bilinear or bicubic interpolation, reconstruct pixels merely by the weighted sum of neighbour ones. To go beyond the local limitation, methods based on non-local mean filtering [24, 35] start to globally search similar patches over the whole LR image.

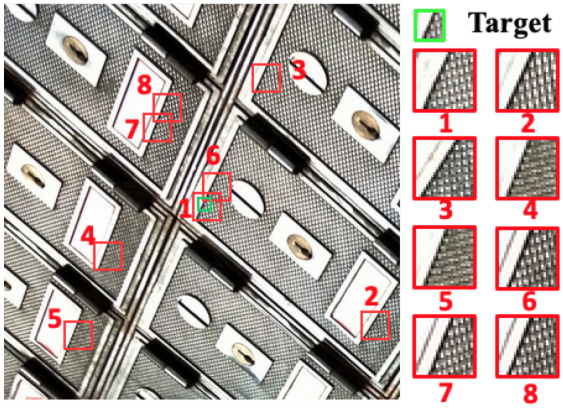

Figure 1. Visualization of most engaged patches captured by our Cross-Scale Non-Local (CS-NL) attention. Cross-scale similarities widely exist in natural images. Multiple high-resolution (HR) patches from the low-resolution (LR) image itself significantly improve target patch super-resolution.

The non-local search for self-similarity can be further extended to *cross-scale* cues. It has been verified that cross-scale patch similarity widely exists in natural images [9, 42]. Intuitively, in addition to non-local pixel-to-pixel matching, pixels can also be matched with larger image patches. The natural cross-scale feature correspondence makes us search high-frequency details directly from LR images, leading to more faithful, accurate and high-quality reconstructions.

Since the first deep learning-based method [4] was proposed, discriminative learning based methods make it possible to use large-scale external image priors for SISR. Compared with traditional methods, they tend to have better

feature representation ability, faster inference speed, end-to-end trainable paradigm [10, 16], and significant performance improvement. To further take the advantages of deep SISR, for several years, efforts [5, 14, 19, 37, 39, 33, 6] have been made on increasing the depth or width of the networks to increase the receptive field or improve the feature representation. However, the essence of the solutions was not changed, but locally finding external similar patches. It yields great limitations of deep SISR. SISR performance was boosted right after the non-local attention modules [2, 20, 38] were proposed. They explored non-local self-similarity property and embedded the non-local modules into the deep network.

What should be the next progress for deep SISR? One intuitive idea is following the traditional methods to explore the non-local *cross-scale* self-similarity in deep networks. Recently, Shocher et al. [25] proposed a zero-shot super-resolution (ZSSR) network to learn the high-frequency details from a pair of down-sampled LR and LR itself using one single test LR image. The essence of ZSSR is an implicit cross-scale patch matching approach using a light-weight network. However, inferring with ZSSR requires additional training time for each new LR image, which is not elegant and efficient enough for practical applications.

Following the successful path of non-local attention modules, in this paper, we are seeking ways of incorporating *cross-scale* non-local attention scheme into the deep SR network. Specifically, we propose a novel Cross-Scale Non-Local (CS-NL) attention module, learning to mine long-range dependencies between LR features to larger-scale HR patches within the same feature map, as shown in Figure 1. After that, we integrate the previous local prior, In-Scale Non-Local (IS-NL) prior and the proposed Cross-Scale Non-Local prior into a Self-Exemplars Mining (SEM) module, and fuse them with multi-branch mutual-projection. Finally, we embed the SEM module into a recurrent framework for image super-resolution task.

In summary, the main contributions of this paper are three-fold:

- The core contribution of the paper is to propose the first Cross-Scale Non-Local (CS-NL) attention module in deep networks for SISR task. We explicitly formulate the pixel-to-patch and patch-to-patch similarities inside the image, and demonstrate that additionally mining cross-scale self-similarities greatly improves the SISR performance.

- We then propose a powerful Self-Exemplar Mining (SEM) cell to fuse information recurrently. Inside the cell, we exhaustively mine all the possible intrinsic priors by combining local, in-scale non-local, and the proposed cross-scale non-local feature correlations, and embrace rich external statistics learned by the network.

- The newly proposed recurrent SR network achieves the state-of-the-art performance on multiple image benchmarks. Extensive ablation experiments further verify the effectiveness of the novel network.

2. Related Works

Self-Similarity in Image SR The fact that small patches tend to recur within and across scale of a same image has been verified for most natural images [9, 42]. Since then, a category of self-similarity based approaches has been extensively developed and achieves promising results. Such algorithms utilize the cross-scale information redundancy of a given image as a unique source for reconstruction without relying on any external examples [7, 8, 9, 13, 23, 26, 31]. In the pioneering work, Glasner *et al.* [9] proposed to jointly exploit repeating patches within and across image scales by integrating the idea of multiple image SR and example-based SR into a unified framework. Furthermore, Freedman *et al.* [7] effectively assumed that similar patches exist in an extremely localized region and thus can greatly reduce computation time. Following this fashion, Yang *et al.* [31] proposed a very fast regression model that focused on only in-place cross-scale similarity. To handle appearance variations in the scene, Huang *et al.* [13] enlarged the internal dictionary by modeling geometric transformations. The idea of internal data repetition has also been applied to solve SR with blur and noisy images [23, 26].

Deep CNNs for Image SR The first work that introduced CNN to solve image SR was proposed by [4], where they interpret the three consecutive convolution layers as corresponding extraction, non-linear mapping and reconstruction step in sparse coding. Kim *et al.* [14] proposed a very deep model VDSR with more than 16 convolution layers benefiting from effective residual learning. To further unleash the power of deep CNNs, Lim *et al.* [19] integrated residual blocks into the SR framework to form a very wide model (EDSR) and a very deep model (MDSR). As the network goes as deep as hundreds of layers, Zhang *et al.* [39] utilized densely connected blocks with global feature fusion to effectively exploit hierarchical features from all intermediate layers. Besides extensive efforts spent on designing wider and deeper structures, algorithms with attention modules [2, 20, 37, 38] were proposed to further enhance representation power of deep CNNs by exploring feature correlations along either spatial or channel dimension.

Non-Local Attention in Deep Networks In recent years, there is an emerging trend of applying non-local attention mechanism to solve various computer vision problems. In general, non-local attention in deep CNNs allows the network to concentrate more on informative areas. Wang *et*

al. [29] initially proposed non-local neural network to seek semantic relationships for high-level tasks, such as image classification and object detection. On the contrary, non-local attention for image restoration is based on non-local similarities prior. Methods, such as NRLN [20], RNAN [37] and SAN [2], incorporate non-local operation into their networks in order to make better use of image structural cues, by considering long-range feature correlations. As such, they achieved considerable performance gain.

However, existing non-local approaches for image restoration only explored feature similarities at the same scale, while ignoring abundant internal LR-HR exemplars across scales, leading to relatively low performance. It is known that the internal HR correspondences contain more relevant high-frequency information and stronger predictive power. To this end, we propose Cross-Scale Non-Local (CS-NL) attention by exploring cross-scale feature correlations.

3. Cross-Scale Non-Local (CS-NL) Attention

In this section, we formulate the proposed cross-scale non-local attention, and compare it with the existing in-scale non-local attention.

In-Scale Non-Local (IS-NL) Attention Non-local attention can explore self-exemplars by summarizing related features from the whole images. Formally, given image feature map X, the non-local attention is defined as

$$Z_{i,j} = \sum_{g,h} \frac{\exp(\phi(X_{i,j}, X_{g,h}))}{\sum_{u,v} \exp(\phi(X_{i,j}, X_{u,v}))} \psi(X_{g,h}), \quad (1)$$

where (i, j), (g, h) and (u, v) are pairs of coordinates of X. $\psi(\cdot)$ is feature transformation function, and $\phi(\cdot, \cdot)$ is correlation function to measure similarity that is defined as

$$\phi(X_{i,j}, X_{g,h}) = \theta(X_{i,j})^T \delta(X_{g,h}), \quad (2)$$

where $\theta(\cdot)$ and $\delta(\cdot)$ are feature transformations. Note that the pixel-wise correlation is measured in the same scale.

Cross-Scale Non-Local (CS-NL) Attention The above formulation can be easily extended to a cross-scale version referring to Figure 2. Instead of measuring the pixel-wise mutual correlation as the in-scale non-local module, the proposed cross-scale attention is designed to measure the correlation between low-resolution pixels and larger-scale patches in the LR image. To super-resolve the LR image, the Cross-Scale Non-Local (CS-NL) attention directly utilizes the patches matched to each pixel within this LR image.

Hence, for super-resolution purposes, cross-scale non-local attention is built upon in-scale attention by finding

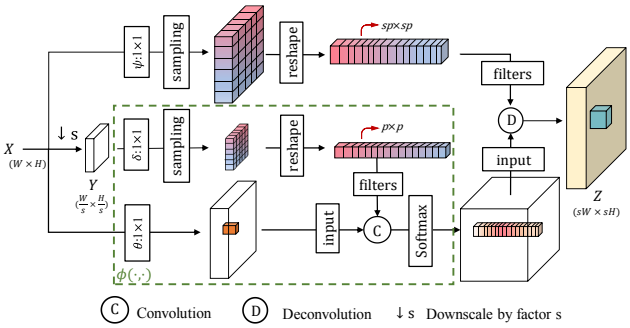

Figure 2. The proposed Cross-Scale Non-Local (CS-NL) attention module. The bottom green box is for patch-level cross-scale similarity-matching. The upper branch shows extracting the original HR patches in LR image.

candidates in features $Y = X \downarrow_s$ downsampled by scaling factor s. The reason to do so is because directly matching pixels with patches using common similarity measurement is infeasible due to spatial dimension difference. So we simply downsample the features to represent the patch as pixel and measure the affinity. Downsampling operation in this paper is bilinear interpolation.

Suppose the input feature map is X $(W \times H)$, to compute pixel-patch similarity, we need to first downsample X to Y $(\frac{W}{s} \times \frac{H}{s})$ and find pixel-wise similarity between X and Y, and finally use corresponding $s \times s$ patches in X to super-resolve pixels in X, thus the output Z will be $sW \times sH$. Cross-scale attention can be adapted from Eq.1 as

$$Z^{s\times s}_{si,sj} = \sum_{g,h} \frac{\exp(\phi(X_{i,j}, Y_{g,h}))}{\sum_{u,v} \exp\left(\phi(X_{i,j}, Y_{u,v})\right)} \psi(X^{s\times s}_{sg,sh}), \quad (3)$$

where $Z^{s\times s}_{si,sj}$ now is the feature patch of size $s \times s$ located at (si, sj). We obtain the weighted-averaged features $Z^{s\times s}_{si,sj}$ directly from the feature patches $X^{s\times s}_{sg,sh}$ extracted from the input feature maps. Intuitively, with the cross-scale attention, we can mine more faithful and richer high-frequency details from the original intrinsic image resources.

Patch-Based Cross-Scale Non-Local Attention Feature-wise affinity measurement can be problematic. First, high-level features are robust to transformations and distortions, that is rotated/distorted low-level patches may yield same high-level features. Take the average pooling as an example, an original region representing a HR *window* and its flipped version have exactly the same high-level features. Therefore, it is likely that many erroneous matches will be synthesized to HR tensors. Besides, adjacent target regions (e.g. $Z^{s\times s}_{si,sj}$ and $Z^{s\times s}_{s(i+1),s(j)}$) are generated in a non-overlapping fashion, possibly creating discontinuous region boundaries artifacts.

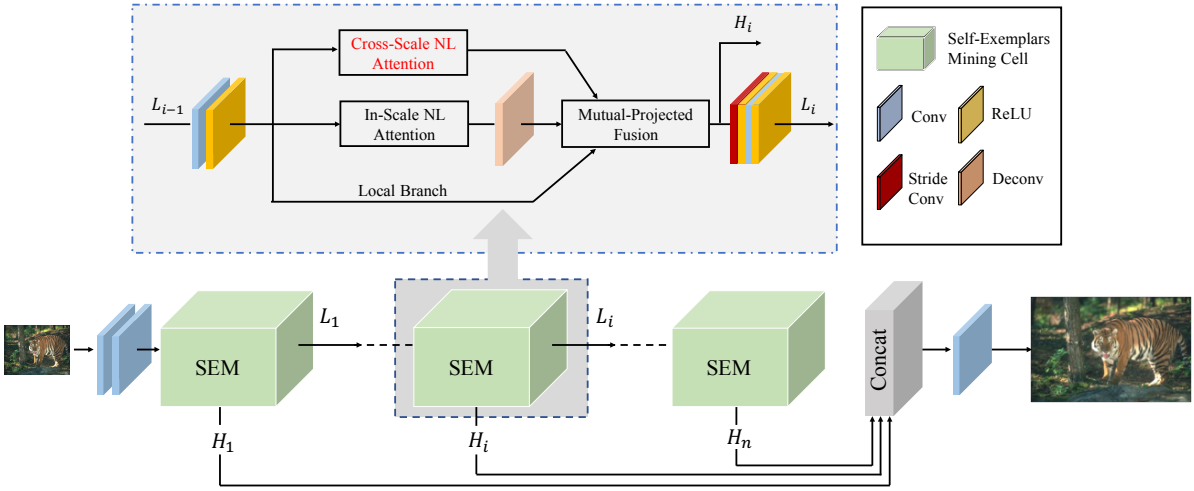

Figure 3. The recurrent architecture with the proposed Self-Exemplars Mining (SEM) cell. Inside SEM, it fuses features learned from a proposed Cross-Scale Non-Local (CS-NL) attention, with others from the In-Scale Non-Local (IS-NL) and the local paths.

Based on the above analysis, we generalize to empirically implement our Cross-Scale Non-Local (CS-NL) attention using another patch-wise matching. Therefore, Eq.3 is generalized to,

$$Z^{sp\times sp}_{si,sj} = \sum_{g,h} \frac{\exp \phi(X^{p\times p}_{i,j}, Y^{p\times p}_{g,h})}{\sum_{u,v} \exp \phi(X^{p\times p}_{i,j}, Y^{p\times p}_{u,v})} \psi(X^{sp\times sp}_{sg,sh}), \tag{4}$$

and Eq.4 will be identical to Eq.3 if $p = 1$. The measured correlations are efficiently extended to patch-level, and regions in the output feature map Z are now densely overlapped due to patch-based matching.

4. Methodology

The proposed network architecture is shown in Figure 3. It is basically a recurrent neural network, with each recurrent cell called Self-Exemplars Mining (SEM) fully integrating local, in-scale non-local, and a newly proposed Cross-Scale Non-Local (CS-NL) priors. In this section, we introduce them in a bottom-up manner.

4.1. CS-NL Attention Module

Figure 2 illustrates the newly-proposed Cross-Scale Non-Local (CS-NL) attention module embedded into the deep networks. As formulated in section 3, we apply a patch-level cross-scale similarity-matching in the CS-NL attention module. Specifically, suppose we are conducting an s-scale super-resolution with the module, given a feature map X of spatial size (W, H), we first bilinearly downsample it to Y with scale s, and match the $p \times p$ patches in X with the downsampled $p \times p$ candidates in Y to obtain the softmax matching score. Finally, we conduct deconvolution on the score by weighted adding the patches of size (sp, sp) extracted from X. The obtained Z of size (sW, sH), will be s times super-resolved than X.

4.2. Self-Exemplars Mining (SEM) Cell

Multi-Branch Exemplars Inside the Self-Exemplars Mining (SEM) cell, we exhaustively mine all the possible intrinsic priors, and embrace rich external image priors. Specifically, we mine the image self-similarities and learn the new information using a multi-branch structure, including the conventional Local (L) and In-Scale Non-Local (IS-NL) branches, and also the newly proposed CS-NL branch.

The local branch, in Figure 3, is a simple identical pathway connecting the convolutional features to the fusion structure. For the IS-NL branch, it contains a non-local attention module adopted from [2] and a deconvolution layer for upscaling the module outputs. The IS-NL module is region-based in this paper. As in [2], we divide the feature maps into region grids, where the inter-dependencies are captured independently in each grid. This reduces the computation burden.

Mutual-Projected Fusion While three-branch structure in SEM generates three feature maps by independently exploiting each information sources from LR images, how to fuse these separate tensors into a comprehensive feature map remains unclear. One possible solution is simply adding or concatenating them together, as widely used in previous works [19, 20, 38, 39]. In this paper, we proposed a mutual-projected fusion to progressively combine features together. The algorithm procedure is illustrated in Figure 4.

To allow the network to concentrate on more informative features, we first compute the residual R_{IC} between two features from IS-NL F_I and CS-NL F_C branch, and then after a single convolution layer $conv$ on R_{IC}, the features

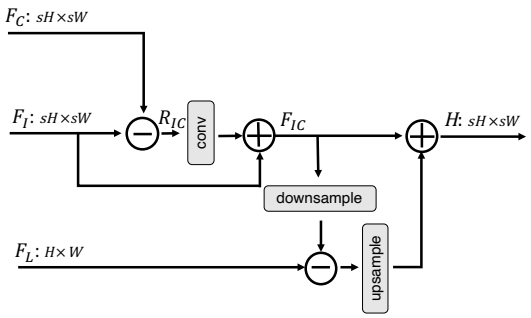

Figure 4. Mutual-projected fusion. Downsample and upsample operations are implemented using stride convolution and stride deconvolution, respectively.

are added back to F_I to obtain F_{IC}.

$$R_{IC} = F_I - F_C, \tag{5}$$

$$F_{IC} = conv(R_{IC}) + F_I. \tag{6}$$

Intuitively, the residual feature R_{IC} represents the details existing in one source while missing in the other. Such inter-residual projection allows the network to focus on only the distinct information between sources while bypassing the common knowledge, thus improves the discriminative ability of the network.

Motivated by the traditional Image SR and recent DBPN [11], we adopt the back-projection approach to incorporate local information to regularize the feature and correct reconstruction errors. Following [11], the final fused feature H is computed by,

$$e = F_L - downsample(F_{IC}), \tag{7}$$

$$H = upsample(e) + F_{IC}, \tag{8}$$

where F_L is the feature maps of the Local branch, $downsample$ is a stride convolution to down-sample F_{IC}, and $upsample$ is a stride deconvolution to upscale the feature maps.

The mutual-projected operation guarantees a residual learning while fusing different feature sources, enabling a more discriminative feature learning compared with trivial adding or concatenating.

4.3. Recurrent Framework

The repeated SEM cells are embedded into a recurrent framework, as shown in Figure 3. At each iteration i, the hidden unit H_i of SEM is directly the fused feature map H, and the output unit L_i is the computed by H_i going through a two-layer CNN. Note that the initial features L_0 are directly computed by the LR image I_{LR} through only two convolutional layers.

Later on, the extracted deep SR features H_i from each iteration i are concatenated together into a wide tensor and mapped to the SR image I_{SR} via one single convolution operation. The network is trained solely with L_1 reconstruction loss.

5. Experiments

5.1. Datasets and Evaluation Metrics

Following [19, 38, 39], we use 800 images from DIV2K [28] dataset to train our models. For testing, we report the performance on five standard benchmark datasets: Set5 [1], Set14 [34], B100 [21], Urban100 [13] and Manga109 [22]. For evaluation, all the SR results are first transformed into YCbCr space and evaluated by PSNR and SSIM [30] metrics on Y channel only.

5.2. Implementation details

We set the recurrence number of SEM as 12 following [20]. For the Cross-Scale Non-Local (CS-NL) attention in SEM, we set patch size $p = 3$ and stride $s = 1$ for dense sampling. We use 3×3 as filter size for all convolution layers except for those in attention blocks where the kernel size is 1×1. The filter size for stride convolution and deconvolution in SEM are set to be equal at each scale, e.g., 6×6, 9×9 and 8×8 for scale factor 2, 3, 4, respectively. All intermediate features have channel $C = 128$ except for those embedded features in attention module, where $C = 64$. The last convolution layer in SEM has 3 convolution filters that transfer a deep SR feature to an RGB image.

During training, we crop 16 images with patch size 48×48 to form a input batch. The training examples are augmented by random rotating $90°$, $180°$, $270°$ and horizontal flipping. To optimize our model, we use ADAM optimizer [15] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and ϵ =1e-8. The initial learning rate is set to 1e-4 and reduced to half every 150 epochs. The training stops at 500 epochs. We implement the model using PyTorch, and train it on Nvidia V100 GPUs.

5.3. Comparisons with State-of-the-arts

To verify the effectiveness of the proposed model, we compare our approach with 11 state-of-the-art methods, which are LapSRN [17], SRMDNF [36], MemNet [27], EDSR [19], DBPN [11], RDN [39], RCAN [37], NLRN[20], SRFBN [18], OISR [12] and SAN [2].

Quantitative Evaluations In Table 1, We report the quantitative comparisons for scale factor $\times 2$, $\times 3$ and $\times 4$. Compared with other methods, our CS-NL-embedded recurrent model achieved the best performance on multiple benchmarks for almost all scaling factors. It worth noting that our model significantly outperforms NLRN, which is the first proposed in-scale non-local network for image restoration.

Table 1. Quantitative results on benchmark datasets. Best and second best results are colored with red and blue.

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LapSRN [17]	×2	37.52	0.9591	33.08	0.9130	31.08	0.8950	30.41	0.9101	37.27	0.9740
MemNet [27]	×2	37.78	0.9597	33.28	0.9142	32.08	0.8978	31.31	0.9195	37.72	0.9740
EDSR [19]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
SRMDNF [36]	×2	37.79	0.9601	33.32	0.9159	32.05	0.8985	31.33	0.9204	38.07	0.9761
DBPN [11]	×2	38.09	0.9600	33.85	0.9190	32.27	0.9000	32.55	0.9324	38.89	0.9775
RDN [39]	×2	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353	39.18	0.9780
RCAN [37]	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
NLRN [20]	×2	38.00	0.9603	33.46	0.9159	32.19	0.8992	31.81	0.9249	–	–
SRFBN [18]	×2	38.11	0.9609	33.82	0.9196	32.29	0.9010	32.62	0.9328	39.08	0.9779
OISR [12]	×2	38.21	0.9612	33.94	0.9206	32.36	0.9019	33.03	0.9365	–	–
SAN [2]	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
CSNLN (ours)	×2	38.28	0.9616	34.12	0.9223	32.40	0.9024	33.25	0.9386	39.37	0.9785
LapSRN [17]	×3	33.82	0.9227	29.87	0.8320	28.82	0.7980	27.07	0.8280	32.21	0.9350
MemNet [27]	×3	34.09	0.9248	30.00	0.8350	28.96	0.8001	27.56	0.8376	32.51	0.9369
EDSR [19]	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
SRMDNF [36]	×3	34.12	0.9254	30.04	0.8382	28.97	0.8025	27.57	0.8398	33.00	0.9403
RDN [39]	×3	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653	34.13	0.9484
RCAN [37]	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
NLRN [20]	×3	34.27	0.9266	30.16	0.8374	29.06	0.8026	27.93	0.8453	-	-
SRFBN [18]	×3	34.70	0.9292	30.51	0.8461	29.24	0.8084	28.73	0.8641	34.18	0.9481
OISR [12]	×3	34.72	0.9297	30.57	0.8470	29.29	0.8103	28.95	0.8680	-	-
SAN [2]	×3	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
CSNLN (ours)	×3	34.74	0.9300	30.66	0.8482	29.33	0.8105	29.13	0.8712	34.45	0.9502
LapSRN [17]	×4	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560	29.09	0.8900
MemNet [27]	×4	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630	29.42	0.8942
EDSR [19]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
SRMDNF [36]	×4	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.09	0.9024
DBPN [11]	×4	32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946	30.91	0.9137
RDN [39]	×4	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028	31.00	0.9151
RCAN [37]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
NLRN [20]	×4	31.92	0.8916	28.36	0.7745	27.48	0.7306	25.79	0.7729	-	-
SRFBN [18]	×4	32.47	0.8983	28.81	0.7868	27.72	0.7409	26.60	0.8015	31.15	0.9160
OISR [12]	×4	32.53	0.8992	28.86	0.7878	27.75	0.7428	26.79	0.8068	-	-
SAN [2]	×4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
CSNLN (ours)	×4	32.68	0.9004	28.95	0.7888	27.80	0.7439	27.22	0.8168	31.43	0.9201

Our method has better performance when the scaling factor is larger. For $\times 4$ settings, our CS-NL embedded model achieves the state-of-the-art PSNR for all the testing benchmarks. In particular, for Urban100 and Manga109 dataset, our model outperforms previous state-of-the-art approaches by 0.4 dB and 0.2 dB, respectively. These datasets contains abundant repeated patterns, such as edges and small corners. Therefore, the superior performance demonstrates the effectiveness of our attention in exploiting internal HR hints. We claim that cross-scale intrinsic priors are indeed effective for a more faithful reconstruction.

Qualitative Evaluations The qualitative evaluations on Urban100 dataset are shown in Figure 5. The proposed model is proven to be more effective for images with repeated high-frequency features like windows, lines, squares, etc. For example, in the figure of building, LR image contains plenty of window features covering long-range of spatial-frequency. Directly utilizing those cross-scale self-exemplars from the images will be significantly better than searching for in-scale features or external patches in the training set. For all the shown examples, our method perceptually out-performs other state-of-the-arts by a large margin.

	EDSR	DBPN	RDN	RCAN	SAN	CSNLN
Para.	43M	10M	22.3M	16M	15.7M	3M
PSNR	38.11	38.09	38.24	38.27	38.31	38.28

Table 2. Model size and performance comparsion on Set5 ($2\times$) .

Model Size Analysis We report the model size and performance for recently competitive SR methods in Table 2. Comparing with others, our model has the least parameters,

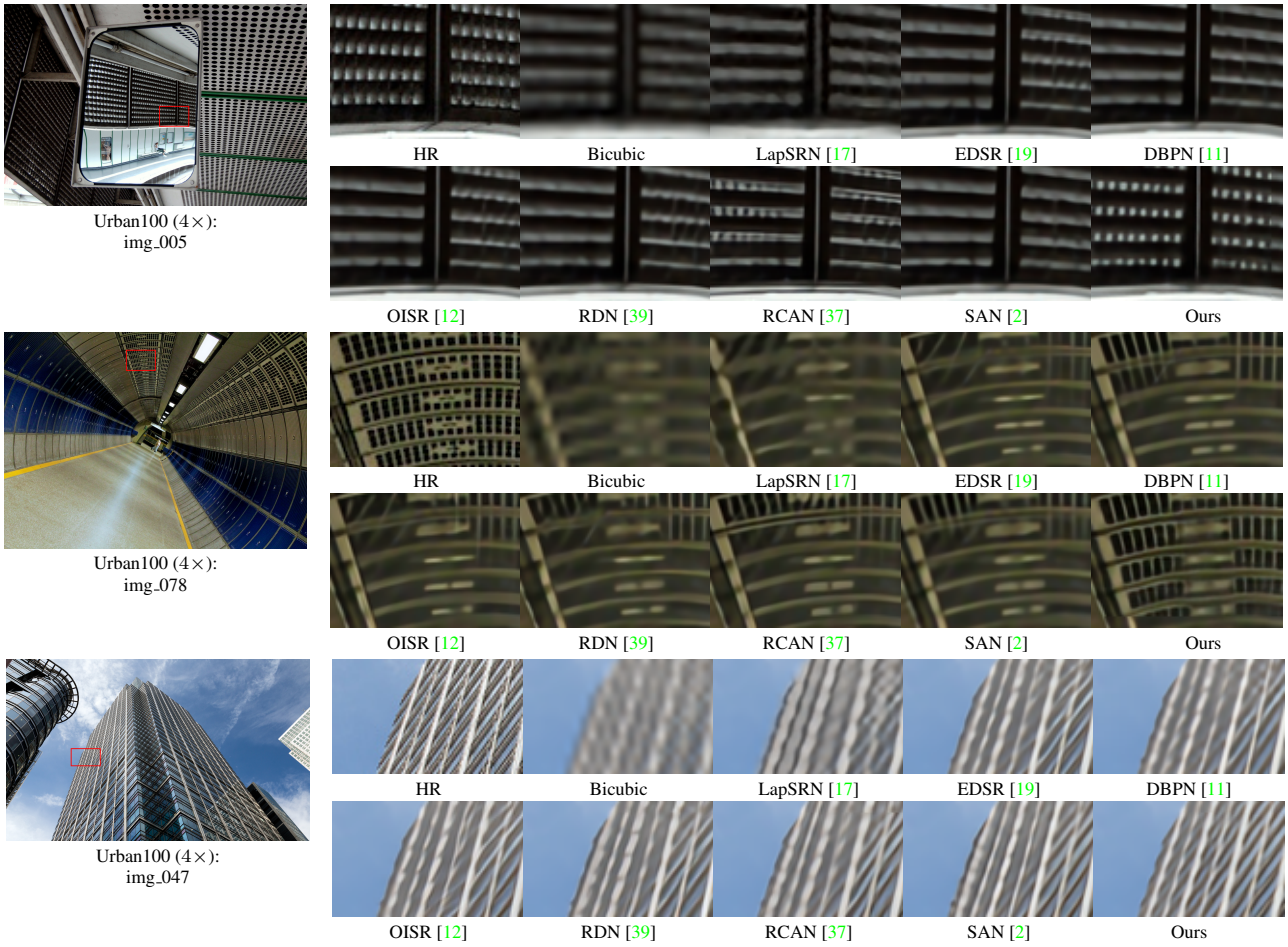

Figure 5. Visual comparison for $4\times$ SR on Urban100 dataset. For all the shown examples, especially the images with repeated edges or structures, our method perceptually out-performs other state-of-the-arts by a large margin.

which only needs 20% parameters of RCAN and SAN, but achieves the second best result. Therefore, our CSNLN obtains better parameter efficiency in comparison with other methods, by effectively mining internal HR hints.

5.4. Ablation Study

Cross-Scale v.s. In-Scale Attention The key difference between our cross-scale non-local attention and the in-scale one is to allow network to benefit from abundant internal HR hints with different scales. To verify it, we visualize its correlation maps on 6 images from Urban100 [13], and compare it with in-scale non-local attention.

As shown in Figure 6, these images contain extensive recurrences of small patterns both within scale and across scale. It is interesting to point out that once the image contains repeated edges, such redundant recurrences are not limited to where high scale patterns appear, but also can be found in-place or even in the area that pattern tends to slightly shrink. For example, the HR appearance of a small corner can be simply found by properly zooming out. All these recurrences contain valuable high frequency information for reconstruction. As shown in Figure 6, the in-scale attention only focuses on pixels with similar intensity. In contrast, our cross-scale non-local attention is able to utilize the abundant repeating structures in the images, demonstrating its effectiveness for exploiting internal HR information.

Self-Exemplars Mining Module To demonstrate the effectiveness of our proposed Self-Exemplars Mining (SEM) module, we construct a baseline model by removing all branches, resulting in a fully convolutional recurrent network (RNN). To keep the total parameters same as other variants, we set 10 convolution layers in the recurrent block. As shown in Table 3, the basic RNN achieves 33.32 dB on Set14 ($\times 2$). Results in first 4 columns demonstrate the effectiveness of individual branch, as each of them brings improvement over the baseline. Furthermore, from last 4 columns, we find that combining these branches achieves the best performance. For example, when cross-scale non-local branch is added, the performance is improved from

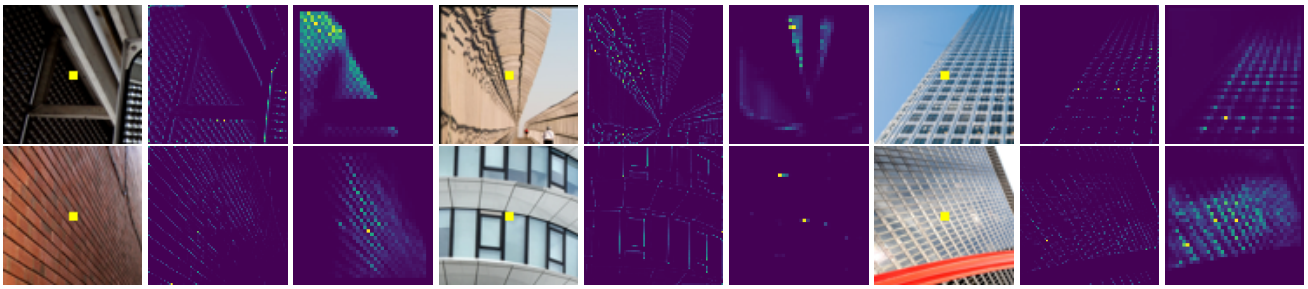

Figure 6. Comparisons of correlation maps of CS-NL attention and IS-NL attention. For each group of three columns, the left one is the input image, the middle one shows the in-scale attention, and the right one depicts the cross-scale attention. one can see that the in-scale attention only focuses on pixels with similar intensity. In contrast, our cross-scale non-local attention is able to utilize the abundant repeated structures in the images, demonstrating its effectiveness for exploiting internal HR information.

Local (L) branch		✓			✓	✓		✓
In-Scale Non-Local (IS-NL) Branch			✓		✓		✓	✓
Cross-Scale Non-Local (CS-NL) branch				✓		✓	✓	✓
PSNR	33.32	33.47	33.52	33.51	33.62	33.64	33.57	**33.74**

Table 3. Ablation study on the branch features in SEM. We report the PSNR results on Set14 (2×) after 200 epochs. With an additional CS-NL branch, the performance becomes 33.74dB compared with the one without CS-NL, 33.62dB.

Attention Patch Size	1×1	3×3	5×5
PSNR	33.67	**33.74**	33.61

Table 4. Effects of patch size for matching.

Fusion	addition	concatenation	Mutual Projection
PSNR	33.69	33.62	**33.74**

Table 5. Comparison of fusion operators.

33.47 dB to 33.64 dB. When both local branch and non-local branch are added to the network, the best performance is achieved by further adding cross-scale non-local branch, resulting in a improvement from 33.62 dB to 33.74 dB.

These facts indicate that the cross-scale correlations learned by our attention can not be captured by either simple convolution or previous in-scale attention module, demonstrating that our CS-NL attention is of crucial importance for fully exploiting information from LR images.

Patch-Based Matching v.s. Pixel-Based Matching In practical implementation, we compute patch-wise correlation rather than pixel-wise correlation. Here we investigate the influence of patch size p in CS-NL attention. We compare the patch size of 1×1, 3×3 and 5×5, where 1×1 is equivalent to pixel-wise matching. As shown in Table 4, the performance peak is at 3×3, which is higher than pixel based matching, indicating that a small patch can serve as a better region descriptor. However, when using a larger patch size, the performance is worse than the pixel-based matching. This is mainly because larger patches mean additional constraint on the content when evaluating similarity, and therefore it becomes harder to find well-matched correspondences. All these results show that it is necessary to choose a proper patch size for effectively computing correlations in CS-NL attention.

Mutual-Projected Fusion We show the effectiveness of our mutual-projected fusion by comparing it with other commonly used fusion strategies, e.g., feature addition and concatenation. As shown in Table 5, it can be found that our projection based fusion obtains the best result. By replacing the addition and concatenation with mutual projection, the performance improves about 0.05 dB and 0.12 dB. These results demonstrate the effectiveness of our fusion module in progressively aggregating information.

6. Conclusion

In this paper, we proposed the first Cross-Scale Non-Local (CS-NL) attention module for image super-resolution deep networks. With the novel module, we are able to sufficiently discover the widely existing cross-scale feature similarities in natural images. Further integrating it with local and the previous in-scale non-local priors, while embracing the abundant external information learned by the network, our recurrent model achieved state-of-the-art performance for multiple benchmarks. Our experiments suggest that exploring cross-scale long-range dependencies will greatly benefit single image super-resolution (SISR) task, and possibly is also promising for general image restoration task.

Acknowledgments This work is in part supported by IBM-Illinois Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network.

References

[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, 2012. 5

[2] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 2, 3, 4, 5, 6, 7

[3] Hasan Demirel and Gholamreza Anbarjafari. Discrete wavelet transform-based satellite image resolution enhancement. *IEEE transactions on geoscience and remote sensing*, 49(6):1997–2004, 2011. 1

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 1, 2

[5] Yuchen Fan, Honghui Shi, Jiahui Yu, Ding Liu, Wei Han, Haichao Yu, Zhangyang Wang, Xinchao Wang, and Thomas S Huang. Balanced two-stage residual networks for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 161–168, 2017. 2

[6] Yuchen Fan, Jiahui Yu, Ding Liu, and Thomas S Huang. Scale-wise convolution for image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2

[7] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):12, 2011. 2

[8] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002. 2

[9] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *Proceedings of the IEEE 12th International Conference on Computer Vision*, pages 349–356, 2009. 1, 2

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 2

[11] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018. 5, 6, 7

[12] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019. 5, 6, 7

[13] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 2, 5, 7

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[17] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 5, 6, 7

[18] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019. 5, 6

[19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2, 4, 5, 6, 7

[20] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *Advances in Neural Information Processing Systems*, pages 1673–1682, 2018. 2, 3, 4, 5, 6

[21] 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 *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 5

[22] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 5

[23] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–952, 2013. 2

[24] Matan Protter, Michael Elad, Hiroyuki Takeda, and Peyman Milanfar. Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Transactions on image processing*, 18(1):36–51, 2008. 1

[25] Assaf Shocher, Nadav Cohen, and Michal Irani. zero-shot super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018. 2

[26] Abhishek Singh and Narendra Ahuja. Super-resolution using sub-band self-similarity. In *Asian Conference on Computer Vision*, pages 552–568. Springer, 2014. 2

[27] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017. 5, 6

[28] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125, 2017. 5

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 3

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[31] Jianchao Yang, Zhe Lin, and Scott Cohen. Fast image super-resolution based on in-place example regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1059–1066, 2013. 2

[32] Haichao Yu, Ding Liu, Honghui Shi, Hanchao Yu, Zhangyang Wang, Xinchao Wang, Brent Cross, Matthew Bramler, and Thomas S Huang. Computed tomography super-resolution using convolutional neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3944–3948. IEEE, 2017. 1

[33] Jiahui Yu, Yuchen Fan, and Thomas Huang. Wide activation for efficient image and video super-resolution. In *Proceedings of the British Machine Vision Conference*, 2019. 2

[34] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 5

[35] Kaibing Zhang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Single image super-resolution with non-local means and steering kernel regression. *IEEE Transactions on Image Processing*, 21(11):4544–4556, 2012. 1

[36] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3271, 2018. 5, 6

[37] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 2, 3, 5, 6, 7

[38] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2, 4, 5

[39] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 2, 4, 5, 6, 7

[40] Yuqian Zhou, Ding Liu, and Thomas Huang. Survey of face detection on low-quality images. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 769–773. IEEE, 2018. 1

[41] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. *arXiv preprint arXiv:2003.04857*, 2020. 1

[42] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *CVPR 2011*, pages 977–984. IEEE, 2011. 1, 2

[43] Wilman WW Zou and Pong C Yuen. Very low resolution face recognition problem. *IEEE Transactions on image processing*, 21(1):327–340, 2011. 1