Image Processing GNN: Breaking Rigidity in Super-Resolution

Yuchuan Tian[1], Hanting Chen[2], Chao Xu[1], Yunhe Wang[2*]
[1] National Key Lab of General AI, School of Intelligence Science and Technology, Peking University. [2] Huawei Noah's Ark Lab.
tianyc@stu.pku.edu.cn, yunhe.wang@huawei.com

Abstract

Super-Resolution (SR) reconstructs high-resolution images from low-resolution ones. CNNs and window-attention methods are two major categories of canonical SR models. However, these measures are rigid: in both operations, each pixel gathers the same number of neighboring pixels, hindering their effectiveness in SR tasks. Alternatively, we leverage the flexibility of graphs and propose the Image Processing GNN (IPG) model to break the rigidity that dominates previous SR methods. Firstly, SR is unbalanced in that most reconstruction efforts are concentrated to a small proportion of detail-rich image parts. Hence, we leverage degree flexibility by assigning higher node degrees to detail-rich image nodes. Then in order to construct graphs for SR-effective aggregation, we treat images as pixel node sets rather than patch nodes. Lastly, we hold that both local and global information are crucial for SR performance. In the hope of gathering pixel information from both local and global scales efficiently via flexible graphs, we search node connections within nearby regions to construct local graphs; and find connections within a strided sampling space of the whole image for global graphs. The flexibility of graphs boosts the SR performance of the IPG model. Experiment results on various datasets demonstrates that the proposed IPG outperforms State-of-the-Art baselines. Codes are available at this link.

1. Introduction

The ill-posed task of Super-Resolution (SR) constructs high-resolution images from low-resolution ones. SR works typically apply Deep Neural Network (DNN) knowledge learned from High-Resolution training images to construct missing details in low-resolution inputs. Existing DNN variants for classical SR tasks include CNNs [16, 24, 50] and Transformers [5, 6, 23]. These measures have proved effective in creating High-Resolution images, and are used in real-world applications like medical imaging [10], remote sensing [15], et cetera.

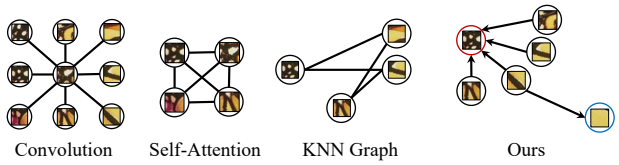

Figure 1. Convolution (left), Self-Attention (middle-left), KNN Graph Aggregation (middle-right), and Graph Aggregation in IPG (right). Compared with other methods, IPG graph aggregation considers the unbalanced nature of SR: detail-rich, high-frequency image nodes (red) have higher node degrees; while flat, low-frequency image nodes (blue) require fewer aggregations.

Despite various measures, it is taken for granted that mainstream SR models treat all pixels in a fairly rigid manner. For instance, as shown in Figure 1, in convolution layers of the CNN-based SR model VDSR [16], the same convolution kernel scans through all pixels of feature maps, *i.e.* each pixel is rigidly designated to communicate with its nearest neighbors; in the Transformer-based model SwinIR [23], all pixels are assigned to equal-sized attention grids for self-attention operations. In both examples, each pixel aggregates information from a fixed number of pixels within a neighborhood of fixed size.

Efforts have been made to overcome the rigidity of convolution and window-attention operations. For example, some recent SR research tries to enhance SR models by enlarging self-attention grids [5, 46] and designing striped global attention windows [6, 22]. But these improvements also introduce a large computation burden. Some other works [12, 26, 52] leverage the flexibility of graphs and propose graph-based methods. With content-based graphs constructed from patch nodes, each image patch could aggregate top-k relevant but distant parts beyond hard local boundaries in convolution and window-attention.

However, these graph measures have limited flexibility because the lopsided nature of SR is overlooked. In SR tasks, only a small proportion of high-frequency pixels require hard reconstruction efforts; the majority of pixels are located in flat, low-frequency regions and thus almost left intact [9]. In response to the imbalance, SR methods should

*Corresponding author

ideally pay more attention to detail-rich regions and care less about flat, detail-poor image parts. However, existing k-Nearest-Neighbor graph-based methods treat all image nodes equally. In other words, all nodes share the same pre-set degree k and the unbalanced nature of SR is not considered. In fact, as we inspect canonical operation paradigms in SR from a graph perspective, we found that degree-equivalence rigidity also manifests in convolution and window-attention: in these paradigms, each pixel on the image aggregates the same number of pixels regardless of image content, and thus, sharing "equal degrees" in graph terms. Equal node degrees that are rigidly assigned to nodes or pixels mismatch their unequal reconstruction demands in SR, thus impacting SR performance.

Trying to break rigidity, we propose the graph-based Image Processing Graph Neural Networks (IPG) model for SR that taps the flexibility potential of graphs. Firstly, to break the degree-equivalence rigidity of convolution, window-attention, and k-Nearest-Neighbor (KNN) graphs, we leverage degree flexibility by proposing a novel degree-variant graph solution based on the unbalanced nature of SR. Specifically, a detail-aware metric is designed to measure the importance of image nodes where larger degrees are assigned to high-frequency nodes. Then, different from graph-based model counterparts, IPG adopts pixels rather than patches as image graph nodes to avoid misalignment issues due to patch rigidity. Finally, in the interest of having both local and global perception without sacrificing efficiency due to large pixel-level search space, we resort to local and global node sampling strategies. In this way, graphs are established efficiently from small pixel subsets and could either focus on local information for detail reconstruction, or span over the whole image for spatially distant but crucial features. With the measures mentioned above, IPG could fully exert graph flexibilities and achieve outstanding performances on SR tasks.

2. Related Work

Recent SR Methods. In canonical approaches to Super-Resolution, manifold previous works phrase SR as a regression problem and is proved effective via applications of CNN models [16, 24, 50, 51]. In these works, a crafted CNN model is applied to the LR image to render it into HR. Since the introduction of Transformers to vision [8], manifold works integrate the self-attention mechanism in Transformers into SR models [4–7, 23, 47, 53]. SwinIR [23] borrows shifted window-attention to SR and demonstrates good performance. Latest SR models include using striped attention windows [6, 22], channel attention [5, 22, 40], permuted attention [53], et cetera.

Despite the fact that manifold SR works follow similar convolution or window-attention methodologies, their SR performance is still confined to the rigidity of ordinary vision models, *i.e.* fixed paradigms for aggregating neighboring pixels and neglecting the unbalanced reconstruction demand among pixels in SR. Hence, we resort to graphs as a flexible way to aggregate information in order to break these sorts of rigidity.

Graphs for Vision. Graphs have been widely applied to vision tasks, including point cloud segmentation [17, 41], human action recognition [42], and image classification [12, 32]. Some previous low-level vision works also view images as graphs. Most works leverage graph to aggregate non-local information, either intra-scale [2, 18, 21, 35, 38] or cross-scale [26, 52]. Other low-level vision works on graphs include cross-layer [27] or inter-image information aggregation [37]. Among these works, edge-conditioned aggregation from neighboring nodes is mostly preferred.

In terms of graph construction from images, a majority of works use k-nearest neighbors [18, 21, 26, 35, 38, 52]. Since all image nodes share the same degree, KNN graphs are not flexible enough; and the unbalanced nature of image nodes in low-level vision tasks is neglected as well. Besides, most graph works regard patches as image nodes [18, 21, 26, 35, 52], disregarding potential patch misalignment issues in low-level vision. As for the scale of graphs, some works [21, 26, 52] search in a global scale to construct graphs; other works [38, 39] construct graphs from a local box. Graphs in these works are not flexible enough in that they fail to take care of both local and global aspects.

3. Method

3.1. Rigidity of Previous SR Methods

Convolution [16, 24, 50] and window-attention [5, 6, 23] are two major pathways for SR model designs. Despite popular usages in SR, these measures are not flexible; their deep-rooted rigidity could hinder SR performance. In one convolution operation, each outputted pixel gathers information within a tiny kernel-sized window; each pixel only has access to its neighboring pixels. For instance, in a standard 3×3 convolution, the perception field of a single pixel is confined to a tiny 3×3 box. All pixels on the image collect information from 8 respective neighbors and themselves. The similar rigidity goes for window attention: though larger window sizes (compared to convolution) are usually adopted, the perception field of window attention is still confined within hard window boundaries. In a 8×8 window-attention scenario, all pixels gather 64 pixels within the window where they belong.

Therefore, beyond rigidity in convolution and window-attention, some works think out of the box and adopt graphs in SR models. Unlike convolution and window-attention where information aggregation is bounded to boxes, these graph-based works are more spatially flexible: each node could aggregate information from its favorite top-k nodes

with loose spatial constraints. In this sense, graph aggregation is not limited to preset rigid patterns; it is more dynamic and scalable compared to convolution and window-attention counterparts.

However, though previous graph-based methods break through hard aggregation boundaries, we hold that the flexibility of graphs is not fully exploited in SR tasks. First and foremost, as shown in Figure 1, all previous methodologies (either graph-based, convolution-based, or window-attention-based) are degree-rigid. SR reconstruction demand is unbalanced among different image parts. But in previous methods, all pixels or nodes on the image are aggregating the same number of pixels or nodes; *i.e.* in graph terms, they share the same aggregation degree. Secondly, previous graph-based works are patch-rigid. Though patches are usually regarded as image nodes, patch aggregation usually requires rigid pixel-wise alignment. Possible misalignment of low-level features in patches makes SR models perform worse. Thirdly, previous graph-based works rigidly use graphs on either global or local scales, but information at both scales is potentially important for SR reconstruction. These rigid aspects are hindering the performance of SR models.

3.2. Graph Construction

To break these rigidities, we construct degree-flexible pixel graphs at local and global scales in the IPG model. In this manner, we could fully exploit the flexibility of graphs and achieve outstanding performance in SR tasks.

Degree Flexibility. Firstly, we try to figure out a degree-flexible graph solution based on the unique "unbalanced property" of SR tasks. SR reconstruction of images features imbalance. As Gou et al. [9] put it, SR is a long-tailed problem where only a small proportion of high-frequency pixels need to be reconstructed; the rest of the image parts need minimal restoration only. However, Gou et al. [9] tries to address the issue by designing training losses rather than rethinking it from the aspect of model designs. In the lopsided SR problem, it is a long-neglected aspect that treating all pixels or parts on the image equivalently is unsuitable and inefficient.

To this end, we opt to assign various node degrees to pixels based on a detail-rich indicator that marks pixels needing more reconstruction efforts. Then we design the detail-rich indicator as follows: given feature map $F \in \mathbb{R}^{H\times W\times C}$ and downsampling ratio s, the detail-rich indicator metric per pixel $D_F \in \mathbb{R}^{H\times W}$ is designated as the absolute difference between the bilinearly downsampled and upsampled feature map and the feature map itself:

$$D_F := \sum_C |F - F_{\downarrow s\uparrow s}|, \tag{1}$$

where s is taken as 2 to avoid severe information loss. We are aware that some explainable SR works [11, 43] propose metrics to measure the importance of a certain part in the output. However, these measures are gradient-based, which requires costly backward passes. In comparison, the proposed metric D_F is operation cheap in that it only requires bilinear interpolation twice. We also provide FLOPs statistics in the Appendix.

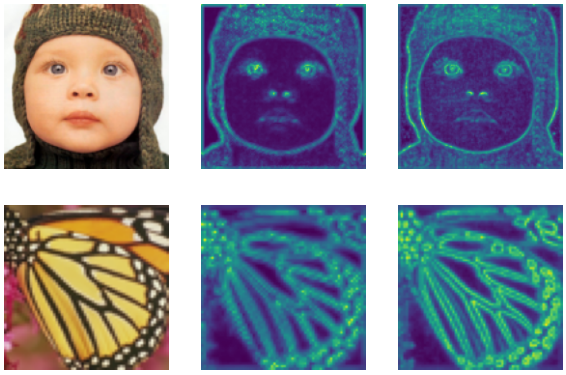

Figure 2. Visualization of the detail-rich indicator D_F in IPG. Original image (left); D_F of the shallowest block (middle); D_F of the deepest block (right). D_F could effectively reflect detail-rich parts on the image.

The overall degree budgets are assigned to each pixels based on D_F. The degree for a pixel node $v \in F$ on the feature map is in proportion to its corresponding pixel value at D_F:

$$\deg(v) \propto D_F(v). \tag{2}$$

The detail-rich indicator D_F in different MGB blocks is visualized. From Figure 2, it can be seen that detail-rich parts are responsive: margins and corners have high D_F while flat color blocks are low in D_F. The visualization reveals that the proposed D_F could effectively reflect high-frequency parts on the image throughout layers.

Pixel Node Flexibility. Then we are faced with the formulation of vertices in image graphs. In previous graph-based vision works [21, 26, 36, 52], graph nodes are typically set as image patches. During graph aggregation, patches are weighted-summed in a pixel-wise manner. However, the forcible element-wise alignment of pixels during aggregation is unsuitable for SR tasks where feature maps have rich low-level features. Object shift and rotation within low-level patches are two major issues causing patch misalignment. As for object shift, locations of objects within patches could vary; location-misaligned patch objects introduce noises. Object rotation is another faulty aspect of patch aggregation. The orientation of patches is disregarded due to rigid rectangular patch shapes during patch alignment.

To avoid the above problems during node aggregation, we hold that finer-grained pixel nodes are better solutions

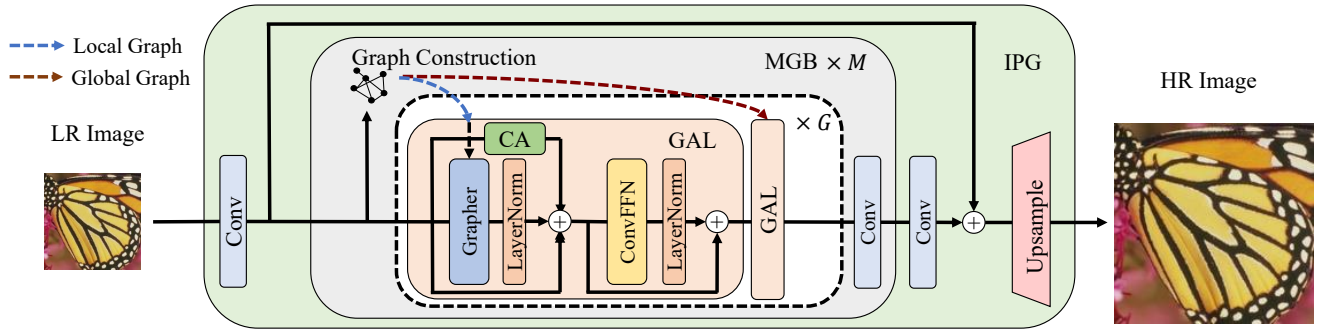

Figure 3. The architectures of IPG, Multiscale Graph-aggregation Blocks (MGB), and Graph Aggregation Layers (GAL). Each IPG consists of M MGB blocks, and each MGB has G GAL layers. Local and global graphs are distributed to GALs alternatively within MGB.

in the low-level vision scenario. Pixel graphs are more flexible: each pixel node could directly find its relevant pixels in aggregation, therefore avoiding pixel misalignment.

However, we admit that compared to pixel graphs, patch graphs could potentially have larger perception fields and are easier to construct due to a smaller number of nodes in total, which means a smaller search space for edge connections. Hence, we need an efficient way to construct a flexible and effective pixel graph.

Space Flexibility. Then we develop the space flexibility of IPG graphs via an efficient search of pixel node connections both locally and globally. We propose that both local and global information are crucial for SR reconstruction: while lossy image parts could rebuild themselves from local neighborhoods, they could also learn from distant but similar features for refinement.

Existing graph-based methods have demonstrated some sort of space flexibility. As shown in Figure 1, compared to convolution and window attention options where a pixel always attends to a fixed neighboring region, graph aggregation could flexibly focus on crucial image parts regardless of space limits. However, existing graph-based works are not spatially flexible enough as they fail to consider both local and global aspects.

In fact, computation considerations are the major concern of a comprehensive connection search. It is costly to construct global graphs by searching through all image nodes. As a remedy, strided sampling is always adopted [21, 26, 52]. But consequently, local regions are neglected. The choice of pixels as graph nodes makes graph construction even more challenging, because node spaces are enlarged further and it is hardly impractical to search through all pixels for graph construction.

For the sake of efficiency, and to gather local-level surrounding features and global-level distant features contributive to detail reconstruction in SR tasks, we use two types of sampling to aggregate both local and global information, shown in Figure 4. For local sampling, the neighborhood surrounding a certain node is chosen as a local-scale search space on which a graph is established; for global sampling, sampled nodes span over the image in a dilated pattern.

Figure 4. An illustration of sampling strategies. The original image with the highlighted image node (left), local sampling (middle), and global sampling (right) are demonstrated. In this way, image graphs could flexibly gather both local and global information in an efficient manner.

In the proposed IPG SR model, both sampling methods are performed in order to construct local and global-scale graphs. For each pixel node, its node connections are searched within the sampling space. In this way, we achieve spatial flexibility in a cost-efficient fashion.

3.3. Graph Aggregation

Now that flexible graphs are constructed, graph aggregation is performed such that each node can communicate with its connected neighbors and use their information for self-refinement in SR. In vision applications of graphs, aggregation in max-pooling [12, 28] or edge-conditioned [21, 26, 36] forms are mostly preferred. We tend to adopt edge-conditioned aggregation [36], as max-pooling could result in significant loss of information from neighboring pixels that are crucial for low-level vision. Since pixel reconstruction in SR heavily relies on abundant neighboring information, edge-conditioned aggregation is adopted as it cares about pixel inter-relations and maintains more neighboring information for effective reconstruction. The edge-conditioned aggregation [21, 26, 36, 52] is mathematically

formulated as follows: at the k^{th} layer in IPG, given node feature $\mathbf{h}^{k-1}$ as input, neighboring node set $\mathcal{N}(v)$, the node v output $\mathbf{h}_v^k$ is calculated as:

$$\mathbf{h}_v^k = \frac{1}{C^k} \sum_{u \in N(v)} \exp\left(f^k\left(u, v\right)\right) \mathbf{h}_u^{k-1}, \tag{3}$$

where $f^k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a parameterized function that measures the correlation between node pair (u, v); $C^k := \sum_{u \in N(v)} \exp\left(f^k\left(u, v\right)\right)$ is a normalizing constant. In our case, cosine similarity is adopted as the correlation metric.

Despite the flexibility of graph aggregation, we are concerned that the spatial information is damaged in the process of graph aggregation: since all nodes are treated equally, the model will grasp little knowledge about node position. Hence, inspired by Liu et al. [25, 29], we add relative position encoding to node features before aggregation to enhance position information.

3.4. Model Architecture

The proposed graph construction and graph aggregation mechanisms are then merged into the effective MetaFormer architecture as the IPG model. The detailed structure of IPG is shown in Figure 3.

The IPG Architecture. The general architecture of IPG follows mainstream SR models [5–7, 23, 53]. As an LR image is inputted into the model, it is first passed to a Conv Layer to extract shallow features. Then the features are passed through a series of Multiscale Graph-aggregation Blocks (MGB) for effective deep feature extraction with the aid of flexible graphs. Each MGB consists of a stack of Graph Aggregation Layers (GAL) where graph aggregation on both local and global scales is performed. Finally, the image is spatially reconstructed via a pixel-shuffle upsampler.

MGB Blocks. Multiscale Graph-aggregation Blocks (MGB) gather both local and global-scale information for effective image SR reconstruction. Both local and global pixel graphs are calculated per block based on block inputs. Block-wise graph calculations enable regular graph updates throughout the model. The construction steps are detailed in Section 3.2, where local and global sampling is performed respectively for local or global graph construction. The two types of graphs (local/global) are then distributed to GALs throughout the block for aggregation operations. Local and global graphs are distributed in a sequentially alternative manner to ensure that both local and global-scale information are sufficiently aggregated.

GAL Layers. The design of Graph Aggregation Layers (GAL) follows the general MetaFormer-like structure [44]. Graphers that are responsible for graph aggregation (as introduced in Section 3.3) are used as token mixers. Grapher gathers local or global information based on the type of graph it receives. Following [5, 22, 53], we leverage efficient Channel Attention (CA) modules and ConvFFN blocks to aid the SR performance of our model.

4. Experiments

4.1. Experiment Setting

We adopt the general training setting of recent SR works [5, 6, 23] for fair comparison. We use DIV2K+Flickr2K (DF2K) as the training set (DIV2K for lightweight models). Cropped patches of size 64×64 are sent into the model. Training data is augmented by random flipping and random rotation of degrees $[0, 90, 180, 270]$. Our model is trained for 500K iterations using the Adam Optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$) at a learning rate of $2e-4$. The training input size is set at 64×64. MultistepLR is adopted where the learning rate is halved at iterations $[250000, 400000, 450000, 475000]$. Batchsize is set as 32.

4.2. Super-Resolution Performance

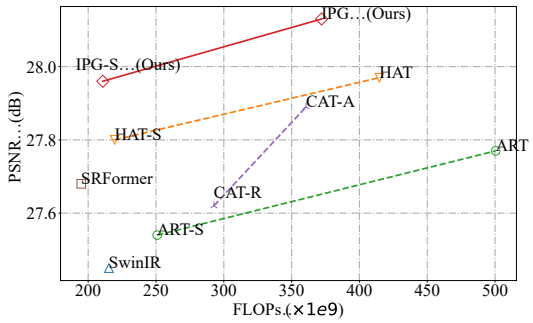

Figure 5. Comparison of the proposed IPG-S and IPG with SR baselines in FLOPs and performance. Urban100×4 in PSNR is reported. Thanks to the flexibility of graphs, IPG could outperform other SR models by >0.1dB at similar FLOPs. Notably, FLOPs only reflect theoretical computation costs instead of actual inference speed.

We compare the performance of our model with various SR baselines in Table 1 on Set5 [3], Set14 [45], BSDS100 [33], Urban100 [13], and Manga109 [34] benchmarks. The chosen baselines are as follows: IGNN [52] is a patch-graph-based SR method; ELAN [49], IPT [4], SwinIR [23], CAT-A [6], ART [47], GRL-B [22], and HAT [5] are transformer-based methods. Performance of SR $\times 2$, $\times 3$, $\times 4$ models are measured by PSNR and SSIM. "+" marks stand for baselines with self-ensemble strategies, which boosts SR performance further. IPG could outcompete other models by large margins. In particular, IPG reaches 28.13dB of PSNR on the Urban100$\times$ 4 task, exceeding existing SOTA by more than 0.1dB with standard training setting. IPG has an outstanding performance in the PSNR versus theoretical FLOPs plot, shown in Figure 5. But notably, IPG runs slower than previous methods for

	Method	SET5 [3]		SET14 [45]		B100 [33]		Urban100 [13]		Manga109 [34]	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2 SR	**IGNN** [52]	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
	ELAN [49]	38.36	0.9620	34.20	0.9228	32.45	0.9030	33.44	0.9391	39.62	0.9793
	IPT [4]	38.37	N/A	34.43	N/A	32.48	N/A	33.76	N/A	N/A	N/A
	SwinIR [23]	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
	EDT [20]	38.45	0.9624	34.57	0.9258	32.52	0.9041	33.80	0.9425	39.93	0.9800
	CAT-A [6]	38.51	0.9626	34.78	0.9265	32.59	0.9047	34.26	0.9440	40.10	0.9805
	ART [47]	38.56	0.9629	34.59	0.9267	32.58	0.9048	34.30	0.9452	40.24	0.9808
	GRL-B* [22]	38.48	0.9627	34.64	0.9265	32.55	0.9045	33.97	0.9437	40.06	0.9804
	HAT* [5]	38.61	0.9630	34.77	0.9266	32.61	0.9053	34.45	0.9465	40.23	0.9806
	IPG (Ours)	38.61	0.9632	34.73	0.9270	32.60	0.9052	34.48	0.9464	40.24	0.9810
	IPG+ (Ours)	38.65	0.9633	34.87	0.9274	32.62	0.9053	34.54	0.9468	40.32	0.9811
×3 SR	**IGNN** [52]	34.72	0.9298	30.66	0.8484	29.31	0.8105	29.03	0.8696	34.39	0.9496
	ELAN [49]	34.90	0.9313	30.80	0.8504	29.38	0.8124	29.32	0.8745	34.73	0.9517
	IPT [4]	34.81	N/A	30.85	N/A	29.38	N/A	29.49	N/A	N/A	N/A
	SwinIR [23]	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
	EDT [20]	34.97	0.9316	30.89	0.8527	29.44	0.8142	29.72	0.8814	35.13	0.9534
	CAT-A [6]	35.06	0.9326	31.04	0.8538	29.52	0.8160	30.12	0.8862	35.38	0.9546
	ART [47]	35.07	0.9325	31.02	0.8541	29.51	0.8159	30.10	0.8871	35.39	0.9548
	GRL-B* [22]	35.05	0.9323	31.00	0.8543	29.49	0.8153	29.83	0.8837	35.24	0.9541
	HAT [5]	35.07	0.9329	31.08	0.8555	29.54	0.8167	30.23	0.8896	35.53	0.9552
	IPG (Ours)	35.10	0.9332	31.10	0.8554	29.53	0.8168	30.36	0.8901	35.53	0.9554
	IPG+ (Ours)	35.19	0.9335	31.14	0.8558	29.57	0.8171	30.48	0.8911	35.65	0.9558
×4 SR	**IGNN** [52]	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
	ELAN [49]	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167	31.68	0.9226
	IPT [4]	32.64	N/A	29.01	N/A	27.82	N/A	27.26	N/A	N/A	N/A
	SwinIR [23]	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
	EDT [20]	32.82	0.9031	29.09	0.7939	27.91	0.7483	27.46	0.8246	32.03	0.9254
	CAT-A [6]	33.08	0.9052	29.18	0.7960	27.99	0.7510	27.89	0.8339	32.39	0.9285
	ART [47]	33.04	0.9051	29.16	0.7958	27.97	0.7510	27.77	0.8321	32.31	0.9283
	GRL-B* [22]	32.90	0.9039	29.14	0.7956	27.96	0.7497	27.53	0.8276	32.19	0.9266
	HAT [5]	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
	IPG (Ours)	33.15	0.9062	29.24	0.7973	27.99	0.7519	28.13	0.8392	32.53	0.9300
	IPG+ (Ours)	33.23	0.9064	29.30	0.7979	28.04	0.7525	28.22	0.8407	32.72	0.9310

Table 1. Comparison of IPG with recent SR methods. Methods with "*" are replicated with standard setting, detailed in the Appendix; items with "+" stands for self-ensemble strategies.

lack of hardware support; IPG costs slightly advantageous peak memory. Apart from the quantitative comparison in PSNR/SSIM metrics, we also provide visual results of difficult details reconstructed by IPG. As shown in Figure 6, the proposed IPG model is compared against recent top-tier SR baselines on the SR $\times 4$ task. Our IPG model could visually outcompete other methods.

4.3. Lightweight SR

We also provide some lightweight variants of IPG. While the architecture of IPG-S and IPG-Tiny remains the same as the base version, their depth and dimension width are significantly reduced for computation-constrained applications. Variant config details are in the Appendix. We choose canonical SR baselines including CARN [1], IMDN [14], LAPAR-A [19], LatticeNet [31], and ESRT [30]; other baselines have MetaFormer architectures, including SwinIR-light [23], ELAN [48], SwinIR-NG [53], and SRFormer-light [53]. IPG could outcompete SR baselines that share similar MetaFormer [44] architectures at similar computation costs. Specifically, IPG-Tiny could achieve more than 0.1dB of PSNR improvement on Urban100 compared with other lightweight baselines.

4.4. Ablation Study

Pixel Nodes vs. Patch Nodes. As claimed in Section 3.2, patches as image graph nodes may suffer in cases like object shift and object rotation. The forcible element-wise alignment of patch pixels could hinder SR performance to a great extent. In order to investigate the negative impact of patch nodes, we provide a patched version of IPG as follows: rather than using pixel nodes, we use 2×2-sized patches as nodes in this patched variant. Other model settings are kept unchanged for fair comparison, including model depths, feature dimensions, MLP ratios, and sampling region size. However, we are aware that the computation costs of patch node aggregation could be different compared to pixel nodes. Hence, we tune the number of aggregation times for the patched variant, to make sure that aggregation of patch nodes has a similar cost as pixel node aggregation. The ablation results are shown in Table 2.

	Set5	Set14	Urban100
2×2-Patched IPG	33.01	29.18	27.81
Pixel IPG	**33.15**	**29.24**	**28.13**

Table 2. Comparison of pixel-node graphs against patch-node graphs in IPG. SR$\times 4$ results are reported.

In Figure 7, we also provide a vision comparison be-

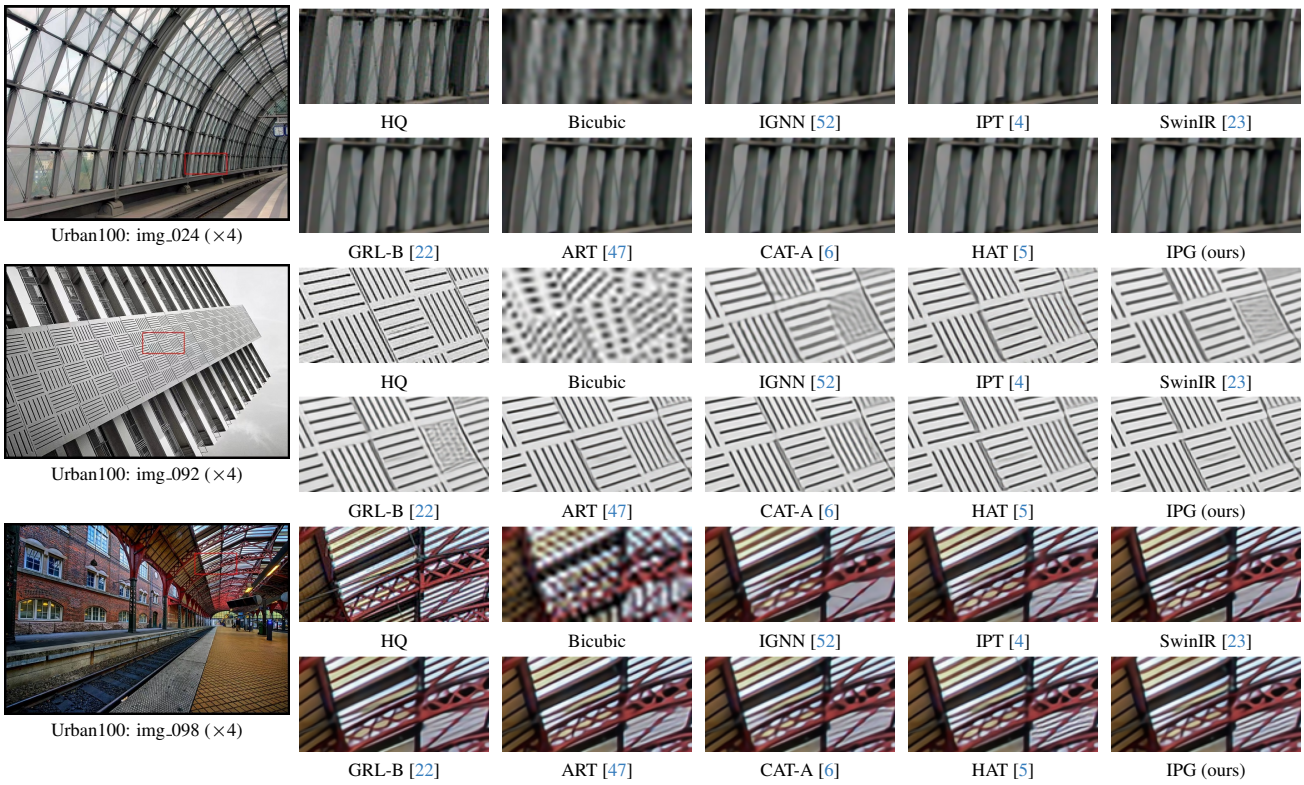

Figure 6. Visual comparison of image SR (×4) with latest SR baselines in challenging cases.

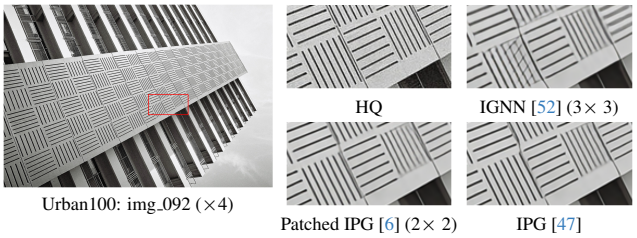

Figure 7. Visual comparison of patch-node graph methods over IPG that uses pixel node graphs.

tween IGNN [52] (3×3 patch), patched IPG (2×2 patch), and IPG (pixel graph). For the IGNN case that uses large patches as nodes, patch alignment is causing trouble: the model is puzzled about the texture orientations, so it makes mistakes during texture reconstruction by aligning pattern-similar but orientation-different patches. For patched IPG, misaligned orientation is causing issues as well: line textures are distorted. In contrast, pixel aggregation in IPG avoids misalignment and performs well.

KNN vs. Degree-flexible Graphs. In canonical graph applications for vision tasks, k-Nearest-Neighbor (KNN) graphs are normally adopted for graph construction. However, KNN still suffers the degree rigidity, as all nodes are connected to a fixed number of neighboring nodes. To improve degree flexibility, we consider the unbalanced demand of pixels in SR tasks and propose a degree-flexible scheme of graph construction (introduced in Section 3.2). We compare KNN and the degree-flexible graph scheme in IPG under the same settings. According to Table 3, Degree-Flex could outcompete the KNN variant of IPG. Interestingly, our Degree-Flex graph could outcompete computation-heavy fully-connected graphs (see Appendix), further implying the effectiveness of our Degree-Flex graphs.

	Set5	Set14	Urban100
KNN	33.09	29.19	28.06
Degree-Flex	**33.15**	**29.24**	**28.13**

Table 3. Comparison of degree-flexible graphs against plain KNN graphs in IPG. SR×4 results are reported.

As shown in Figure 8, we visually compare the connected regions of detail-poor and detail-rich pixels at the local scale. Detail-poor pixels are usually on flat, changeless parts of the image; these pixels require minimal node connection for reconstruction. On the contrary, detail-rich pixel nodes are usually on edges and corners, which are assigned higher node degrees and thus a larger region of perception. In this sense, the detail-rich indicator metric D_F could endow IPG with degree flexibility.

Local/Global Graphs. In the design of IPG, we aggregate information from both local and global scales based on local and global pixel node sampling. Though numerous SR works highlight the importance of either local or non-local

	Method	Training Dataset	Params	FLOPs	SET5 [3]		SET14 [45]		B100 [33]		Urban100 [13]		Manga109 [34]	
					PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2 SR	**CARN** [1]	DIV2K	1592K	222.8G	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
	IMDN [14]	DIV2K	694K	158.8G	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	38.88	0.9774
	LAPAR-A [19]	DF2K	548K	171G	38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283	38.67	0.9772
	LatticeNet [31]	DIV2K	756K	169.5G	38.15	0.9610	33.78	0.9193	32.25	0.9005	32.43	0.9302	N/A	N/A
	ESRT [30]	DIV2K	751K	-	38.03	0.9600	33.75	0.9184	32.25	0.9001	32.58	0.9318	39.12	0.9774
	SwinIR-light [23]	DIV2K	910K	244G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
	ELAN [49]	DIV2K	621K	203G	38.17	0.9611	33.94	0.9207	32.30	0.9012	32.76	0.9340	39.11	0.9782
	SwinIR-NG [7]	DIV2K	1181K	274.1G	38.17	0.9612	33.94	0.9205	32.31	0.9013	32.78	0.9340	39.20	0.9781
	SRFormer-light [53]	DIV2K	853K	236G	38.23	0.9613	33.94	0.9209	32.36	0.9019	32.91	0.9353	39.28	0.9785
	IPG-Tiny (Ours)	DIV2K	872K	245.2G	38.27	0.9616	34.24	0.9236	32.35	0.9018	33.04	0.9359	39.31	0.9786
×3 SR	**CARN** [1]	DIV2K	1592K	118.8G	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
	IMDN [14]	DIV2K	703K	71.5G	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519	33.61	0.9445
	LAPAR-A [19]	DF2K	594K	114G	34.36	0.9267	30.34	0.8421	29.11	0.8054	28.15	0.8523	33.51	0.9441
	LatticeNet [31]	DIV2K	765K	76.3G	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538	N/A	N/A
	ESRT [30]	DIV2K	751K	-	34.42	0.9268	30.43	0.8433	29.15	0.8063	28.46	0.8574	33.95	0.9455
	SwinIR-light [23]	DIV2K	918K	111G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
	ELAN [49]	DIV2K	629K	90.1G	34.61	0.9288	30.55	0.8463	29.21	0.8081	28.69	0.8624	34.00	0.9478
	SwinIR-NG [7]	DIV2K	1190K	114.1G	34.64	0.9293	30.58	0.8471	29.24	0.8090	28.75	0.8639	34.22	0.9488
	SRFormer-light [53]	DIV2K	861K	105G	34.67	0.9296	30.57	0.8469	29.26	0.8099	28.81	0.8655	34.19	0.9489
	IPG-Tiny (Ours)	DIV2K	878K	109.0G	34.64	0.9292	30.61	0.8470	29.26	0.8097	28.93	0.8666	34.30	0.9493
×4 SR	**CARN** [1]	DIV2K	1592K	90.9G	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
	IMDN [14]	DIV2K	715K	40.9G	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
	LAPAR-A [19]	DF2K	659K	94G	32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871	30.42	0.9074
	LatticeNet [31]	DIV2K	777K	43.6G	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	N/A	N/A
	ESRT [30]	DIV2K	751K	-	32.19	0.8947	28.69	0.7833	27.69	0.7379	26.39	0.7962	30.75	0.9100
	SwinIR-light [23]	DIV2K	930K	63.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
	ELAN [49]	DIV2K	640K	54.1G	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
	SwinIR-NG [7]	DIV2K	1201K	63.0G	32.44	0.8980	28.83	0.7870	27.73	0.7418	26.61	0.8010	31.09	0.9161
	SRFormer-light [53]	DIV2K	873K	62.8G	32.51	0.8988	28.82	0.7872	27.73	0.7422	26.67	0.8032	31.17	0.9165
	IPG-Tiny (Ours)	DIV2K	887K	61.3G	32.51	0.8987	28.85	0.7873	27.73	0.7418	26.78	0.8050	31.22	0.9176

Table 4. Comparison of IPG-Tiny with recent lightweight SR methods. Under comparable Params and FLOPs, IPG-Tiny could outcompete lightweight SR baselines.

Figure 8. Visual demonstration of parts connected to a pixel node. Compared with detail-poor nodes (left), detail-rich nodes (right) requires a larger amount of aggregation due to difficulty in reconstruction.

aggregation, we demonstrate that both aspects are actually crucial for SR performance. As shown in Table 5, we perform experiments that uses local or global-scale graphs only under identical settings, and compare them with the original version of IPG. The optimal performance is reached when both local and global aggregation are performed.

	Set5	Set14	Urban100
Local Only	33.08	29.19	28.10
Global Only	33.04	29.11	27.59
Local + Global	**33.15**	**29.24**	**28.13**

Table 5. Ablations of local and global graph usages on SR×4.

5. Conclusion

Existing SR measures are rigid both in terms of space and degree. In order to break rigidity in SR, the paper proposes IPG – an SR model that fully leverages the flexibility of graphs. Firstly, we are aware that SR is a task where detail-rich image parts require lopsided reconstruction efforts. Hence, we designed a degree-varying scheme for graph construction rather than k-nearest neighbors. Then we take pixels rather than patches for graph aggregation. To reach the maximum amount of spatial flexibility, nodes are sampled both on local and global scales. IPG reaches State-of-the-art performance compared with SR baselines.

Acknowledgement. This work is supported by the National Key R&D Program of China under Grant No.2022ZD0160300 and the National Natural Science Foundation of China under Grant No.62276007. We gratefully acknowledge the support of MindSpore, CANN and Ascend AI Processor used for this research.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, 2018. 6, 8

[2] Qiqi Bao, Bowen Gang, Wenming Yang, Jie Zhou, and Qingmin Liao. Attention-driven graph neural network for deep face super-resolution. *IEEE Transactions on Image Processing*, 31:6455–6470, 2022. 2

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012. 5, 6, 8

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 2, 5, 6, 7

[5] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 1, 2, 5, 6, 7

[6] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Cross aggregation transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:25478–25490, 2022. 1, 2, 5, 6, 7

[7] Haram Choi, Jeongmin Lee, and Jihoon Yang. N-gram in swin transformers for efficient lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2071–2081, 2023. 2, 5, 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Yuanbiao Gou, Peng Hu, Jiancheng Lv, Hongyuan Zhu, and Xi Peng. Rethinking image super resolution from long-tailed distribution learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14327–14336, 2023. 1, 3

[10] Hayit Greenspan. Super-resolution in medical imaging. *The computer journal*, 52(1):43–63, 2009. 1

[11] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9199–9208, 2021. 3

[12] Kai Han, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu. Vision gnn: An image is worth graph of nodes. *Advances in Neural Information Processing Systems*, 35:8291–8303, 2022. 1, 2, 4

[13] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 5, 6, 8

[14] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM MM*, 2019. 6, 8

[15] Kui Jiang, Zhongyuan Wang, Peng Yi, Guangcheng Wang, Tao Lu, and Junjun Jiang. Edge-enhanced gan for remote sensing image superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5799–5812, 2019. 1

[16] 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. 1, 2

[17] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4558–4567, 2018. 2

[18] Hunsang Lee, Hyesong Choi, Kwanghoon Sohn, and Dongbo Min. Cross-scale knn image transformer for image restoration. *IEEE Access*, 11:13013–13027, 2023. 2

[19] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. In *NeurIPS*, 2020. 6, 8

[20] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv:2112.10175*, 2022. 6

[21] Yao Li, Xueyang Fu, and Zheng-Jun Zha. Cross-patch graph convolutional network for image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4651–4660, 2021. 2, 3, 4

[22] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023. 1, 2, 5, 6, 7

[23] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1, 2, 5, 6, 7, 8

[24] 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. 1, 2

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[26] Ziyu Liu, Ruyi Feng, Lizhe Wang, Wei Han, and Tieyong Zeng. Dual learning-based graph neural network for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 1, 2, 3, 4

[27] Ziyu Liu, Ruyi Feng, Lizhe Wang, Wei Han, and Tieyong Zeng. Dual learning-based graph neural network for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 2

[28] Ziyu Liu, Ruyi Feng, Lizhe Wang, Wei Han, and Tieyong Zeng. Dual learning-based graph neural network for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 4

[29] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 5

[30] Zhisheng Lu, Hong Liu, Juncheng Li, and Linlin Zhang. Efficient transformer for single image super-resolution. *arXiv:2108.11084*, 2021. 6, 8

[31] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *ECCV*, 2020. 6, 8

[32] Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, and Yun Fu. Image as set of points, 2023. 2

[33] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001. 5, 6, 8

[34] 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*, 2017. 5, 6, 8

[35] Chong Mou and Jian Zhang. Graph attention neural network for image restoration. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2

[36] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3693–3702, 2017. 3, 4

[37] Tomasz Tarasiewicz, Jakub Nalepa, and Michal Kawulok. A graph neural network for multiple-image super-resolution. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1824–1828. IEEE, 2021. 2

[38] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Image denoising with graph-convolutional neural networks. In *2019 IEEE international conference on image processing (ICIP)*, pages 2399–2403. IEEE, 2019. 2

[39] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deep graph-convolutional image denoising. *IEEE Transactions on Image Processing*, 29:8226–8237, 2020. 2

[40] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. Omni aggregation networks for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22387, 2023. 2

[41] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 2

[42] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[43] Anni Yu and Yu-Bin Yang. Learning to explain: a gradient-based attribution method for interpreting super-resolution networks. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 3

[44] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. *CoRR*, abs/2111.11418, 2021. 5, 6

[45] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, 2010. 5, 6, 8

[46] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution, 2023. 1

[47] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. *arXiv preprint arXiv:2210.01427*, 2022. 2, 5, 6, 7

[48] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European Conference on Computer Vision*, pages 649–667. Springer, 2022. 6

[49] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. *arXiv:2203.06697*, 2022. 5, 6, 8

[50] 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. 1, 2

[51] 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

[52] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. *Advances in neural information processing systems*, 33:3499–3509, 2020. 1, 2, 3, 4, 5, 6, 7

[53] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. *arXiv preprint arXiv:2303.09735*, 2023. 2, 5, 6, 8