iHAST: Integrating Hybrid Attention for Super-Resolution in Spatial Transcriptomics

Xi Li
xil43@uci.edu.cn
Ziheng Duan
zihend1@uci.edu
Yi Dai
ydai12@uci.edu
Siwei Xu
s.xu@uci.edu
Jing Zhang*
zhang.jing@uci.edu

Computer Science Dept.
University of California, Irvine
Irvine, USA

Abstract

Spatial transcriptomics (ST) technologies have transformed genomic research by facilitating spatially-resolved gene expression profiling, offering unprecedented opportunities to explore cellular communication and organization. However, most existing ST assays are constrained to resolving cell clusters or multicellular structures, resulting in the amalgamation of signals from multiple cells and obscuring the spatial dynamics of diverse cell populations. To address this limitation, we conceptualize ST data as a specific type of image, where cells represent pixels and genes are akin to channels. Consequently, we propose a novel Hybrid Attention Transformer framework, named iHAST , to enhance the spatial resolution of ST data. In support of this framework, we have curated an extensive ST dataset comprising over 800 examples from diverse cell types and sequencing technologies for training and testing purposes. Subsequently, we conducted rigorous evaluations of our approach and consistently observed performance enhancements surpassing those achieved by state-of-the-art methods, underscoring the robustness and generalizability of our iHAST model across a diverse array of biological contexts and experimental conditions.

1 Introduction

Spatial transcriptomics (ST) technology has transformed genomic research by enabling spatially-resolved gene expression profiling within the native context of tissues [29][27][33][37]. This breakthrough provides an unparalleled opportunity to unravel the spatial heterogeneity of gene expression, offering a comprehensive depiction of cellular communication and organization. However, despite its promise, many existing ST assays primarily resolve cell clusters or multicellular structures rather than individual cells, leading to the amalgamation of signals from multiple, potentially diverse cell types within a single measurement point[25][37].

Such aggregation may obscure the spatial dynamics of diverse cell populations, significantly impacting various downstream analyses, such as cell-type identification, characterization of cellular functions, and investigation of cell-to-cell communication. Therefore, it is essential to design novel methods to enhance the resolution of ST data.

In response to the challenges associated with improving the resolution of ST data, we have developed a novel framework for resolution enhancement, named iHAST , drawing inspiration from super-resolution (SR) techniques within the realm of computer vision **Fig.2**. The code for iHAST is publicly available at `https://github.com/aicb-ZhangLabs/iHAST`. Specifically, iHAST treats each ST dataset as a specialized form of image data, where individual ST spots are analogized to pixels, and their corresponding gene expression values are represented as extended RGB channels. Subsequently, iHAST employs downsampling of the ST data to create high-low resolution pairs to train a Hybrid Attention Transformer model and predict high-resolution gene expression patterns. Consequently, the trained iHAST model demonstrates the capability further to enhance the spatial resolution of existing ST data inputs. In contrast to prevailing SR methods utilized in conventional image analyses, we have devised two distinct modules tailored to address the specific challenges inherent in ST data, as outlined in our contributions detailed below.

- **Wavelet Transform:** The substantial dynamic range of gene expressions, encompassing multiple orders of magnitude, presents a notable obstacle in achieving precise expression prediction during the resolution enhancement procedure. In response to these challenges, iHAST incorporates a Wavelet Transform module to mitigate oversmoothing effects and bolster the encoding of high-frequency information. This module substantially improves the encoding of fine details, a critical aspect for precise prediction of sparse gene expressions and preservation of cell-type diversity.

- **Advanced Positional Encoding:** We propose a Sinusoidal and Linear Positional Encoding strategy that enhances spatial awareness within our network architecture. This improvement enables a more profound comprehension of both local and global cell-cell interactions, thereby facilitating detailed biological analyses and extending insights beyond conventional local observations

- **Curate Large ST Dataset and Achieve SOTA Performance:** We curated a comprehensive dataset comprising over 800 diverse datasets obtained through three prominent streamline technologies, encompassing approximately 30 distinct complex tissues. Subsequently, we conducted rigorous evaluations of our approach using over 100 ST datasets, consistently observing performance enhancements surpassing those achieved by state-of-the-art methods. This extensive validation process underscores the robustness and generalizability of our iHAST model across a diverse array of biological contexts and experimental conditions.

2 Related Work

2.1 Spatial transcriptomics

ST assays emerged as a transformative technology in genomics, enabling concurrent gene expression profiling within the spatial confines of tissues, thereby furnishing crucial insights into cellular organization, communication, and function (Rao et al., 2021[29]). This methodology plays a pivotal role in elucidating essential biological processes, such as cell differentiation, communication, tissue structure, and the tissue microenvironment. While traditional technologies like in situ hybridization (ISH)-based methods (e.g., smFISH[2], MERFISH[3], seqFISH[9]) and in situ sequencing (ISS)-based methods (e.g., FISSEQ[18], STARmap[35]) have significantly contributed to our understanding, they often exhibit limited throughput and scalability, thereby constraining their utility in large-scale studies or high-resolution mapping of gene expressions.

Recent advancements in barcode-based ST technology[30][34][22][26] enhance throughput and scalability, allowing for the profiling of thousands of genes across defined tissue regions. However, a drawback of this technology is its limited spatial resolution compared to single-cell techniques, which can lead to the mixing of signals from different cell types at a single measurement point. Thus, developing new computational and experimental methods to improve ST data resolution remains crucial.

2.2 Existing methods to boost the spatial resolution of ST data

Numerous computational methods have been proposed to tackle challenge of gene-expression prediction and resolution enhancement from diverse perspectives. For example, several research groups have devised computational algorithms aimed at dissecting spatially resolved gene expression data by estimating the contributions of individual cell types [24] [23][8]. These methods commonly rely on accurate reference profiles for each cell type, which can pose challenges, particularly in heterogeneous tissues. Moreover, these methods may encounter difficulties in accurately distinguishing closely related cell types or accommodating spatial expression variations, potentially leading to inaccuracies in cell type estimation.

Subsequently, several seminal studies highlighted the wealth of information on cell types and transcriptomic profiles contained within histology images[12][38]. These studies employed image processing techniques to align high-resolution histology images with spatial transcriptomics (ST) data, facilitating the integration of spatial context and cellular morphology [28]. The fusion of these modalities enables a finer resolution of gene expressions within tissues. However, it is worth noting that histology images may be absent in numerous applications, particularly outside the realm of cancer research, posing persistent challenges in enhancing the resolution of ST data.

2.3 Existing SR methods for traditional image analysis

Deep learning has significantly impacted various fields, including SR in computer vision [7] [39] [36] [19] [20]. Dong pioneered single image SR with SRCNN, employing a basic CNN architecture [6]. Kim introduced VDSR, a deeper network with residual learning, surpassing SRCNN's performance [13]. Zhang enhanced SR quality with RDN, integrating residual and dense connections [40]. Lim contributed EDSR, a 64-layer residual network [21]. Lai developed Lap-SRN, focusing on multiresolution and efficiency [15, 16]. To address parameter efficiency, Kim proposed DRCN [14]. Tai extended this concept in DRRN and MemNet, combining recursive learning and weight sharing [31, 32]. Ledig introduced SR-GAN for visually enhanced SR [17]. Haris's DBPN offered iterative up-sampling/down-sampling layers for better contextual information capture [10]. These advancements signify the integration of deep learning in SR enhancement.

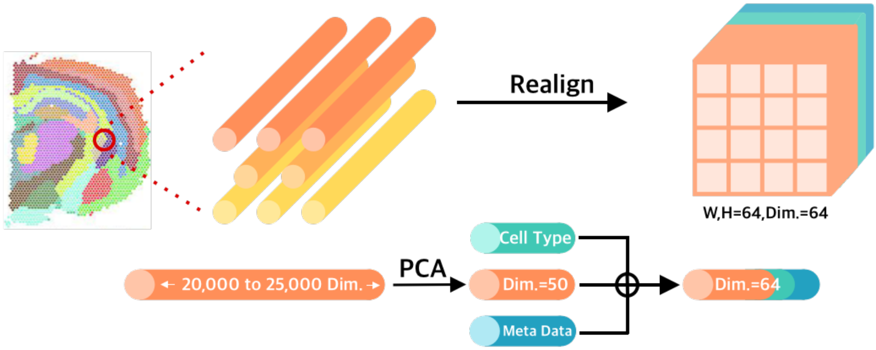

Figure 1: **Data preprocessing workflow** Gene expression profiles are reduced in dimensionality using PCA, combined with cell-type annotations, and rearranged from a honeycomb-like layout to a bitmap-like representation, enabling a computer vision process to enhance resolution and facilitate accurate cell-type prediction.

3 Method

3.1 Large-scale ST dataset curation

We collected a total of 877 datasets from various spatial transcriptomics technology platforms, including 10x Visium, StereoSeq, and SlideSeq. These datasets encompass a large range of tissues, such as brain, embryo, and kidney.

3.2 Uniform ST Data Preprocessing

We adhered to the standard data preprocessing pipelines by Seurat[1], selecting highly expressed and variable genes for normalization using recommended parameters. After reducing gene expression dimensions to 50 components through Principal Component Analysis (PCA) for computational efficiency, we addressed the challenge of uniform coverage in ST data, typically sampled in a hexagonal pattern. To align with visual models, we transformed this hexagonal structure into a regular 64x64 grid, repositioning data points to fit standard image processing frameworks while preserving gene expression accuracy.

For the training process, we employed a preprocessed dataset comprising 877 samples with cell-type annotations. Given the distinct gene expression profiles associated with various cell types, we incorporated cell-type labels as pivotal information in our transformer model. Consequently, we amalgamated cell-type labels and metadata (consisting of 13 dimensions) with the 50-dimensional PCA result to generate a 64-dimensional representation for each spot (pixel). Subsequently, the data underwent preprocessing into 64x64x64 instances and was downsampled to 32x32x64 to construct the paired High-Low resolution data pair. This structured representation, coupled with cell-type annotations, facilitates the acquisition of intricate spatial patterns and relationships by our deep learning model, essential for accurate resolution enhancement.

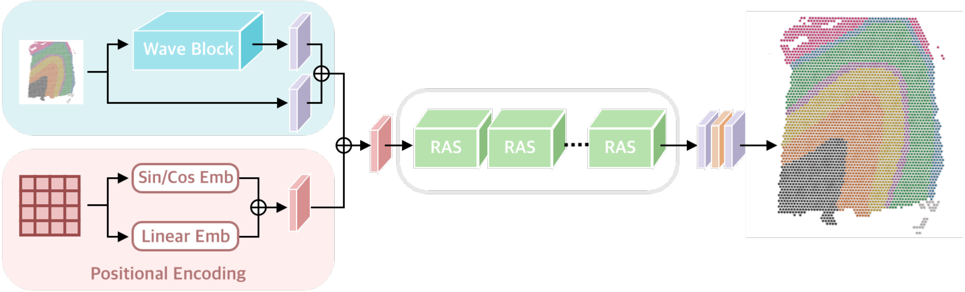

Figure 2: Overall Architecture: The processed data is first passed through two feature extraction modules: the Wavelet Transform Module and the Positional Encoding Module. After these modules extract and process the features, their outputs are concatenated and fed into six Residual Attention Sets (RAS). Finally, the reconstruction phase compiles these enhanced features to produce an output with improved resolution.

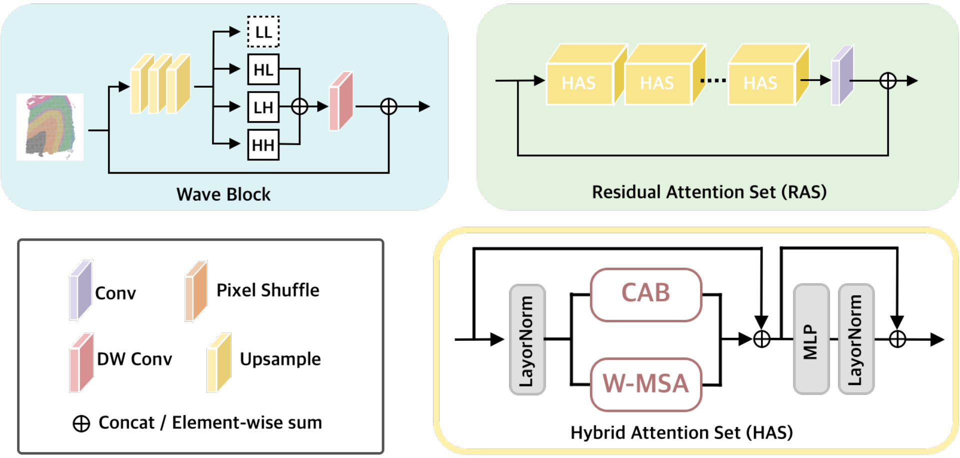

Figure 3: The structure of Wavelet Transform module, Residual Attention Set and Hybrid Attention Set.

3.3 Over all architecture

As shown in **Fig.** 2, our architecture integrates three components: a Wavelet Transformation Module, a Positional Encoding Module, and a Hybrid Attention Transformer Module. By leveraging the Wavelet Transformation Module, iHAST adeptly captures high-frequency details such as textures and edges, significantly augmenting the diversity of the output. Concurrently, the Positional Encoding Module bolsters the architecture's ability to assimilate and process global contextual information. This enriched input is then processed by the Transformer Module, facilitating deeper learning and refinement. Drawing inspiration from the state-of-the-art methodologies in the SR domain, particularly the HAT [5][4] [20]framework and SwinIR [20] framework.

3.4 Wavelet transform Module

Unlike traditional image data, gene expression values within cells typically encompass several orders of magnitude to accommodate diverse cellular functions and responses to various stimuli, posing significant challenges in the super-resolution task. Conventional approaches often yield results that appear excessively smooth or blurred, lacking the sharpness and clarity of the original high-resolution counterparts. Therefore, we designed the wavelet transform module to augment the extraction of multi-level high-frequency information from ST data, thereby markedly enhancing detail and texture retention in super-resolution tasks.

Given a discrete signal $x[n]$, the Haar transform computes the approximation coefficients $a[n]$ and detail coefficients $d[n]$ using the following equations:

$$a[n] = \frac{1}{\sqrt{2}}(x[2n]+x[2n+1]), \quad d[n] = \frac{1}{\sqrt{2}}(x[2n]-x[2n+1]). \tag{1}$$

For a two-dimensional image, the process is applied first along rows and then along columns (or vice versa), resulting in four sub-bands for each level of decomposition: LL (approximation), LH (horizontal details), HL (vertical details), and HH (diagonal details).

By discarding the LL (Approximation Coefficients) as shown in fig. 3, which contain predominantly low-frequency information, and focusing on the concatenation of the remaining high-frequency components, we construct a comprehensive high-frequency feature map that captures intricate image details more effectively.

Subsequently, the extracted high-frequency feature map undergoes processing through a depth-wise (DW) convolution layer, which further refines the feature representation. This refined high-frequency feature map is then concatenated with the original up-sampled image, reintegrating the enhanced details back into the image structure.

3.5 Positional encoding block

Given that ST data inherently contains the coordinates of each spot, perfectly aligning with the concept of positional encoding, we have developed a novel position encoding module to better represent this coordinate information. iHAST propose a novel position encoding module that leverages the complementary strengths of sine/cosine and linear embeddings. This dual-embedding approach comprehensively captures spatial information, promoting a rich and positionally-aware feature representation. Concatenation of the encoded features is followed by a depth-wise convolution (DW Conv) to refine spatial information and enhance local context integration.

Sinusoidal positional embedding The Sinusoidal positional encoding for a position p and dimension d is defined as follows:

$$PE(p,2i) = \sin\left(\frac{p}{10000^{2i/D}}\right), \quad PE(p,2i+1) = \cos\left(\frac{p}{10000^{2i/D}}\right). \tag{2}$$

where $PE(p,2i)$ and $PE(p,2i+1)$ are the positional encodings at position p for the ith dimension, D is the total number of dimensions, and i ranges from 0 to $D/2-1$.

Linear positional embedding Linear positional embedding assigns a unique embedding to each position in a sequence. For a given position p, the linear positional embedding is

$$LPE(p) = E_p \tag{3}$$

where $LPE(p)$ denotes the linear positional embedding for position p, and E_p represents the embedding vector associated with position p. The vectors E_p are parameters learned during the training process of the model. Furthermore, we seamlessly integrate the output with the results of a wavelet transform. This leverages multi-resolution analysis, capturing both global and local dependencies within the image.

3.6 Residual Attention Module

Similar to the use of Residual Hybrid Attention Groups (RHAG) and Hybrid Attention Blocks (HAB) from HAT [5][4] framework, our model integrates these components into its architecture, designated as RAS and HAS respectively. However, we have strategically omitted the Overlapping Cross-Attention Block (OCAB) component and made structural adjustments to HAB to better suit our objectives.

Initially, a 3×3 depthwise(DW) convolution layer $H_{\text{Conv}}(\cdot)$ processes the enhanced low-resolution input image $I_{LQ} \in \mathbb{R}^{H \times W \times C_{\text{in}}}$, which includes wavelet features and spatial coordinates, to extract shallow features $F_0 \in \mathbb{R}^{H \times W \times C}$ as:

$$F_0 = H_{SF}(I_{LQ}), \tag{4}$$

To bridge the gap from low to high-dimensional space for pixel tokens, the architecture initially employs a DW convolution layer that processes the input image. Subsequently, deep features, denoted as $F_{DF} \in \mathbb{R}^{H \times W \times C}$, are derived through N_1 sets of the Residual Attention Set (RAS) blocks followed by an additional 3×3 convolution. The process is defined by

$$F_{DF} = H_{\text{Conv}}(F_N), \tag{5}$$

where F_i is the output of the i-th RAS block, expressed as: $F_i = H_{RAS_i}(F_{i-1}), \quad \text{for} \quad i = 1, 2, \ldots, N$. We define $\quad i$ as 6, representing the total number of Residual Attention Set (RAS) blocks used in the process. Deep and shallow features are amalgamated through a global residual connection, culminating in the reconstruction of the high-quality image output. This is articulated as:

$$I_{HQ} = H_{\text{Rec}}(F_0 + F_{DF}), \tag{6}$$

where pixel-shuffle is utilized for super-resolution or dual convolutions for tasks requiring equivalent resolution. The principal elements, termed HAS, comprise N_2 iterations of the hybrid attention set (HAS) followed by a 3×3 convolution layer, integrated with a residual connection.

4 Experiment

4.1 Training setting

The dataset used in this study includes various tissue samples, with the specific dataset name and version noted. The tasks include classification, regression, and generation, evaluated primarily using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). The model architecture features an input size of 32×32, an output size of 64×64, and 64 in-channels. A pixel shuffle with a 2x upscaling factor is used for upsampling, and the embedding dimension is set to 180. The training setup includes a batch size of 32, an initial learning rate of 2.00×10^{-4}, and a predefined adjustment strategy. The Adam optimizer is used, with L2 regularization, over 20,000 epochs. The hardware is an NVIDIA GeForce RTX 3090, and the software framework is PyTorch, version 1.12.

4.2 Super-resolution gene expression prediction

Quantitative Results. We tested our model on three distinct datasets from different spatial transcriptomics technologies: 10x Visium, StereoSeq, and SlideSeq. These datasets represent diverse approaches to capturing ST data, each with unique challenges in resolution and quality. By evaluating our model across these platforms, we provide a comprehensive assessment of its robustness and effectiveness in enhancing spatial resolution across different biological contexts.

Table 1 presents the quantitative comparison of our approach at iteration 60,000 with state-of-the-art SR methods including HAT[4, 5], SwinIR[20], IMDN[11], and RRDB[36]. Our results significantly surpass these established methods across all datasets, due in part to optimizations specifically tailored for handling the high-dimensional nature of biological data. This targeted enhancement enables our model to more effectively resolve the intricate details necessary for accurate gene expression analysis in spatial transcriptomics.

Model	10xVisium		StereoSeq		SlideSeq	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM	PSNR(dB)	SSIM
iHAST	60.38	0.9936	55.23	0.9873	61.83	0.9962
HAT[5]	46.91	0.9302	43.48	0.8707	47.02	0.9365
SWINIR[20]	58.01	0.9868	52.34	0.9662	57.44	0.9863
IMDN[11]	51.18	0.1410	49.86	0.2477	49.83	0.2603
RRDB[36]	52.03	0.9668	46.13	0.9495	47.59	0.9528

Table 1: quantitative comparison with SR method.

Table 2 shows the quantitative comparison of our approach at iteration 20,000 with the relative gene-prediction methods : DIST [41]. The DIST work approach also utilizes graph networks to model spatial transcriptomics (ST) data with the objective of enhancing data resolution. However, DIST focuses on training and transferring models using individual pairs of data, thereby facilitating parallel learning across extensive ST datasets. This strategy enables effective scaling but may limit the model's ability to generalize across varying tissue types and conditions due to the isolated handling of data pairs.

Method	psnr(dB)	SSIM
iHAST	60.38	0.9936
DIST [41]	55.54	0.9863

Table 2: Quantitative comparison with relative biology method method.

4.3 Ablation Study

To assess the impact of individual components in our iHAST model, we conducted an ablation study, as summarized in Table 3. We compared the full model's performance with variants excluding the wavelet transform module (w/o Wave) and the positional encoding module (w/o coord).

Removing the wavelet transform resulted in lower PSNR and SSIM, highlighting its role in preserving high-frequency details. Similarly, excluding positional encoding reduced performance, underscoring its importance in enhancing spatial awareness for modeling long-term cell-cell interactions. These results confirm the critical roles of both modules in achieving high-quality super-resolution.

Model	10xVisium		StereoSeq		SlideSeq	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM	PSNR(dB)	SSIM
iHAST	60.38	0.9936	55.23	0.9873	61.83	0.9962
w/o Wave	53.10	0.9803	47.31	0.9344	53.17	0.9815
w/o Coord	55.11	0.9791	42.36	0.7887	59.38	0.9932

Table 3: Ablation Study.

We explored the impact of additional metadata and cell-type information on our model's performance using the 10xVisium dataset, as these were not available in the other two datasets. Specifically, we evaluated the model under three conditions: using all data ("iHAST"), excluding cell-type information ("w/o Cell-type"), and omitting metadata ("w/o Metadata"). The results highlight the crucial roles of both cell-type information and metadata in enhancing model performance.

Method	psnr(dB)	SSIM
iHAST	60.38	0.9936
w/o Cell-type	45.37	0.9085
w/o Meta data	31.02	0.6694

Table 4: Ablation Study in data preprocessing of 10xVisium data.

The results, as shown in Table 4, highlight the significant impact of both cell-type annotations and metadata on the model's ability to reconstruct high-quality gene-expression. The drastic decrease in both PSNR and SSIM scores upon the removal of these components emphasizes their pivotal role in our framework.

5 Conclusion

In this study, we introduced iHAST , a novel super-resolution framework inspired by computer vision techniques aimed at enhancing the resolution of ST data. Our approach treats ST spots as pixels with gene expression values as extended RGB channels and employs a Hybrid Attention Transformer architecture to synthesize a higher resolution output. Pioneering the application of super-resolution techniques in this domain, this method integrates global, local, high-frequency, and positional features using wavelet transformations and advanced encoding strategies, resulting in significant improvements in detail preservation and spatial awareness, and achieving state-of-the-art (SOTA) performance. The findings provide deeper insights into tissue architecture and cellular interactions, laying the groundwork for future advancements in achieving single-cell resolution in gene expression prediction.

References

[1] Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 2023. doi: 10.1038/s41587-023-01767-y. URL https://doi.org/10.1038/s41587-023-01767-y.

[2] Jingxun Chen, David McSwiggen, and Elçin Ünal. Single molecule fluorescence in situ hybridization (smfish) analysis in budding yeast vegetative growth and meiosis. *JoVE (Journal of Visualized Experiments)*, (135):e57774, 2018.

[3] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

[4] Xiangyu Chen, Xintao Wang, Wenlong Zhang, Xiangtao Kong, Yu Qiao, Jiantao Zhou, and Chao Dong. Hat: Hybrid attention transformer for image restoration. *arXiv preprint arXiv:2309.05239*, 2023.

[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.

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014.

[7] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016.

[8] Rui Dong and Guo-Cheng Yuan. Spatialdwls: accurate deconvolution of spatial transcriptomic data. *Genome biology*, 22(1):145, 2021.

[9] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568 (7751):235–239, 2019.

[10] 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.

[11] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019.

[12] Yuran Jia, Junliang Liu, Li Chen, Tianyi Zhao, and Yadong Wang. Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1):bbad464, 2024.

[13] 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.

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016.

[15] 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.

[16] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2599–2613, 2018.

[17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[18] Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C Ferrante, Richard Terry, Brian M Turczyk, Joyce L Yang, Ho Suk Lee, John Aach, et al. Fluorescent in situ sequencing (fisseq) of rna for gene expression profiling in intact cells and tissues. *Nature protocols*, 10(3):442–458, 2015.

[19] Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer-based image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021.

[20] 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.

[21] 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.

[22] Yang Liu, Mingyu Yang, Yanxiang Deng, Graham Su, Archibald Enninful, Cindy C Guo, Toma Tebaldi, Di Zhang, Dongjoo Kim, Zhiliang Bai, et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell*, 183(6): 1665–1681, 2020.

[23] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications*, 14(1):1155, 2023.

[24] Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature biotechnology*, 40(9):1349–1359, 2022.

[25] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.

[26] Christopher R Merritt, Giang T Ong, Sarah E Church, Kristi Barker, Patrick Danaher, Gary Geiss, Margaret Hoang, Jaemyeong Jung, Yan Liang, Jill McKay-Fleisch, et al. Multiplex digital spatial profiling of proteins and rna in fixed tissue. *Nature biotechnology*, 38(5):586–599, 2020.

[27] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature methods*, 19(5):534–546, 2022.

[28] Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv*, pages 2021–11, 2021.

[29] Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

[30] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[31] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[32] 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.

[33] Luyi Tian, Fei Chen, and Evan Z Macosko. The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 41(6):773–782, 2023.

[34] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, Joshua Gould, et al. High-density spatial transcriptomics arrays for in situ tissue profiling. *bioRXiv Preprint at https://doi. org/10.1101/563338*, 2018.

[35] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, 2018.

[36] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[37] Cameron G Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14(1):68, 2022.

[38] Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022.

[39] 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.

[40] 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.

[41] Yanping Zhao, Kui Wang, and Gang Hu. Dist: spatial transcriptomics enhancement using deep learning. *Briefings in Bioinformatics*, 24(2):bbad013, 2023.