MTVNet: Multi-Contextual Transformers for Volumes – Network for Super-Resolution with Long-Range Interactions

August Leander Høeg*[1], Sophia W. Bardenfleth[1], Hans Martin Kjer[1], Tim B. Dyrby[1,2], Vedrana Andersen Dahl[1], and Anders Dahl[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU)
[2]Danish Research Centre for Magnetic Resonance (DRCMR), Copenhagen University Hospital Hvidovre
aulho@dtu.dk

Abstract

Recent advances in transformer-based models have led to significant improvements in 2D image super-resolution. However, leveraging these advances for volumetric super-resolution remains challenging due to the high memory demands of self-attention mechanisms in 3D volumes, which severely limit the receptive field. As a result, long-range interactions, one of the key strengths of transformers, are underutilized in 3D super-resolution. To investigate this, we propose MTVNet, a volumetric transformer model that leverages information from expanded contextual regions at multiple resolution scales. Here, coarse resolution information from boarder context regions is carried on to inform the super-resolution prediction of a smaller area. Using transformer layers at each resolution, our coarse-to-fine modeling limits the number of tokens at each scale and enables attention over larger regions than previously possible. We compare our method, MTVNet, against state-of-the-art models on five 3D datasets. Our results show that expanding the receptive field of transformer-based methods yields significant performance gains on high-resolution 3D data. While CNNs outperform transformers on low-resolution data, transformer-based methods excel on high-resolution volumes with exploitable long-range dependencies, with our MTVNet achieving state-of-the-art performance. Our code is available at https://github.com/AugustHoeg/MTVNet.

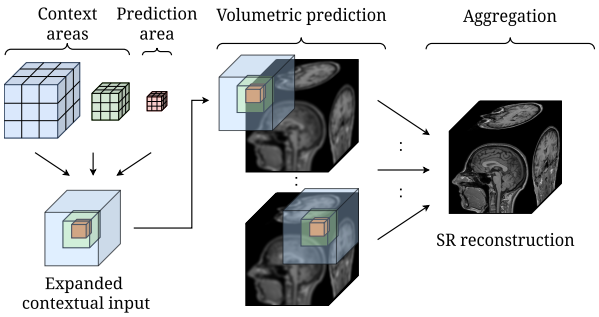

Figure 1. Overview of MTVNet that is informed by a large contextual volume processed at multiple resolution scales for predicting SR in the center volume.

1 Introduction

In recent years, super-resolution (SR) and other vision tasks have seen significant improvements via usage of vision transformers (ViTs). Although ViTs achieve state-of-the-art performance in 2D SR [1–4], few studies have attempted applying ViTs for volumetric SR. Part of the success of ViTs is their increased receptive field compared to Convolutional Neural Networks (CNNs), enabling inferences based on broader image context [5]. Based on experiences from 2D SR, it is logical to assume ViTs will outperform CNNs in the domain of 3D data. However, in volumetric SR, ViTs are challenged by the cubic growth in tokens required to process larger 3D image contexts. Although window-based attention allieviates the quadratic complexity of self-attention [6], the complexity of 3D data still limits the receptive field of volumetric SR models. Because of this disadvantage, the performance gap of CNNs vs. ViT-based architectures for volumetric SR has yet to be fully understood.

Several works have studied enhancement of 3D medical data such as MRI (magnetic resonance imaging) and CT (computed tomography) by upscaling slices independently [7–11]. While such methods circumvent the complexity issues of volumetric SR, not fully considering the 3D context reduces performance and risks slice discontinuities [12–16].

Current brain MRI benchmark datasets for evaluating volumetric SR are relatively low-resolution [17], limiting the benefits of a larger receptive field. Advancements in medical imaging technology enable higher spatial resolution [18], resulting in larger volumes where volumetric SR can benefit from long-range contextual information. Given the potential of SR in clinical settings and the increasing interest in applications like multi-resolution synchrotron imaging [19], there is a need for volumetric SR methods designed for high-resolution (HR) 3D data.

Aside increasing contextual information in volumetric SR, recent studies in 2D SR have shown that the window-based attention mechanism of the

*Corresponding Author.

Proceedings of the 7th Northern Lights Deep Learning Conference (NLDL), PMLR 307, 2026.
© 2026 August Leander Høeg, Sophia W. Bardenfleth, Hans Martin Kjer, Tim B. Dyrby, Vedrana Andersen Dahl, & Anders Dahl. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

Swin-Transformer [6] is not ideal for capturing relationships across distant image regions. Using Local Attribution Mapping (LAM), Chen et al. [2] showed that strengthening long-range information exchange can lead to significant performance gains. Similarly, recent ViT research has focused on modeling long-range interactions to increase performance [20, 21].

To address the limitations caused by self-attention, we present MTVNet, a volumetric SR approach that leverages multi-contextual information from regions beyond the prediction area (see fig. 1) and employs hierarchical attention to enhance long-range information propagation. This approach builds on the assumption that regions closest to the prediction area provide the most critical contextual information, while more distant regions contribute less. Consequently, we design a coarse-to-fine feature extraction and image tokenization scheme that allocates less compute to regions further from the prediction area, enabling larger volumetric inputs without exceeding GPU memory. Inspired by FasterViT [21], we introduce a hierarchical attention mechanism for volumetric image processing, improving modelling of long-range interactions to enhance SR performance.

Finally, MTVNet enables us to investigate the performance gap between CNNs and ViT-based methods for volumetric SR. We compare MTVNet with convolutional and ViT-based SR methods in both 2D and 3D across low-resolution brain MRI data and high-resolution CT data. Extensive experiments show that on low-resolution MRI datasets, CNNs outperform ViTs due to their stronger ability to model local image dependencies. This suggests that the architectural advantages of ViTs only emerge in high-resolution data, where long-range contextual information becomes more important. Conversely, in high-resolution data, we find ViT-based methods achieve superior performance, with our MTVNet leveraging broader contextual input to achieve state-of-the-art performance. These findings highlight that the relative performance of ViT-based methods in 3D highly depends on input resolution.

2 Related Work

2.1 Learning-based super-resolution

The benefits of learning-based SR over classical interpolation were first shown by SRCNN [22]. Several CNN-based models have since been proposed to improve performance and efficiency [23–27].

Despite the success of CNNs, many vision tasks, including image classification [5, 6, 20, 21], object detection [28–31], and segmentation [32–36] have seen improvements using vision transformers. In 2D SR, SwinIR [1] demonstrated the potential of ViTs over CNN-based models by incorporating the Swin Transformer [6] in a residual network scheme. Building upon the success of SwinIR, Chen et al. [2, 37] proposed cross attention of overlapping window partitions and channel attention mechanisms to enable activation of more input pixels. Recently, Hsu et al. [4] suggested combining Swin transformer layers and gating mechanisms in a densely-connected structure [38, 39] to alleviate information bottlenecks. Although these methods achieve state-of-the-art performance in 2D SR, the increased complexity of 3D data makes them difficult to transfer directly to 3D, except when applied slice-wise.

2.2 Super-resolution for 3D volumes

Super-resolution of 3D volumes finds motivation in clinical applications, where workflows are highly dependent on the interpretation of fine-grained structures that are often undersampled during routine acquisitions. SR enhancement of these structures enables improved diagnostic sensitivity and treatment planning through more precise delineation of organs and lesions. SR reconstructions from LR scans allow shorter acquisition times and alleviate requirements for scanner hardware replacement, enabling increased scanner throughput and accessibility. In CT, SR reduces patient health risks by allowing lower radiation scan protocols without compromising image quality [12, 13, 15–17]. Recognizing these benefits, several 3D SR methods have been proposed, including slice-wise and volumetric approaches.

Slice-wise methods predict each slice independently, enabling support for deeper architectures but neglecting cross-slice information, potentially causing discontinuities in slice predictions. Volumetric SR methods fully utilize the context in 3D, increasing computational complexity but enabling better performance thanks to improved inter-plane modelling [12–16]. Inspired by SRCNN [22] and SRGAN [13], Pham et al. [40] and Chen et al. [13] proposed volumetric adaptations of convolutional SR models and demonstrated the potential of volumetric SR over slice-wise approaches. Research in volumetric SR has since grown rapidly and several methods have been proposed to improve efficiency and performance [12, 14, 15, 41–46]. These approaches are similar to 2D SR, only they aim to improve the image quality along all dimensions of a volumetric image instead. However, other approaches including axial SR models [47–49] have been proposed to increase the slice count of low-resolution MRI volumes while preserving in-plane resolution. To alleviate the limitation of fixed upscaling factors, arbitrary scale SR based on Implicit Neural Representation [45, 50, 51] have been proposed. Multi-contrast volumetric models [49, 51] that leverage information from multiple MRI modalities (T1- and T2-weighted images) have also been proposed. Recent advances in ViTs have also inspired volumetric SR methods.

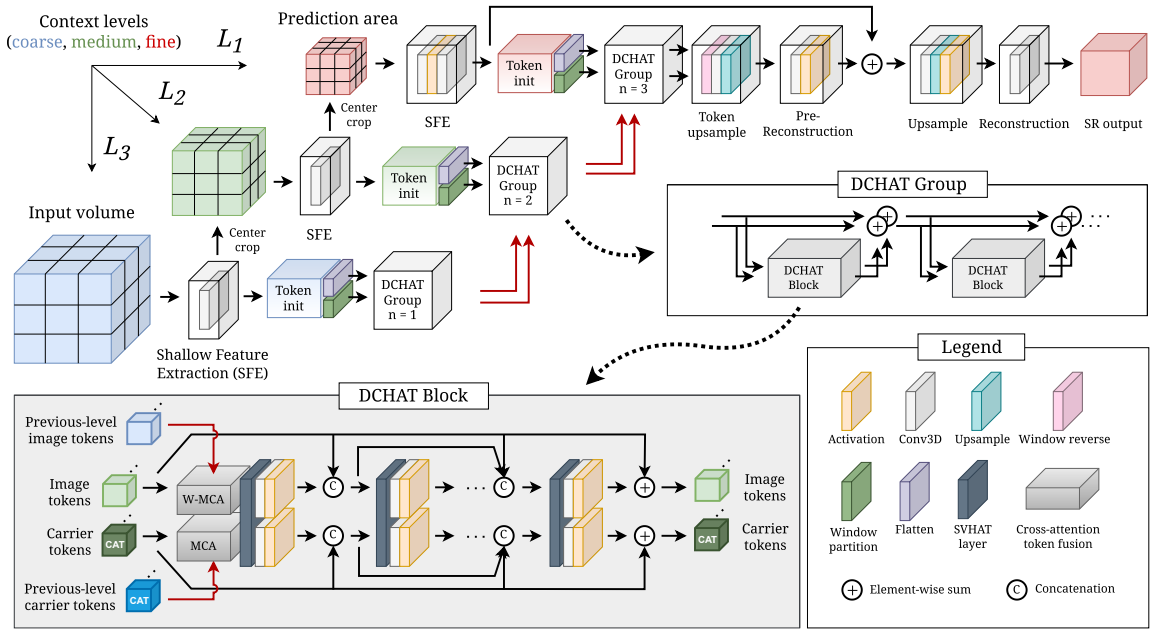

Figure 2. Illustration of MTVNet and the structure of DCHAT block and DCHAT group. Our proposed architecture consists of up to three network stages L_1, L_2, L_3 of multi-contextual volumetric image processing. In each succeeding network stage of MTVNet, we tokenize spatial subsets of the input volume using progressively smaller 3D patch sizes, resulting in both coarse- and fine-grained feature extraction. The depth of subsequent DCHAT groups increases from $n = 1$ to 3 DCHAT blocks towards the last network stage, which produces the SR prediction. Image tokens and carrier tokens from preceding network stages (red arrows) are fused into later stages using multi-head cross-attention (MCA) and window-based multi-head cross-attention (W-MCA).

SuperFormer [16] merged feature embeddings and volume embeddings using a volumetric transformer-based network structure similar to SwinIR [1]. Also inspired by SwinIR, Ji et al. [49] implemented a transformer-based GAN (generative adversarial network) for axial SR using residual swin transformer blocks [1, 6]. The CFTN model [52] used 3D residual channel attention blocks [26] and transformers to capture global cross-scale dependencies between multi-scale feature embeddings. Li et al. [51] proposed a 2D slice-wise multi-modal arbitrary scale SR model featuring a rectangle-window cross-attention transformer to model long-range dependencies.

2.3 ViT enhancements

With the increasing usage of ViTs across image tasks, several works have sought to address the scalability of self-attention. Liu et al. [53] proposed SwinV2, featuring improved normalization and a more robust attention mechanism using cosine similarly. EfficientFormer [54] proposed a lightweight ViT architecture featuring efficient attention mechanisms to achieve competitive accuracy and inference speeds. In CrossViT [20], multi-scale tokenization and efficient cross-attention mechanisms were used to extract and fuse feature representations at different image scales. In connection with scaling ViTs to higher input resolution, several works have suggested augmenting local attention to improve long-range interactions while maintaining efficiency. Twins [55] combined local attention and globally sub-sampled attention to improve efficiency and capture both local and long-range dependencies. RegionViT [56] suggested combining attention between local and regional tokens for conveying global information between attention windows, improving long-range interaction and efficiency. Similarly, FasterViT [21] proposed a hybrid CNN/ViT architecture featuring hiearchical attention mechanisms using local tokens and specialized carrier tokens. These works find natural applicability for volumetric image tasks due to the high data complexity. For instance, FINE [57] used global attention using memory tokens for improved 3D segmentation performance. Yet, to our knowledge, MTVNet is the first to leverage these concepts for volumetric SR.

3 Methods

3.1 Network architecture

The architecture of MTVNet consists of up to three network stages L_1, L_2, L_3 marked by respectively red, green and blue in fig. 2. Stages L_3 and L_2 extract features from regions beyond the SR prediction area and merges them into L_1. These features serve as a prior for stage L_1, producing a SR output conditioned on the surrounding image context.

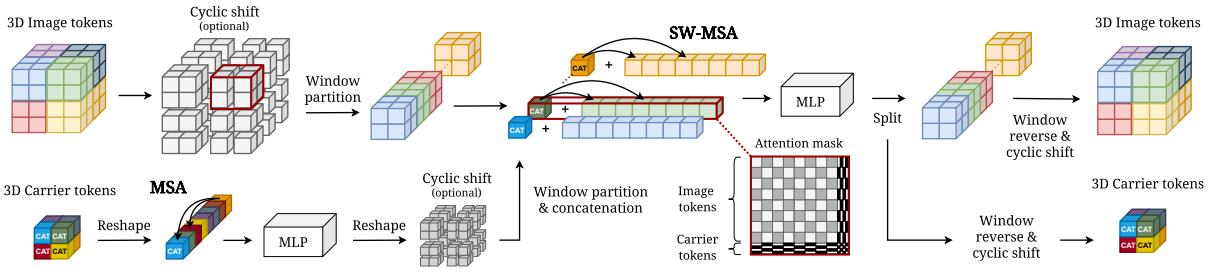

Figure 3. Illustration of our proposed SVHAT. Carrier tokens first undergo full self-attention (MSA) before being concatenated with the tokens of their respective local attention window. Shifted windowed self-attention (SW-MSA) is then performed within each attention window, with the carrier tokens enabling global information exchange between neighboring windows. Attention masking is used to drop information exchange between non-adjacent image tokens and carrier tokens, where grey and black areas indicate masked regions.

Shallow Feature Extraction. Each network stage performs shallow feature extraction (SFE) using $3 \times 3 \times 3$ convolutional layers, producing shallow feature embeddings $\mathcal{F}_{\text{SFE}} \in \mathbb{R}^{C_{\text{SFE}} \times H \times W \times D}$, where C_{SFE} is the feature dimension. The feature map of each stage's SFE module is spatially center-cropped and passed as input to the next, producing more complex features in subsequent stages.

Token initialization. During token initialization, shallow features are projected and tokenized using differently-sized 3D image patches. Stages L_1, L_2 and L_3 use progressively larger patch sizes, covering broader context regions using the same number of tokens or less. Specifically, image token embeddings (ITEs) in each stage are produced by applying a convolution with a stride and kernel size of $p_i \times p_i \times p_i$, where $i \in \{1, 2, 3\}$ is the stage level. These tokens are then projected into vector embeddings of length C_{emb} before being partitioned into 3D local attention windows of $M \times M \times M$ tokens. MTVNet employs specialized carrier tokens (CATs) which summarize the features within each local attention window. In each stage, carrier tokens are generated by applying a convolution with a stride and kernel size of $\lfloor \frac{M}{N_{cat}} \rfloor \times \lfloor \frac{M}{N_{cat}} \rfloor \times \lfloor \frac{M}{N_{cat}} \rfloor$ to the image tokens, yielding ${N_{cat}}^3$ carrier tokens of embedding length C_{emb} for each attention window.

Deep Feature Extraction. For deep feature extraction in each stage, MTVNet employs groups of dense-connected hierarchical attention (DCHAT) blocks. Each group consists of $n \in \{1, 2, 3\}$ DCHAT blocks connected in a residual scheme, as shown in fig. 2. Cross-attention [58] is applied to merge the image tokens and carrier tokens of each DCHAT group into the subsequent stage's DCHAT group, facilitating propagation of multi-scale information.

Reconstruction. In the final stage, token upsampling is performed using deconvolution, transforming token embeddings back into the image space. These features are then refined in a pre-reconstruction stage before being fused with the shallow features from stage L_1 through a long skip-connection. The fused features are then upsampled using a 3D pixel-shuffle layer [59]. We employ a 3D pre-convolution layer initialized according to the ICNR method described in [60] to prevent checkerboard artifacts.

3.2 DCHAT block

For efficient extraction of volumetric image features, we propose a DCHAT block, see fig. 2. Inspired by DRCT [4], our DCHAT block employs a densely connected structure of volumetric transformer layers, LeakyReLU activations, and convolutions. To preserve the feature space of image tokens and carrier tokens, we process each token set using separate skip connections and convolutions. Additionally, we match the embedding dimension of all token embeddings throughout each block to equally promote learning of progressively complex features. As in DRCT [4], we utilize $1 \times 1 \times 1$ convolutions as gating mechanisms between transformer layers to filter redundant features, enabling direct feature transition between DCHAT blocks.

3.3 SVHAT layer

Inspired by FasterViT [21], we implement SVHAT (shifting volumetric hierarchical attention transformer) layer. Like FasterViT, SVHAT adopts the same use of specialized carrier tokens, which serve to summarize and propagate information between local attention windows. First, full attention of all carrier tokens enables global information exchange between attention window summaries. Then, each set of local window tokens is concatenated with their carrier tokens, and windowed attention is applied jointly, allowing carrier tokens to convey information from other windows. This alternating attention procedure efficiently transfers global information between local attention windows to improve information flow, see fig. A.2 in appendix which illustrates the intuition behind carrier tokens. To further enhance this, we reintroduce the notion of shifted-window attention from [6], see fig. 3. Before window partitioning, 3D cyclic-shifting is performed to allow the attention

of tokens in neighboring windows. To account for the presence of carrier tokens, we shift image tokens and carrier tokens by $\lfloor \frac{M}{2} \rfloor$ and $\lfloor \frac{N_{cat}}{2} \rfloor$ voxels, respectively, conserving the alignment of the token spaces. Attention masking is applied to drop interactions between non-adjacent image/carrier tokens.

We compute attended carrier token embeddings $\mathbf{x}_{\text{cat}}^{L,t}$ at network level L and transformer layer t as:

$$\begin{aligned} \hat{\mathbf{x}}_{\text{cat}}^{L,t} &= \mathbf{x}_{\text{cat}}^{L,t-1} + \gamma_1 \, \text{MSA}\left(\text{LN}\left(\mathbf{x}_{\text{cat}}^{L,t-1}\right)\right), \\ \mathbf{x}_{\text{cat}}^{L,t} &= \hat{\mathbf{x}}_{\text{cat}}^{L,t} + \gamma_2 \, \text{MLP}\left(\text{LN}\left(\hat{\mathbf{x}}_{\text{cat}}^{L,t}\right)\right), \end{aligned} \tag{1}$$

where γ_1, γ_2 are learnable channel-wise scaling factors, MSA denotes multi-headed self-attention [58], LN is Layer Normalization [61], and MLP is the multi-layer perceptron.

Next, we compute the attention of image tokens and carrier tokens using windowed self-attention, see eq. (2). Carrier tokens are window partitioned and concatenated with their corresponding set of local window tokens to produce sequences of $M^3 + {N_{\text{cat}}}^3$ tokens for each window. Window-attended tokens $\mathbf{x}_{\mathbf{w}}^{L,t+1}$ are computed using post-normed shifted window self-attention (SW-MSA) [53] as:

$$\begin{aligned} \mathbf{x}_{\mathbf{w}}^{L,t} &= [\mathbf{x}^{L,t-1},\ \mathbf{x}_{\text{cat}}^{L,t}] \\ \hat{\mathbf{x}}_{\mathbf{w}}^{L,t+1} &= \mathbf{x}_{\mathbf{w}}^{L,t} + \text{LN}\left(\text{SW-MSA}\left(\mathbf{x}_{\mathbf{w}}^{L,t}\right)\right) \\ \mathbf{x}_{\mathbf{w}}^{L,t+1} &= \hat{\mathbf{x}}_{\mathbf{w}}^{L,t+1} + \text{LN}\left(\text{MLP}\left(\hat{\mathbf{x}}_{\mathbf{w}}^{L,t+1}\right)\right) \end{aligned} \tag{2}$$

The carrier tokens and image tokens are then separated for compatibility with later SVHAT layers.

Prior to the attention mechanisms described in eq. (1) and eq. (2), SVHAT uses multi-head cross-attention (MCA) layers to facilitate information exchange across network stages L_1, L_2, L_3. Each cross-attention layer implements a two-layer MLP to ensure dimension compatibility between cross-scale token sequences. Then, MCA is applied to capture relationships between tokens from current and previous network stages. Exploiting the compactness of the carrier token space, we compute cross-attended carrier tokens $\mathbf{x}_{\text{cross, cat}}^{L}$ using full MCA:

$$\mathbf{x}_{\text{cross, cat}}^{L} = \text{LN}\left(\text{MCA}\left(\mathbf{x}_{\text{cat}}^{L,t-1}, \text{MLP}\left(\mathbf{x}_{\text{cat}}^{L-1}\right)\right)\right), \tag{3}$$

where $\mathbf{x}_{\text{cat}}^{L-1}$ denotes the final set of carrier tokens from the previous network stage. A similar window-based multi-head cross-attention (W-MCA) mechanism is used for capturing relationships between image tokens, see equation 4. The cross-attended image tokens $\mathbf{x}_{\text{cross}}^{L}$ are computed as:

$$\mathbf{x}_{\text{cross}}^{L} = \text{LN}\left(\text{W-MCA}\left(\mathbf{x}^{L,t-1}, \text{MLP}\left(\mathbf{x}^{L-1}\right)\right)\right), \tag{4}$$

where $\mathbf{x}^{L-1}$ denote the final set of image tokens from the previous network stage. Finally, the cross-attended token embeddings are fused by addition:

$$\begin{aligned} \mathbf{x}_{\text{cat}}^{L,t-1} &= \bar{\mathbf{x}}_{\text{cat}}^{L,t-1} + \mathbf{x}_{\text{cross, cat}}^{L} \\ \mathbf{x}^{L,t-1} &= \bar{\mathbf{x}}^{L,t-1} + \mathbf{x}_{\text{cross}}^{L} \end{aligned} \tag{5}$$

Here, $\bar{\mathbf{x}}^{L,t-1}$ and $\bar{\mathbf{x}}_{\text{cat}}^{L,t-1}$ denote image- and carrier tokens before fusion. For more details on the functionality of SVHAT, see appendix A.

4 Experiments

Datasets. We use four public MRI datasets and one CT-based dataset to train/evaluate our proposed MTVNet: The Human Connectome Project (HCP) 1200 Subjects dataset [62], the IXI dataset[1], the Brain Tumor Segmentation Challenge (BraTS) 2023 [63–66] and Kirby 21 [67]. All scans were acquired using 1.5T-3T MRI platforms with a volume size of $\leq 320^3$ voxels. Finally, we use the Femur Archaeological CT Superresolution (FACTS) dataset [68], which includes 12 registered 3D volume pairs of archaeological femur bones scanned using clinical-CT and micro-CT. The FACTS dataset consists of large volumes ($\sim 2000^3$ voxels) featuring detailed trabecular bone structures. Two SR tasks are considered using this dataset: In FACTS-Synth, we use downsampled micro-CT images as the LR model input, while FACTS-Real instead uses the clinical-CT images. Refer to appendix B for additional details.

Models. We evaluate the SR performance of MTVNet against 2D models RCAN [26] and HAT [2, 37], as well as six volumetric models: mDCSRN [15], EDDSR [44], MFER [46], RRDBNet3D [27], SuperFormer [16], and ArSSR [45]. We adapt mDCSRN and SuperFormer, originally designed to restore images degraded by 3D k-space truncation [13, 16], by extending them with the upsampling module from MTVNet. We use the authors' suggested upsampling for the remaining models.

Training. We train all models from scratch on each dataset for 100K iterations on a single A100 80GB GPU. For ArSSR, we collate $N = 8000$ randomly sampled HR/LR point pairs from 15 patches per batch. The rest use batch size 4 for MRI or 5 for CT data. The LR patch size is set to $32 \times 32 \times 32$ or 32×32 in case of 2D. MTVNet L_2 and L_3 with two and three stages use patch sizes $64 \times 64 \times 64$ and $128 \times 128 \times 128$, respectively. All models are optimized using ADAM [69] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use a multi-step learning rate scheduler, halving the learning rate once after 50k, 70k, 85k, and 95k iterations. All model parameters are optimized using pure L1 loss. HR/LR pairs are generated using volumetric blurring followed by downsampling via linear interpolation. In FACTS-Real, we use downsampled clinical-CT images as the LR input and the micro-CT images as the HR reference.

Evaluation. We reconstruct all test samples from each dataset using strided aggregation of SR patch predictions. Patch predictions are tiled using an overlap of $4 \times s$ voxels where s is the upscaling fac-

[1] https://brain-development.org/ixi-dataset/

	FACTS-Synth Dataset						FACTS-Real Dataset					
	Scale 4×			Scale 3×			Scale 4×			Scale 3×		
2D methods	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE
RCAN [26]	27.88	.8940	.1952	31.03	.9336	.1359	20.51	.3554	.5391	20.72	.3870	.5548
† HAT [37]	28.05	.8951	.1924	31.15	.9334	.1355	20.54	.3686	.5343	20.63	.4242	.5614
3D methods	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE
ArSSR [45]	28.83	.8998	.1779	30.78	.9284	.1459	20.88	.3871	.4881	20.68	.3980	.5767
EDDSR [44]	29.86	.9109	.1620	33.22	.9451	.1104	20.62	.3531	.4815	19.84	.3499	.5223
MFER [46]	29.48	.9094	.1646	32.50	.9420	.1179	21.58	.4708	.4080	21.64	.4671	.4096
mDCSRN [15]	29.77	.9099	.1624	33.23	.9460	.1090	21.31	.4078	.4765	21.37	.4259	.4922
† SuperFormer [16]	30.46	.9175	.1481	33.47	.9480	.1055	20.93	.3491	.4846	21.40	.4038	.4463
RRDBNet3D [27]	29.78	.9120	.1584	33.21	.9442	.1093	21.64	.4670	.4022	21.91	.4775	.4019
† MTVNet	31.57	.9303	.1313	33.91	.9502	.1020	21.52	.4576	.4061	21.74	.4633	.4051

	HCP 1200 Dataset			IXI Dataset			BraTS 2023 Dataset			Kirby 21 Dataset		
	Scale 4×			Scale 4×			Scale 4×			Scale 4×		
2D methods	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE
RCAN [26]	32.61	.8812	.1593	28.52	.8367	.1768	33.47	.9306	.1505	33.06	.8978	.2256
† HAT [37]	32.39	.8770	.1640	28.42	.8331	.1791	33.20	.9266	.1551	31.18	.8561	.2940
3D methods	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE	PSNR	SSIM	NRMSE
ArSSR [45]	27.90	.8118	.2810	24.22	.7204	.3060	22.96	.3182	.4437	31.82	.8632	.2775
EDDSR [44]	30.12	.8335	.2174	25.22	.7394	.2597	32.66	.9169	.1686	33.51	.8946	.2244
MFER [46]	33.40	.8933	.1484	25.23	.7611	.2576	34.76	.9430	.1309	35.68	.9307	.1719
mDCSRN [15]	33.46	.8941	.1470	29.50	.8558	.1622	34.76	.9431	.1308	35.26	.9255	.1806
† SuperFormer [16]	33.70	.8982	.1430	29.89	.8679	.1545	34.60	.9400	.1333	35.85	.9341	.1675
RRDBNet3D [27]	34.31	.9092	.1331	30.27	.8793	.1488	35.20	.9486	.1242	36.27	.9376	.1598
† MTVNet	34.04	.9046	.1374	30.16	.8754	.1502	35.16	.9477	.1250	35.97	.9355	.1654

Table 1. Quantitative comparison of state-of-the-art 2D/volumetric SR models on datasets FACTS-Synth, FACTS-Real, HCP 1200, IXI, BraTS 2023, and Kirby 21. The best performance metrics PSNR ↑ / SSIM ↑ / NRMSE ↓ are highligthed in red, and second best in blue. Transformer-based methods are marked with a † symbol.

tor, then smoothed using a Hanning window. Performance metrics Peak-Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Normalized Root Mean Square Error (NRMSE) are computed slice-wise in the axial direction and averaged over all samples, excluding slices where the foreground area occupies less than 20%.

4.1 Implementation details

All MTVNet configurations use a learning rate of 2e−4 without weight decay. For MRI data, we use MTVNet L_2, as we found two stages to be enough to cover most whole scans. In the high-resolution FACTS dataset, we use MTVNet L_3 with three network stages. Each DCHAT block has 6 SVHAT layers. The number of shallow features C_{SFE} and embedding features C_{emb} are set to 128, with skip-connection features $C_{\text{skip}} = 64$. In MTVNet L_3, patch sizes are $p_1 = 2$, $p_2 = 4$, $p_3 = 8$; in L_2, $p_1 = 2$, $p_2 = 4$. The attention window size is $M = 8$, with $N_{cat} = 4$ carrier tokens. To reduce memory, we halve feature channels in MTVNet, mDCSRN, SuperFormer, and RRDBNet3D before upsampling.

4.2 Quantitative results

Table 1 compares MTVNet with eight SOTA SR models: RCAN, HAT, ArSSR, EDDSR, MFER, mDCSRN, SuperFormer, and RRDBNet3D. Across all brain MRI datasets (HCP 1200, IXI, BraTS 2023, and Kirby 21), MTVNet achieves competitive results. We observe the CNN-based RRDBNet3D slightly outperforming the ViT-based MTVNet and SuperFormer on brain MRI, while in 2D, RCAN similarly surpasses the newer ViT-based HAT. This trend suggests that for low-resolution data, CNNs outperform ViTs, contradicting earlier findings [16]. We reason that the advantage of CNNs in these datasets stems from a combination of low image resolution and local image dependencies being predominant, limiting the benefits of the broader receptive field offered by ViTs.

In the high-resolution FACTS dataset, where we can leverage the multi-contextual architecture of our proposed method, we observe several new trends: in FACTS-Synth, ViT-based methods surpass CNN-based architectures in both 2D and 3D, with MTVNet outperforming all methods by a large margin. Compared with SuperFormer, MTVNet improves PSNR by 0.44dB–1.11dB, and by 0.70dB–1.79dB over RRDBNet3D, illustrating that added contextual information yields significant gains in high-resolution volumetric SR. In FACTS-Real, where clinical CT images serve as LR input, the best results are achieved by RRDBNet3D, MFER and MTVNet, despite the similarity to FACTS-Synth. We hypothesize this stems from the domain shift between micro-CT and clinical-CT, which weakens

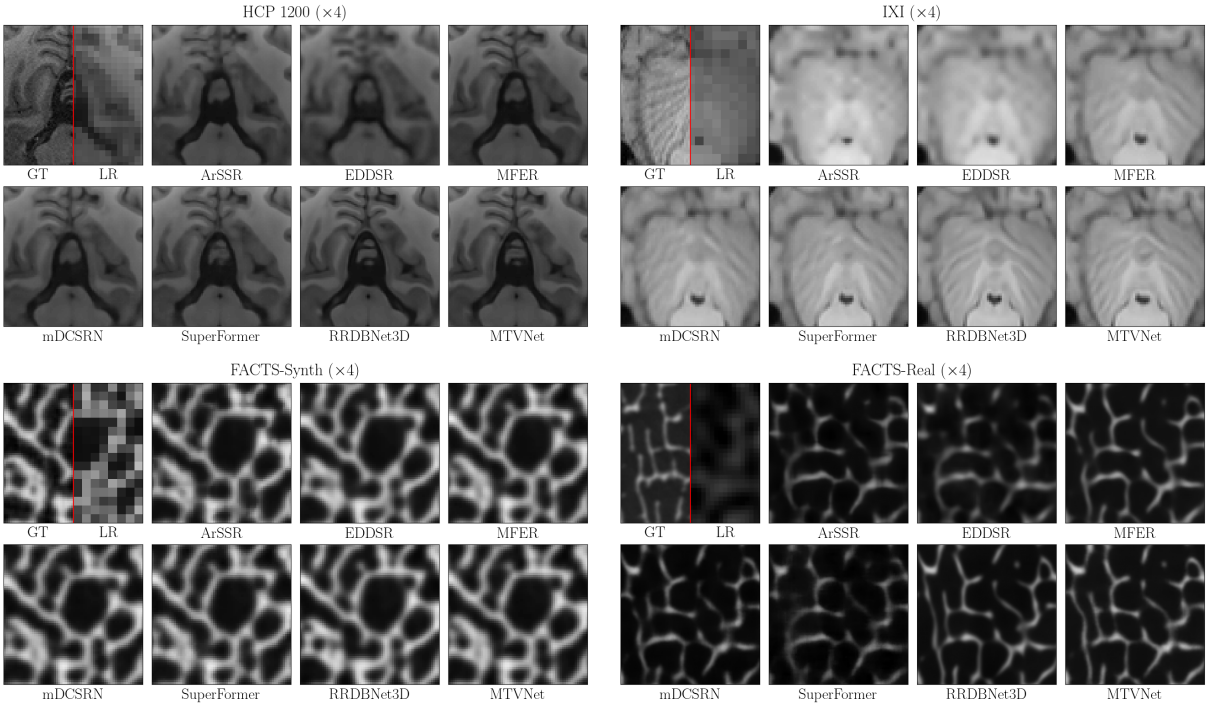

Figure 4. Visual comparisons of SR outputs on HCP 1200, IXI, FACTS-Synth, and FACTS-Real at 4× upscaling. Ground truth (GT) and LR inputs are shown in the top-left, separated by a red line.

long-range dependencies that would otherwise benefit ViTs. Overall, these results indicate that CNN vs. ViT performance depends strongly on both input resolution and on the underlying 3D image structure.

4.3 Qualitative results and 3D LAM

Fig. 4 shows a visual comparison of SR predictions on scale 4× for HCP 1200, IXI, FACTS-Synth, and FACTS-Real. We find that MTVNet produces faithful reconstructions of structures and patterns across all datasets. Compared with ArSSR, EDDSR, MFER, mDCSRN, and SuperFormer, our MTVNet produces notably sharper features while producing similar results as RRDBNet3D. In the Brain MRI datasets HCP 1200 and IXI, we find that many methods struggle to reconstruct anatomical details while RRDBNet3D and our MTVNet produce the clearest results. Refer to appendix E for more comparisons.

Next, we investigate how volumetric SR models leverage surrounding image context using LAM [70]. We extend LAM into 3D to visualize context usage in volumetric SR predictions. Fig. 5 shows log-scaled LAM activations on FACTS-Synth at ×4 upscaling, where higher intensities indicate stronger voxel contributions towards the region marked by the red box. The blue box highlights the SR prediction area, which is constant across methods. Although no predictions are computed outside this area, MTVNet can use information from these regions via its contextual stages. To quantify context usage, we compute the Diffusion Index (DI) [70]. We report the mean DI across slices for volumetric SR methods and per-slice DI for 2D methods. Examples from FACTS-Synth show that MTVNet, with contextual stages, enables broader leverage of input context than competing methods. Additional LAM comparisons are provided in appendix C. To study the importance of context across datasets, we compute average DI scores over 50 random 3D patch samples from HCP 1200 and FACTS-Synth, see fig. 6. DI scores are generally lower in HCP 1200, especially for stronger models, suggesting that context is less critical in brain MRI. Conversely, SR models achieve higher DI in FACTS-Synth, with MTVNet achieving the highest average DI among all methods.

4.4 Ablation experiments

We perform ablation of the features of MTVNet, including carrier tokens and contextual network stages across MRI and CT data. Table 2 shows a quantitative comparison on BraTS 2023 and FACTS-Synth using ×4 upscaling. Using BraTS 2023, replacing the baseline Swin transformer layers [6] with our SVHAT layers using carrier tokens results in modest performance gains across all metrics. Using FACTS-Synth, increasing the number of context levels of MTVNet results in significant performance improvements across all metrics. Compared with MTVNet L_1, adding an extra level of context increases PSNR by 0.44dB. Similarly, using three contextual network stages further improves PSNR by 1.1dB.

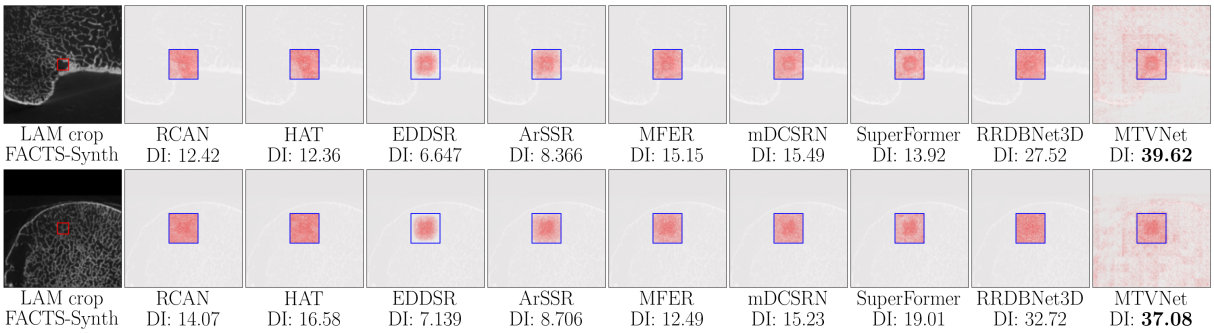

Figure 5. LAM comparisons of SR models using FACTS-Synth at ×4 upscaling. The blue box marks the prediction area for SR, which is the same for all methods. The highest DI ↑ is highlighted in **bold**.

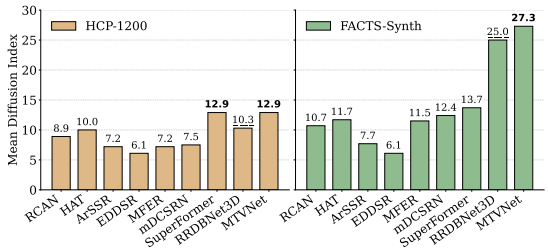

Figure 6. DI score averages across 50 random 3D samples from HCP 1200 (×4) and FACTS-Synth (×4).

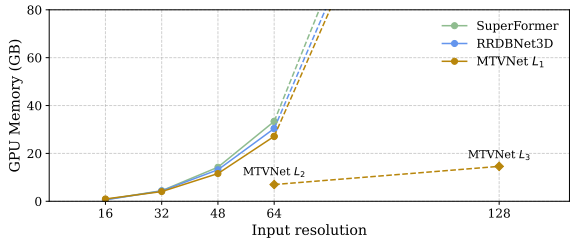

Figure 7. GPU memory usage across 3D patch resolutions. Contextual stages in MTVNet enable resolutions of 128^3 and higher without exceeding VRAM limits.

Method	#Params	CATs	PSNR/SSIM/NRMSE
Baseline	19.7M	✗	35.00 / .9460 / .1273
MTVNet L_1	34.6M	✓	35.05 / .9467 / .1265
Method	**#Params**	**Levels**	**PSNR/SSIM/NRMSE**
MTVNet L_1	34.6M	1	30.03 / .9168 / .1550
MTVNet L_2	109.0M	2	30.47 / .9211 / .1452
MTVNet L_3	138.2M	3	31.57 / .9303 / .1313

Table 2. Ablation on the effect of carrier tokens and context levels. Best metrics are underlined.

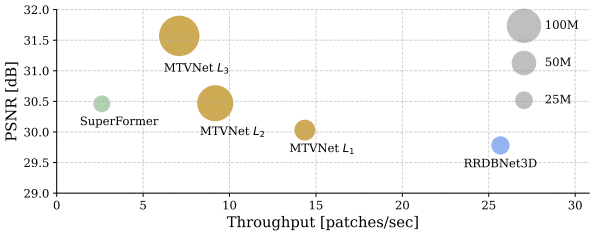

Figure 8. PSNR vs. Throughput using FACTS-Synth (4×). Throughput is measured on a NVIDIA H100 GPU using a batch size of 1.

4.5 Memory footprint of MTVNet

Fig. 7 shows the memory footprint of SuperFormer, RRDBNet3D, and MTVNet across volumetric input resolutions. Memory footprint is measured as the peak GPU memory usage for one forward and backward pass using a batch size of 1. With one network stage, MTVNet L_1 exhibits better memory scaling than SuperFormer and RRDBNet3D. Provided the prediction area is fixed to 32^3, adding contextual network stages allows processing of input sizes far exceeding the capabilities of other architectures.

Fig. 8 shows PSNR vs. throughput of SuperFormer, RRDBNet3D, and MTVNet on FACTS-Synth using 4× upscaling. MTVNet achieves SOTA performance while maintaining a higher throughput than SuperFormer, despite having more parameters.

5 Conclusion

In this work, we present MTVNet, a ViT-based method for volumetric SR tailored for high-resolution 3D data. Our method overcomes the challenge of limited contextual information through a multi-contextual network structure with a coarse-to-fine feature extraction and tokenization scheme, enabling processing of larger input sizes than competing methods. We model long-range dependencies by combining global and window-based attention to exchange information in a larger input volume.

We compare MTVNet against 2D and volumetric SR approaches across several data domains, including brain MRI data and high-resolution CT data. Based on extensive experiments, we find that CNN-based models outperform ViT-based models in certain 3D data domains. CNN-based SR models are especially effective in low-resolution 3D volumes where the receptive field of transformers cannot be leveraged as effectively. Nevertheless, our proposed MTVNet with extra contextual processing layers outperforms all other models in high-resolution 3D data with long-range image dependencies.

Acknowledgments

Research reported in this publication is supported by the Infrastructure for Quantitative AI-based Tomography (QUAITOM) supported by the Novo Nordisk Foundation (Grant number NNF21OC0069766) and the Multiscale label-free 3D x-ray imaging: Visualizing cells and tissue architecture simultaneously (Xtreme-CT) supported by the Novo Nordisk Foundation (Grant number grant NNF22OC0077698).

References

[1] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. "SwinIR: Image Restoration Using Swin Transformer". In: *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2021, pp. 1833–1844. DOI: 10.1109/ICCVW54120.2021.00210.

[2] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong. "Activating More Pixels in Image Super-Resolution Transformer". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 22367–22377. DOI: 10.1109/CVPR52729.2023.02142.

[3] S.-C. Chu, Z.-C. Dou, J.-S. Pan, S. Weng, and J. Li. "HMANet: Hybrid Multi-Axis Aggregation Network for Image Super-Resolution". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2024, pp. 6257–6266. DOI: 10.1109/CVPRW63382.2024.00629.

[4] C.-C. Hsu, C.-M. Lee, and Y.-S. Chou. "DRCT: Saving Image Super-Resolution away from Information Bottleneck". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2024, pp. 6133–6142. DOI: 10.1109/CVPRW63382.2024.00618.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. " Swin Transformer: Hierarchical Vision Transformer using Shifted Windows ". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2021, pp. 9992–10002. DOI: 10.1109/ICCV48922.2021.00986. URL: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986.

[7] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. de Marvao, S. Cook, D. O'Regan, and D. Rueckert. "Multi-input cardiac image super-resolution using convolutional neural networks". In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19*. Springer. 2016, pp. 246–254. DOI: https://doi.org/10.1007/978-3-319-46726-9_29.

[8] S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu, F. Liang, D. Feng, and D. Liang. "Accelerating magnetic resonance imaging via deep learning". In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. 2016, pp. 514–517. DOI: 10.1109/ISBI.2016.7493320.

[9] X. Jiang, Y. Xu, P. Wei, and Z. Zhou. "CT Image Super Resolution Based On Improved SRGAN". In: *2020 5th International Conference on Computer and Communication Systems (ICCCS)*. 2020, pp. 363–367. DOI: 10.1109/ICCCS49078.2020.9118497.

[10] H. Xia, N. Cai, H. Wang, Y. Mao, H. Wang, J. Li, and P. Wang. "Brain MR image super-resolution via a deep convolutional neural network with multi-unit upsampling learning". In: *Signal, Image and Video Processing* 15 (2021), pp. 931–939. DOI: https://doi.org/10.1007/s11760-020-01817-x.

[11] L. Song, Q. Wang, T. Liu, H. Li, J. Fan, J. Yang, and B. Hu. "Deep robust residual network for super-resolution of 2D fetal brain MRI". In: *Scientific reports* 12.1 (2022), p. 406. DOI: https://doi.org/10.1038/s41598-021-03979-1.

[12] Y. Chen, F. Shi, A. G. Christodoulou, Y. Xie, Z. Zhou, and D. Li. "Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2018, pp. 91–99. DOI: https://doi.org/10.1007/978-3-030-00928-1_11.

[13] Y. Chen, Y. Xie, Z. Zhou, F. Shi, A. Christodoulou, and D. Li. "Brain MRI super resolution using 3D deep densely connected neural networks". In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 739–742. DOI: 10.1109/ISBI.2018.8363679.

[14] C.-H. Pham, C. Tor Díez, H. Meunier, N. Bednarek, R. Fablet, N. Passat, and F. Rousseau. "Multiscale brain MRI super-resolution using deep 3D convolutional networks". In: *Computerized Medical Imaging and Graphics* 77 (Aug. 2019). DOI: 10.1016/j.compmedimag.2019.101647.

[15] Y. Chen, A. G. Christodoulou, Z. Zhou, F. Shi, Y. Xie, and D. Li. "MRI Super-Resolution with GAN and 3D Multi-Level DenseNet: Smaller, Faster, and Better". In: *arXiv preprint arXiv:2003.01217* (2020). DOI: 10.48550/arXiv.2003.01217.

[16] C. Forigua, M. Escobar, and P. Arbelaez. "SuperFormer: Volumetric Transformer Architectures for MRI Super-Resolution". In: *Simulation and Synthesis in Medical Imaging*. Ed. by C. Zhao, D. Svoboda, J. M. Wolterink, and M. Escobar. Cham: Springer International Publishing, 2022, pp. 132–141. ISBN: 978-3-031-16980-9. DOI: https://doi.org/10.1007/978-3-031-16980-9_13.

[17] Z. Ji, B. Zou, X. Kui, J. Liu, W. Zhao, C. Zhu, P. Dai, and Y. Dai. "Deep learning-based magnetic resonance image super-resolution: a survey". In: *Neural Computing and Applications* (2024), pp. 1–28. DOI: https://doi.org/10.1007/s00521-024-09890-w.

[18] E. Wehrse, L. Klein, L. T. Rotkopf, W. Stiller, M. Finke, G. G. Echner, C. Glowa, S. Heinze, C. H. Ziener, H. P. Schlemmer, et al. "Ultra-high resolution whole body photon counting computed tomography as a novel versatile tool for translational research from mouse to man". In: *Zeitschrift für Medizinische Physik* 33.2 (2023), pp. 155–167. DOI: https://doi.org/10.1016/j.zemedi.2022.06.002.

[19] C. Walsh, P. Tafforeau, W. Wagner, D. Jafree, A. Bellier, C. Werlein, M. Kühnel, E. Boller, S. Walker-Samuel, J. Robertus, et al. "Imaging intact human organs with local resolution of cellular structures using hierarchical phase-contrast tomography". In: *Nature methods* 18.12 (2021), pp. 1532–1541. DOI: 10.1038/s41592-021-01317-x.

[20] C.-F. Chen, Q. Fan, and R. Panda. "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 357–366. DOI: 10.1109/ICCV48922.2021.00041.

[21] A. Hatamizadeh, G. Heinrich, H. Yin, A. Tao, J. M. Alvarez, J. Kautz, and P. Molchanov. "FasterViT: Fast Vision Transformers with Hierarchical Attention". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=kB4yBiNmXX.

[22] C. Dong, C. C. Loy, K. He, and X. Tang. "Image Super-Resolution Using Deep Convolutional Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.2 (2016), pp. 295–307. DOI: 10.1109/TPAMI.2015.2439281.

[23] C. Dong, C. C. Loy, and X. Tang. "Accelerating the Super-Resolution Convolutional Neural Network". In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe, J. Matas, N. Sebe, and M. Welling. Springer, 2016, pp. 391–407. ISBN: 978-3-319-46475-6. DOI: https://doi.org/10.1007/978-3-319-46475-6_25.

[24] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. " Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network ". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, July 2017, pp. 105–114. DOI: 10.1109/CVPR.2017.19. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.19.

[25] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. "Enhanced Deep Residual Networks for Single Image Super-Resolution". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017, pp. 1132–1140. DOI: 10.1109/CVPRW.2017.151.

[26] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. "Image Super-Resolution Using Very Deep Residual Channel Attention Networks". In: *Computer Vision – ECCV 2018*. Ed. by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss. Cham: Springer International Publishing, 2018, pp. 294–310. ISBN: 978-3-030-01234-2. DOI: https://doi.org/10.1007/978-3-030-01234-2_18.

[27] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks". In: *Computer Vision – ECCV 2018 Workshops*. Ed. by L. Leal-Taixé and S. Roth. Springer, 2019, pp. 63–79. ISBN: 978-3-030-11021-5. DOI: https://doi.org/10.1007/978-3-030-11021-5_5.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[29] Z. Gao, L. Wang, B. Han, and S. Guo. "AdaMixer: A Fast-Converging Query-Based Object Detector". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 5354–5363. DOI: 10.1109/CVPR52688.2022.00529.

[30] B. Roh, J. Shin, W. Shin, and S. Kim. "Sparse DETR: Efficient End-to-End Object Detection with Learnable Sparsity". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=RRGVCN8kjim.

[31] T. Shehzadi, K. A. Hashmi, D. Stricker, and M. Z. Afzal. "Object detection with transformers: A review". In: *arXiv preprint arXiv:2306.04670* (2023). DOI: 10.48550/arXiv.2306.04670.

[32] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li. "TransBTS: Multimodal Brain Tumor Segmentation Using Transformer". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert. Cham: Springer International Publishing, 2021, pp. 109–119. ISBN: 978-3-030-87193-2. DOI: https://doi.org/10.1007/978-3-030-87193-2_11.

[33] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation". In: *Proceedings of the European Conference on Computer Vision Workshops(ECCVW)*. 2022. DOI: https://doi.org/10.1007/978-3-031-25066-8_9.

[34] Y. Gao, M. Zhou, D. Liu, and D. Metaxas. "A Multi-scale Transformer for Medical Image Segmentation: Architectures, Model Efficiency, and Benchmarks". In: *arXiv preprint arXiv:2203.00131* (Feb. 2022). DOI: 10.48550/arXiv.2203.00131.

[35] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by A. Crimi and S. Bakas. Cham: Springer International Publishing, 2022, pp. 272–284. ISBN: 978-3-031-08999-2. DOI: https://doi.org/10.1007/978-3-031-08999-2_22.

[36] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, et al. "TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers". In: *Medical Image Analysis* 97 (2024), p. 103280. DOI: 10.1016/j.media.2024.103280.

[37] X. Chen, X. Wang, W. Zhang, X. Kong, Y. Qiao, J. Zhou, and C. Dong. "HAT: Hybrid Attention Transformer for Image Restoration". In: *arXiv preprint arXiv:2309.05239* (2023). DOI: 10.48550/arXiv.2309.05239.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

[39] T. Tong, G. Li, X. Liu, and Q. Gao. "Image Super-Resolution Using Dense Skip Connections". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 4809–4817. DOI: 10.1109/ICCV.2017.514.

[40] C.-H. Pham, A. Ducournau, R. Fablet, and F. Rousseau. "Brain MRI super-resolution using deep 3D convolutional networks". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. 2017, pp. 197–200. DOI: 10.1109/ISBI.2017.7950500.

[41] I. Sánchez and V. Vilaplana. "Brain MRI super-resolution using 3D generative adversarial networks". In: *arXiv preprint arXiv:1812.11440* (2018). DOI: 10.48550/arXiv.1812.11440.

[42] J. Du, L. Wang, Y. Liu, Z. Zhou, Z. He, and Y. Jia. "Brain MRI Super-Resolution Using 3D Dilated Convolutional Encoder–Decoder Network". In: *IEEE Access* 8 (2020), pp. 18938–18950. DOI: 10.1109/ACCESS.2020.2968395.

[43] W. Lu, Z. Song, and J. Chu. "A novel 3D medical image super-resolution method based on densely connected network". In: *Biomedical Signal Processing and Control* 62 (2020), p. 102120. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2020.102120. URL: https://www.sciencedirect.com/science/article/pii/S1746809420302706.

[44] L. Wang, J. Du, A. Gholipour, H. Zhu, Z. He, and Y. Jia. "3D dense convolutional neural network for fast and accurate single MR image super-resolution". In: *Computerized Medical Imaging and Graphics* 93 (2021), p. 101973. ISSN: 0895-6111. DOI: https://doi.org/10.1016/j.compmedimag.2021.101973. URL: https://www.sciencedirect.com/science/article/pii/S0895611121001221.

[45] Q. Wu, Y. Li, Y. Sun, Y. Zhou, H. Wei, J. Yu, and Y. Zhang. "An Arbitrary Scale Super-Resolution Approach for 3D MR Images via Implicit Neural Representation". In: *IEEE Journal of Biomedical and Health Informatics* 27.2 (2023), pp. 1004–1015. DOI: 10.1109/JBHI.2022.3223106.

[46] H. Li, Y. Jia, H. Zhu, B. Han, J. Du, and Y. Liu. "Multi-level feature extraction and reconstruction for 3D MRI image super-resolution". In: *Computers in Biology and Medicine* 171 (2024), p. 108151. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2024.108151. URL: https://www.sciencedirect.com/science/article/pii/S001048252400235X.

[47] R. Ge, G. Yang, C. Xu, Y. Chen, L. Luo, and S. Li. "Stereo-Correlation and Noise-Distribution Aware ResVoxGAN for Dense Slices Reconstruction and Noise Reduction in Thick Low-Dose CT". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan. Cham: Springer International Publishing, 2019, pp. 328–338. ISBN: 978-3-030-32226-7. DOI: https://doi.org/10.1007/978-3-030-32226-7_37.

[48] L. Wang, H. Zhu, Z. He, Y. Jia, and J. Du. "Adjacent slices feature transformer network for single anisotropic 3D brain MRI image super-resolution". In: *Biomedical Signal Processing and Control* 72 (2022), p. 103339. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2021.103339. URL: https://www.sciencedirect.com/science/article/pii/S1746809421009368.

[49] Z. Ji, B. Zou, X. Kui, J. Liu, W. Zhao, C. Zhu, P. Dai, and Y. Dai. "Deep learning-based magnetic resonance image super-resolution: a survey". In: *Neural Computing and Applications* 36.21 (July 2024), pp. 12725–12752. ISSN: 1433-3058. DOI: 10.1007/s00521-024-09890-w. URL: https://doi.org/10.1007/s00521-024-09890-w.

[50] J. Zhu, C. Tan, J. Yang, G. Yang, and P. Lio'. "Arbitrary Scale Super-Resolution for Medical Images". In: *International Journal of Neural Systems* 31.10 (2021). PMID: 34304719, p. 2150037. DOI: 10.1142/S0129065721500374. URL: https://doi.org/10.1142/S0129065721500374.

[51] G. Li, L. Zhao, J. Sun, Z. Lan, Z. Zhang, J. Chen, Z. Lin, H. Lin, and W. Xing. "Rethinking Multi-Contrast MRI Super-Resolution: Rectangle-Window Cross-Attention Transformer and Arbitrary-Scale Upsampling". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 21173–21183. DOI: 10.1109/ICCV51070.2023.01941.

[52] W. Zhang, L. Wang, W. Chen, Y. Jia, Z. He, and J. Du. "3d Cross-Scale Feature Transformer Network for Brain Mr Image Super-Resolution". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 1356–1360. DOI: 10.1109/ICASSP43922.2022.9746092.

[53] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo. "Swin Transformer V2: Scaling Up Capacity and Resolution". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 11999–12009. DOI: 10.1109/CVPR52688.2022.01170.

[54] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren. "Efficientformer: Vision transformers at mobilenet speed". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 12934–12949.

[55] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen. "Twins: Revisiting the design of spatial attention in vision transformers". In: *Advances in neural information processing systems* 34 (2021), pp. 9355–9366.

[56] C.-F. Chen, R. Panda, and Q. Fan. "RegionViT: Regional-to-Local Attention for Vision Transformers". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=T_V3uLix7V.

[57] L. Themyr, C. Rambour, N. Thome, T. Collins, and A. Hostettler. "Memory transformers for full context and high-resolution 3D Medical Segmentation". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2022, pp. 121–130.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[59] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Con-

volutional Neural Network". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1874–1883. DOI: 10.1109/CVPR.2016.207.

[60] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi. "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize". In: (July 2017). DOI: 10.48550/arXiv.1707.02937.

[61] J. Ba, J. Kiros, and G. Hinton. "Layer Normalization". In: *ArXiv e-prints* (July 2016). DOI: 10.48550/arXiv.1607.06450.

[62] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, and K. Ugurbil. "The WU-Minn Human Connectome Project: An overview". In: *NeuroImage* 80 (2013). Mapping the Connectome, pp. 62–79. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2013.05.041. URL: https://www.sciencedirect.com/science/article/pii/S1053811913005351.

[63] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024. DOI: 10.1109/TMI.2014.2377694.

[64] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos. "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific Data* 4 (Sept. 2017). DOI: 10.1038/sdata.2017.117.

[65] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos. "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection". In: *The cancer imaging archive* 286 (2017). DOI: 10.7937/K9/TCIA.2017.GJQ7R0EF.

[66] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. Kitamura, S. Pati, et al. "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv 2021". In: *arXiv preprint arXiv:2107.02314* (2021). DOI: 10.48550/arXiv.2107.02314.

[67] B. A. Landman, A. J. Huang, A. Gifford, D. S. Vikram, I. A. L. Lim, J. A. Farrell, J. A. Bogovic, J. Hua, M. Chen, S. Jarso, S. A. Smith, S. Joel, S. Mori, J. J. Pekar, P. B. Barker, J. L. Prince, and P. C. van Zijl. "Multi-parametric neuroimaging reproducibility: A 3-T resource study". In: *NeuroImage* 54.4 (2011), pp. 2854–2866. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2010.11.047. URL: https://www.sciencedirect.com/science/article/pii/S1053811910015259.

[68] S. W. Bardenfleth, V. A. Dahl, C. Villa, G. Kazakia, and A. B. Dahl. "Superresolution of Real-World Multiscale Bone CT Verified with Clinical Bone Measures". In: *Medical Image Understanding and Analysis*. Ed. by M. H. Yap, C. Kendrick, A. Behera, T. Cootes, and R. Zwiggelaar. Cham: Springer Nature Switzerland, 2024, pp. 160–173. ISBN: 978-3-031-66958-3. DOI: https://doi.org/10.1007/978-3-031-66958-3_12.

[69] D. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations* (Dec. 2014).

[70] J. Gu and C. Dong. "Interpreting Super-Resolution Networks with Local Attribution Maps". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 9195–9204. DOI: 10.1109/CVPR46437.2021.00908.

A Details of SVHAT layer

An overview of our proposed SVHAT layer featuring separate attention branches for carrier tokens and image tokens is illustrated in fig. A.1. The first branch (red) follows the attention procedure from FasterViT [21], whereas the second branch follows the procedure from SwinV2 [53] with post-normalization. We use multi-head cross attention (MCA) and window-based multi-head cross attention (W-MCA) to merge tokens from previous network stages before computing attention in each branch. Embedding dimensions from previous network stages are matched using a small MLP.

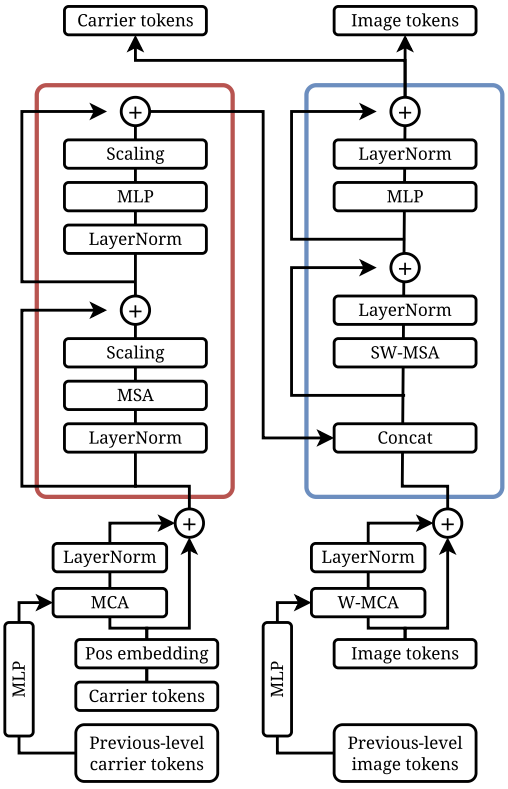

Figure A.1. Overview of our proposed SVHAT layer that captures global and local token dependencies using separate attention branches for carrier tokens (red) and image tokens (blue).

An illustration of the functionality of carrier tokens is provided in fig. A.2, showing both full CAT attention and local window attention with CATs.

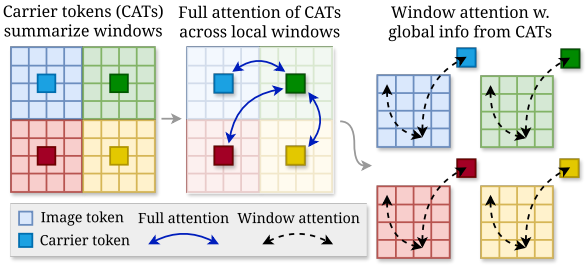

Figure A.2. Illustration of the carrier token (CAT) attention mechanism proposed by Hatamizadeh et al. [21] and adopted by SVHAT. Carrier tokens convey global information between local attention windows to enrich long-range information propagation.

B Dataset details

Human Connectome Project
The Human Connectome Project (HCP) 1200 Subjects Data Release [62] includes structural MRI scans from 1113 healthy subjects acquired using a 3T scanning platform. We use the T1-weighted images, featuring an isotropic resolution of 0.7 mm and a matrix size of $320 \times 320 \times 256$. Following [15, 16], the dataset is split into 780 subjects for training, and 111 each for validation, evaluation, and testing. Performance evaluation is conducted on the test set.

Information eXtraction from Images
The Information eXtraction from Images (IXI) dataset contains multi-modality MRI data (PD-, T1-, and T2-weighted) from 600 healthy subjects scanned with one 3T and two 1.5T scanning platforms. We use 581 T1-weighted scans, of which 507 have a resolution of $0.9375 \times 0.9375 \times 1.2$ mm and a matrix size of $256 \times 256 \times 150$, while the remaining 74 have a similar resolution but a matrix size of $256 \times 256 \times 146$. The dataset is split into 500 subjects for training, 6 for validation, and 75 for testing, with evaluation performed on the test set.

Brain Tumor Segmentation Challenge 2023
For the Brain Tumor Segmentation Challenge (BraTS) 2023, we use 1470 T1-weighted skull-stripped MRI scans of glioma patients, standardized to an isotropic resolution of 1 mm and a matrix size of $240 \times 240 \times 155$. We use the dataset split provided by the challenge, which allocates 1251 subjects for training and 219 for validation, with evaluation performed on the validation set.

Kirby 21
The Kirby 21 dataset includes multi-modality MRI scans from healthy individuals with no history of neurological conditions. We use the 42 T2-weighted images, which have a resolution of $1 \times 0.9375 \times 0.9375$ mm and a matrix size of $180 \times 256 \times 256$. The data is split into 37 images for training (KKI-06 to KKI-42) and 5 for testing (KKI-01 to KKI-05).

Femur Archaeological CT Superresolution
The Femur Archaeological CT Superresolution (FACTS) dataset comprises 12 archaeological proximal femurs scanned with clinical-CT and micro-CT platforms [68]. Clinical-CT scans have a resolution of $0.21 \times 0.21 \times 0.4$ mm, while micro-CT scans have a resolution of $58 \times 58 \times 58$ μm. Clinical-CT volumes are registered and linearly interpolated to match the micro-CT matrix size. The dataset is split into 10 images for training and 2 (f_002 and f_138) for testing.

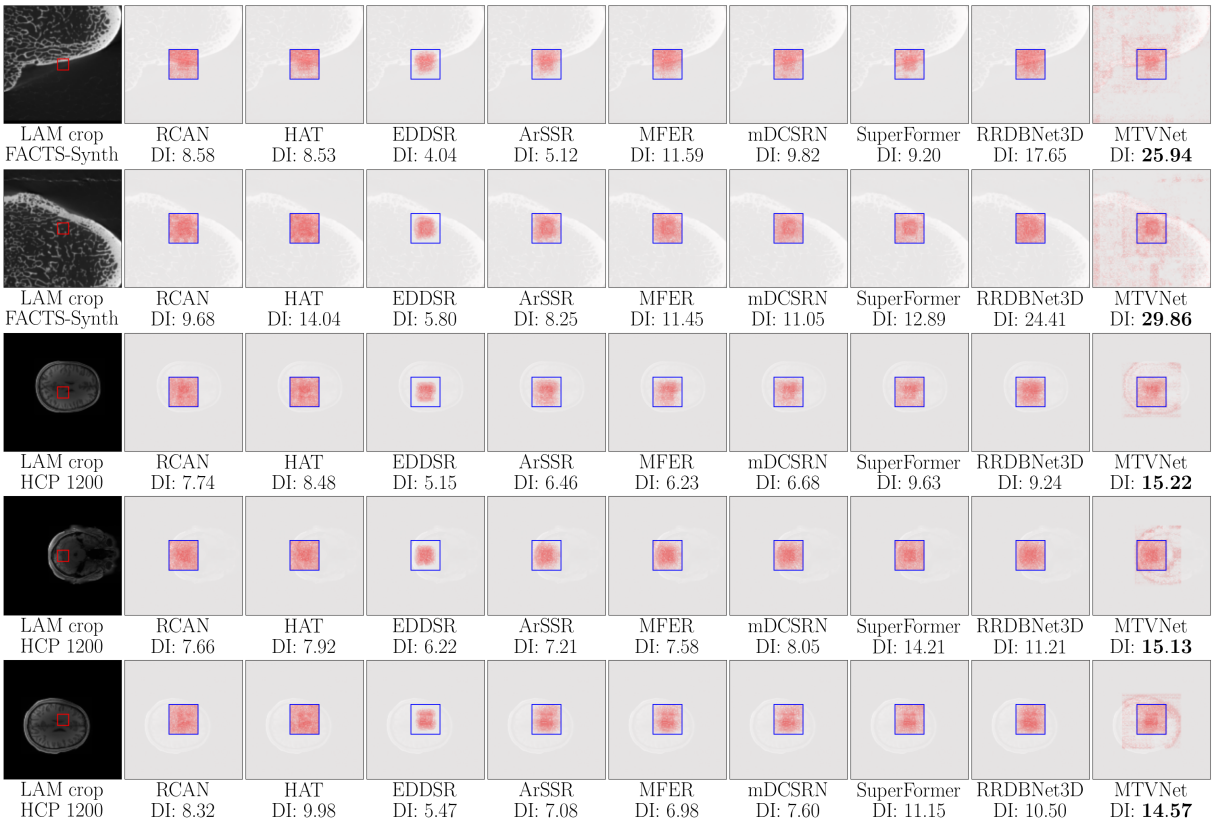

Figure B.1. LAM comparisons of SR models using FACTS-Synth and HCP 1200 at ×4 upscaling. The blue box shows the SR prediction area, which is kept constant across all methods. The highest DI ↑ is highlighted in **bold**.

C More LAM comparisons

Additional visual comparisons of the LAM method using FACTS-Synth and the HCP 1200 dataset at ×4 upscaling are shown in figure B.1. Similar to FACTS-Synth, we find that MTVNet incorporates information from the surrounding image context when trained using the HCP 1200 dataset. Despite this, we observe the CNN-based RRDBNet3D achieving better performance, suggesting that image context is less critical in brain MRI data.

D MTVNet hyperparameters

Tab. E.1 shows an overview of the GPU memory usage, throughput and parameter count of MTVNet using different hyperparameter configurations. Memory usage is measured as the maximum GPU memory required for a single forward and backward pass using a batch size of 1. The throughput in patches/sec is measured using a batch size of 1 assuming 4× upscaling. All configurations use MTVNet L_3 with three network stages. The number of blocks in table E.1 denotes the total number of DCHAT blocks used, with 6 blocks corresponding to a depth of (1, 2, 3) for network stages (L_1, L_2, L_3), 9 blocks corresponding to network depths (2, 3, 4), and 12 blocks corresponding to network depths (3, 4, 5).

Parameter	Memory usage	Throughput	#Params
C_{skip}			
64*	10.47 GB	7.79 patches/sec	138.3M
96	10.83 GB	7.70 patches/sec	226.6M
128	11.24 GB	7.50 patches/sec	340.3M
N_{cat}			
4*	10.47 GB	7.79 patches/sec	138.3M
2	10.31 GB	7.75 patches/sec	141.0M
1	10.37 GB	7.90 patches/sec	163.0M
#Blocks			
6*	10.47 GB	7.79 patches/sec	138.3M
9	10.98 GB	5.51 patches/sec	198.2M
12	11.49 GB	4.25 patches/sec	258.1M
C_{emb}			
128*	10.47 GB	7.79 patches/sec	138.3M
192	10.03 GB	7.47 patches/sec	194.5M
256	10.30 GB	7.34 patches/sec	261.3M

Table E.1. Overview of memory usage, throughput and no. of parameters using different hyperparameter configurations of MTVNet L_3. Baseline parameters of MTVNet L_3 are highlighted with an asterisk.

E More visual comparisons

Additional visual comparisons of SR predictions for HCP 1200, IXI, BraTS 2023, Kirby 21, FACTS-Synth, and FACTS-Real are shown in fig. E.1.

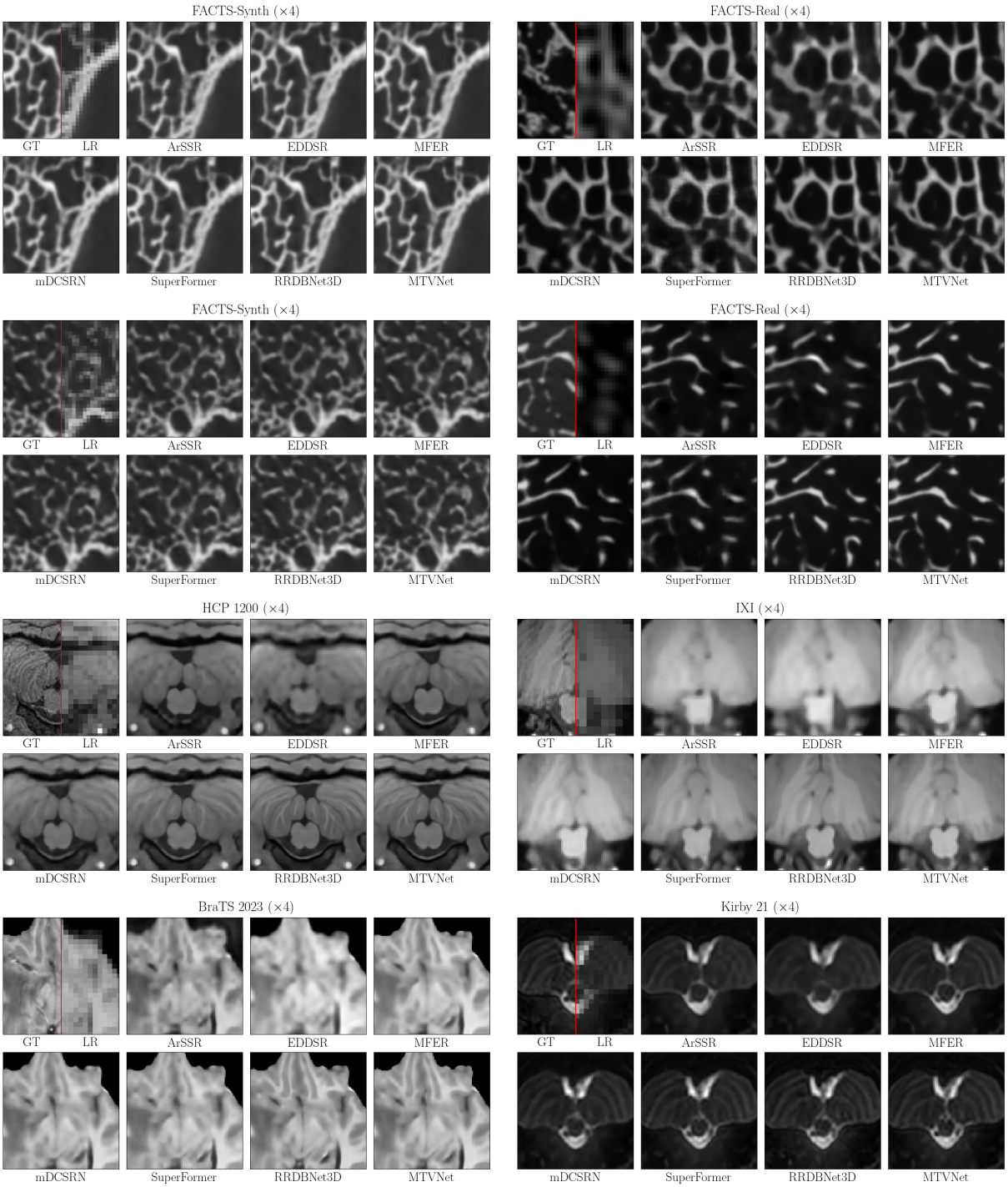

Figure E.1. Visual comparisons of SR outputs on HCP 1200, IXI, BraTS 2023, Kirby 21, FACTS-Synth, and FACTS-Real at 4× upscaling. Ground truth (GT) and LR inputs are shown in the top-left, separated by a red line.