VoDaSuRe: A Large-Scale Dataset Revealing Domain Shift in Volumetric Super-Resolution

Supplementary Material

1. VoDaSuRe dataset overview

A detailed overview of all 16 samples in the VoDaSuRe dataset is provided in Tab. 1, including volume shapes, slice splits (when applicable), voxel sizes, and scanning devices. The table lists only the physically acquired scans (HR/LR) and the registered LR volumes. Additional downsampled pyramid levels produced during OME-Zarr conversion are omitted for clarity. Note that the voxel sizes of the LR, and registered LR scans differ slightly, as the voxel size of the acquired LR scans did not exactly match the desired $4\times$ resolution difference compared with HR. This discrepancy is accounted for during the registration procedure.

Sample selection. To ensure a diverse set of structural characteristics, we intentionally include materials with varying degrees of microstructural complexity. We chose wood samples due to their well-organized tubular structures, as well as MDF and cardboard for their more chaotic arrangements of layers and fibers. We also chose to incorporate bone samples (femur, vertebrae, and animal bone) to have volumes with smoother structures typically seen in medical imaging datasets of clinical volumes. The finest microstructures in wood, MDF, and cardboard samples lie near the resolution limit of the LR scans but are clearly visible in the HR scans. This design choice ensures meaningful super-resolution scenarios where relevant structural details are partially lost in the LR input.

Stitching & reconstruction. All scans are reconstructed using the standard software provided with each scanner. Similarly, stitching of multiple vertical scans is performed using the native stitching tools of the respective devices.

2. VoDaSuRe preprocessing

Intensity matching. During the curation of VoDaSuRe, we observed notable differences in intensity distributions between all acquired LR and HR scans. For LR scans, the reduced cone-beam dispersion of the CT setup resulted in increased detector counts, improved signal-to-noise ratio, and higher contrast compared with HR acquisition. In some scans, the effect of region-of-interest scanning in high resolution (the scan region surrounded by material that is not accounted for in the reconstruction) resulted in small intensity differences between HR and LR scans in regions furthest from the rotational axis. To mitigate this, we applied intensity matching of registered LR slices to downsampled HR slices. Fig. 1 shows the effect of intensity matching using bamboo. The LR slice appears noticeably brighter with stronger contrast than the HR slice, which is also reflected in the intensity histograms of the two slices. After intensity matching, the intensity profile of the LR slice matches that of the HR slice but retains the same structural information.

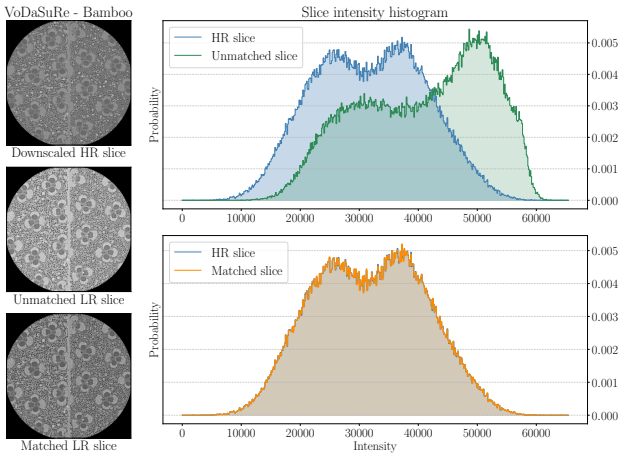

Figure 1. Visualization of the intensity matching procedure used in VoDaSuRe. The intensity distribution of registered LR slices is adjusted to match the distribution of downsampled HR slices.

We initially attempted to match the intensities of registered LR slices directly to HR slices, but found that this led to unrealistic intensity scaling. The HR slices contain a significantly larger proportion of high-intensity voxels due to their higher resolution, whereas these details are spatially averaged in the LR scans. Consequently, direct HR–LR matching causes the LR slices to become oversaturated. To avoid this, we first downsample the HR slices to the LR voxel size and then perform intensity matching. This downsampling suppresses high-frequency content while maintaining overall intensity statistics, resulting in more stable and physically meaningful intensity alignment.

Registration. Fig. 6 illustrates the accuracy of the HR-LR registration after intensity matching. Cropped regions from corresponding HR and registered LR volumes highlight the expected loss in microstructural detail. To assess spatial alignment, we create checkerboard visualizations and absolute difference images. The checkerboard images confirm the continuity of structures across the HR and registered LR volumes and demonstrates the effectiveness of our registration procedure. Similarly, the absolute difference images of the HR, and bicubic-interpolated LR slices validate the alignment, and also reveal the high-frequency information absent in the LR volumes.

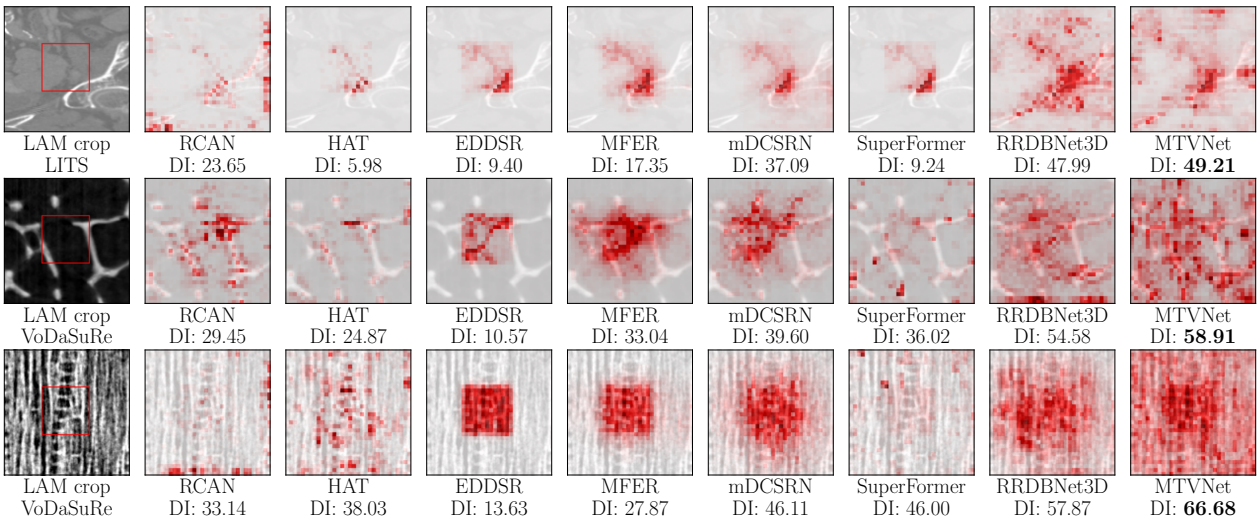

Figure 2. LAM comparisons of SR models. Top row: example from CTSpine1K, middle and bottom row: examples from VoDaSuRe. The highest DI ↑ is highlighted in **bold**.

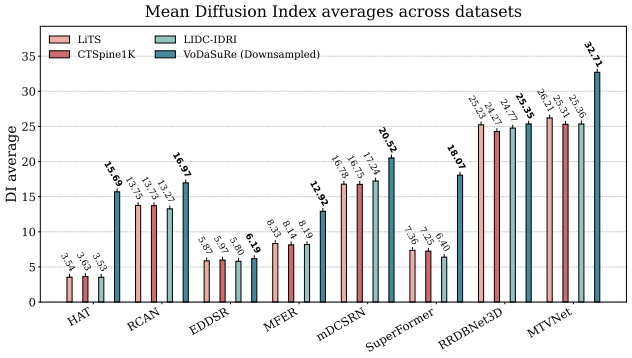

Figure 3. Diffusion index (DI) averages using datasets CTSpine1K, LiTS and LIDC-IDRI for all SR models. The highest DI ↑ scores for each dataset are highlighted in **bold**.

3. LAM analysis

To assess the degree of contextual dependency of SR predictions across datasets, we employ Local Attribution Mapping (LAM) [15]. Using LAM, we compare the spread of input voxel attributions for SR models trained on datasets with fine microstructures, e.g. VoDaSuRe, and models trained on medical data with smoother variations. Fig. 2 shows slice-averaged LAM results at scale $4\times$, where regions of higher intensities indicate stronger pixel/voxel contributions. We also report the slice-wise average Diffusion Index (DI) [15] as an estimate for overall context usage. Examples show that all models leverage broader involvement of input voxels in VoDaSuRe. To quantify this effect, we evaluate all SR models on 100 randomly sampled 3D patches from CTSpine1K, LiTS, LIDC-IDRI, and VoDaSuRe, and calculate the average DI of all models across all patches, see Fig. 3. We observe consistently higher DI across all methods, meaning SR models rely on broader spatial context in VoDaSuRe compared with CTSpine1K, LiTS and LIDC-IDRI. This suggests that long-range information is more important in VoDaSuRe than in medical imaging datasets, where models rely more on local image context. In particular, we observe ViT-based methods HAT, SuperFormer and MTVNet exhibiting noticeably greater increases in diffusion index using VoDaSuRe compared with CNN-based methods. Despite this, we did not find a correlation in performance, as the CNN-based RRDBNet3D was the overall strongest baseline in both medical datasets and VoDaSuRe.

4. OME-Zarr dataloader

Fig. 4 shows our data loading pipeline. We instantiate N worker processes that concurrently load volumetric patches from disk, with each worker using multiple threads that each maintain their own data queues to avoid contention. After loading and augmentation, patches are stored in the respective thread's data queue. During runtime, the main process collates batches of patches from all worker processes to maintain data throughput. This way, our pipeline scales to extremely large datasets, as full volumes are never held in system memory. Each OME-Zarr store in VoDaSuRe contains multiple resolution levels. By sampling patches from corresponding regions at different levels, we conveniently generate LR–HR pairs. The resolution gap between pyramid levels defines the SR scale, with each step yielding a $2\times$ difference. Our implementation is fully PyTorch-compatible and integrates seamlessly with training frameworks that use volumetric patch-based sampling for tasks such as segmentation, classification, and detection.

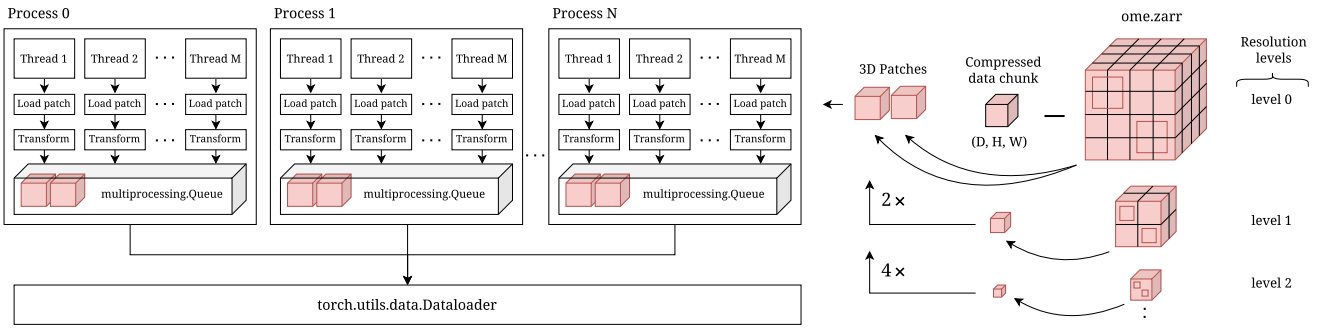

Figure 4. Illustration of the data loading pipeline for VoDaSuRe based on the OME-Zarr data format.

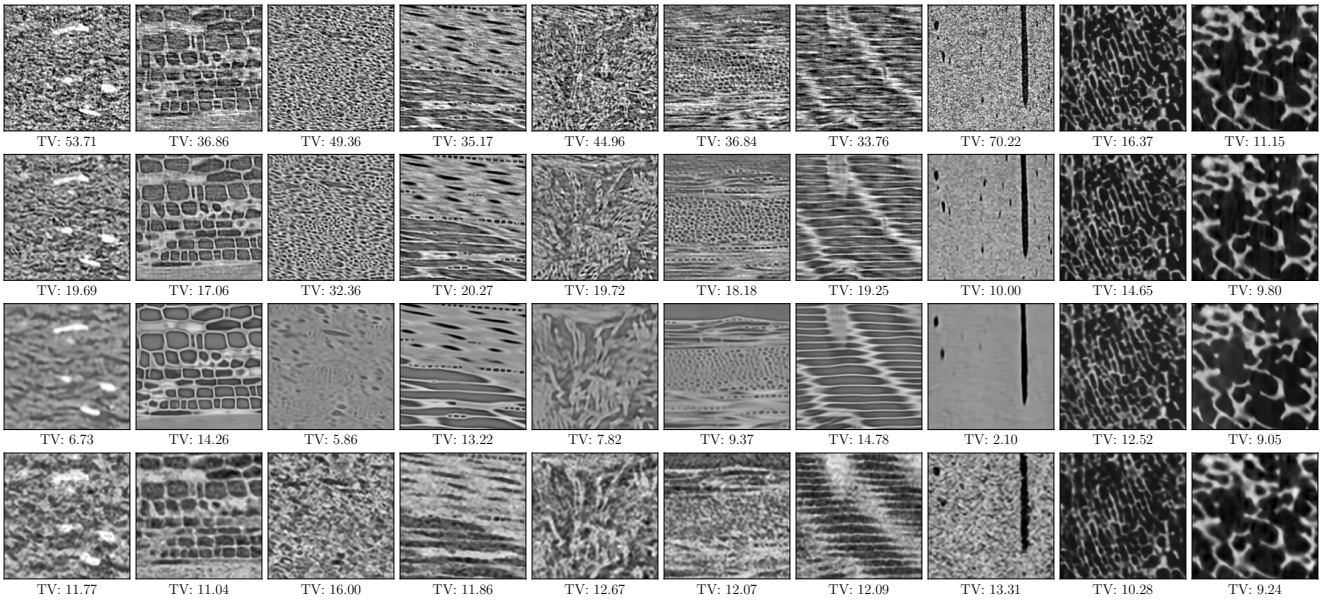

Figure 5. Visualizations from VoDaSuRe. From top to bottom: HR data, model predictions using downsampled LR data, model predictions using real LR data, and model predictions obtained by training on downsampled data but evaluating using real LR data input. All outputs are obtained at $4\times$ upscaling using RRDBNet3D. Total variation (TV) is shown for each slice.

5. Additional visualizations

Fig. 5 shows additional visualizations of SR model predictions using different training and evaluation data configurations from VoDaSuRe at scale $4\times$. Using downsampled LR data for training but real LR input data for evaluation results in distorted model predictions, highlighting the difference between the two data domains. Fig. 7 provides a showcase of orthogonal image slices from VoDaSuRe, including HR, registered LR and unregistered LR slices. Images are normalized for the purpose of visualization.

6. Training time

Tab. 2 summarizes the average training time of SR models at scale $4\times$. Training time is measured as the time to complete 100K training iterations averaged across all datasets.

7. Evaluation metrics and frequency analysis

We report PSNR, SSIM, NRMSE and LPIPS for quantitative evaluation and include total variation (TV) as an indicator of spatial smoothing. While TV captures reductions in local variation in model predictions, it does not distinguish between the removal of noise and the loss of meaningful high-frequency structure. Therefore, our interpretation of TV is done together with visual inspection, which clearly illustrates the characteristic smoothing effect observed when training on real LR data. To further analyze frequency characteristics, we additionally compute power spectrum visualizations and radial frequency profiles for three slice examples from VoDaSuRe, see Fig. 8. As spatial frequency increases, we find that SR predictions derived from scanned LR data exhibit faster decline in signal power compared with SR predictions derived from downsampled images.

Sample name	Scan	Volume shape (D×H×W)	Slice split (train/test)	Voxel size [μm]	Scanning device	Data size
Bamboo	High-resolution	$5440 \times 1920 \times 1920$	4960 / 480	$1.671 \times 1.671 \times 1.671$	Zeiss Versa 520	37.4 GB
	Low-resolution	$3520 \times 1920 \times 1920$	-	$6.637 \times 6.637 \times 6.637$		24.2 GB
	Registered	$1360 \times 480 \times 480$	1240 / 120	$6.684 \times 6.684 \times 6.684$		597.7 MB
Cardboard	High-resolution	$5120 \times 1920 \times 1920$	4640 / 480	$2.031 \times 2.031 \times 2.031$	Zeiss Versa 520	35.2 GB
	Low-resolution	$3360 \times 1920 \times 1920$	-	$8.017 \times 8.017 \times 8.017$		23.1 GB
	Registered	$1280 \times 480 \times 480$	1160 / 120	$8.124 \times 8.124 \times 8.124$		562.5 MB
Cypress	High-resolution	$5440 \times 1920 \times 1920$	4960 / 480	$1.671 \times 1.671 \times 1.671$	Zeiss Versa 520	37.4 GB
	Low-resolution	$1920 \times 1920 \times 1920$	-	$6.636 \times 6.636 \times 6.636$		13.2 GB
	Registered	$1360 \times 480 \times 480$	1240 / 120	$6.684 \times 6.684 \times 6.684$		597.7 MB
Elm	High-resolution	$5440 \times 1920 \times 1920$	4960 / 480	$1.671 \times 1.671 \times 1.671$	Zeiss Versa 520	37.4 GB
	Low-resolution	$3520 \times 1920 \times 1920$	-	$6.637 \times 6.637 \times 6.637$		24.2 GB
	Registered	$1360 \times 480 \times 480$	1240 / 120	$6.684 \times 6.684 \times 6.684$		597.7 MB
MDF	High-resolution	$3680 \times 1920 \times 1920$	3200 / 480	$1.671 \times 1.671 \times 1.671$	Zeiss Versa 520	25.3 GB
	Low-resolution	$3520 \times 1920 \times 1920$	-	$6.637 \times 6.637 \times 6.637$		24.2 GB
	Registered	$920 \times 480 \times 480$	800 / 120	$6.685 \times 6.685 \times 6.685$		404.3 MB
Ox bone	High-resolution	$4960 \times 1920 \times 1920$	4480 / 480	$1.199 \times 1.199 \times 1.199$	Zeiss Versa 520	34.1 GB
	Low-resolution	$1920 \times 1920 \times 1920$	-	$4.798 \times 4.798 \times 4.798$		13.2 GB
	Registered	$1240 \times 480 \times 480$	1120 / 120	$4.796 \times 4.796 \times 4.796$		544.9 MB
Oak	High-resolution	$5440 \times 1920 \times 1920$	4960 / 480	$1.671 \times 1.671 \times 1.671$	Zeiss Versa 520	37.4 GB
	Low-resolution	$3200 \times 1920 \times 1920$	-	$6.637 \times 6.637 \times 6.637$		22.0 GB
	Registered	$1360 \times 480 \times 480$	1240 / 120	$6.684 \times 6.684 \times 6.684$		597.7 MB
Larch	High-resolution	$5120 \times 1920 \times 1920$	4640 / 480	$1.669 \times 1.669 \times 1.669$	Zeiss Versa 520	35.2 GB
	Low-resolution	$3200 \times 1920 \times 1920$	-	$6.637 \times 6.637 \times 6.637$		22.0 GB
	Registered	$1280 \times 480 \times 480$	1160 / 120	$6.674 \times 6.674 \times 6.674$		562.5 MB
Femur 15	High-resolution	$1600 \times 1280 \times 1920$	Train	$58 \times 58 \times 58$	Nikon XT H 225	7.3 GB
	Low-resolution	$600 \times 600 \times 600$		$232 \times 232 \times 232$		412.0 MB
	Registered	$400 \times 320 \times 480$		$232 \times 232 \times 232$		117.2 MB
Femur 21	High-resolution	$1280 \times 1600 \times 1760$	Train	$58 \times 58 \times 58$	Nikon XT H 225	6.7 GB
	Low-resolution	$600 \times 600 \times 600$		$232 \times 232 \times 232$		412.0 MB
	Registered	$320 \times 400 \times 440$		$232 \times 232 \times 232$		107.4 MB
Femur 74	High-resolution	$1120 \times 1760 \times 1600$	Train	$58 \times 58 \times 58$	Nikon XT H 225	5.9 GB
	Low-resolution	$600 \times 600 \times 600$		$232 \times 232 \times 232$		412.0 MB
	Registered	$280 \times 440 \times 400$		$232 \times 232 \times 232$		94.0 MB
Femur 01	High-resolution	$960 \times 1440 \times 1600$	Test	$58 \times 58 \times 58$	Nikon XT H 225	4.1 GB
	Low-resolution	$600 \times 600 \times 600$		$232 \times 232 \times 232$		412.0 MB
	Registered	$240 \times 360 \times 400$		$232 \times 232 \times 232$		65.9 MB
Vertebrae A	High-resolution	$1920 \times 1920 \times 1920$	Train	$22 \times 22 \times 22$	Nikon XT H 225	13.2 GB
	Low-resolution	$800 \times 960 \times 640$		$88 \times 88 \times 88$		937.5 MB
	Registered	$480 \times 480 \times 480$		$88 \times 88 \times 88$		210.9 MB
Vertebrae B	High-resolution	$1920 \times 1920 \times 1920$	Train	$22 \times 22 \times 22$	Nikon XT H 225	13.2 GB
	Low-resolution	$800 \times 960 \times 640$		$88 \times 88 \times 88$		937.5 MB
	Registered	$480 \times 480 \times 480$		$88 \times 88 \times 88$		210.9 MB
Vertebrae C	High-resolution	$1920 \times 1920 \times 1920$	Train	$22 \times 22 \times 22$	Nikon XT H 225	13.2 GB
	Low-resolution	$960 \times 800 \times 960$		$88 \times 88 \times 88$		1.4 GB
	Registered	$480 \times 480 \times 480$		$88 \times 88 \times 88$		210.9 MB
Vertebrae D	High-resolution	$1920 \times 1920 \times 1920$	Test	$22 \times 22 \times 22$	Nikon XT H 225	13.2 GB
	Low-resolution	$960 \times 800 \times 960$		$88 \times 88 \times 88$		1.4 GB
	Registered	$480 \times 480 \times 480$		$88 \times 88 \times 88$		210.9 MB

Table 1. Overview of VoDaSuRe, including sample names, volume shapes, slice splits for training and testing, voxel sizes and scanning devices. For vertebrae and femur samples, we reserve whole scans for training/test, while remaining scans are split into training/test slices.

Method	RCAN	HAT	EDDSR	SuperFormer	MFER	mDCSRN	MTVNet	RRDBNet3D
No. of parameters	15.6M	20.8M	0.8M	20.4M	1.7M	1.7M	67.0M	26.1M
Avg. training time	9.47 h	5.64 h	8.09 h	32.85 h	50.65 h	7.76 h	31.05 h	16.48 h

Table 2. Average training time of SR methods at $4\times$ upscaling. The measured times is the average time to complete 100K training iterations across datasets CTSpine1K, LiTS, LIDC-IDRI and VoDaSuRe.

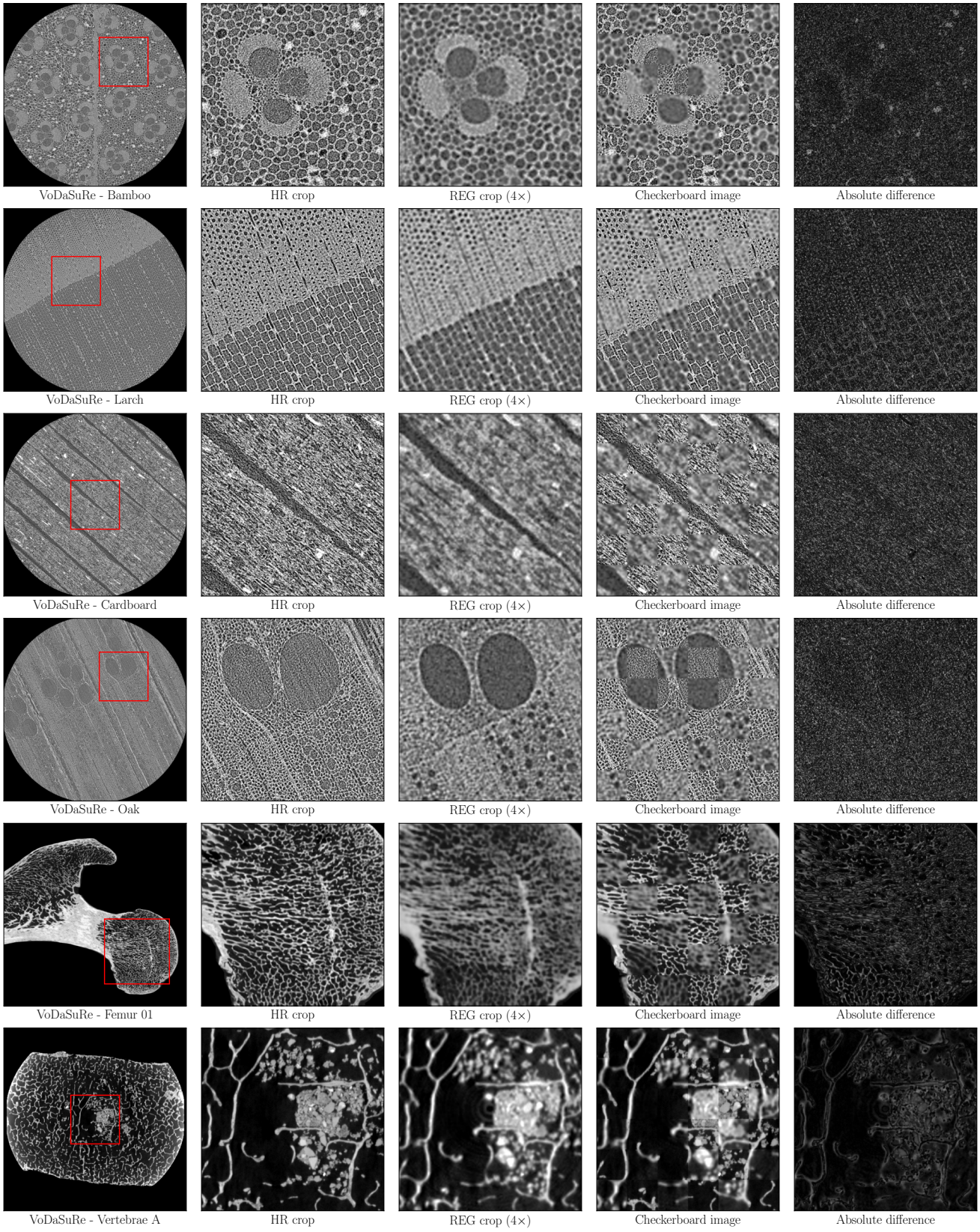

Figure 6. Evaluation of HR-LR registrations in VoDaSuRe. From left to right: Full HR slice, cropped HR slice, cropped registered LR slice, checkerboard image, and absolute difference image between HR and interpolated LR slice.

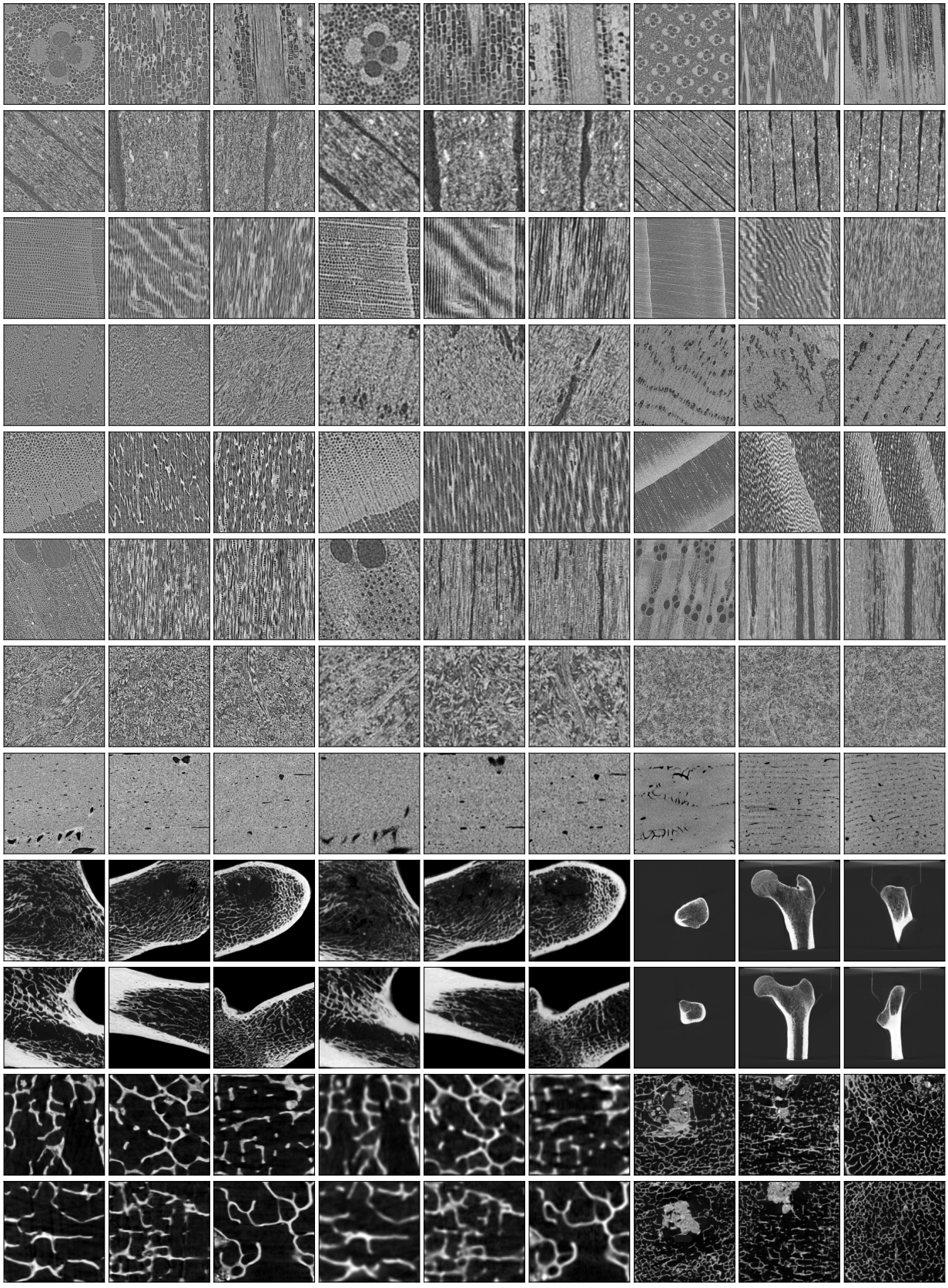

Figure 7. Orthogonal slices from VoDaSuRe, including high-resolution (left), registered (middle) and unregistered LR slices (right).

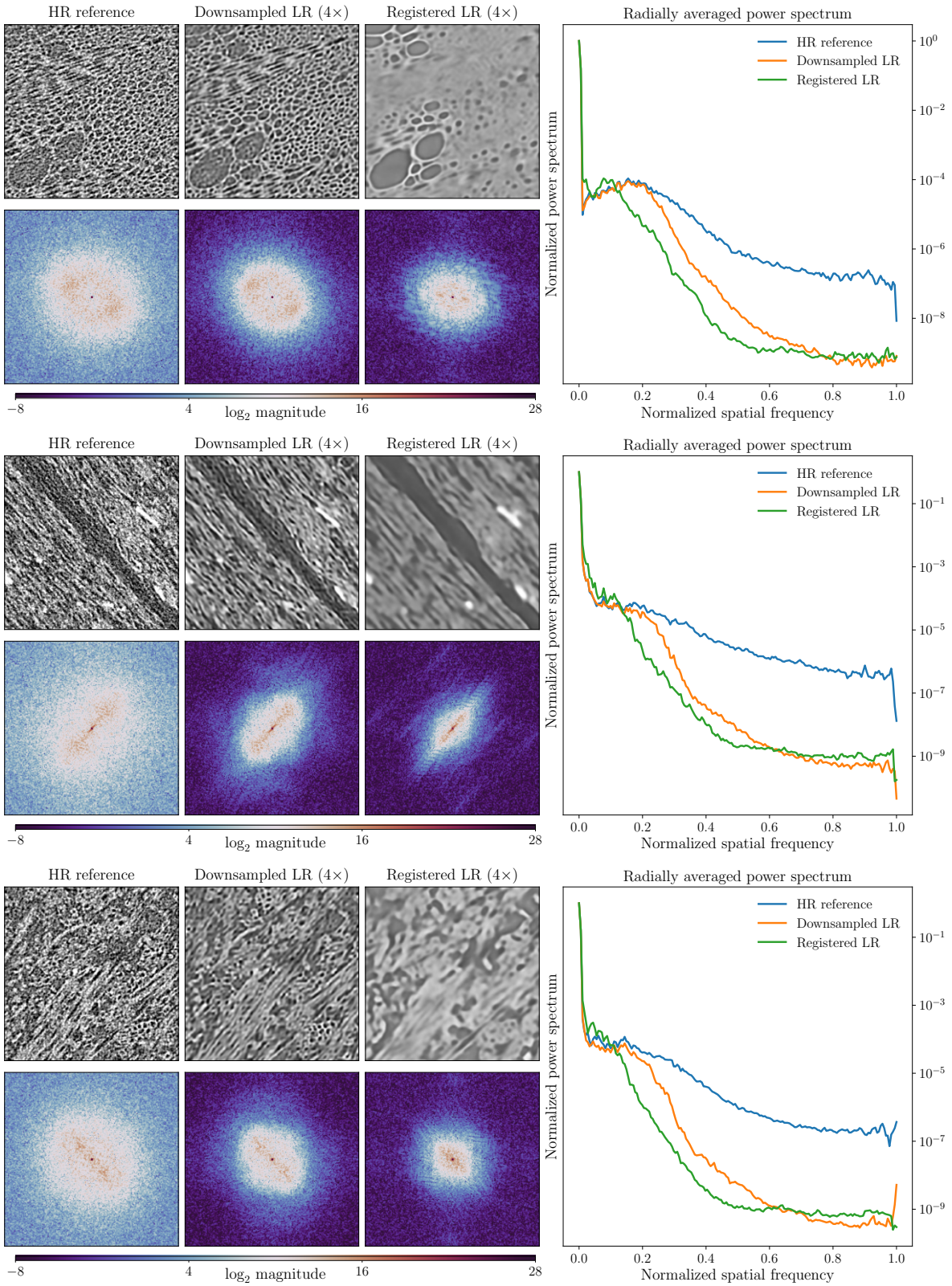

Figure 8. Comparison of spatial frequency distributions for HR images, and SR predictions using downsampled and real LR images from VoDaSuRe using RRDBNet3D at scale $4\times$. The bottom row shows the $\log_2$ power spectra computed from the FFTs of the corresponding images in the top row. Radially averaged power profiles (right) show the relative distribution of power as a function of spatial frequency.