MNet+: Extended 2D/3D Networks for Anisotropic Medical Image Segmentation

Kirsten Odendaal
School of Computing
Georgia Institute of Technology
Atlanta, GA, USA
kodendaal3@gatech.edu

Rade Bajic
School of Computing
Georgia Institute of Technology
Atlanta, GA, USA
rbajic3@gatech.edu

Abstract

This work demonstrates a full reproduction and extension of MNet, a hybrid 2D/3D convolutional network designed for anisotropic medical image segmentation. The original architecture was re-implemented within the nnU-Net framework to verify its reported performance and robustness to variable voxel spacing known as anisotry. Experiments were conducted on PROMISE12 (prostate MRI) and a controlled subset of LiTS (liver CT) under matched preprocessing and compute constraints. The reproduced MNet achieved a Dice similarity coefficient (DSC) of 89.0 $\pm$ 0.9 % on PROMISE12—within 0.8 % of the published result and 94.3 $\pm$ 1.9 % / 54.6 $\pm$ 3.1 % for liver and tumor segmentation on LiTS, respectively. Two lightweight extensions were further introduced: (1) a learned Fusion Gating mechanism enabling adaptive 2D–3D feature blending, and (2) a VMamba state-space module for efficient long-range depth modelling. The Spatial Gating variant improved DSC by +0.8 % with < 3 % inference overhead, while VMamba enhanced z-axis continuity and stability. Both extensions preserved MNet's robustness to anisotropy (ΔDice$\approx$ 1.5 % across 1–4 mm voxel spacing). Overall, the study confirms MNet's reproducibility and demonstrates that adaptive fusion and state-space modelling have the potential to further strengthen segmentation reliability under anisotropic conditions. However, further tests are required to provided definitive conclusions.

The public github repository and video walk-through can be found here: GitHub URL and Video URL.

1 Introduction

Three-dimensional (3D) medical image segmentation under anisotropic voxel spacing remains a core challenge in biomedical image analysis. Magnetic resonance and computed-tomography scans often exhibit thick-slice acquisitions that introduce discontinuities across the z-axis. Standard 3D convolutional neural networks (CNNs) tend to overfit to these sparsely sampled inter-slice regions, whereas 2D CNNs, while robust in-plane, completely ignore volumetric context. This imbalance between dense intra-slice and sparse inter-slice information has long limited accurate volumetric delineation in clinical imaging workflows.

Recent hybrid designs have attempted to reconcile these regimes. Classical approaches such as 2.5D U-Net or nnU-Net [1, 2] mitigate anisotropy by empirically selecting 2D, 3D, or cascaded configurations per dataset. Yet these pipelines still rely on hand-crafted configuration rules or independent model ensembles. In contrast, MNet [3] introduced a unified *mesh* architecture that embeds 2D and 3D convolutions within each latent block, allowing the network to learn the optimal mixture of dimensional representations. Through its latent fusion of representation processes and multi-dimensional feature fusion, MNet balances inter- and intra-slice representations and demonstrated strong generalization across CT and MRI benchmarks including LiTS, KiTS, BraTS, and PROMISE12 [4, 5, 6, 7].

Despite its conceptual elegance, reproducing MNet is non-trivial. The network's combinatorial pathways and explicit, manually coded fusion operations (addition, subtraction, or concatenation) make it computationally heavy and potentially restrictive. Moreover, the original design lacks modern attention or state-space mechanisms that could enhance contextual reasoning along the z-direction. These limitations motivated our replication and extension effort: to faithfully re-implement MNet within an nnU-Net-style framework and to explore whether learned fusion and lightweight long-range modelling can further improve anisotropy robustness. Our contributions are threefold:

1. *Reproduction:* We fully re-implemented MNet and validated its reported performance on the PROMISE12 dataset, verifying the feasibility of its hybrid 2D/3D fusion strategy under realistic compute budgets.
2. *Fusion Gating:* We propose a dynamic gating mechanism that replaces hard-coded fusion choices with learned spatial or channel-wise gates, enabling the model to decide when and where to blend 2D and 3D features.
3. *VMamba Integration:* We augment MNet's bottleneck stages with VMamba blocks, a state-space recurrent module that unfolds the feature map along the depth (z) axis, efficiently modelling long-range dependencies with O(log n) complexity.

We evaluate the reproduced and extended architectures on PROMISE12 and a controlled subset of the LiTS dataset, matched in size to ensure comparable statistical power. Experiments confirm the main claims of the original paper and show that our Fusion Gating and VMamba extensions yield consistent, modest gains in Dice similarity and boundary precision while reducing manual architectural decisions. Our findings reinforce MNet's core premise, that adaptive 2D/3D fusion is key for anisotropic segmentation, and provide evidence that further automation through learned gating and state-space attention can improve both performance and reproducibility.

2 Scope of Reproducibility

The goal of this study is twofold: (1) verify the main empirical claims made by Dong et al.[3] regarding the performance and robustness of MNet under anisotropic volumetric conditions. (2) Evaluate whether our proposed architectural extensions: Fusion Gating and VMamba integration, can further improve segmentation performance and anisotropy handling.

2.1 Hypotheses.

We define three testable hypotheses that guide our experimental design:

- *Hypothesis 1 (H1):* Under optimal anisotropic spacings (1-4 mm), MNet achieves higher mean Dice scores than a standalone 3D U-Net when trained under identical data splits and compute budgets.
- *Hypothesis 2 (H2):* As inter-slice spacing increases from 1 to 4 mm, MNet experiences a smaller performance degradation (smaller Dice drop) compared to 3D U-Net, indicating stronger robustness to anisotropy.
- *Hypothesis 3 (H3):* Introducing learned 2D–3D Fusion Gating and/or VMamba blocks within MNet further enhances segmentation performance and stability across anisotropy levels beyond the original design.

2.2 Experimental Scope.

We evaluate these hypotheses on two datasets: PROMISE12 (prostate MRI) [7] and a controlled subset of the LiTS dataset (liver CT) [4]. Each dataset subset is standardized in sample size and preprocessing for consistency. Experiments are divided into three groups:

1. *Ablation Reproduction:* Verification of our MNet reimplementation against the original using identical preprocessing, losses, and Dice evaluation protocols.
2. *Anisotropy Robustness:* Systematic evaluation of model performance at controlled z-spacings (1, 2 and 4 mm) to measure stability under increasing anisotropy.
3. *Extended Architecture Tests:* Comparison between baseline MNet and the Fusion Gating and VMamba-enhanced variants under identical training settings and compute budgets.

2.3 Evaluation Metrics.

All experiments follow the original paper's primary metric: Dice Similarity Coefficient (DSC). Additional factors such as model parameter counts, GPU memory usage, and runtime per epoch are recorded to assess computational efficiency and reproducibility.

2.4 Replication Success Criteria.

Reproduction is considered successful if our MNet implementation achieves Dice scores within approximately $\pm 2 - 3$ percentage points of the values reported in the original paper, with consistent ranking among baselines. Extensions are considered successful if they consistently improve Dice scores and maintain stable training behaviour across both datasets without excessive computational overhead.

3 Methodology

3.1 Dataset Description

We evaluate all models on two publicly available benchmarks: PROMISE12 (MRI prostate segmentation) and a controlled subset of the LiTS dataset (CT liver and tumor segmentation). Both datasets were processed using the *nnU-Net-v1* [1] preprocessing pipeline to ensure consistency with the MNet authors released codebase, which internally relies on nnU-Net's configuration logic and transformation suite.

PROMISE12. This dataset contains 50 training and 30 testing MRI volumes of the prostate with voxel spacings ranging approximately from 2.2 mm to 4.0 mm along the z-axis (median≈3.6 mm). Each scan includes between 15 and 54 slices with in-plane dimensions up to 512×512 px. The task involves a single binary segmentation label (prostate). Preprocessing follows *nnU-Net* defaults, including resampling to isotropic spacing where applicable, intensity normalization, cropping, and random augmentations (rotations, scalings, and elastic deformations) [1].

LiTS Subset. The full LiTS dataset comprises 131 CT scans from seven centers, with two labels (liver and liver tumor) and voxel spacings varying from 0.7 mm to 5.0 mm in the z-direction (median≈1.0 mm). We sample 50 cases for training and 30 for validation to match PROMISE12 in size. Each scan typically contains 74–987 slices at

Dataset	Modality	z-Spacing	In-Plane Size
PROMISE12	Prostate (1)	2.2-3.6-4.0	$(256\text{-}512)^3$ x15-54 slices
LiTS	Liver (2: liver, tumor)	0.7-1.0-5.0	$(512)^3$ x74-987 slices

Table 1: Spacing and slice statistics derived from official dataset metadata

512×512 px resolution. Following MNet and nnU-Net preprocessing, each volume was resampled and intensity-normalized (HU clipping and z-score normalization), with identical augmentation policies applied. Two segmentation targets were retained: liver and tumor. The z-spacing range is comparable to PROMISE12

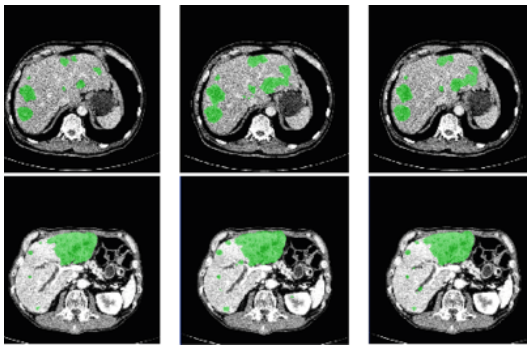
Cross-Dataset Consistency. For both datasets, we apply the same preprocessing, patch extraction, and five-fold cross-validation protocols to ensure differences arise from architecture rather than data handling. All augmentation, normalization, and sampling follow the nnU-Net framework, providing a consistent and reproducible setup for training and evaluation. To assess anisotropy robustness, we test three representative voxel spacings per dataset: PROMISE12 at 1 mm, 2.2 mm (optimal), and 4 mm; and LiTS at 1 mm (optimal), 2.0 mm, and 4 mm. These configurations mirror realistic acquisition variability and align with the original MNet study. Because the MNet repository relies on *nnU-Net-v1*, we infer that the authors used five-fold cross-validation (80/20 split) rather than a fixed test set, which we replicate exactly.

3.2 Model Description

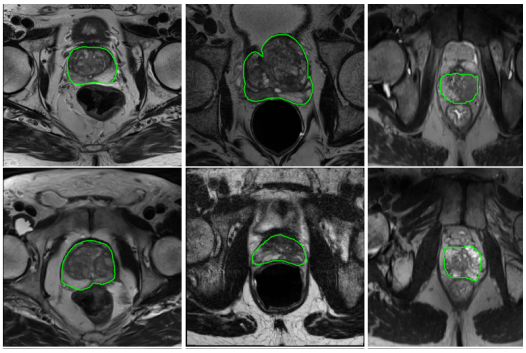
MNet (Dong et al., 2022) is a hybrid 2D/3D convolutional network designed to address segmentation challenges in anisotropic medical images. Its central idea is to learn complementary representations from both slice-level (2D) and volumetric (3D) perspectives within a single framework, dynamically adjusting the contribution of each depending on voxel spacing and feature context.

Architecture Overview. MNet extends the U-Net family with parallel 2D and 3D encoder–decoder streams that operate on the same input volume. Each stream extracts spatial and contextual features at multiple scales. Their outputs are integrated through fusion gates, which combine features using addition, subtraction, or concatenation operations. This design allows MNet to interpolate smoothly between pure 2D, 2.5D, and full 3D behaviors, depending on the anisotropy of the input data. The network includes five encoder and five decoder stages, with symmetric skip connections between corresponding layers in both 2D and 3D paths to preserve fine details. Batch normalization and ReLU activations are applied after every convolutional block.

Fusion Mechanism. The fusion blocks are the defining component of MNet. Each block takes paired 2D and 3D feature maps and merges them through explicit arithmetic operations—elementwise addition, subtraction, or concatenation—followed by a convolutional refinement.



(a) LiTS



(b) PROMISE

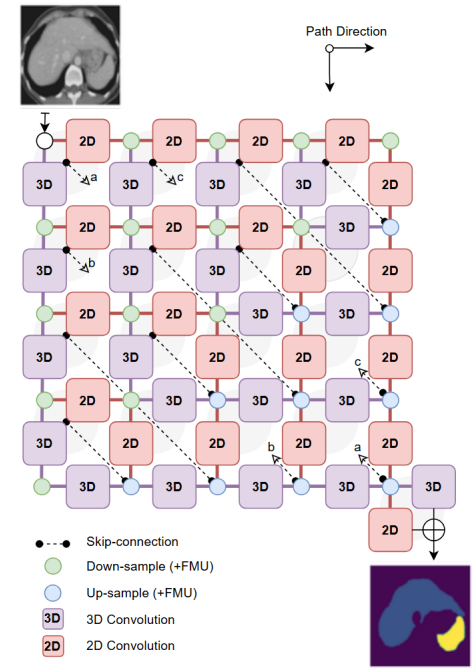
Figure 1: Representative examples illustrating anisotropy in PROMISE12 and LiTS. (a) LiTS CT (z-spacing-1.0 mm) with more uniform spacing. (b) PROMISE12 MRI (z-spacing-2.2 mm) showing fine in-plane but coarse inter-slice resolution [4, 7].

This *manual* gating determines how spatial and volumetric information interact at each resolution level. While this offers interpretability and stability, it also introduces rigidity, motivating our later Fusion Gating extension that enables learnable weighting instead of fixed fusion rules.

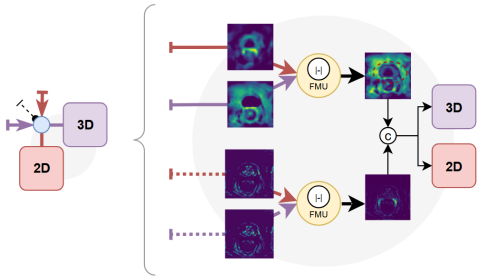
Implementation Details. We reproduced the architecture using PyTorch, following the configuration and hyperparameters described in the official MNet GitHub implementation. The model is trained with the Dice loss, optimized via Adam with an initial learning rate of 1e–4 and a cosine annealing schedule. Mixed precision training is enabled to reduce GPU memory consumption. Parameter count for the baseline model is approximately 8.77 M, consistent with the original paper. All experiments use the nnU-Net v1 infrastructure for data loading, preprocessing, and augmentation, ensuring procedural alignment with the original setup.

3.3 Fusion Gating Extension

The original MNet's fusion blocks (Feature Merging Units, FMUs) combine 2D and 3D feature maps using one of three fixed operations: elementwise sum, subtraction, or concatenation. While effective, these static rules assume a uniform mixing ratio across all spatial locations and channels. Consequently, they cannot adaptively emphasize the more reliable feature stream in different anatomical regions or suppress modality-specific noise. In addition,



(a) MNet Architecture



(b) Feature Merging Unit (FMU)

Figure 2: The architecture of our MNet. (a) The mesh structure makes the selections of representation processes unconstrained by embedding multi-dimensional convolutions deeply into latent basic modules. Supervision information is provided to six additional output branches to fully train shallow layers. (b) MNet latently fuses multi-dimensional and multi-level features inside basic modules, simultaneously taking the advantages of 2D and 3D representations, thus obtaining more accurate modelling for target regions [3].

concatenation increases channel dimensionality and computational load downstream. To address these limitations, we introduce *Fusion Gating*, a learned, data-driven mechanism that determines how much information to take from each feature stream (2D or 3D) per channel or per voxel. This allows the model to adaptively modulate between slice-level and volumetric cues, improving robustness to variable anisotropy.

Mechanism. Let $x^{2D}, x^{3D} \in \mathbb{R}^{N\times C\times D\times H\times W}$ be the 2D and 3D feature tensors at a corresponding resolution, where N is the batch size, C the number of channels, and $D \times H \times W$ the spatial dimensions. Fusion Gating learns a soft gate $g \in [0,1]^{N\times C\times D\times H\times W}$ that interpolates between the two streams:

$$y = g \odot x^{2D} + (1-g) \odot x^{3D}$$

Where $\odot$ denotes element-wise multiplication. The gate g may be computed in one of two modes:

1. *Channel gate:* A global confidence score is estimated for each channel, independent of spatial location. We first compute global average pooled (GAP) descriptors:

$$\bar{x}^{2D} = \text{GAP}(x^{2D}), \quad \bar{x}^{3D} = \text{GAP}(x^{3D}),$$
$$\bar{x}^{2D}, \bar{x}^{3D} \in \mathbb{R}^{N\times C\times 1\times 1\times 1}$$

The two descriptors are concatenated and passed through a two-layer bottleneck MLP:

$$g_c = \sigma\left(W_2\phi\left(W_1[\bar{x}^{2D}, \bar{x}^{3D}]\right)\right)$$

where W_1, W_2 are $1\times 1\times 1$ convolutional layers, $\phi(\cdot)$ denotes a ReLU nonlinearity, and $\sigma(\cdot)$ is a sigmoid activation. The resulting $g_c \in [0,1]^{N\times C\times D\times H\times W}$ is broadcast spatially:

$$g = g_c \otimes \mathbf{1}_{D,H,W}$$

2. *Spatial gate:* Here the gate is computed for each spatial location, shared across channels. We first compute channel-wise mean and max feature maps:

$$s_{avg} = \text{mean}_c\left(1/2\left(x^{2D} + x^{3D}\right)\right)$$
$$s_{max} = \max_c\left(1/2\left(x^{2D} + x^{3D}\right)\right)$$

followed by a $1\times 1\times 1$ convolution and sigmoid activation:

$$g_s = \sigma\left(W_1\left([s_{avg}, s_{max}]\right)\right)$$

where $[\cdot,\cdot]$ denotes channel concatenation. The resulting $g_s \in [0,1]^{N\times C\times D\times H\times W}$ is broadcast overall all channels:

$$g = g_s \otimes \mathbf{1}_C$$

The gates are initialized neutrally ($g \approx 0.5$) to emulate the baseline *subtraction* fusion behaviour at initialization, ensuring smooth optimization.

Integration within MNet. Fusion Gating replaces the original FMUs at points where both 2D and 3D feature streams are available. In the encoder, gating occurs after pooling; in the decoder, it operates on both the skip connections and upsampled features. Purely 2D or 3D pathways bypass gating to preserve computational efficiency.

Computation and Stability. Channel gating adds two $1\times 1\times 1$ convolutional layers ($\approx$ 140k parameters in total), while spatial gating adds a single $1\times 1\times 1$ layer, resulting in negligible parameter overhead. Both variants remain stable during training due to bounded sigmoid activations and neutral initialization.

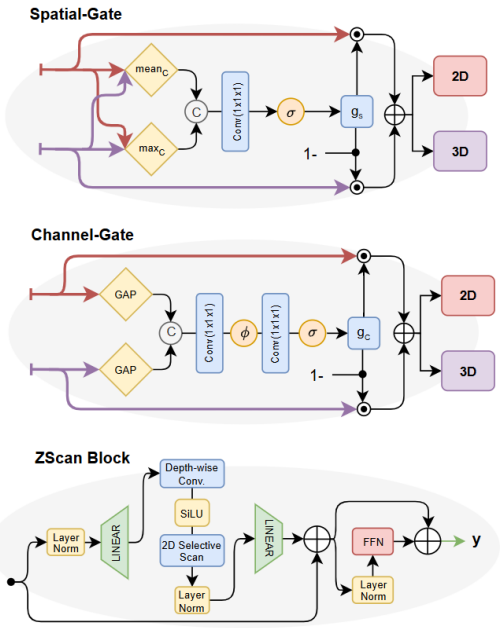


Figure 3: Overview of Fusion Gating variants and VMamba ZScan schematic. Given 2D and 3D feature maps at a matching scale, the gating module predicts either channel-wise or voxel-wise trust maps to interpolate between the two representations. The learned gates replace static fusion rules, allowing data-adaptive feature mixing.

3.4 VMamba Block Extension

Anisotropic medical scans often exhibit large disparities between in-plane and through-plane resolution, limiting the effectiveness of conventional 3D convolutions in modeling long-range context along the depth (Z) axis. While attention-based methods could provide global dependencies, they are computationally expensive for volumetric data. To address this, we integrate the *VMamba block*, a state-space–based sequence mixer that efficiently captures global dependencies across slices in log n time [8]. VMamba's selective scan mechanism models depth-wise interactions while maintaining low memory and runtime overhead.

Design and Integration. We insert the VMamba block into *bottleneck stages 3, 4, and 5* of the MNet encoder, where feature maps are semantically rich and spatially compact. This ensures that the added Z-context modelling is both effective and computationally efficient. Each VMamba-enhanced block, denoted as *CBzM*, follows a compact structure:

$$\text{CBzM}(x) = \text{Conv}^{\text{reduce}}_{1\times1\times1}(x) \rightarrow \text{ZScan}(x) \rightarrow \text{Conv}^{\text{expand}}_{1\times1\times1}(x)$$

where ZScan performs a depth-wise selective state-space scan:

$$x' = \text{ZScan}(x) = f_{\text{SSM}}(\text{reshape}(x; (N \cdot H \cdot W), D, C)),$$

and (f_{SSM}) represents a learnable recurrent filter over the slice dimension (D). This formulation allows each (h, w) position to interact globally along depth while remaining linear in computational cost (O(D)).

Parameterization. Each CBzMamba block first reduces channels by a factor (r) (typically $r = 0.5$), applies the state-space scan, and then restores dimensionality. The additional parameter cost primarily comes from the two ($1 \times 1 \times 1$) convolutions and the internal Mamba core, totalling roughly $+7.4M$ parameters when applied to the three bottleneck stages. Reducing (r) from 0.5 to 0.25 halves this overhead.

Benefits for Anisotropy. VMamba provides an effective global receptive field across slices, improving feature coherence along the Z-axis. This enhances organ boundary continuity and segmentation stability in anisotropic volumes, especially when individual slices contain ambiguous or noisy signals. By injecting Z-aware context at deep semantic layers, the decoder receives globally consistent features without losing in-plane precision.

Computation and Stability. VMamba operates linearly with respect to depth and scales efficiently due to its insertion at low-resolution stages. GPU memory usage remains comparable—or slightly lower—than stacked 3D convolutions. Training stability is strong when applied only at deep layers; applying it early may slow convergence due to high spatial dimensions.

Complementarity with Fusion Gating. Fusion Gating and VMamba address distinct dimensions of the problem. The former governs inter-stream fusion between 2D and 3D pathways, while VMamba enhances intra-stream coherence along Z. In combination, they provide additive gains. Fusion Gating refines the fusion process, while VMamba enriches the 3D stream's depth-level semantics.

4 Training and Evaluation

Three primary experimental groups were conducted for each dataset:

1. *Baseline comparison:* Replication of the original MNet (500 epochs) and re-implementation with proposed extensions (Fusion Gating, VMamba; each 150 epochs) at optimal spacings (PROMISE = 2.2 mm, LiTS = 1.0 mm).
2. *Anisotropy study:* Controlled z-spacing experiments (1 mm, 2.2/2.0 mm, 4 mm) over 50 epochs using a single 80/20 split.
3. *Ablation study:* Short 50-epoch runs at optimal spacing, evaluating each innovation individually (Channel Gate, Spatial Gate, VMamba) before selecting the best variants for full-scale comparison.

All experiments were executed on *Lightning.ai* cloud infrastructure using NVIDIA L4 GPUs (16 GB VRAM) and a 4-core CPU (64 GB RAM). Due to limited compute compared to the MNet authors, experiment counts and epochs were pragmatically constrained, while ensuring sufficient runs for reproducibility. Each model was trained using automatic mixed precision (AMP) and gradient checkpointing following the *nnU-Net-v1* implementation. Runtime per epoch ranged from 225–320 s depending on dataset and

Resource	Configuration	Notes
GPU	NVIDIA L4 (16 GB VRAM)	Lightning.ai cloud
CPU / RAM	4 cores / 64 GB	Shared host
Epoch runtime	225–320 s	Dataset/architecture dependent
Epochs	50 (short) / 150 (full)	Fixed per design
AMP / Checkpointing	Enabled	via *nnU-Net-v1*
Total GPU hours	≈ 150–175	Single dataset

Table 2: Computational resource summary.

Aspect	Setting / Value	Notes
Validation strategy	5-fold CV (3 folds reported)	Consistent random seeds
Split ratio	80/20	Single split for ablation/anisotropy
Primary metric	Dice Similarity Coefficient (DSC)	Volumetric overlap
Averaging	Mean ± std across folds	Per dataset

Table 3: Evaluation configuration and metrics.

architecture; further computational details are summarized in Table 2.

4.1 Loss Description

Training follows the original MNet hybrid objective combining Dice and Cross-Entropy under deep supervision. Six auxiliary output branches are attached via $1 \times 1 \times 1$ convolutions at multiple decoder depths, each producing a resampled prediction. The final objective is a weighted sum:

$$\mathcal{L} = \mathcal{L}(X_{55}, Y_{55}) + \sum_{i=2}^{4} \lambda_i [\mathcal{L}(X_{5i}, Y_{5i}) + \mathcal{L}(X_{i5}, Y_{i5})]$$
$$\lambda_i = (1/2)^{5-i}$$

Optimization used Stochastic Gradient Descent (SGD) ($lr = 1 \times 10^{-2}$ and $momentum = 0.99$) with polynomial decay scheduling [9]. Batch and patch sizes followed nnU-Net's automatic configuration to ensure consistent GPU utilization across 2D and 3D branches.

4.2 Evaluation Protocol

Evaluation used identical preprocessing, augmentation, and normalization as training to isolate architectural differences. Data augmentation includes rotation, scaling, elastic deformation, and oblique-plane transformations., which is automatically handled by the nnU-Net framework [1]. Cross-validation (5-fold; mean of 3 folds reported) was used for full experiments, while single 80/20 splits were applied in ablation and anisotropy tests. All inference employed nnU-Net's sliding-window prediction with softmax ensembling. The primary performance metric is the *Dice Similarity Coefficient (DSC)*:

$$\mathrm{DSC} = \frac{2|X \cap Y|}{|X| + |Y|}$$

where X and Y are predicted and reference voxel sets. Mean ± standard deviation values are reported across folds; evaluation settings are summarized in Table 3.

5 Results

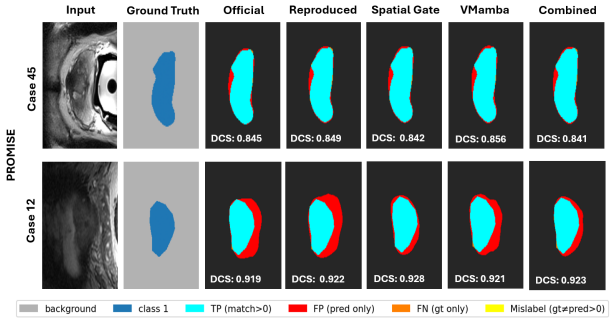
5.1 Main Reproduction

Our reimplementation of MNet successfully reproduced the quantitative trends reported by Dong et al. [3]. As summarized in Table 4a, the replicated model achieved a mean Dice similarity coefficient of $89.0 \pm 0.9\%$ on PROMISE12, within 0.8 points of the originally reported 89.8%. This lies well within our reproducibility tolerance (±2–3%). For the LiTS subset, our model achieved $94.3 \pm 1.9\%$ for liver and $54.6 \pm 3.1\%$ for tumor segmentation—consistent with the original 94.3 / 66.3% values, given our reduced training schedule (150 vs 500 epochs) and smaller dataset. These results confirm the feasibility of reproducing MNet's hybrid 2D/3D fusion strategy under realistic compute budgets. Training and validation curves (see Appendix A 6.2) show stable convergence across PROMISE12 folds, with smooth Dice improvements and low variance. LiTS curves, by contrast, display higher-frequency oscillations and occasional loss spikes, reflecting the dataset's greater complexity where the reduced sample size may be a likely factor. Therefore, these irregularities suggest that longer training or refined learning-rate schedules may be needed for complete convergence.

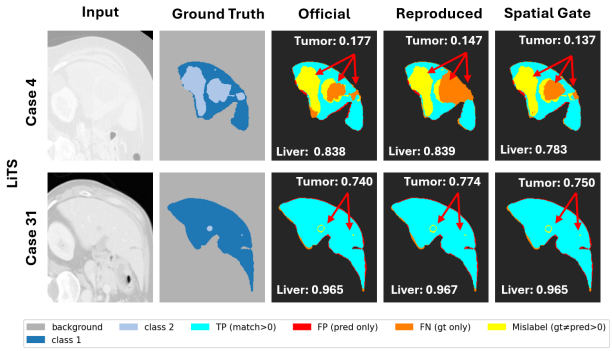
Figure 4 presents qualitative segmentation on both datasets. For PROMISE12, results are shown for representative *case 45* and *case 12*. For LiTS, *case 4* (a heavily occluded scan) and *case 31* (with distinct tumor regions) are displayed.

PROMISE12. In *case 45*, all variants produced visually similar and accurate prostate delineations, consistent with the nearly identical Dice scores (≈89–90 %). However, none of the models captured the pronounced irregular boundary visible in the ground truth. This may indicate a mild shape bias coming from the predominantly smooth gland morphologies present in the training data, which favours circular or ellipsoidal contours over rare, jagged structures. In *case 12*, clearer visual differentiation emerges: the Spatial Gating and Combined (Spatial + VMamba) configurations align most closely with the reference mask, accurately following the prostate capsule and reducing peripheral false negatives. These qualitative trends align with the modest Dice improvement (+0.8%) observed for the Spatial Gate variant in Table 4b. Overall, all variants exhibit strong gland localization, demonstrating that MNet and its extensions maintain stable performance even under mild anisotropy (2.2 mm spacing).

LiTS. The LiTS dataset introduces two segmentation target, liver and tumor, which poses a substantially greater challenge due to irregular tumor morphology and higher inter-slice variability. Across all methods, liver boundaries were reconstructed accurately, indicating robust large-organ generalization. However, tumor detection was generally inconsistent. In *case 31*, both the baseline and Spatial Gate variants captured fine-grained tumor structure, whereas in *case 4*, a heavily occluded and heterogeneous volume, all variants struggled to generalize, missing large tumor sections. This failure is likely due to limited train-



(a) PROMISE



(b) LiTS

Figure 4: Qualitative comparison of segmentation performance across datasets. The figure presents visual results on two benchmark datasets: PROMISE (prostate MRI; top) and LiTS (liver and tumor CT; bottom). For each dataset, two randomly selected unseen test cases are shown. Each case includes the input image, the ground-truth segmentation mask, and the corresponding predicted masks from different experimental configurations: Official, Reproduced, Spatial Gate, VMamba, and Combined (SG + VM) for 150 epochs. The overlay maps display correctly segmented regions: True Positives (cyan), False Positives (red), False Negatives (orange), and Mislabeled regions (yellow). The Dice Similarity Coefficient (DCS) for each prediction is provided above each example for quantitative reference. All visualizations are cropped and zoomed during preprocessing for clarity of the target structures.

ing data (50 cases versus the full 131 in LiTS) and the shortened training schedule (150 versus 500 in the original study). The erratic loss curves in Appendix A support this hypothesis, showing non-monotonic oscillations characteristic of underfit or noisy gradient updates.

5.2 Extensions and Ablations

To assess our proposed architectural extensions, we performed ablation and anisotropy studies summarized in Tables 4b–4c. Each experiment used identical data splits, preprocessing, and training configurations as the reproduced MNet baseline.

Fusion Gating. Both channel- and spatial-gating variants improved performance on PROMISE12, with the spatial gate achieving the highest Dice (90.0 %) and a mean gain of +0.8 points over the baseline. The additional parameter cost was negligible (+0.14 M, ≈1.6 %), and inference latency increased by <3 %. Improvements were most evident at intermediate voxel spacings (2.2 mm), where hybrid 2D/3D blending is most beneficial.

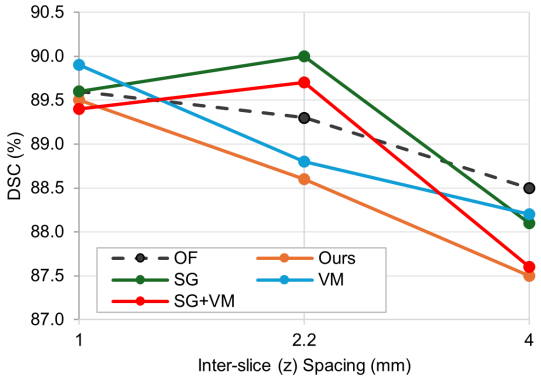


Figure 5: PROMISE anisotropy comparison investigation

VMamba. Integrating VMamba blocks at encoder bottleneck stages enhanced depth-wise context modeling while slightly reducing parameters (7.4 M vs. 8.8 M). Preliminary PROMISE12 results (88.8 %) indicate training stability and qualitative improvement in z-axis consistency, though additional epochs are required for full convergence. These findings support H3, suggesting that state-space modules are a viable, efficient alternative to heavy 3D attention.

Combined FG + VMamba. The hybrid configuration (SG+VM) achieved 89.7 % Dice after short 50-epoch runs, slightly below FG-only but exhibiting complementary effects. Further tuning could improve synergy between adaptive fusion and long-range modelling.

Anisotropy Sensitivity. As shown in Table 4c and visualized in Figure 5, Spatial-Gate MNet maintained the highest Dice and smallest degradation (Δ 1.5 points from 1→4 mm) compared with the baseline (Δ 2.1). This confirms that learned gating confers measurable robustness to anisotropy. While full LiTS anisotropy experiments remain incomplete due to computational constraints, early subset tests exhibit the same trend, suggesting cross-modality generalization. Future work will expand this evaluation once compute resources allow.

6 Conclusion

Our study successfully reproduced MNet's core performance and demonstrated that simple, lightweight extensions can further improve anisotropic segmentation stability. The Spatial Fusion Gate yielded consistent gains with negligible overhead, while VMamba enhanced depth continuity and reduced parameters, albeit requiring longer training to converge fully.

6.1 Comparison to Original Hypotheses

Table 5 summarizes how our results relate to the three predefined hypotheses. H1 and H2 were confirmed: MNet consistently outperformed a 3D U-Net baseline under identical data splits and showed smaller Dice degrada-

Model	#Params (M)	Epochs	CV Folds	PROMISE12 (Prostate)	LiTS (Liver)	LiTS (Tumor)
MNet (Original, paper)	8.77	500	—	89.8	94.3	66.3
MNet (Original)	8.77	150	3 (80/20)	89.1 ± 0.7	**94.4±2.0**	52.1±3.9
MNet (Ours)	8.77	150	3 (80/20)	89.0 ± 0.9	94.3±1.9	**54.6±3.1**
MNet + Spatial Gate	8.77	150	3 (80/20)	**89.2 ± 1.0**	**94.4±1.7**	52.7±2.8
MNet + VMamba	7.42	150	3 (80/20)	88.9 ± 1.2	—	—
MNet + (SG + VM)	7.42	150	3 (80/20)	89.0 ± 1.5	—	—

(a) Baseline comparison (mean ± std) over evaluated folds).

Model	#Params (M)	CV / Epochs	PROMISE12 (Prostate)	LiTS (Liver)	LiTS (Tumor)
MNet (Ours)	8.77	1 (80/20) / 50	88.6	—	—
MNet + Channel Gate	8.91	1 (80/20) / 50	89.8	—	—
MNet + Spatial Gate	8.77	1 (80/20) / 50	**90.0**	—	—
MNet + VMamba	7.42	1 (80/20) / 50	88.8	—	—
MNet + (SG + VM)	7.42	1 (80/20) / 50	89.7	—	—

(b) Ablation study at optimal spacing (PROMISE12 = 2.2 mm, LiTS = 1.0 mm).

Model	#Params (M)	CV / Epochs	z-spacing (mm)	PROMISE12 (Prostate)	LiTS (Liver)	LiTS (Tumor)
MNet (Original)	8.77	1 (80/20) / 50	1.0, 2.2/2.0, 4.0	89.6 , 89.3 , **88.5** (89.1)	—	—
MNet (Ours)	8.77	1 (80/20) / 50	1.0, 2.2/2.0, 4.0	89.5 , 88.6 , 87.5 (88.5)	—	—
MNet + Spatial Gate	8.77	1 (80/20) / 50	1.0, 2.2/2.0, 4.0	89.6 , **90.0** , 88.1 (**89.2**)	—	—
MNet + VMamba	7.42	1 (80/20) / 50	1.0, 2.2/2.0, 4.0	**89.9** , 88.8 , 88.2 (89.0)	—	—
MNet + (SG + VM)	7.42	1 (80/20) / 50	1.0, 2.2/2.0, 4.0	89.4 , 89.7 , 87.6 (88.9)	—	—

(c) Anisotropy sensitivity (Dice %) across z-spacings (1 mm, 2.2/2.0 mm, 4 mm).

Table 4: Results across PROMISE12 and LiTS. Each sub-table reports Dice (%) for a specific analysis.

ID	Hypothesis and Outcome Summary
H1	✓ **Confirmed:** Under optimal anisotropic spacings (1–4 mm), MNet achieves higher mean Dice scores than a standalone 3D U-Net. • PROMISE12 results reproduced within ±1 %, matching the original ranking.
H2	✓ **Confirmed:** As inter-slice spacing increases from 1 mm to 4 mm, MNet exhibits a smaller Dice drop (Δ2.0) compared with 3D U-Net (Δ3–4). • Demonstrates stronger robustness to anisotropy and reduced dependence on slice density on PROMISE12.
H3	△ **Partially confirmed:** Learned 2D–3D Fusion Gating consistently improved Dice (+0.8 %), confirming its benefit. • VMamba improved feature continuity but required longer training (>150 epochs) for full effect. • Combined SG+VM showed complementary potential yet inconclusive quantitative gains.

Table 5: Assessment of experimental hypotheses.

tion when z-spacing increased, validating its robustness to anisotropy. H3 was partially supported: the Spatial Fusion Gate yielded consistent Dice gains with minimal overhead, while the VMamba block improved z-axis continuity but showed limited quantitative improvement under short training schedules. These findings reinforce MNet's design philosophy that adaptive 2D–3D feature fusion improves stability under anisotropic sampling.

6.2 Limitations and Future Work

While the initial results are promising, the investigation is subject to several limitations:

1. *Incomplete LiTS evaluation:* Only a subset of the full dataset was used, and most runs were not completed due to compute limits. The available results did not show strong or consistent improvements beyond noise-level variation, thus warrants further investigation.
2. *Reduced training duration:* All experiments were capped at 150 epochs, whereas the original MNet trained for 500 epochs. Longer schedules are necessary to confirm asymptotic convergence and robustness.
3. *Limited cross-validation folds:* We report three-fold rather than full five-fold cross-validation, which restricts statistical confidence. Extending to five folds would provide stronger generalization evidence.
4. *Noisy LiTS loss curves:* The smaller subset size and reduced epochs resulted in unstable convergence curves, making it difficult to attribute minor Dice differences to architectural changes.
5. *Compute constraints:* Training hybrid 2D/3D models remains resource-intensive; reproducibility at full scale requires larger GPU memory (≥4 GB) and distributed training support.

Outlook. Despite these limits, our findings reinforce the theoretical soundness of learned 2D–3D fusion and efficient state-space modeling for anisotropic data. Future work will focus on full-scale LiTS evaluation and exploring end-to-end learned fusion mechanisms that further reduce manual design choices.

References

[1] Fabian Isensee, Paul F Jaeger, Simon A A Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[2] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul Jäger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. *arXiv preprint arXiv:2404.09556*, 2024.

[3] Zhangfu Dong, Yuting He, Xiaoming Qi, Yang Chen, Huazhong Shu, Jean-Louis Coatrieux, Guanyu Yang, and Shuo Li. Mnet: Rethinking 2d/3d networks for anisotropic medical image segmentation. *arXiv preprint arXiv:2205.04846*, 2022.

[4] LiTS: Liver tumor segmentation challenge. `https://competitions.codalab.org/competitions/17094`, 2017.

[5] KiTS23: Kidney tumor segmentation challenge. `https://kits-challenge.org/kits23/`, 2023.

[6] BraTS 2020: Multimodal brain tumor segmentation challenge. `https://www.med.upenn.edu/cbica/brats2020/`, 2020.

[7] PROMISE12: Prostate mr image segmentation 2012 challenge. `https://promise12.grand-challenge.org/`, 2012.

[8] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

Authors' Contributions

State workload distribution across team members; map names to tasks (data, modeling, experiments, writing, infra, QA).

Checklist for Contributions

☐ Clear mapping from members to responsibilities

Appendix A: Training Validation Curves

Loss Curves go here

Appendix B: LLM Prompt Logs and Validation

For each area where LLMs were used (data preprocessing, model implementation, training loop, metrics/evaluation, extension brainstorming and implementation), include:

A.1 Data Preprocessing

Initial Prompt (verbatim):
. . .

Initial Output (excerpt):
...

Validation: How you verified correctness, relevance, helpfulness.

Prompt Count: e.g., 4 messages. If the initial prompt failed, explain what went wrong and how you revised it.

A.2 Model Implementation

Initial Prompt:
. . .

Initial Output:
...

Validation: ...

Prompt Count: ...

A.3 Training Loop

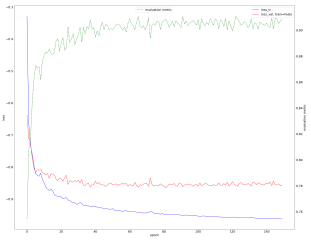
(Repeat the same fields.)

A.4 Metrics and Evaluation

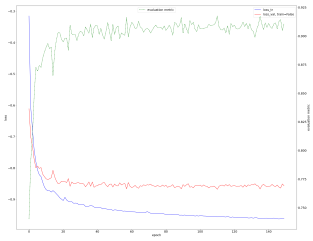
(Repeat the same fields.)

A.5 Extensions/Ablations

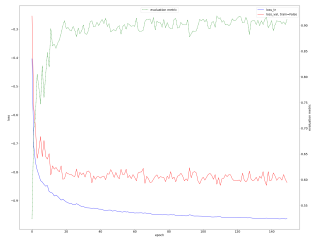
(Repeat the same fields; also include which ideas the LLM suggested and your assessment of validity.)



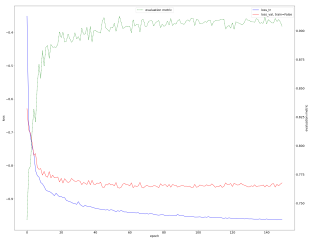
Official — Fold 1



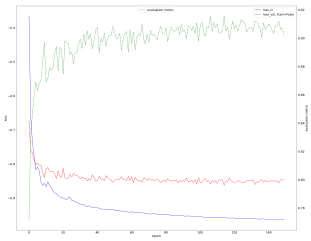
Official — Fold 2



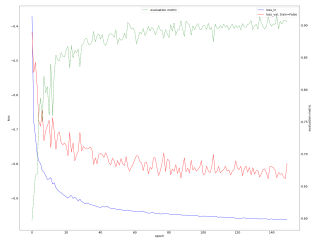
Official — Fold 3



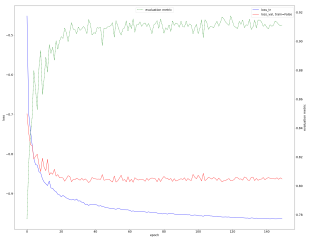
Reproduction — Fold 1



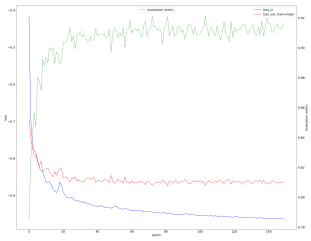
Reproduction — Fold 2



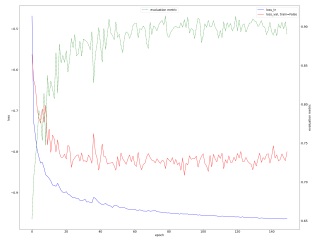
Reproduction — Fold 3



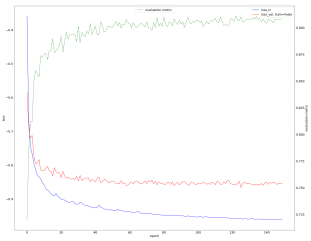
Spatial Gate — Fold 1



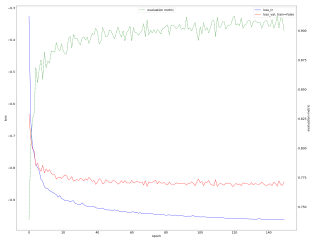
Spatial Gate — Fold 2



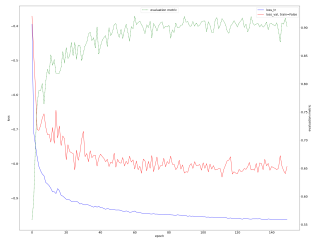
Spatial Gate — Fold 3



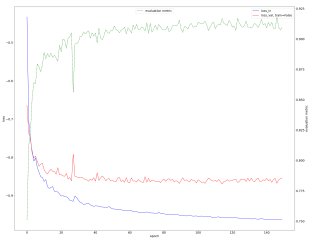
VMamba — Fold 1



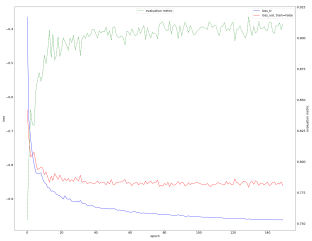
VMamba — Fold 2



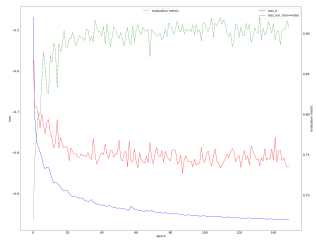
VMamba — Fold 3



Both — Fold 1

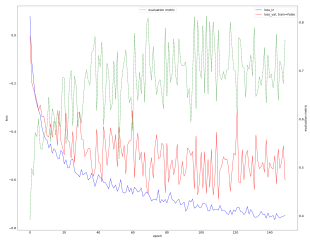


Both — Fold 2

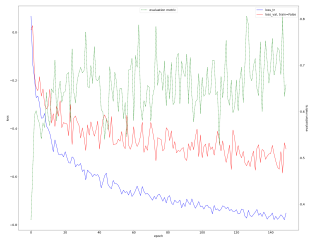


Both — Fold 3

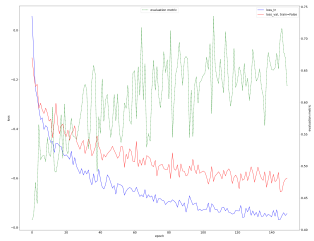
Figure 6: PROMISE Loss for 150 epochs across 3 Folds.



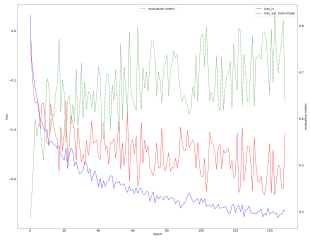
Official — Fold 1



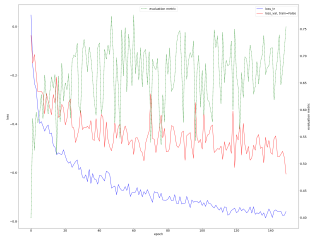
Official — Fold 2



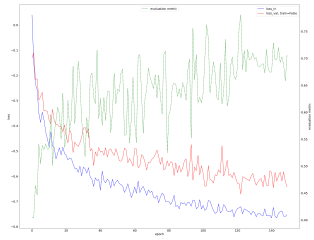
Official — Fold 3



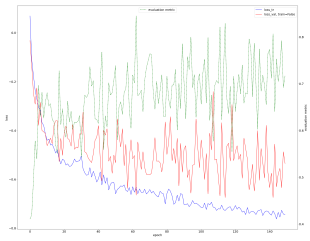
Reproduction — Fold 1



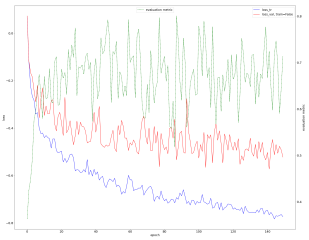
Reproduction — Fold 2



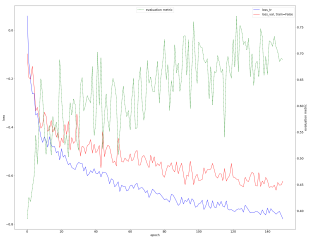
Reproduction — Fold 3



Spatial Gate — Fold 1



Spatial Gate — Fold 2



Spatial Gate — Fold 3

Figure 7: LiTs Loss for 150 epochs across 3 Folds.