Spatially-Adaptive Gradient Re-parameterization for 3D Large Kernel Optimization

Ho Hin Lee[1], Quan Liu[2], Shunxing Bao[1], Yuankai Huo[1], Bennett A. Landman[1]

[1]Department of Computer Science, Vanderbilt University, Nashville, TN, USA; [2]Accenture

ICML International Conference On Machine Learning

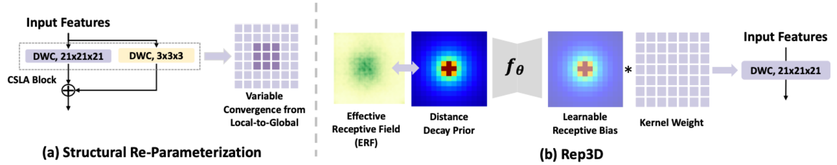
Why Are Large 3D Kernels Hard to Optimize?

Large-kernel CNNs offer an efficient, transformer-free way to capture the large effective receptive fields needed for volumetric medical image segmentation. By scaling depth-wise convolutions to large kernels such as **21 × 21 × 21**, they can model broader 3D context without the high cost of adapting Vision Transformers to dense prediction. However, simply making the kernel larger does not guarantee better performance:

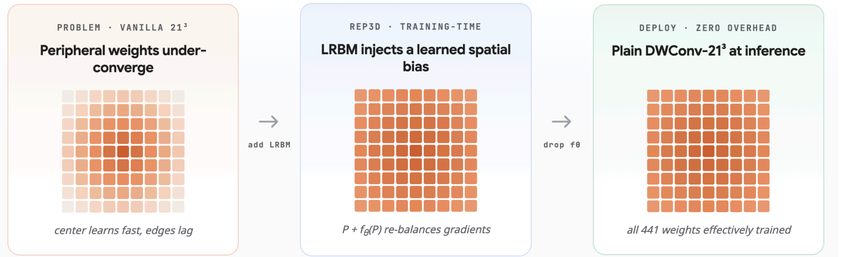
1) Performance saturates or degrades with naïve kernel scaling: Although larger kernels increase the theoretical receptive field, segmentation accuracy often stops improving once encoder kernels become too large.

2) Effective receptive fields are spatially biased: Central kernel elements converge faster and capture local cues, while peripheral elements learn more slowly, limiting effective local-to-global feature learning.

3) Existing re-parameterization methods are not ideal for 3D: Structural re-parameterization methods in natural images use multi-branch designs, such as CSLA blocks, to implicitly create spatially varying learning rates. However, this adds training complexity and does not transfer cleanly to memory-intensive 3D volumetric models.



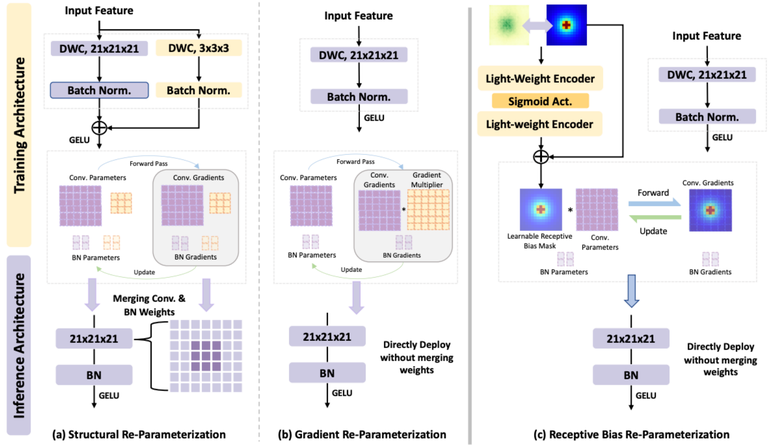
Our Goal: Learn a spatially adaptive optimization bias that makes large 3D kernels practical to train, while scaling up the effective receptive field efficiently for 3D representation learning.



Contributions

1. **We reveal a spatial optimization bias in large 3D kernels**
   - Central kernel elements receive **stronger updates**, while peripheral elements converge **more slowly**.
2. **We propose Rep3D for stable large-kernel 3D segmentation**
   - Rep3D enables effective training of **21 × 21 × 21** depth-wise convolution kernels, provides a scalable way to capture large 3D context
3. **We introduce Low-Rank Receptive Bias Modeling**
   - A learnable mask is generated to adaptively scales kernel elements during training, which turns ERF-inspired local-to-global learning behavior into a learnable optimization prior.

Rep3D: Intuitions



What Does Multi-Branch Re-Parameterization Really Optimize?

Structural re-parameterization trains a multi-branch block but deploys it as a single large convolution. We revisit this design from an optimization perspective. A multi-branch block with a large and small convolution branch can be written as:

$$Y_{multi-branch} = \alpha_L(X \star W_L) + \alpha_S(X \star W_S)$$

Because convolution is linear, the two branches can be merged into one equivalent large kernel:

$$W' = \alpha_L W_L + \alpha_S Pad(W_S)$$

Key Observations Here: The Kernel Center Learns Faster

During training:

- Elements inside the small-kernel region receive gradients from **both branches**
- Peripheral elements receive gradients only from the **large branch**

Thus, the equivalent large kernel has spatially **different update strength**:

$$\lambda_{eff}(P) = \begin{cases} \alpha_L\eta_L + \alpha_S\eta_S, & \mathbb{P} \in \Omega_S \\ \alpha_L\eta_L, & \mathbb{P} \in \Omega_L/\Omega_S \end{cases}$$

Where Ω_S is the central small-kernel region and Ω_L is the full large-kernel support.

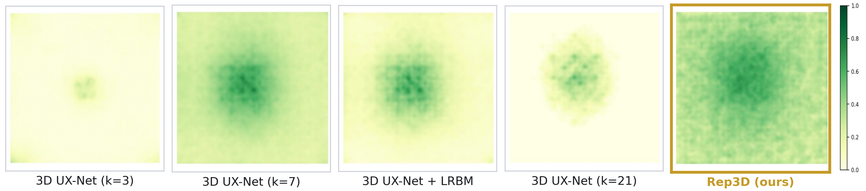
Theoretical Insights for Efficient Large Kernel Learning:

Multi-branch re-parameterization implicitly creates a **spatially non-uniform learning rate**:

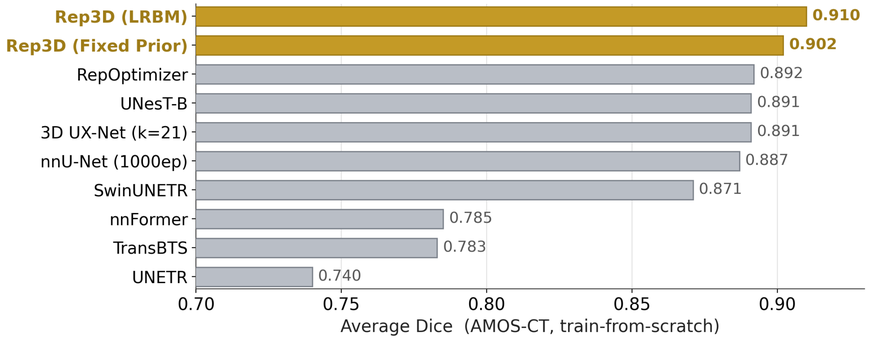
- **Center kernel elements** converge faster, while **peripheral elements** converge slower
- Which matches the local-to-global diffusion pattern of effective receptive fields

Rep3D generalizes this idea: instead of relying on a fixed auxiliary branch, we learn a spatial modulation mask that adaptively controls the optimization strength of each 3D kernel element.

3D Kernels Effective Receptive Field Behavior



Results: Multi-Scale (Organ to Tumor) Segmentation



Still #1 vs. the latest SOTA comparing mean Dice across of all anatomical semantics all five benchmarks:

Method	AMOS-CT	AMOS-MRI	KiTS	Panaceas & Tumor	Hepatic & Tumor
nnU-Net (1000 ep)	0.887	0.847	0.706	0.703	0.660
ResEnc nnU-Net	0.892	0.850	0.711	0.706	0.661
STU-Net-H	0.900	0.848	0.707	0.712	0.648
MedNeXt	0.897	0.856	0.720	0.713	0.663
U-Mamba	0.904	0.859	0.729	0.716	0.665
Rep3D (Ours)	**0.910**	**0.864**	**0.736**	**0.723**	**0.674**

What have we learnt from Rep3D?

1) **Large 3D kernels are powerful, but not automatically effective:** Simply increasing kernel size expands the theoretical receptive field but can lead to saturated or even degraded optimization.
2) **Multi-branch design provides an insight of converging different speeds across kernel elements .** Central regions learn faster and capture local structure, while peripheral regions are harder to optimize for global context.
3) **Adapt learnable spatial bias directly during weight optimization.** Instead of relying on fixed multi-branch design, a learnable receptive-bias prior knowledge can be generated to guide local-to-global kernel optimization.