XConv: Low-memory stochastic backpropagation for convolutional layers

Anonymous authors
Paper under double-blind review

Abstract

Training convolutional neural networks at scale demands substantial memory, largely due to storing intermediate activations for backpropagation. Existing approaches—such as checkpointing, invertible architectures, or gradient approximation methods like randomized automatic differentiation—either incur significant computational overhead, impose architectural constraints, or require non-trivial codebase modifications. We propose XConv, a drop-in replacement for standard convolutional layers that addresses all three limitations: it preserves standard backpropagation, imposes no architectural constraints, and integrates seamlessly into existing codebases. XConv exploits the algebraic structure of convolutional layer gradients, storing highly compressed activations and approximating weight gradients via multi-channel randomized trace estimation. We establish convergence guarantees and derive error bounds for the proposed estimator, showing that the variance of the resulting gradient errors is comparable to that of stochastic gradient descent. Empirically, XConv achieves performance comparable to exact gradient methods across classification, generative modeling, super-resolution, inpainting, and segmentation—with gaps that narrow as the number of probing vectors increases—while reducing memory usage by a factor of two or more and remaining computationally competitive with optimized convolution implementations.

1 Introduction

While transformer-based architectures have achieved remarkable success across numerous domains, their quadratic computational complexity with respect to sequence length poses significant scalability challenges (Vaswani et al., 2017; Tay et al., 2022). This has renewed interest in convolutional architectures (Younesi et al., 2024), with recent works demonstrating that carefully designed convolutional neural networks (CNNs) can match or exceed transformer performance in certain settings (Mao et al., 2021; Liu et al., 2022; Cui et al., 2023b;a; Woo et al., 2023; Wang et al., 2023; Liu et al., 2023; Hou et al., 2024; Ding et al., 2024). As such, convolutional layers continue to form a key component of current neural network designs (Ronneberger et al., 2015; Ding et al., 2022; Liu et al., 2022; Wang et al., 2024; Ma et al., 2024b). However, scaling up CNNs remains challenging due to two primary factors: (i) while the computational demands during forward evaluation are relatively modest, significant computational resources are needed during training (Griewank & Walther, 2000; Ronneberger et al., 2015; Chen et al., 2016); (ii) training requires storing intermediate activations for backpropagation, creating a memory bottleneck that becomes particularly acute when scaling to higher-dimensional data. While several recent works have addressed the computational complexity of CNN training through architectural innovations (Howard et al., 2017; Liu et al., 2022; Chen et al., 2023a;b; Vasu et al., 2023), the memory bottleneck associated with storing activations remains a critical challenge.

Existing methods for addressing this memory bottleneck include checkpointing approaches (Griewank & Walther, 2000; Chen et al., 2016; Beaumont et al., 2019; Shah et al., 2021; Feng & Huang, 2021; He & Yu, 2023; Korthikanti et al., 2023; Beaumont et al., 2024; Hong et al., 2025), which recompute activations during the backward pass, yielding exact gradients but incurring significant computational overhead and requiring careful integration with automatic differentiation to ensure correct gradient flow; invertible network architectures (Gomez et al., 2017; Haber & Ruthotto, 2017; Jacobsen et al., 2018; Hascoet et al., 2023; Ulidowski, 2023; Orozco et al., 2024; Zhang & Zhang, 2024; Zhao et al., 2024a), which enable activations

to be recovered from outputs but impose strict architectural constraints that limit representation power; and approximation methods that compute unbiased gradient estimates via techniques such as randomized automatic differentiation (RAD) (Oktay et al., 2021) which intervenes in the computational graph, zeroth-order methods (Malladi et al., 2023) that bypass backpropagation entirely by perturbing parameters, approximate backpropagation (Yang et al., 2024b) which modifies the backward pass with shared approximations, or direct feedback alignment (DFA) (Nøkland, 2016; Han & Yoo, 2019; Wang et al., 2019; Nøkland & Eidnes, 2019; Launay et al., 2020; Frenkel et al., 2021; Refinetti et al., 2021; Nakajima et al., 2024; Yang et al., 2024a) which bypasses the backward pass by routing error signals through fixed random matrices directly to each layer. Collectively, each family of methods trades one constraint for another: checkpointing sacrifices computation for exact gradients, invertible architectures sacrifice design flexibility, and approximation-based approaches require non-trivial codebase modifications, specialized framework support, or changes to the training pipeline. A method that reduces memory while preserving standard backpropagation, imposing no architectural constraints, and integrating seamlessly into existing codebases remains an open challenge.

Our work is based on the premise that exact computations are often not needed—a viewpoint advocated in the field of randomized linear algebra (Tropp et al., 2019; Martinsson & Tropp, 2020), and more recently in parametric machine learning (Oktay et al., 2021), where it has been argued that spending computational resources on exact gradients is unnecessary when stochastic optimization is used. A similar argument was used earlier in the context of parameter estimation with partial differential equation constraints (Haber et al., 2012; Aravkin et al., 2012; van Leeuwen & Herrmann, 2014). Building on this premise, we propose **XConv**, a memory-efficient training approach for convolutional layers that addresses all three limitations simultaneously. XConv preserves standard backpropagation—unlike DFA or zeroth-order methods—imposes no architectural constraints—unlike invertible networks—and requires no changes to the computational graph or training pipeline—unlike RAD or checkpointing. This is achieved by exploiting the specific algebraic structure of convolutional layer gradients: we store highly compressed versions of layer activations and approximate weight gradients using multi-channel randomized trace estimation, enabling XConv to serve as a drop-in replacement for standard 2D and 3D convolutional layers in any existing architecture while remaining computationally competitive with optimized convolution implementations.

By means of relatively straightforward algebraic manipulations, we write the gradient with respect to a convolution weight in terms of the trace of a matrix formed from the outer product of the convolutional layer input and the backpropagated residual, combined with a shift operation. Next, we approximate this trace with an unbiased randomized trace estimation technique (Hutchinson, 1989; Avron & Toledo, 2011; Roosta-Khorasani & Ascher, 2015; Martinsson & Tropp, 2020; Meyer et al., 2021) for which we prove convergence and derive theoretical error bounds by extending recent theoretical results (Cortinovis & Kressner, 2022). We show that exact convolution gradients are not strictly necessary: randomized gradient estimates yield noise comparable in scale to stochastic optimization noise while reducing memory consumption, with accuracy that improves systematically with the number of probing vectors.

Contributions. The main contributions of this work are:

- We propose XConv, a drop-in replacement for standard convolutional layers that approximates gradients via multi-channel randomized trace estimation, enabling memory reduction with minimal implementation overhead.
- We establish convergence guarantees and derive theoretical error bounds for the proposed estimator, extending existing results to non-symmetric matrices.
- We empirically demonstrate that XConv achieves performance comparable to exact gradient methods across classification, generative modeling, super-resolution, inpainting, and segmentation, with accuracy that improves systematically with the number of probing vectors, while meaningfully reducing memory consumption.

Outline. Section 2 reformulates the gradient of convolution layers as a trace and establishes convergence guarantees for its randomized approximation. Section 3 presents the XConv algorithm and describes how it

integrates into existing architectures as a drop-in replacement. Section 4 evaluates gradient fidelity, memory savings, and downstream task performance across a range of architectures and applications.

2 Theory

We cast the action of convolutional layers into a framework that exposes the underlying linear algebra, allowing gradients with respect to convolution weights to be identified as traces of a matrix. These traces can then be approximated by randomized trace estimation (Avron & Toledo, 2011), which greatly reduces the memory footprint with competitive computational overhead. We derive expressions for the single-channel case, prove convergence and error bounds by extending existing results to non-symmetric matrices, and then generalize to multi-channel convolutions. The multi-channel case calls for a new type of sparse probing to minimize crosstalk between channels, for which we also derive accuracy bounds.

Single channel case Let us start by writing the action of a single channel convolutional layer as follows

$$\mathbf{Y} = \mathbf{W}\mathbf{X} \in \mathbb{R}^{N\times B}, \quad \text{where} \quad \mathbf{W} = \sum_{i=1}^{n_w} w_i \mathbf{T}_{k(i)}, \tag{1}$$

and N, B, n_w are the number of pixels, batch size, and number of convolution weights (K^2 for a K by K kernel), respectively. For the i^{th} weight w_i, the convolutions themselves correspond to applying a circular shift with offset $k(i)$, denoted by $\mathbf{T}_{k(i)}$, followed by multiplication with the weight. Given this expression for the action of a single-channel convolutional layer, expressions for the gradient with respect to weights can easily be derived by using the chain rule and standard linear algebra manipulations (Petersen & Pedersen, 2008)—i.e., we have

$$\begin{aligned}\frac{\partial}{\partial w_i} f(\mathbf{W}\mathbf{X}) &= \operatorname{tr}\left(\left(\frac{\partial f(\mathbf{W}\mathbf{X})}{\partial \mathbf{W}}\right)^\top \frac{\partial \mathbf{W}}{\partial w_i}\right) \\ &= \operatorname{tr}\left(\left(\delta\mathbf{Y}\mathbf{X}^\top\right)^\top \mathbf{T}_{k(i)}^\top\right) = \operatorname{tr}\left(\mathbf{X}\delta\mathbf{Y}^\top\mathbf{T}_{-k(i)}\right), \quad i = 1,\dots,n_w.\end{aligned} \tag{2}$$

This expression for the gradient with respect to the convolution weights corresponds to computing the trace—i.e., the sum of the diagonal elements denoted by $\operatorname{tr}(\mathbf{A}) = \sum_i A_{ii}$, of the outer product between the residual collected in $\delta\mathbf{Y}$ and the layer's input $\mathbf{X}$, after applying the shift. The latter corresponds to a right circular shift along the columns.

Computing estimates for the trace through the action of matrices—i.e., without access to entries of the diagonal, is common practice in the emerging field of randomized linear algebra (Tropp et al., 2019; Martinsson & Tropp, 2020). Going back to the seminal work by Hutchinson (Hutchinson, 1989; Meyer et al., 2021), unbiased matrix-free estimates for the trace of a matrix exist involving probing with random vectors $\mathbf{z}_j$, $j = 1,\dots,r$, with r the number of probing vectors and $\mathbb{E}(\mathbf{z}_j\mathbf{z}_j^\top) = \mathbf{I}$ with $\mathbf{I}$ the identity matrix. Under this assumption, unbiased randomized trace estimates can be derived from

$$\operatorname{tr}(\mathbf{A}) = \operatorname{tr}\left(\mathbf{A}\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top\right]\right) = \mathbb{E}\left[\mathbf{z}^\top\mathbf{A}\mathbf{z}\right] \approx \frac{1}{r}\sum_{j=1}^{r}\left[\mathbf{z}_j{}^\top\mathbf{A}\mathbf{z}_j\right]. \tag{3}$$

By combining equation 2 with the above unbiased estimator for the trace, we arrive at the following approximation for the gradient with respect to the convolution weights:

$$\delta w_i \approx \frac{1}{r}\sum_{j=1}^{r}\left(\mathbf{z}_j^\top\mathbf{X}\right)\left(\delta\mathbf{Y}^\top\mathbf{T}_{-k(i)}\mathbf{z}_j\right), \quad i = 1,\dots,n_w. \tag{4}$$

From this expression the memory savings during the forward pass are obvious since $\overline{\mathbf{X}} = \mathbf{Z}^\top\mathbf{X}$, where $\overline{\mathbf{X}} \in \mathbb{R}^{r\times B}$ with $r \ll N$. However, convergence rate guarantees were only established under the additional assumption that $\mathbf{A}$ is positive semi-definite (PSD, Kaperick (2019)). While the outer product $\mathbf{X}\delta\mathbf{Y}^\top\mathbf{T}_{-k(i)}$

we aim to probe here is not necessarily PSD, improving upon recent results by Cortinovis & Kressner (2022), we show that the condition of PSD can be relaxed to asymmetric matrices by a symmetrization procedure that does not change the trace. More precisely, we show in the following proposition that the gradient estimator in equation 4 is unbiased and converges to the true gradient as $r \to \infty$ with a rate of about $r^{-1/2}$ (for details of the proof, we refer to the Appendix A.2).

Proposition 1. *Let $\mathbf{A} \in \mathbb{R}^{N\times N}$ be a square matrix and let the probing vectors be i.i.d. Gaussian with 0 mean and unit variance. Then for any small number $\delta > 0$, with probability $1-\delta$, we have*

$$\left| \frac{1}{r} \sum_{i=1}^{r} \left[\mathbf{z}_i^\top \mathbf{A} \mathbf{z}_i \right] - \operatorname{tr}(\mathbf{A}) \right| \leq \frac{4\|\mathbf{A}\|_2}{r} \log \frac{2}{\delta} + \frac{2\|\mathbf{A}\|_F}{\sqrt{r}} \log^{1/2} \frac{2}{\delta}.$$

Imposing a small probability of failure δ means the $\log \frac{2}{\delta}$ term in the upper bound is large, which implies that neither term in the upper bound is dominating for all the r values. Depending on which term is dominant, the range of r can be divided into two regimes, the small r regime and the large r regime. In the small r regime, the first term dominates, and the error decays as $1/r$. In the large r regime, the second term dominates and the error decays as $1/\sqrt{r}$. The phase transition happens when r is about $4\|\mathbf{A}\|_2^2/\|\mathbf{A}\|_F^2 \log(2/\delta) \equiv \frac{4}{\rho}\log(2/\delta)$, where $\rho \equiv \|\mathbf{A}\|_F^2/\|\mathbf{A}\|_2^2$ is known as the effective rank, which reflects the rate of decay of the singular values of $\mathbf{A}$. We see that as r increases, the larger the effective rank is, the earlier the phase transition occurs, after which the decay rate of the error will slow down. Before discussing details of the proposed algorithm, let us first extend the above randomized trace estimator to multi-channel convolutions.

Multi-channel case In general, convolutional layers involve several input and output channels. In that case, the output of the m^{th} channel can be written as

$$\mathbf{Y}^m = \sum_{n=1}^{C_{\text{in}}} \mathbf{W}^{n,m} \mathbf{X}^n, \quad \text{where} \quad \mathbf{W}^{n,m} = \sum_{i=1}^{n_w} w_i^{n,m} \mathbf{T}_{k(i)} \tag{5}$$

for $n = 1, \ldots, C_{\text{in}},\ m = 1, \ldots, C_{\text{out}}$ with C_{in}, C_{out} the number of input and output channels and $w_i^{n,m}$ the i^{th} weight between the n^{th} input and m^{th} output channel. In this multi-channel case, the gradients consist of the single channel gradient for each input/output channel pair, i.e., $\delta w_i^{n,m} = \operatorname{tr}(\mathbf{X}^n(\delta \mathbf{Y}^m)^\top \mathbf{T}_{-k(i)})$.

While randomized trace estimation can in principle be applied to each input/output channel pair independently, we propose to treat all channels simultaneously to further improve computational performance and memory use. Let the outer product of the $(n,m)^{\text{th}}$ input/output channel be $\mathbf{A}^{n,m}$, i.e., $\mathbf{A}^{n,m} = (\mathbf{X}^n(\delta \mathbf{Y}^m)^\top \mathbf{T}_{-k(i)})^\top$, computing $\delta w_i^{n,m}$ means estimating $\operatorname{tr}(\mathbf{A}^{n,m})$. To save memory, instead of probing each $\mathbf{A}^{n,m}$, we probe the stacked matrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}^{1,1} & \cdots & \mathbf{A}^{1,C_{\text{in}}} \\ \vdots & & \vdots \\ \mathbf{A}^{C_{\text{out}},1} & \cdots & \mathbf{A}^{C_{\text{out}},C_{\text{in}}} \end{pmatrix}$$

by r length $C_{\text{in}} \times N$ probing vectors stored in $\mathbf{Z} \in \mathbb{R}^{NC_{\text{in}}\times r}$, and estimate each $\operatorname{tr}(\mathbf{A}^{m,n})$ via the following estimators

$$G^{m,n}(\mathbf{A}) := \frac{1}{r} \sum_{j=1}^{r} \mathbf{z}_{n,j}^\top \left(\mathbf{A}\mathbf{z}_j\right)_m, \tag{6}$$

where $(\cdot)_m$ extracts the m^{th} block from the input vector. That is to say, we simply stack the input and residual, yielding matrices of size $(N \times C_{\text{in}}) \times B$ and $(N \times C_{\text{out}}) \times B$ whose outer product $\mathbf{X}\delta\mathbf{Y}^\top$ (i.e., $\mathbf{A}^\top$ of the $\mathbf{A}$ in equation 6) is no longer necessarily square. To estimate the trace of each $N \times N$ sub-block, in equation 6, we (*i*) probe the full outer product from the right with r probing vectors $\mathbf{z}_j$ of length $(N \times C_{\text{in}})$; (*ii*) reshape the resulting matrix into a tensor of size (N, C_{out}, B) while the probing matrix is shaped into a tensor of size (N, C_{in}, B) (i.e., separate each block of $\mathbf{A}z_j$), and (*iii*) probe each individual block again from the left. This leads to the desired gradient collected in a $C_{\text{in}} \times C_{\text{out}}$ matrix. We refer to Figure 1a, which illustrates this multi-channel randomized trace estimation. After (*i*), we only need to save $\overline{\mathbf{X}} = \mathbf{Z}^\top \mathbf{X}$ in memory rather than $\mathbf{X}$ that leads to a memory reduction by a factor of $\frac{NC_{\text{in}}}{r}$.

Unfortunately, the improved memory use and computational performance boost of the above multi-channel probing reduces the accuracy of the randomized trace estimation because of crosstalk amongst the channels. Since this crosstalk is random, the induced error can be reduced by increasing the number of probing vectors r, but this will go at the expense of more memory use and increased computation. To avoid this unwanted overhead, we introduce a new type of random probing vectors that minimizes the crosstalk by again imposing $\mathbb{E}(\mathbf{z}\mathbf{z}^\top) = \mathbf{I}$ but now on the multi-channel probing vectors that consist of multiple blocks corresponding to the number of input/output channels.

Explicitly, we draw each $\mathbf{z}_{n,j}$, the n^{th} block of the j^{th} probing vector, according to

$$\mathbf{z}_{n,j} \sim \begin{cases} \mathcal{N}(\mathbf{0}, \mathbf{I}_N) & \text{with probability } p_n \\ 0 & \text{with probability } 1 - p_n \end{cases}. \tag{7}$$

For different values of (n, j), the $\mathbf{z}_{n,j}$'s are drawn independently with a predefined probability p_n of generating a nonzero block. Compared to conventional (Gaussian) probing vectors (see Figure 1b top left), these multi-channel probing vectors contain sparse non-zero blocks (see Figure 1b top right), which reduces the crosstalk (juxtapose with second row of Figure 1b). It can be shown that crosstalk becomes less when $p_n \downarrow 0$ and $r \to \infty$.

Given probing vectors drawn from equation 7, we have to modify the scaling factor of the multi-channel randomized trace estimator equation 6 to ensure it is unbiased,

$$\widetilde{G}^{n,m}(\mathbf{A}) := \frac{1}{\text{nnz}(\mathbf{Z}_n)} \sum_{j=1}^{r} \mathbf{z}_{n,j}^\top \left(\mathbf{A}\mathbf{z}_j\right)_m, \tag{8}$$

where $\text{nnz}(\mathbf{Z}_n)$ is the number of non-zero columns in block n. We prove the following convergence result for this estimator (the proof can be found in Appendix A.2).

Theorem 1 (Succinct version)**.** *Let $p = \min_n p_n$, r be the number of probing vectors. For any small number $\delta > 0$, with probability at least $1 - \delta - 3C_{in}e^{-rp^2/2}$, we have for any $n = 1, \ldots, C_{in}$ and $m = 1, \ldots, C_{out}$,*

$$\left|\widetilde{G}^{n,m}(\mathbf{A}) - \text{tr}(\mathbf{A}^{n,m})\right| \leq c \cdot \frac{\frac{1}{\sqrt{p_n}}\|\mathbf{A}^{n,m}\|_F + \sum_{j=1, j\neq n}^{C_{in}} \sqrt{\frac{p_j}{p_n}}\|\mathbf{A}^{n,j}\|_F}{\sqrt{r}} \log^{1/2} \frac{C_{out}C_{in}}{\delta},$$

where c is an absolute constant and C_{in} and C_{out} are the numbers of input and output channels.

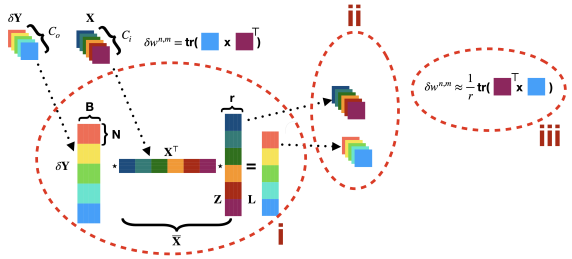
Theorem 1 provides convergence guarantee for our special multi-channel simultaneous probing procedure. Similar to Proposition 1, Theorem 1 in its original form (supplementary material) also has a two-phase behaviour. So the discussion under Proposition 1 applies here. For simplification of presentation, we only presented the bound for the large r regime in this succinct version. Still, we can see that the error bound for estimating $\text{tr}(\mathbf{A}^{n,m})$ not only depends on the norm of the current block $\mathbf{A}^{n,m}$, but also other blocks in that row, which is expected since we simultaneously probe the entire row instead of each block individually for memory efficiency. Admittedly, due to technical difficulties, we cannot theoretically show that decreasing the sampling probability p decreases the error. Nevertheless, we observe better performance in the numerical experiments.

3 Stochastic optimization with multi-channel randomized trace estimation

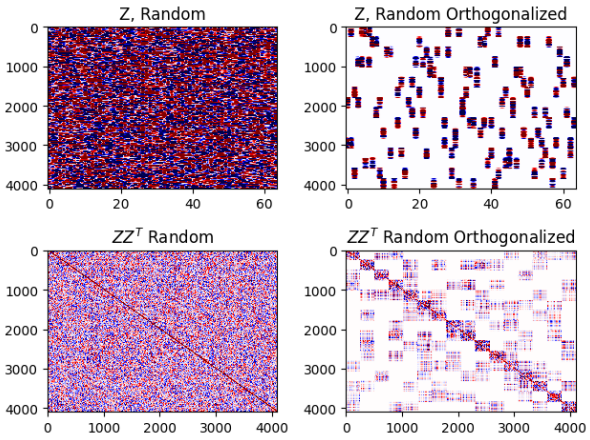
Given the expressions for approximate gradient calculations and bounds on their error, we now present the XConv algorithm and demonstrate its use as a drop-in replacement in existing convolutional architectures.

3.1 Low-memory stochastic backpropagation

The key point of the randomized trace estimator in Equation equation 8 is that it allows for on-the-fly compression of the state variables during the forward pass. For a single convolutional layer $\mathbf{Y} = \mathbf{conv}(\mathbf{X}, \mathbf{w})$



(a) Multi-channel randomized trace estimation. The three steps of the algorithm estimate the trace of a sub-block of the outer product.



(b) Probing matrices and corresponding approximation of the identity for $C_{\text{in}} = 16$, $C_{\text{out}} = 16$, $r = 64$, and $N = 256$. The orthogonalization step by zeroing out blocks (compare plots in top row) leads to a near block diagonal approximation of the identity with much less crosstalk between the different channels (compare plots in second row).

Figure 1: Multi-channel probing illustration and effect of orthogonalization. (a) Schematic of the three-step trace estimation procedure for a single sub-block. (b) Probing matrices $\mathbf{Z}$ and their Gram matrices $\mathbf{Z}^\top\mathbf{Z}$ before (left) and after (right) block orthogonalization, showing reduced cross-channel interference.

with input $\mathbf{X}$ and convolution weights $\mathbf{w}$, our approximation involves three simple steps, namely **(1)** probing of the state variable $\overline{\mathbf{X}} = \mathbf{Z}^\top\mathbf{X}$, **(2)** matrix-free formation of the outer product $\mathbf{L} = \overline{\mathbf{X}}\delta\mathbf{Y}^\top$, and **(3)** approximation of the gradient via $\delta w_i = \frac{1}{r}\operatorname{tr}(\mathbf{L}\mathbf{T}_{-k(i)}\mathbf{Z})$, $i = 1, \ldots, n_w$. These three steps lead to substantial memory reductions even for a relatively small image of size 32×32 ($N = 1024$ pixels) and $C_{\text{in}} = 16$. In that case, our approach leads to a memory reduction by a factor of $\frac{NC_{\text{in}}}{r} = 2^{14-\gamma}$ for $r = 2^\gamma$. For $\gamma = 7$ this leads to roughly $100\times$ memory saving. Because the probing vectors are generated on the fly, we only need to allocate memory for $\overline{\mathbf{X}}$ during the forward pass as long as we also store the state s of the random generator. During backpropagation, we initialize the state, generate the probing vectors, followed by applying a shift and product by $\mathbf{L}$. These steps are summarized in Algorithm 1. This simple algorithm provides a highly compressed estimate of the true gradient with respect to its weights.

Algorithm 1 Low-memory approximate gradient convolutional layer. The random seed s and random probing matrix $\mathbf{Z}$ are independently redrawn for each layer and training iteration.

Forward pass:

1. Forward convolution $\mathbf{Y} = \mathbf{conv}(\mathbf{X}, \mathbf{w})$
2. Draw a new random seed s and probing matrix $\mathbf{Z}[s]$
3. Compute and save $\overline{\mathbf{X}} = \mathbf{Z}^\top[s]\mathbf{X} \in \mathbb{R}^{r\times B}$
4. Store $\overline{\mathbf{X}}, s$

Backward pass:

1. Load random seed s and probed forward $\overline{\mathbf{X}}$
2. Redraw probing matrix $\mathbf{Z}[s]$ from s
3. Compute backward probe $\mathbf{L} = \overline{\mathbf{X}}\delta\mathbf{Y}^\top$
4. Compute gradient $\delta w_i = \frac{1}{r}\operatorname{tr}(\mathbf{L}\mathbf{T}_{-k(i)}\mathbf{Z}[s])$

(a) Directly instantiating XConv layers.

```
class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        self.conv1 = Xconv2D(
            1, 32, 3, 32, 1, padding=1
        )

        self.conv2 = Xconv2D(
            32, 64, 3, 32, 1, padding=1
        )
        self.dropout1 = nn.Dropout(0.25)
        self.dropout2 = nn.Dropout(0.5)
        self.fc2 = nn.Linear(128, 10)
```

(b) Converting an existing model in-place.

```
from pyxconv.utils import convert_net

model = nn.Sequential(
    nn.Conv2d(3, 64, 3, padding=1),
    nn.BatchNorm2d(64),
    nn.ReLU()
)

convert_net(
    model,
    ps=256,
    xmode="independent"
)
```

Figure 2: XConv can be integrated either by directly instantiating XConv layers (left) or by converting existing convolutional models using a single API call (right).

3.2 Ease of Integration

Figure 2 illustrates the ease of integrating XConv: it implements Algorithm 1 internally but fully encapsulates the complexity within the layer abstraction. We also provide an adaptive version that selectively disables XConv in deep layers with small spatial extent, where activation storage is inexpensive.

4 Experiments

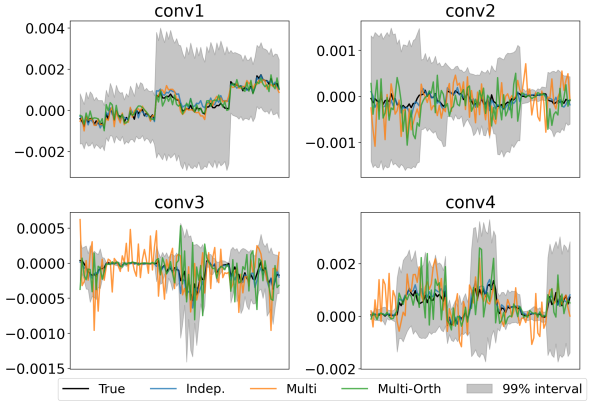
We now systematically analyze the consequences of replacing standard convolutional layers with XConv on gradient fidelity, memory consumption, and computational overhead. We first perform a layer-by-layer analysis of gradient deviation, then introduce Average Gradient Error (AGE) as a global diagnostic metric to quantify aggregate gradient fidelity across entire models. We then relate these findings to peak memory usage and wall-clock benchmarks, enabling a comprehensive evaluation of the accuracy–memory–compute trade-offs induced by XConv.

4.1 Minibatch versus randomized trace estimation errors

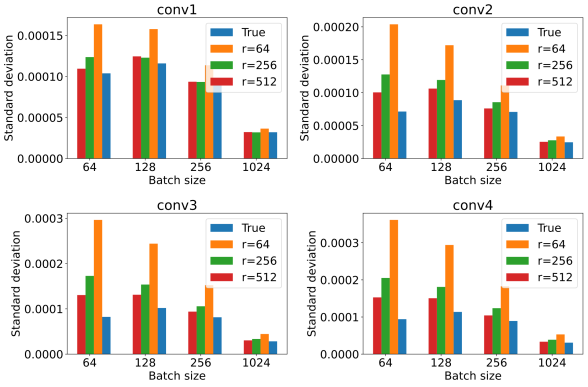
Simply stated, stochastic optimization involves gradients that contain random errors known as gradient noise. As long as this noise is not too large and independent for different gradient calculations, algorithms such as stochastic gradient descent, where gradients are computed for randomly drawn minibatches, converge under certain conditions. In addition, the presence of gradient noise helps the algorithm to avoid bad local minima, which arguably leads to better generalization of the trained network (Neelakantan et al., 2015; Huang et al., 2020). Therefore, as long as the batch size is not too large, one can expect the trained network to perform well.

We argue that the same applies to stochastic optimization with gradients approximated by (multi-channel) randomized trace estimation as long as the errors behave similarly. In a setting where memory comes at a premium this means that we can expect training to be successful for gradient noise with similar variability. To this end, we conduct an experiment where the variability of $5 \times 5 \times C_{\text{in}} \times C_{\text{out}}$ convolution weights is calculated for the true gradient for different randomly drawn minibatches of size $B = 128$. We do this for a randomly initialized image classification network designed for the CIFAR-10 dataset (for network details, see Table 5 in Appendix A.5).

For comparison, approximate gradients are also calculated for randomized trace estimates obtained by probing independently ("Indep." in blue), multi-channel ("Multi" in orange), and multi-channel with orthogonalization ("Multi-Ortho" in green). The batch sizes are for a fixed number of probing vectors of $r = 2048$ selected such that the total memory use is the same as for the true gradient calculations. From the plots in Figure 3a, we



(a) Randomized trace estimation of the gradient of our randomly initialized CNN for the CIFAR-10 dataset. While gradient noise is present, its magnitude is reduced by the orthogonalization when weights are small.



(b) Standard deviation of the gradients w.r.t. the weights for each of the four convolutional layers in the neural network. The standard deviation is computed over 40 mini-batches randomly drawn from the CIFAR-10 dataset.

Figure 3: Gradient analysis on CIFAR-10. (a) Per-weight gradient estimates for four convolutional layers under different probing strategies. (b) Standard deviation of gradients across 40 mini-batches for varying batch sizes and probing vectors r.

observe that as expected the independent probing is close to the true gradient followed by the more memory efficient multi-channel probing with and without orthogonalization. While all approximate gradients are within the 99% confidence interval, the orthogonalization has a big effect when the gradients are small (see conv3).

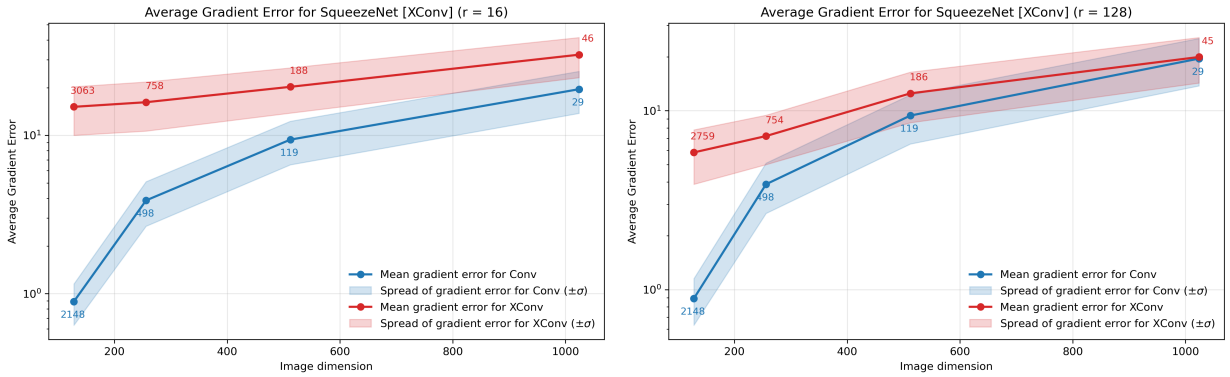
To better understand the interplay between batch size $B \in \{64, 128, 256, 1024\}$ and probing vectors $r \in \{64, 256, 512\}$, we estimate the standard deviation from 40 randomly drawn minibatches. As expected, the standard deviations increase for smaller batch size and fewer probing vectors. However, since XConv's smaller memory footprint allows for larger batch sizes, the variability can be controlled within a given memory budget. Figure 3b shows that the standard deviation reduces with increasing batch size, mirroring the behaviour of stochastic gradient descent. This confirms that the approximation noise introduced by XConv does not dominate mini-batch noise, reaffirming that exact gradient computations are unnecessary for stable training.

4.2 Average Gradient Error

Comparing approximate and exact gradients in deep networks is nontrivial: gradients vary across layers, scales, and architectures, and per-layer discrepancies do not directly reflect the cumulative effect on optimization. To enable architecture-level comparison, we introduce a global diagnostic metric—Average Gradient Error (AGE)—that quantifies the aggregate deviation between gradients estimated by standard convolution and randomized trace estimation across an entire model. AGE measures the magnitude of gradient noise induced by approximation and stochasticity, aggregated across mini-batches, and is intended as a diagnostic of gradient fidelity rather than a predictor of generalization.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^K$ denote a dataset of K samples, $\ell(f_\theta(x_i), y_i)$ a per-example loss, and $L(\theta) = \frac{1}{K}\sum_{i=1}^K \ell(f_\theta(x_i), y_i)$ the empirical risk with full gradient $\mathbf{g}(\theta) = \nabla_\theta L(\theta)$. We partition $\mathcal{D}$ into $M = \lfloor K/B \rfloor$ mini-batches $\{\mathcal{B}_b\}_{b=1}^M$ of size B, each yielding a mini-batch gradient $\mathbf{g}^{(b)}(\theta) = \frac{1}{B}\sum_{i\in\mathcal{B}_b} \nabla_\theta \ell(f_\theta(x_i), y_i)$ so that $\mathbf{g}(\theta) = \frac{1}{M}\sum_{b=1}^M \mathbf{g}^{(b)}(\theta)$. The Average Gradient Error is then defined as

$$\mathrm{AGE}(\theta) = \frac{1}{M}\sum_{b=1}^{M} \left\| \mathbf{g}(\theta) - \mathbf{g}^{(b)}(\theta) \right\|_2^2. \tag{9}$$



(a) Probing vector r=16

(b) Probing vector r=128

Figure 4: Average gradient error vs. image dimension for SqueezeNet. Results are shown for probing vectors $r \in \{16, 128\}$ over 10 runs. Bold numbers indicate the maximum batch-size that fits in memory for each method (blue: standard convolution, red: XConv) under a fixed 16 GB memory budget. Across all probing-vector settings, XConv consistently permits larger batch sizes than standard convolution. The AGE is higher for XConv, but the gap reduces progressively with increasing image dimension N, becoming small at $N = 1024$ for probing vector $r = 128$.

For the randomized trace estimation, the mini-batch gradient given by $\mathbf{g}^{(b)}$ contains the stochastic mini-batch and approximation noise. For both the standard convolution and randomized trace estimation cases, AGE is measured relative to the same reference full-gradient $\mathbf{g}(\theta)$. We use AGE to compare gradient fidelity across models and approximation strategies, isolating the contribution of randomized trace estimation from stochastic mini-batch noise.

4.2.1 SqueezeNet

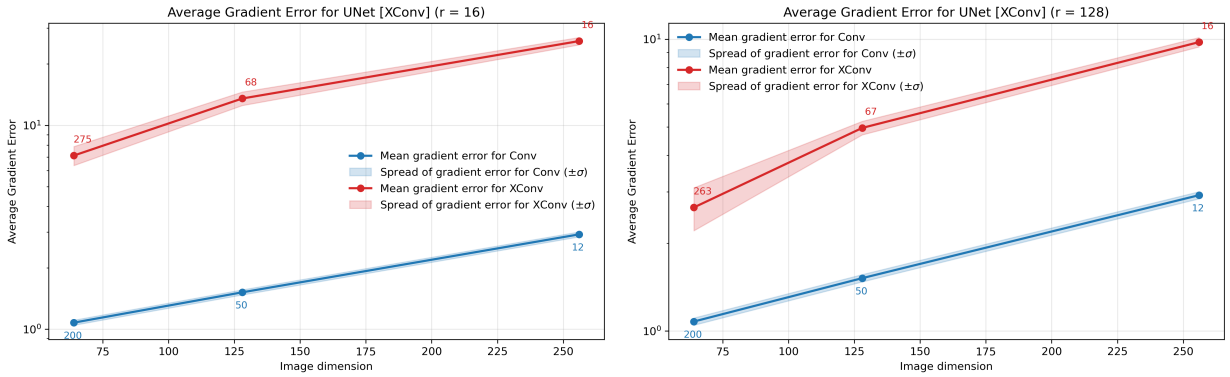
We begin our analysis with SqueezeNet (Iandola et al., 2016), a lightweight convolutional architecture designed to achieve competitive accuracy with significantly fewer parameters. SqueezeNet constitutes a challenging test case for XConv: its reduced spatial aggregation limits natural averaging effects that can otherwise mitigate stochastic gradient noise, potentially amplifying the impact of gradient approximation.

We evaluate AGE across probing vectors $r \in \{16, 32, 64, 128\}$ and image dimensions $N \in \{128, 256, 512, 1024\}$. Results for $r \in \{16, 128\}$ are reported in Figure 4, with additional probing vector configurations provided in Appendix A.3. Under a fixed 16 GB memory budget, XConv consistently permits larger batch sizes than standard convolution across the evaluated probing-vector configurations, reflecting its lower peak memory footprint. For example, at $r = 16$ and image dimension $N = 1024$, XConv supports approximately 1.6× larger batch sizes than standard convolution (46 vs. 29), with similar or greater gains at lower resolutions.

Despite the compact architecture and limited spatial smoothing, we observe a consistent reduction in AGE as both the probing vectors r and image dimension N increase. In particular, the AGE gap between XConv and standard convolution decreases monotonically with increasing image resolution, becoming small at the largest tested scale ($N = 1024$) for $r = 128$. These results indicate that even in parameter-efficient architectures, gradient deviations introduced by XConv diminish systematically with increased probing capacity and spatial resolution.

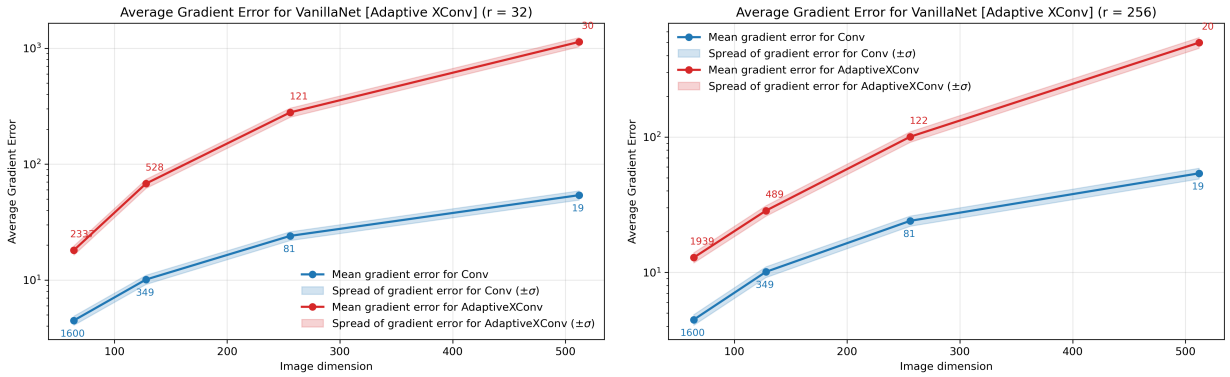
4.2.2 U-Net

We continue our analysis on U-Net (Ronneberger et al., 2015), a convolutional encoder–decoder architecture with skip-connections that has become a core building block in score-based diffusion models. Its deep hierarchical structure and large intermediate activations make it a particularly challenging and practically relevant setting for evaluating XConv.



(a) Probing vector r=16 (b) Probing vector r=128

Figure 5: Average gradient error vs. image dimension for U-Net. Results are shown for probing vectors (a) $r = 16$ and (b) $r = 128$, over 10 runs. Numbers at each data point indicate the maximum batch size that fits in memory for each method (blue: standard convolution, red: XConv) under a fixed memory budget. XConv increases AGE relative to Conv, but remains within one order of magnitude across all image dimensions. The gap narrows as r increases, indicating that approximation noise can be controlled by adjusting the number of probing vectors.



(a) Probing vector r=32 (b) Probing vector r=256

Figure 6: Average gradient error vs. image dimension for VanillaNet. Results are shown for probing vectors $r \in \{32, 256\}$ over 10 runs. Numbers at each data point indicate the maximum batch size that fits in memory for each method (blue: standard convolution, red: Adaptive XConv) under a fixed memory budget. Due to adaptive selection of approximate layers, the AGE gap relative to Conv does not follow a monotonic trend with increasing r.

Figure 5 reports the AGE for image dimensions $N = \{64, 128, 256, 512\}$ and probing vectors $r = \{16, 128\}$. Additional probing vector configurations can be found in Appendix A.3. Across all resolutions, XConv increases AGE relative to exact convolutional gradients; however, the increase remains bounded within a single order of magnitude.

AGE decreases systematically with an increasing number of probing vectors r. In contrast to lightweight architectures such as SqueezeNet, U-Net's extensive skip connections and large activations lead to higher gradient approximation error, though the error remains bounded and decreases with r. This behavior suggests that XConv can maintain adequate gradient fidelity even in deep, memory-intensive encoder–decoder architectures.

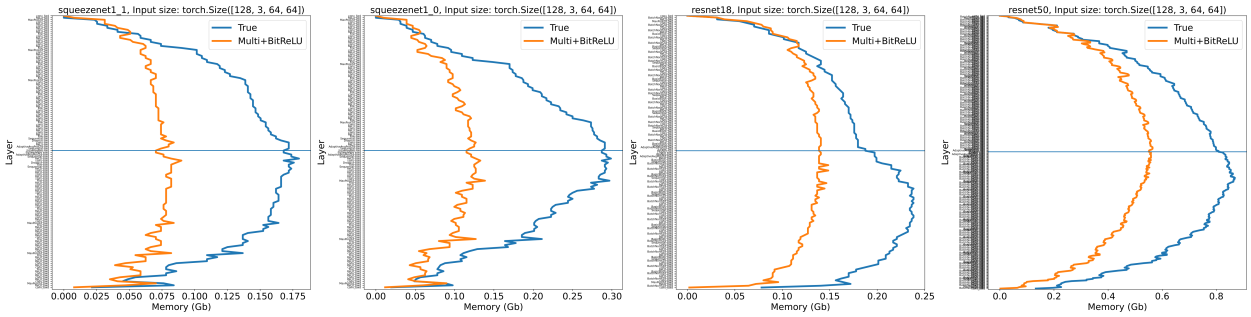


Figure 7: Layer-by-layer memory usage for a single gradient step. Standard convolution (blue) and XConv (orange) are compared for SqueezeNet 1.1, SqueezeNet 1.0, ResNet-18, and ResNet-50 (left to right) at a fixed input size. Memory savings of 2× or more are achievable depending on the ratio of convolutional to non-convolutional layers.

4.2.3 VanillaNet

We complete our analysis on VanillaNet (Chen et al., 2023a), a minimalistic architecture with shallow feature maps and small activation footprints. Since aggressive gradient approximation is neither necessary nor uniformly beneficial in such lightweight networks, we employ an adaptive variant of XConv that selectively applies approximation to layers with the largest activation footprints.

We evaluate AGE across probing vectors $r \in \{16, 32, 128, 256\}$ and image dimensions $N \in \{64, 128, 256, 512\}$. Results for $r = 32$ and $r = 256$ are shown in Figure 6 with additional results provided in Appendix A.3.

Unlike the previous settings, AGE does not exhibit a strictly monotonic dependence on r. This behavior is expected: increasing the number of probing vectors reduces estimator variance only in layers where XConv is active, while layers computed with exact gradients contribute no approximation error.

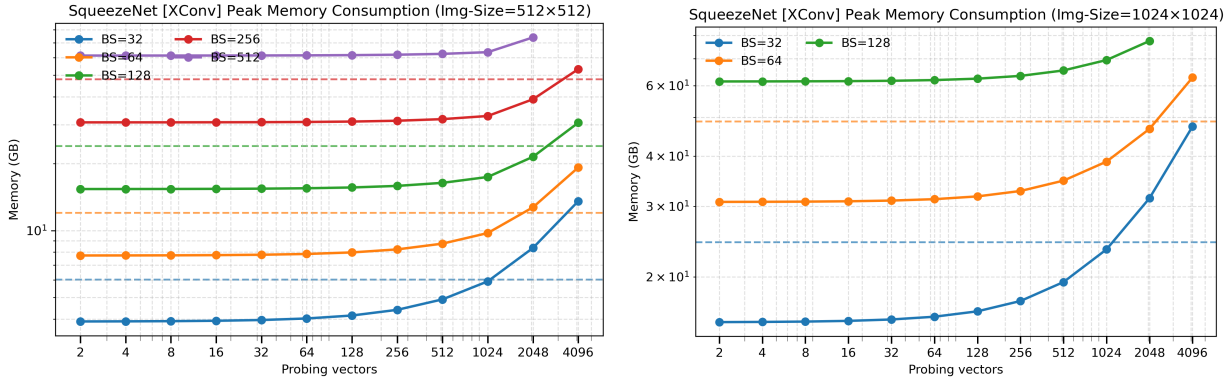
This highlights a key design principle: gradient approximation need not be applied uniformly. In lightweight networks, selectively approximating only the most memory-intensive layers suffices, while preserving exact gradients elsewhere stabilizes optimization. This stands in contrast to deeper architectures, where increasing r consistently improves gradient fidelity across the network.

4.3 Overall effective memory savings

Gradient fidelity alone does not determine practical utility—XConv must also translate into meaningful end-to-end memory savings. We compare peak memory consumption between otherwise identical models differing only in whether standard convolutional layers are replaced with XConv.

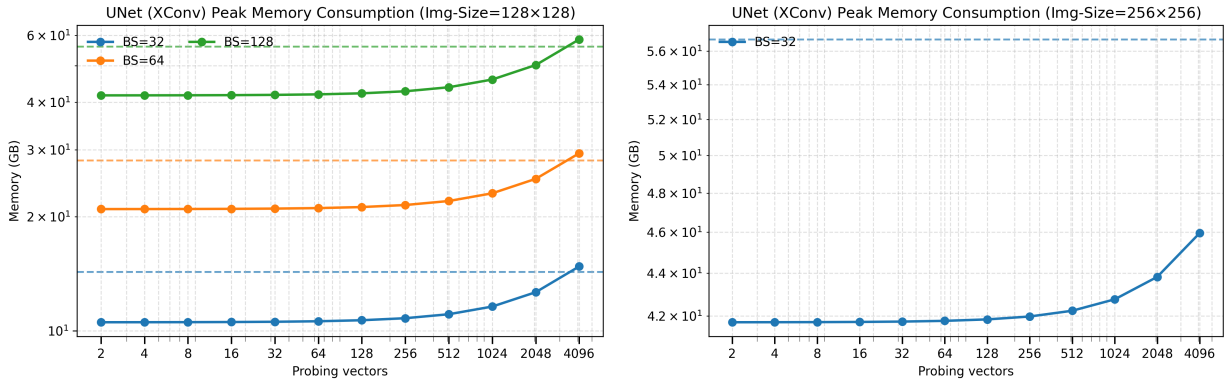
Approximate gradient calculations with multi-channel randomized trace estimation can lead to significant memory savings within convolutional layers. Because these layers operate in conjunction with other network layers such as `ReLU` and batchnorms, the overall effective memory savings depend on the ratio of pure convolutional and other layers and on the interaction between them. This is especially important for layers such as `ReLU`, which rely on the next layer to store the state variable during backpropagation. Unfortunately, that approach no longer works because our low-memory convolutional layer does not store the state variable. However, this situation can be remedied easily by only keeping track of the signs (Oktay et al., 2021).

To assess the effective memory savings of multi-channel trace estimation, we include in Figure 7 layer-by-layer comparisons of memory usage for different versions of the popular SqueezeNet (Iandola et al., 2016) and ResNet (He et al., 2016). The memory use for the conventional implementation is plotted in blue and our implementation in orange. The results indicate that memory savings by a factor of two or more are achievable, which can allow for larger batch sizes or increases in the width/depth of the network. As expected, the savings depend on the ratio of CNN versus other layers.



(a) Image Dimension N=512 (b) Image Dimension N=1024

Figure 8: Peak memory curves for SqueezeNet. The horizontal dashed line denotes the true gradient's memory consumption. For image dimension $N \in \{512, 1024\}$, peak memory for the true gradient is infeasible to compute for several batch sizes. Overall, peak memory usage increases with the number of probing vectors, and under a fixed memory budget, randomized trace estimation supports larger batch sizes.



(a) Image Dimension N=128 (b) Image Dimension N=256

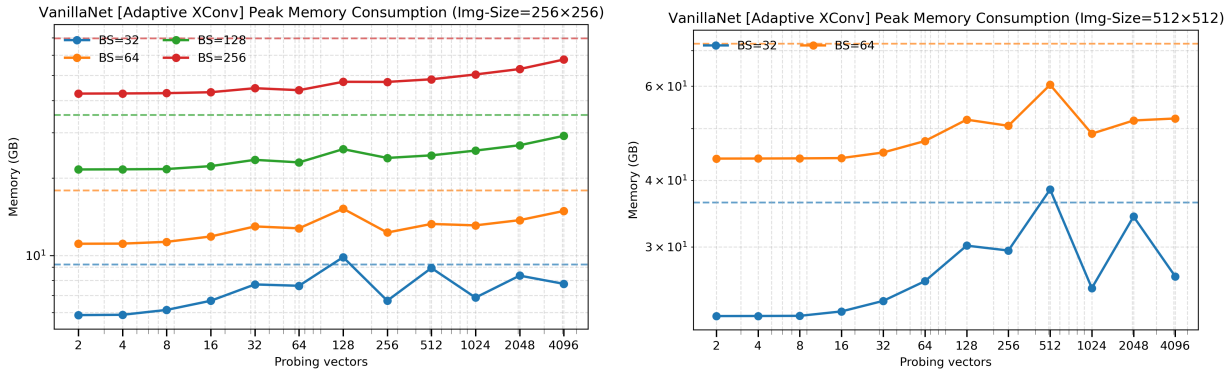
Figure 9: Peak memory curves for U-Net. The horizontal dashed line denotes the true gradient's memory consumption. Overall, peak memory usage increases with the number of probing vectors, and under a fixed memory budget, randomized trace estimation supports larger batch sizes.

We extend this memory analysis to entire architectures: SqueezeNet, U-Net and VanillaNet. Peak memory is the limiting factor for feasible batch size and image resolution during training and therefore serves as the primary metric of interest.

Figure 8 reports the peak memory consumption of SqueezeNet across varying batch sizes and probing vectors at a fixed image resolution of $N = 512$ and $N = 1024$. The corresponding results for lower image resolutions are provided in Figure 24 in the appendix. In these settings, XConv enables training regimes—either higher batch sizes or higher spatial resolutions—that would otherwise be infeasible under standard convolution due to memory constraints.

Results for U-Net at image resolutions $N \in \{128, 256\}$ are shown in Figure 9. Results for lower resolutions can be found in Figure 24. Despite U-Net's extensive skip connections and deep encoder–decoder structure, XConv reduces peak memory consumption across the tested probing vectors and batch sizes.

Finally, Figure 10 presents results for VanillaNet. Owing to its shallow architecture and heterogeneous layer-wise memory profile, we employ an adaptive variant of XConv that selectively applies approximation to layers. As a result, peak memory does not vary monotonically with the number of probing vectors. Nevertheless,



(a) Image Dimension N=256

(b) Image Dimension N=512

Figure 10: Peak memory curves for VanillaNet. The horizontal dashed line denotes the true gradient's memory consumption. Overall, peak memory usage increases with the number of probing vectors. Due to selective application of randomized trace estimation, the memory does not exhibit a monotonic dependence on probing vectors r.

across a broad range of probing budgets, XConv remains a memory-efficient alternative, confirming that selective approximation can yield meaningful savings.

4.4 Wall-clock benchmarks

Ideally, reducing the memory footprint during training should not come at the expense of significant computational overhead. To ensure this is indeed the case, we implemented the multi-channel randomized trace estimation optimized for CPUs in Julia (Bezanson et al., 2017) and for GPUs in PyTorch (Paszke et al., 2019). Implementation details and benchmarks are included in Appendix A.4.

Our benchmarking experiments show competitive performance on both CPUs, against NNLib (Innes et al., 2018; Innes, 2018), and GPUs, against optimized CUDA implementations. On CPUs, we observe speedups of up to 10× over the standard *im2col* (Chellapilla et al., 2006) implementation for large images and large batch sizes, provided the number of probing vectors remains relatively small. On GPUs, we remain competitive with CuDNN kernels (Chetlur et al., 2014) and occasionally outperform them, though there is room for further optimization. In all cases, there is a slight decrease in computational throughput when the number of channels increases. Overall, approximate gradient calculations with multi-channel randomized trace estimation substitute expensive convolutions between the input and output channels with a relatively simple combination of matrix-free actions of the outer product on random probing vectors on the right and dense linear matrix operations on the left (cf. equation 8 and Algorithm 1).

The wall-clock benchmark results along with the hardware description can be found in Figures 11 and 12.

4.5 Quantitative performance of XConv

Having quantified XConv's memory savings and gradient fidelity, we now evaluate whether these approximations meaningfully affect end-to-end model performance. We study classification, generative modeling, inpainting, super-resolution, and segmentation to determine whether the impact is consistent across fundamentally different learning objectives.

4.5.1 Classification

MNIST dataset We design a simple neural network and evaluate classification performance on the MNIST dataset, varying the batch size B and the number of probing vectors r. Implementations in both Julia and Python are evaluated.

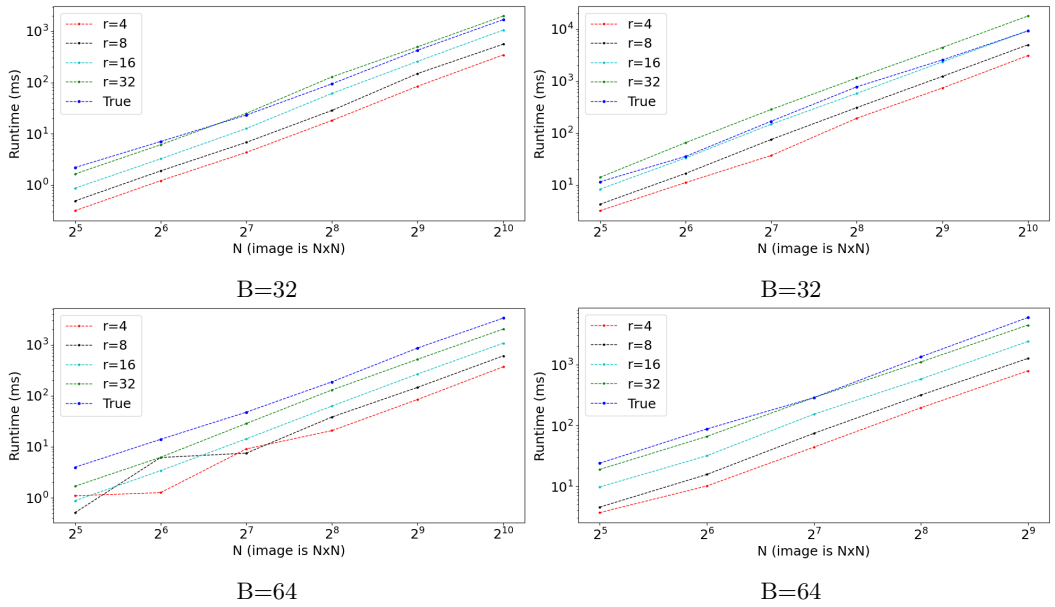


Figure 11: CPU benchmark on an *Intel(R) Xeon(R) CPU E3-1270 v6 @ 3.80GHz* node. The left column contains the runtimes for 4 channels and the right column for 32 channels. For large images and batch sizes, our implementation achieves speedups of up to 10×.

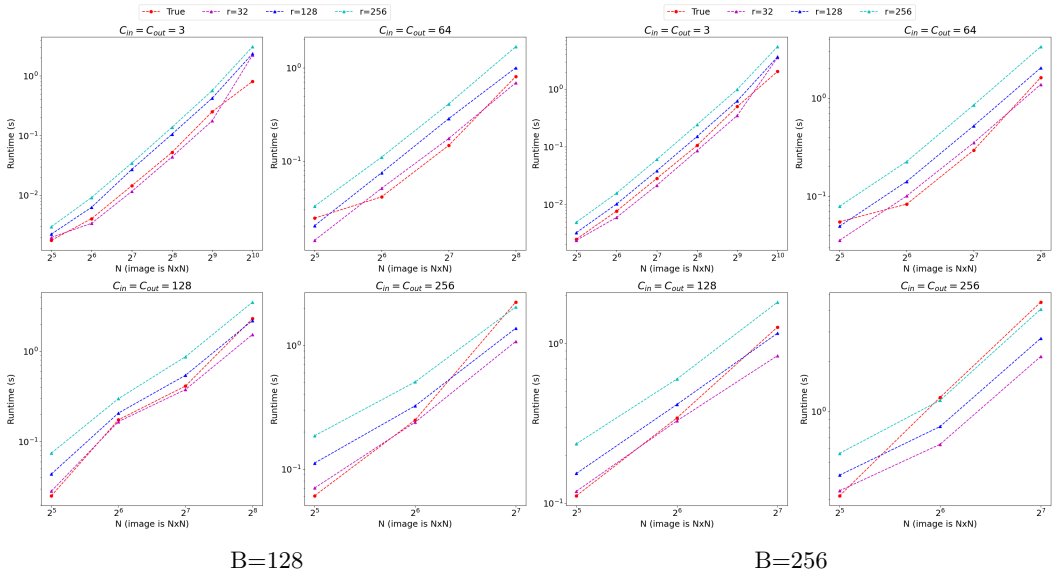


Figure 12: GPU benchmark on a *Tesla K80* (Azure NC6 instance) for a single gradient for varying batch sizes B, image sizes N and number of channels $C_{\text{in}} = C_{\text{out}}$. For larger scale problems we perform as well, if not better, than state-of-the-art CuDNN kernels. Additional GPU benchmarks for smaller batch sizes can be found in Figure 22.

We use the network architectures (detailed in Table 3 and 4 of Appendix A.5 for Julia and PyTorch, with training parameters listed in Appendix A.6) for varying batch sizes $B \in \{64, 128, 256\}$ and number of probing vectors $r \in \{2, 16, 64, 256\}$. The network test accuracies for the Julia implementation, where the default convolutional layer implementation is replaced by **XConv.jl**, are listed in Table 1 for the default implementation and for our implementation where gradients of the convolutional layers are replaced by our

Table 1: Test accuracy for varying batch sizes B and number of probing vectors r on the MNIST dataset.

	$B = 64$	$B = 128$	$B = 256$
True	0.9905	0.9898	0.9901
$r = 2$	0.9625	0.9692	0.9745
$r = 16$	0.9753	0.9803	0.9823
$r = 64$	0.9777	0.9723	0.9782
$r = 256$	0.9718	0.9706	0.9791

approximations. The results show that our low-memory implementation remains competitive even for a small number of probing vectors, yielding a memory saving of about 2.5×. We note that accuracy does not increase monotonically with r; the small fluctuations across probing vector settings are within the range of training stochasticity and do not indicate systematic degradation.

We obtained the results listed in Table 1 with the ADAM (Kingma & Ba, 2015) optimization algorithm. To further test robustness, we also train with stochastic line searches (SLS, Vaswani et al. (2019)), which set learning rate parameters automatically at the cost of an extra gradient calculation. The full accuracy-vs.-epoch curves (Figure 25 in Appendix A.9) confirm that the induced randomness does not adversely affect the line searches: XConv achieves competitive accuracy for $r \geq 16$ across all batch sizes $B \in \{64, \dots, 2048\}$, yielding a 2.5× memory reduction.

CIFAR-10 dataset We now compare standard stochastic gradient descent with approximate gradient methods based on multi-channel randomized trace estimation on the CIFAR-10 dataset.

To ensure a fair comparison, we fix the memory usage between the regular gradient, and the approximate gradients obtained by probing independently ("Indep." in green with $r = 32$), multi-channel ("Multi." in orange with $r = 256$), and multi-channel with orthogonalization ("Multi-Ortho" in red with $r = 256$). The batch size for the approximate gradient examples is increased from $B = 128$ to $B = 256$ to reflect the smaller memory footprint. Results for the training/testing loss and accuracy are included in Figure 13. The following observations can be made from these plots. First, there is a clear gap between the training/testing loss for the true and approximate gradients. This gap is also present in the training/testing accuracy, albeit relatively small. However, doubling the batch size effectively halves the training runtime.

4.5.2 Generative modeling

While supervised classification provides a controlled setting to assess discriminative performance, generative modeling via diffusion places greater emphasis on optimization dynamics and is therefore more sensitive to gradient approximation. In this section, we evaluate whether the approximate gradients induced by XConv alter the generative performance of U-Net-based diffusion models.

We consider a U-Net-based diffusion model trained on the MNIST dataset and evaluate the performance under varying numbers of probing vectors $r \in \{32, 64, 128, 256\}$. All experiments use a fixed batch size of 768, a learning rate of 10^{-3}, and are trained for 200 epochs.

Training and validation loss curves (Appendix A.10, Figure 26) confirm that, despite approximation error in gradient computation, XConv-based models exhibit training dynamics that closely track those of the exact convolution baseline across all tested probing vectors.

The qualitative generation results are shown in Figure 14. Across all probing vectors, XConv-based models produce samples that are visually similar to those generated by the standard convolutional baseline.

These qualitative findings are consistent with the quantitative Fréchet Inception Distance (FID) results, reported in Figure 15. For a large number of probing vectors ($r = 256$), XConv achieves FID scores close to those of standard convolution. Notably, these results hold despite AGE increasing by up to an order of magnitude (Figure 20), suggesting that U-Net-based diffusion models can tolerate moderate levels of gradient approximation error, particularly when the number of probing vectors is sufficiently large.

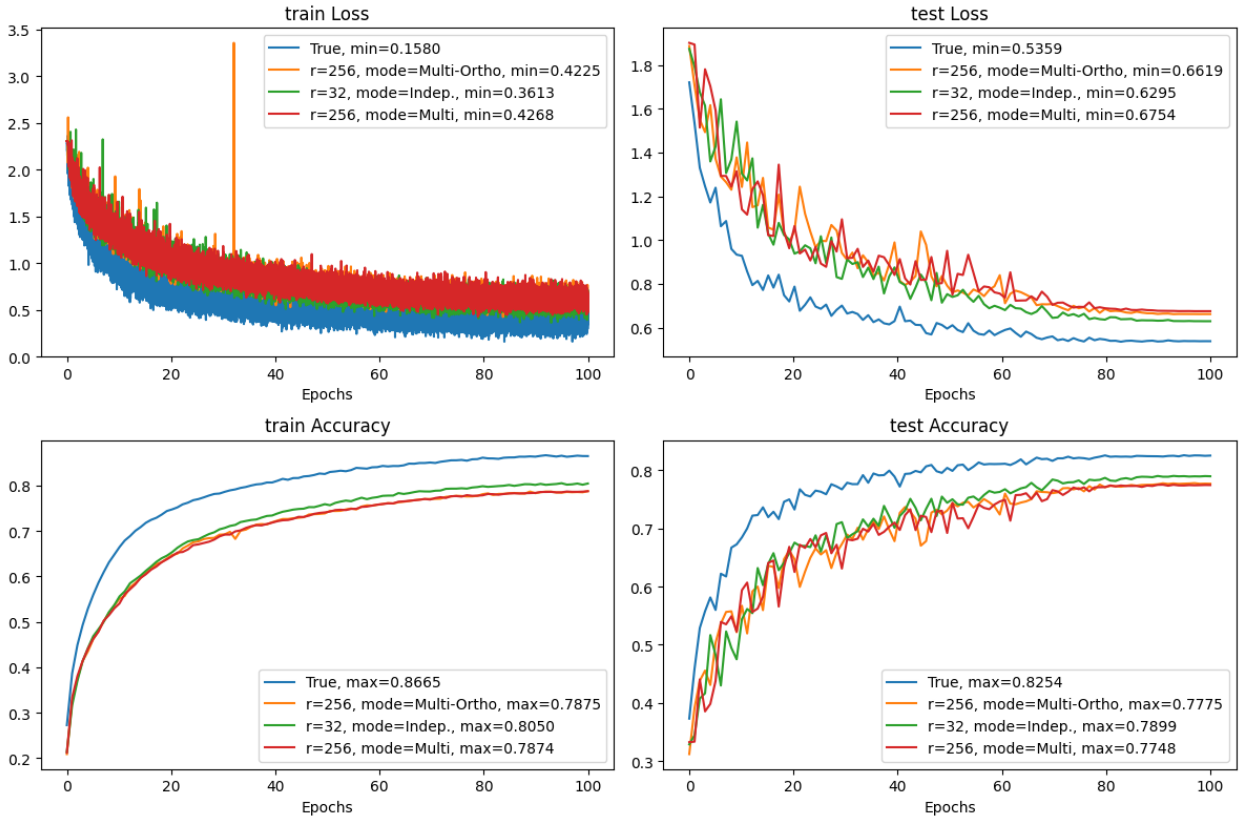
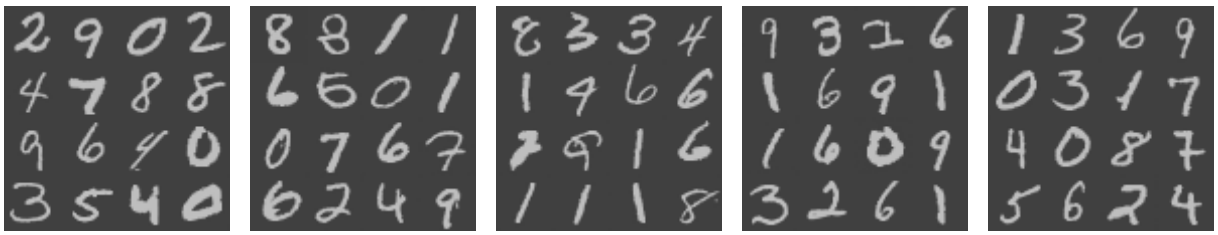


Figure 13: CIFAR-10 training with equivalent memory budget comparing standard convolution (B=128) to XConv with independent probing (r=32, green), multi-channel (r=256, orange), and multi-channel orthogonalized probing (r=256, red) at B=256. Top row: training and test loss. Bottom row: training and test accuracy after 100 epochs.



standard $r = 32$ $r = 64$ $r = 128$ $r = 256$

Figure 14: Generated samples from the standard network and XConv. Generated samples remain visually consistent across varying probing vectors r.

These experiments suggest that U-Net-based diffusion models can accommodate moderate gradient approximation while largely preserving convergence and sample quality, provided the number of probing vectors is sufficiently large.

4.5.3 Super-resolution and inpainting

To evaluate performance on inverse problems, we adopt the Deep Image Prior (DIP) framework (Ulyanov et al., 2018), which leverages the inductive bias of convolutional networks to recover structure from corrupted observations without external training data. DIP thus provides a direct test of whether XConv preserves this

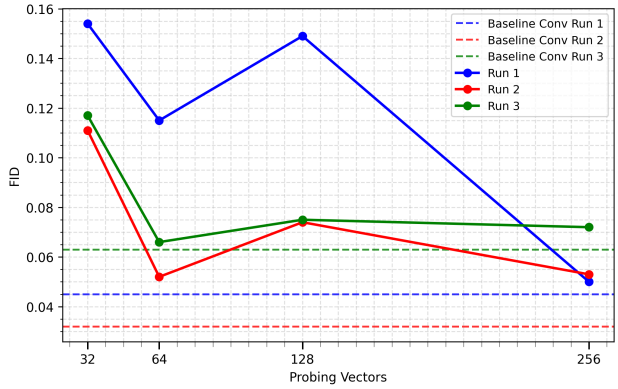


Figure 15: Fréchet Inception Distance (FID; lower is better) across three runs for the U-Net diffusion model on MNIST. Dashed lines denote FID scores of the standard convolution baseline. At $r = 256$, XConv achieves FID comparable to standard convolution.

Original Image | Low-Res Image | DIP | DIP ($r = 256$)

Figure 16: Qualitative super-resolution results comparing Conv and XConv. The reconstructions obtained with XConv are visually similar to those of standard DIP.

implicit regularization for super-resolution and inpainting. Figure 23a in the appendix illustrates the tradeoff between reconstruction quality and peak memory consumption for DIP-based super-resolution. We observe that $r = 256$ represents a favorable operating point: it substantially improves reconstruction quality over smaller r, while still maintaining a clear memory advantage over exact convolution. As expected, increasing r improves reconstruction fidelity at the cost of higher memory usage, revealing a clear accuracy–memory tradeoff.

Based on this tradeoff, we select $r = 256$ for subsequent experiments. For XConv-based experiments, we reduce the learning rate to 3×10^{-4} to improve optimization stability under stochastic gradient approximation; no additional task-specific hyperparameter tuning is performed. Qualitative super-resolution results on the Set5 dataset Bevilacqua et al. (2012), shown in Figure 16, indicate that replacing convolution layers with XConv yields visually comparable reconstructions.

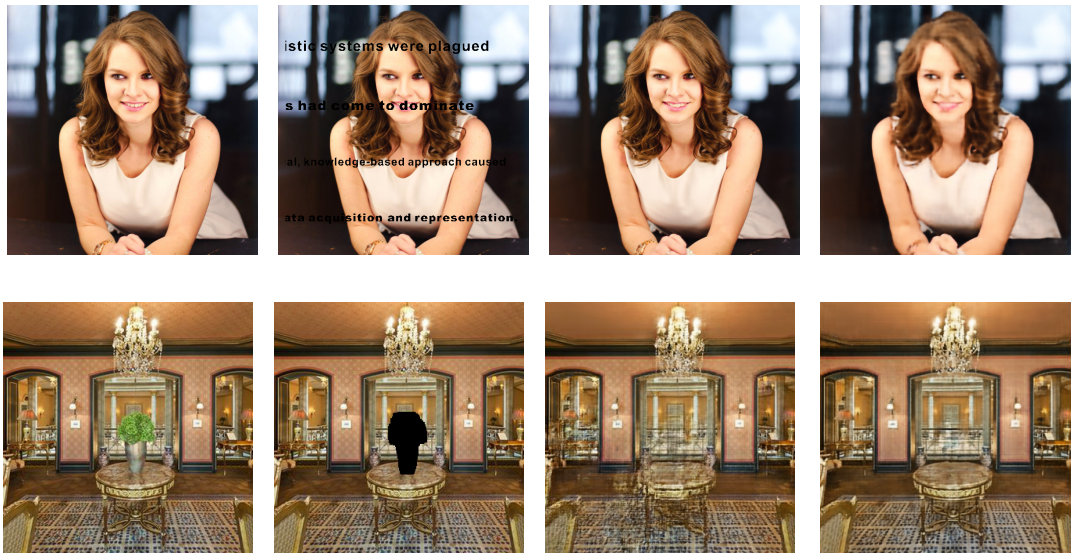


Figure 17: Qualitative inpainting results comparing Conv and XConv. XConv achieves visually similar inpainting results to convolution.

Figure 23b in the appendix reports peak memory usage for the inpainting task. Across a broad range of probing vectors ($r \leq 512$), XConv consistently requires less memory than standard convolution. Using the same probing vectors ($r = 256$), qualitative inpainting results in Figure 17 further suggest that XConv produces visually similar reconstructions without introducing obvious artifacts. Together, these results indicate that the implicit regularization exploited by DIP is largely retained under approximate gradient computation, while offering reductions in memory consumption.

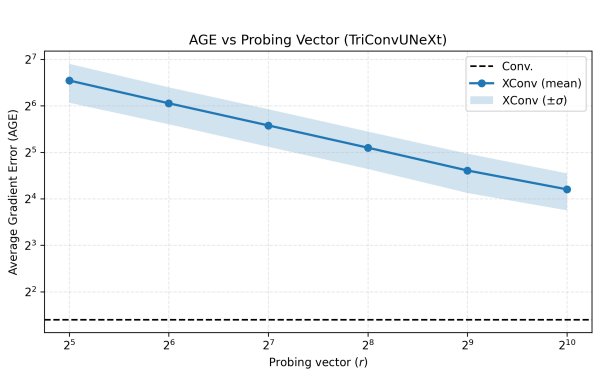
4.5.4 Segmentation

We evaluate gland segmentation (GlaS dataset (Sirinukunwattana et al., 2017), 165 histological images at 256×256 resolution) using TriConvUNeXt (Ma et al., 2024a), a lightweight model combining dilated and deformable convolutions. Dense pixel-wise prediction is highly sensitive to gradient quality, making segmentation a stringent test for XConv. All standard convolution layers are replaced with XConv while keeping all other hyperparameters fixed.

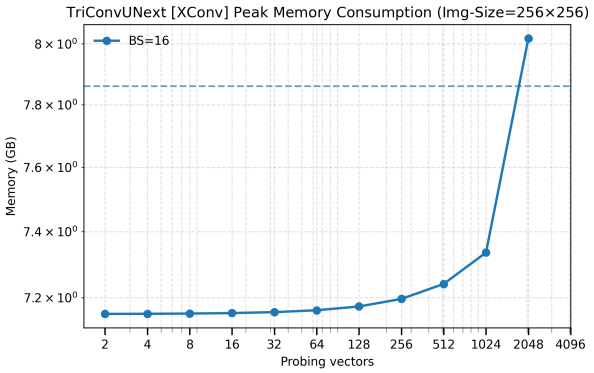
Figure 18b reports the peak memory usage as a function of the number of probing vectors r at a fixed image resolution, while Figure 18a reports the corresponding average gradient error (AGE). Increasing the number of probing vectors r consistently reduces both the magnitude and variance of AGE, but leads to rapidly increasing memory consumption beyond moderate probing vector budgets. Notably, the marginal improvement in gradient fidelity diminishes for $r > 1024$, while memory usage continues to grow sharply.

Based on this tradeoff, we select $r = 1024$ for segmentation experiments as it provides a favorable compromise with feasible memory overhead. Quantitative results in Table 2 show that XConv achieves segmentation performance close to standard convolution, with Dice similarity coefficient (DICE) and accuracy (ACC) differing by less than 1%. Qualitative predictions in Figure 19 are consistent with these quantitative findings.

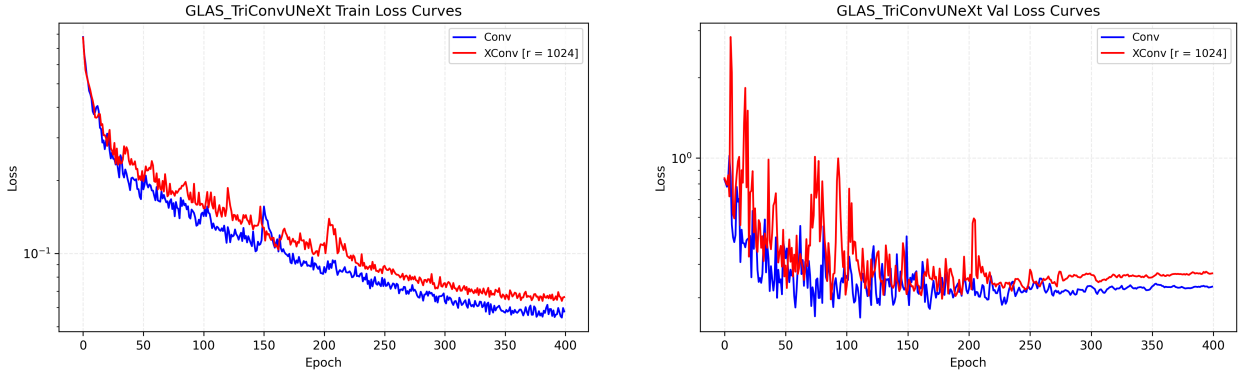
These results suggest that, even for dense prediction tasks at high spatial resolution, XConv can maintain adequate optimization fidelity while reducing memory requirements.



(a) AGE vs. probing vectors (r)



(b) Peak memory vs. probing vectors (r).



(c) Train (left) and validation (right) loss curves

Figure 18: AGE (top-left), peak memory consumption (top-right), and train-validation loss curves of the TriConvUNeXt model. The horizontal dashed line in both top graphs denotes standard convolution. Increasing the number of probing vectors r reduces AGE while increasing memory consumption. The train-validation loss curves show that XConv follows similar training dynamics as standard convolution.

Probing Vector (r)	**DICE**	**ACC**
standard convolution	0.905	0.904
1024	0.900	0.898

Table 2: Quantitative segmentation results on the GlaS dataset with TriConvUNeXt. DICE measures the overlap between the ground-truth and predicted segmentation masks; higher values indicate better performance. Results comparing standard convolution to XConv with probing vectors ($r = 1024$). Despite approximate gradients, XConv achieves segmentation accuracy within 1% of Conv, consistent with the qualitative results in Figure 19.

5 Related work

The memory pressure of backpropagation has motivated several lines of work. Checkpointing approaches (Griewank & Walther, 2000; Beaumont et al., 2019; Korthikanti et al., 2023) recompute activations during the backward pass, yielding exact gradients at the cost of additional computation. Invertible neural networks (Haber & Ruthotto, 2017; Jacobsen et al., 2018; Hascoet et al., 2023; Orozco et al., 2024) recover activations from outputs but impose architectural constraints that may limit expressiveness. Approximate arithmetic (Gupta et al., 2015; Xi et al., 2023) and memory-efficient optimizers (Zhao et al., 2024b) reduce memory through lower numerical precision or projected gradient updates, while local learning methods (Saiku-

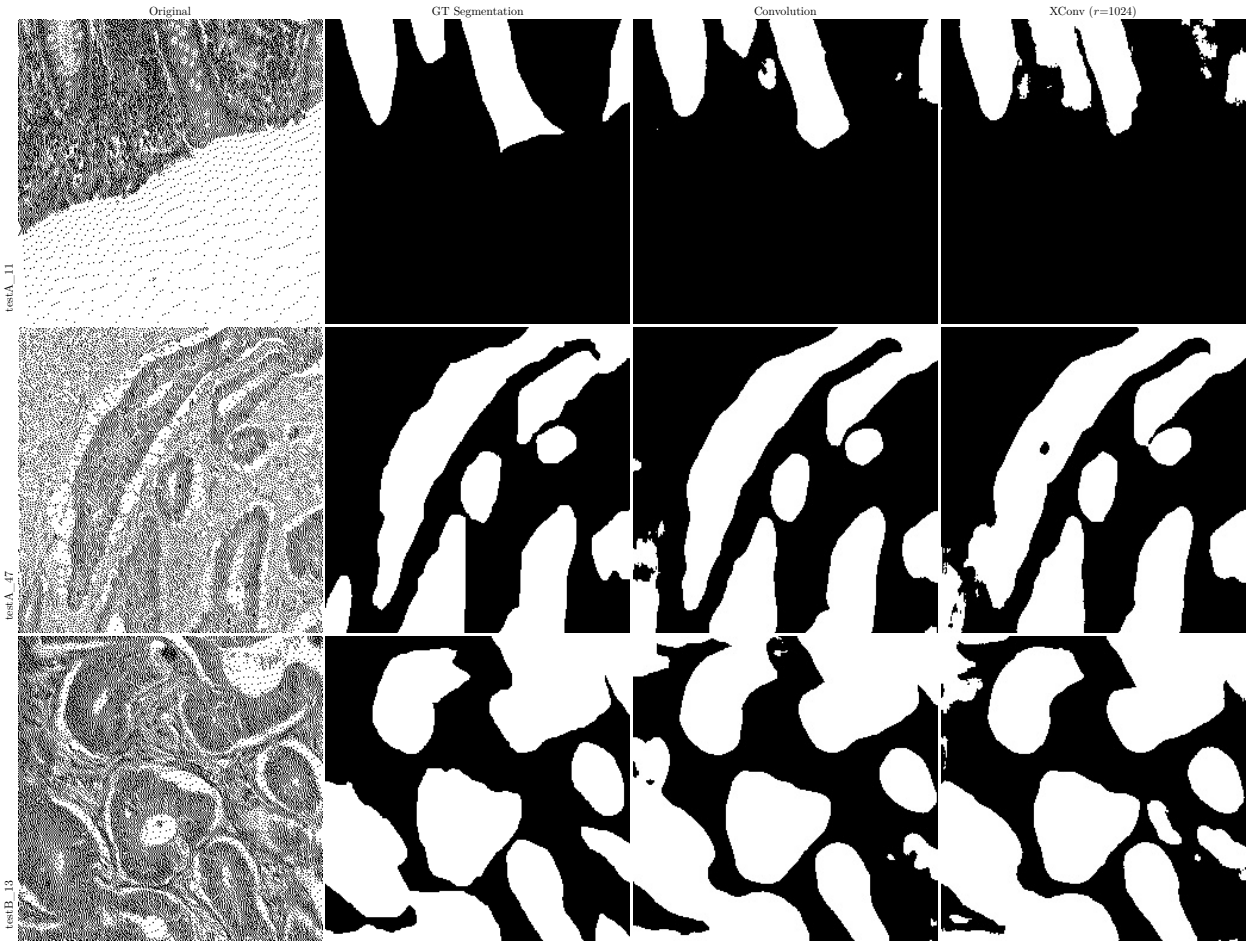


Figure 19: Qualitative segmentation results on the GlaS dataset comparing standard convolution and XConv. Each row corresponds to a different test image, while columns show the original image, ground-truth segmentation map, convolution, and XConv (r=1024) predictions. At a sufficiently large number of probing vectors, XConv produces segmentations that are visually similar to those obtained with standard convolution.

mar & Varghese, 2024) break inter-layer dependencies but require non-trivial architectural modifications. Direct feedback alignment (Nøkland, 2016; Han & Yoo, 2019; Frenkel et al., 2021) replaces backpropagated gradients with random projections. Randomized linear algebra techniques (Avron & Toledo, 2011; Martinsson & Tropp, 2020; Meyer et al., 2021) provide unbiased gradient approximations—a desirable property for stochastic optimization (Neelakantan et al., 2015). Among these, randomized automatic differentiation (RAD) (Oktay et al., 2021) is closest to our work but requires intervention in the computational graph. In contrast, XConv exploits the specific algebraic structure of convolutional layer gradients, enabling a simpler approach that acts as a drop-in replacement for 2D and 3D convolutional layers in existing frameworks.

6 Conclusion and future work

We introduced XConv, a memory-efficient convolutional layer that approximates gradients via multi-channel randomized trace estimation. By storing only compressed activations, XConv reduces the memory footprint while remaining computationally competitive with standard implementations. The approach comes with convergence guarantees and error bounds, and our experiments show that networks trained with XConv achieve performance close to exact gradient methods across diverse tasks, with accuracy that improves systematically as the number of probing vectors increases. The degree of memory savings and accuracy depends on the architecture and task, with typical memory reductions of 2× or more. These properties, combined with

recent architectural innovations that reduce computational complexity (Howard et al., 2017; Liu et al., 2022; Chen et al., 2023a;b) and advances in specialized photonic hardware for randomized probing (Saade et al., 2016), open directions for scaling CNNs to higher-dimensional data such as video representation learning and other 3D applications. More broadly, the principle of approximating weight gradients via randomized trace estimation is not limited to convolutions—extending this approach to attention layers, where the memory footprint of stored activations is similarly prohibitive, is a promising direction for future work.

Declaration of AI usage. Claude (Anthropic) was used to improve the academic tone and technical clarity of the writing, correct grammatical errors and enhance readability, and ensure consistent formatting of bibliographic entries. All scientific content, including methodology, theoretical results, and experimental design, was produced entirely by the authors.

References

Aleksandr Y. Aravkin, Michael P. Friedlander, Felix J. Herrmann, and Tristan van Leeuwen. Robust inversion, dimensionality reduction, and randomized sampling. *Mathematical Programming*, 134(1):101–125, 2012.

Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM*, 58(2):8:1–8:34, 2011.

Olivier Beaumont, Lionel Eyraud-Dubois, Julien Herrmann, Alexis Joly, and Alena Shilova. Optimal checkpointing for heterogeneous chains: How to train deep neural networks with limited memory. *arXiv preprint arXiv:1911.13214*, 2019.

Olivier Beaumont, Lionel Eyraud-Dubois, Julien Herrmann, Alexis Joly, and Alena Shilova. Optimal re-materialization strategies for heterogeneous chains: How to train deep neural networks with limited memory. *ACM Transactions on Mathematical Software*, 50(2):10:1–10:38, 2024.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, 2012.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

Kumar Chellapilla, Sidd Puri, and Patrice Simard. High performance convolutional neural networks for document processing. In *Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.

Hanting Chen, Yunhe Wang, Jianyuan Guo, and Dacheng Tao. VanillaNet: The power of minimalism in deep learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7050–7064, 2023a.

Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, et al. Run, don't walk: Chasing higher FLOPS for faster neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12021–12031, 2023b.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, et al. cuDNN: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014.

Alice Cortinovis and Daniel Kressner. On randomized trace estimates for indefinite matrices with an application to determinants. *Foundations of Computational Mathematics*, 22:875–903, 2022.

Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Focal network for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13001–13011, 2023a.

Yuning Cui, Wenqi Ren, Sining Yang, Xiaochun Cao, and Alois Knoll. IRNeXt: Rethinking convolutional network design for image restoration. In *International Conference on Machine Learning*, pp. 6545–6564, 2023b.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11963–11975, 2022.

Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, et al. UniRepLKNet: A universal perception large-kernel ConvNet for audio, video, point cloud, time-series and image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5513–5524, 2024.

Jianwei Feng and Dong Huang. Optimal gradient checkpoint search for arbitrary computation graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11433–11442, 2021.

Charlotte Frenkel, Martin Lefebvre, and David Bol. Learning without feedback: Fixed random learning signals allow for feedforward training of deep neural networks. *Frontiers in Neuroscience*, 15:629892, 2021.

Aidan N. Gomez, Mengye Ren, Raquel Urtasun, and Roger B. Grosse. The reversible residual network: Backpropagation without storing activations. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Andreas Griewank and Andrea Walther. Algorithm 799: Revolve: An implementation of checkpointing for the reverse or adjoint mode of computational differentiation. *ACM Transactions on Mathematical Software*, 26(1):19–45, 2000.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning*, pp. 1737–1746, 2015.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1): 014004, 2017.

Eldad Haber, Matthias Chung, and Felix J. Herrmann. An effective method for parameter estimation with PDE constraints with multiple right hand sides. *SIAM Journal on Optimization*, 22(3):739–757, 2012.

Donghyeon Han and Hoi-Jun Yoo. Efficient convolutional neural network training with direct feedback alignment. *arXiv preprint arXiv:1901.01986*, 2019.

Tristan Hascoet, Quentin Febvre, Yasuo Ariki, and Tetsuya Takiguchi. Reversible designs for extreme memory cost reduction of CNN training. *EURASIP Journal on Image and Video Processing*, 2023:1, 2023.

Horace He and Shangdi Yu. Transcending runtime-memory tradeoffs in checkpointing by being fusion aware. In *Proceedings of Machine Learning and Systems*, volume 5, pp. 414–427, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Ding-Yong Hong, Tzu-Hsien Tsai, Ning Wang, Pangfeng Liu, and Jan-Jan Wu. GPU memory usage optimization for backward propagation in deep network training. *Journal of Parallel and Distributed Computing*, 199:105053, 2025.

Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2Former: A simple transformer-style ConvNet for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 8274–8283, 2024.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

W. Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K. Terry, et al. Understanding generalization through visualizations. In *NeurIPS Workshop on "I Can't Believe It's Not Better!"*, pp. 87–97, 2020.

M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics – Simulation and Computation*, 18(3):1059–1076, 1989.

Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv preprint arXiv:1602.07360*, 2016.

Michael Innes, Elliot Saba, Keno Fischer, Dhairya Gandhi, Marco Concetto Rudilosso, et al. Fashionable modelling with Flux. *arXiv preprint arXiv:1811.01457*, 2018.

Mike Innes. Flux: Elegant machine learning with Julia. *Journal of Open Source Software*, 3(25):602, 2018.

Jörn-Henrik Jacobsen, Arnold W. M. Smeulders, and Edouard Oyallon. i-RevNet: Deep invertible networks. In *International Conference on Learning Representations*, 2018.

Bryan Kaperick. *Diagonal Estimation with Probing Methods*. PhD thesis, Virginia Polytechnic Institute and State University, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Vijay Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, et al. Reducing activation recomputation in large transformer models. In *Proceedings of Machine Learning and Systems*, volume 5, 2023.

Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback alignment scales to modern deep learning tasks and architectures. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9346–9360, 2020.

Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, et al. More ConvNets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. In *International Conference on Learning Representations*, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, et al. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Chao Ma, Yuan Gu, and Ziyang Wang. TriConvUNeXt: A pure CNN-based lightweight symmetrical network for biomedical image segmentation. *Journal of Imaging Informatics in Medicine*, 37:2311–2323, 2024a.

Xu Ma, Xiyang Dai, Yue Bai, Yizhou Wang, and Yun Fu. Rewrite the stars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5694–5703, 2024b.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, et al. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mingyuan Mao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, et al. Dual-stream network for visual recognition. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25346–25358, 2021.

Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

Raphael A. Meyer, Cameron Musco, Christopher Musco, and David P. Woodruff. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms*, pp. 142–155, 2021.

Mitsumasa Nakajima, Yongbo Zhang, Katsuma Inoue, Yasuo Kuniyoshi, Toshikazu Hashimoto, et al. Reservoir direct feedback alignment: Deep learning by physical dynamics. *Communications Physics*, 7(1):411, 2024.

Arvind Neelakantan, Luke Vilnis, Quoc V. Le, Ilya Sutskever, Lukasz Kaiser, et al. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Arild Nøkland and Lars Hiller Eidnes. Training neural networks with local error signals. In *International Conference on Machine Learning*, pp. 4839–4850, 2019.

Deniz Oktay, Nick McGreivy, Joshua Aduol, Alex Beatson, and Ryan P. Adams. Randomized automatic differentiation. In *International Conference on Learning Representations*, 2021.

Rafael Orozco, Philipp Witte, Mathias Louboutin, Ali Siahkoohi, Gabrio Rizzuti, et al. InvertibleNetworks.jl: A Julia package for scalable normalizing flows. *Journal of Open Source Software*, 9(99):6554, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019.

K. B. Petersen and M. S. Pedersen. The Matrix Cookbook, 2008. Version 20081110.

Maria Refinetti, Stéphane d'Ascoli, Ruben Ohana, and Sebastian Goldt. Align, then memorise: The dynamics of learning with feedback alignment. In *International Conference on Machine Learning*, pp. 8925–8935, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

Farbod Roosta-Khorasani and Uri Ascher. Improved bounds on sample size for implicit matrix trace estimators. *Foundations of Computational Mathematics*, 15(5):1187–1212, 2015.

A. Saade, F. Caltagirone, I. Carron, L. Daudet, A. Drémeau, et al. Random projections through multiple optical scattering: Approximating kernels at the speed of light. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6215–6219, 2016.

Dhananjay Saikumar and Blesson Varghese. NeuroFlux: Memory-efficient CNN training using adaptive local learning. In *Proceedings of the Nineteenth European Conference on Computer Systems*, pp. 999–1015, 2024.

Aashaka Shah, Chao-Yuan Wu, Jayashree Mohan, Vijay Chidambaram, and Philipp Kraehenbuehl. Memory optimization for deep networks. In *International Conference on Learning Representations*, 2021.

Korsuk Sirinukunwattana, Josien P. W. Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, et al. Gland segmentation in colon histology images: The GlaS challenge contest. *Medical Image Analysis*, 35:489–502, 2017.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):109:1–109:28, 2022.

Joel A. Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Streaming low-rank matrix approximation with an application to scientific simulation. *SIAM Journal on Scientific Computing*, 41(4):A2430–A2463, 2019.

Irek Ulidowski. Saving memory space in deep neural networks by recomputing: A survey. In *International Conference on Reversible Computation*, pp. 89–105, 2023.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018.

Tristan van Leeuwen and Felix J. Herrmann. 3D frequency-domain seismic inversion with controlled sloppiness. *SIAM Journal on Scientific Computing*, 36(5):S192–S217, 2014.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. MobileOne: An improved one millisecond mobile backbone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7907–7917, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, et al. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. RepViT: Revisiting mobile CNN from ViT perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15909–15920, 2024.

Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, et al. InternImage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14408–14419, 2023.

Ziheng Wang, Sree Harsha Nelaturu, and Saman Amarasinghe. Accelerated CNN training through gradient approximation. In *2nd Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications*, pp. 31–35, 2019.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, et al. ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Haocheng Xi, Changhao Li, Jianfei Chen, and Jun Zhu. Training transformers with 4-bit integers. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Heesung Yang, Soha Lee, and Hyeyoung Park. Direct feedback learning with local alignment support. *IEEE Access*, 12:81388–81397, 2024a.

Yuchen Yang, Yingdong Shi, Cheems Wang, Xiantong Zhen, Yuxuan Shi, et al. Reducing fine-tuning memory overhead by approximate and memory-sharing backpropagation. In *International Conference on Machine Learning*, 2024b.

Abolfazl Younesi, Mohsen Ansari, Mohammadamin Fazli, Alireza Ejlali, Muhammad Shafique, et al. A comprehensive survey of convolutions in deep learning: Applications, challenges, and future trends. *IEEE Access*, 12:41180–41218, 2024.

Hong Zhang and Yu Zhang. Memory-efficient reversible spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16759–16767, 2024.

Chen Zhao, Shuming Liu, Karttikeya Mangalam, Guocheng Qian, Fatimah Zohra, et al. Dr2Net: Dynamic reversible dual-residual networks for memory-efficient finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15835–15844, 2024a.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, et al. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *International Conference on Machine Learning*, 2024b.

A Appendix

A.1 Implementation and code availability

Our probing algorithm is implemented both in Julia, using `LinearAlgebra.BLAS` on CPU and `CUDA.CUBLAS` on GPU for the linear algebra computations, and in PyTorch using standard linear algebra utilities. The Julia interface is designed so that preexisting networks can be reused as we are overloading `rrule` (see ChainRulesCore.jl) to switch easily between the conventional true gradient (NNlib.jl) and ours. The PyTorch implementation defines a new layer that can be swapped for the conventional convolutional layer, `torch.nn.Conv2d` or `torch.nn.Conv3d`, in any network using the `convert_net` utility function. Both implementations support 2D and 3D convolutions, making XConv readily applicable to volumetric data such as medical imaging and video.

The software and examples will be made available under an MIT license upon publication.

A.2 Proofs of Proposition 1 and Theorem 1

For a square matrix $\mathbf{A}$, let $G(\mathbf{A})$ be the trace estimator:

$$G(\mathbf{A}) = \frac{1}{r}\sum_{j=1}^{r} \mathbf{z}_j^\top \mathbf{A}\mathbf{z}_j$$

where $\mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ are i.i.d. Gaussian vectors. We now prove the proposition and theorem stated in Section 2.

A.2.1 Proof of Proposition 1

We restate Proposition 1 here.

Proposition 2. *Let $\mathbf{A} \in \mathbb{R}^{N\times N}$ be a square matrix. Then for any small number $\delta > 0$, with probability $1-\delta$,*

$$|G(\mathbf{A}) - \mathrm{tr}(\mathbf{A})| \le \left(\frac{4\|\mathbf{A}\|_2}{r}\log\frac{2}{\delta} + \frac{2\|\mathbf{A}\|_F}{\sqrt{r}}\log^{1/2}\frac{2}{\delta}\right).$$

The proof uses the following result on trace estimation of symmetric matrices.

Lemma 2 (Theorem 5 of Cortinovis & Kressner (2022))**.** *Let $\mathbf{B} \in \mathbb{R}^{N\times N}$ be symmetric. Then*

$$P(|G(\mathbf{B}) - \mathrm{tr}(\mathbf{B})| \ge \epsilon) \le 2\exp\left(-\frac{r\epsilon^2}{4\|\mathbf{B}\|_F^2 + 4\epsilon\|\mathbf{B}\|_2}\right)$$

for all $\epsilon > 0$.

Proof of Proposition 2. For a symmetric matrix $\mathbf{B}$, Lemma 2 immediately implies that for any small number $\delta > 0$, with probability $1-\delta$,

$$|G(\mathbf{B}) - \mathrm{tr}(\mathbf{B})| \le \frac{4\|\mathbf{B}\|_2}{r}\log\frac{2}{\delta} + \frac{2\|\mathbf{B}\|_F}{\sqrt{r}}\log^{1/2}\frac{2}{\delta}.$$

Now for our asymmetric $\mathbf{A}$, let $\mathbf{B} = \frac{\mathbf{A}+\mathbf{A}^\top}{2}$. Then $G(\mathbf{A}) = G(\mathbf{B})$, $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{B})$, $\|\mathbf{B}\|_2 \le \|\mathbf{A}\|_2$, and $\|\mathbf{B}\|_F \le \|\mathbf{A}\|_F$, then the proposition follows. □

A.2.2 Preparation lemmas for Theorem 1

Lemma 3. *Let $\mathbf{A} \in \mathbb{R}^{N\times N}$ be a square matrix, $\mathbf{z}_j, \mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ be random Gaussian vectors for $j = 1, \ldots, r$, and all the $\mathbf{x}_j$ and $\mathbf{z}_j$ are independent of each other. Then for any $\delta > 0$, with probability $1-\delta$,*

$$\left|\frac{1}{r}\sum_{j=1}^{r}\mathbf{z}_j^\top\mathbf{A}\mathbf{x}_j\right| \le c\left(\frac{\|\mathbf{A}\|_2}{r}\log\frac{2}{\delta} + \frac{\|\mathbf{A}\|_F}{\sqrt{r}}\log^{1/2}\frac{2}{\delta}\right)$$

where c is some absolute constant independent of r.

Proof of Lemma 3. Set $T := \frac{1}{r}\sum_{j=1}^{r} \mathbf{z}_j^\top \mathbf{A}\mathbf{x}_j$. For each summand, we have

$$\mathbf{z}_j^\top \mathbf{A}\mathbf{x}_j = \mathbf{z}_j^\top \mathbf{U}\mathbf{S}\mathbf{V}^\top \mathbf{x}_j = \tilde{\mathbf{z}}_j^\top \mathbf{S}\tilde{\mathbf{x}}_j = \sum_t s_t \tilde{\mathbf{z}}_j[t]\tilde{\mathbf{x}}_j[t] \equiv \sum_t s_t f_{j,t}, \tag{10}$$

where the first equality used the singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, in the second equality, we defined $\tilde{\mathbf{z}}_j = \mathbf{U}^\top \mathbf{z}_j$ and $\tilde{\mathbf{x}}_j = \mathbf{V}^\top \mathbf{x}_j$, which are still Gaussian. In the third equality, we used s_t to denote the t^{th} diagonal entry of $\mathbf{S}$ and $\tilde{\mathbf{z}}_j[t]$ and $\tilde{\mathbf{x}}_j[t]$ to denote the t^{th} entry of $\tilde{\mathbf{z}}_j$ and $\tilde{\mathbf{x}}_j$, respectively. In the last equality, we defined $f_{j,t} := \tilde{\mathbf{z}}_j[t]\tilde{\mathbf{x}}_j[t]$. Since $\mathbf{z}_j$ and $\mathbf{x}_j$ are i.i.d., so are $f_{j,t}$. And since $f_{j,t}$ are products of independent sub-Gaussian random variables, they obey the sub-exponential distribution, i.e.,

$$\|f_{j,t}\|_{\psi_1} \le \|\tilde{\mathbf{z}}_j[t]\|_{\psi_2}\|\tilde{\mathbf{x}}_j[t]\|_{\psi_2} = c^2,$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm and $\|\cdot\|_{\psi_2}$ the sub-Gaussian norm. We also used the property that there is a constant c, such that for any σ, a Gaussian variable $a \sim \mathcal{N}(0,\sigma^2)$ has a sub-Gaussian norm $\|a\|_{\psi_2} \le c\sigma$, and this property is applied on $\tilde{\mathbf{z}}_j[t]$ and $\tilde{\mathbf{x}}_j[t]$ who are both $\mathcal{N}(0,1)$ variables due to the rotation invariance of Gaussian vectors.

Applying the Bernstein inequality Vershynin (2018) to $T = \frac{1}{r}\sum_{j,t} s_t f_{j,t}$, we obtain

$$P(|T| \ge \tilde{t}) \le e^{-c' \min\{\frac{r\tilde{t}^2}{4\|\mathbf{A}\|_F^2}, 4\frac{r\tilde{t}}{\|\mathbf{A}\|_2}\}},$$

where c' is some absolute constant. Letting δ be the right-hand-side probability, the above implies

$$P\left(|T| \ge c\left(\frac{\|\mathbf{A}\|_2}{r}\log\frac{2}{\delta} + \frac{\|\mathbf{A}\|_F}{\sqrt{r}}\log^{1/2}\frac{2}{\delta}\right)\right) \le \delta,$$

with some constant c. Then the lemma is proved. □

Lemma 4. *Let $\mathbf{A} \in \mathbb{R}^{N\times N}$ be a square matrix, $\mathbf{z}_j, \mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ be random Gaussian vectors for $j = 1,\dots,r$, and all the $\mathbf{x}_j$'s and $\mathbf{z}_j$'s are independent of each other. Let $\mathbf{y}_j$ be the random vector that equals $\mathbf{x}_j$ with probability p and equals 0 with probability $1-p$. Then for any $\delta > 0$ with probability over $1 - \delta - 2e^{-rp^2/2}$,*

$$\left|\frac{1}{r}\sum_{j=1}^{r} \mathbf{z}_j^\top \mathbf{A}\mathbf{y}_j\right| \le c\left(\frac{\|\mathbf{A}\|_2}{r}\log\frac{2}{\delta} + \frac{\sqrt{p}\|\mathbf{A}\|_F}{\sqrt{r}}\log^{1/2}\frac{2}{\delta}\right),$$

where c is some absolute constant independent of r.

Proof. The proof is very similar to that of Lemma 3. Set $T := \frac{1}{r}\sum_{j=1}^{r} \mathbf{z}_j^\top \mathbf{A}\mathbf{y}_j$. Let $g_j = 1_{\{\mathbf{y}_j \neq \mathbf{0}\}}$ be the indicator function of whether $\mathbf{y}_j$ is non-zero. Then clearly $\mathbf{y}_j = \mathbf{x}_j g_j$. For each summand, we have

$$\mathbf{z}_j^\top \mathbf{A}\mathbf{y}_j = \mathbf{z}_j^\top \mathbf{U}\mathbf{S}\mathbf{V}^\top \mathbf{x}_j g_j = \tilde{\mathbf{z}}_j^\top \mathbf{S}\tilde{\mathbf{x}}_j g_j = \sum_t s_t \tilde{\mathbf{z}}_j[t]\tilde{\mathbf{x}}_j[t] g_j \equiv g_j \sum_t s_t f_{j,t}, \tag{11}$$

where in the first equality, we used the singular value decomposition, $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, and in the second equality, we defined $\tilde{\mathbf{z}}_j = \mathbf{U}^\top \mathbf{z}_j$ and $\tilde{\mathbf{x}}_j = \mathbf{V}^\top \mathbf{x}_j$. In the third equality, we used s_t to denote the t^{th} diagonal entry of $\mathbf{S}$ with $\tilde{\mathbf{z}}_j[t]$ and $\tilde{\mathbf{x}}_j[t]$ denoting the t^{th} entry of $\tilde{\mathbf{z}}_j$ and $\tilde{\mathbf{x}}_j$, respectively. In the last equality, we defined $f_{j,t} := \tilde{\mathbf{z}}_j[t]\tilde{\mathbf{x}}_j[t]$. From Lemma 3, $f_{j,t}$ follow sub-exponential distributions, i.e., $\|f_{j,t}\|_{\psi_1} \le c^2$.

Conditional on g_j, applying the Bernstein inequality to $\tilde{T} := \frac{1}{\sum_j 1_{\{g_j \neq 0\}}} \sum_{j,t} s_t f_{j,t} g_j$, we obtain

$$P(|\tilde{T}| \geq \tilde{t}) \leq e^{-c' \min\left\{ \frac{\tilde{t}^2 \sum_i 1_{\{g_i \neq 0\}}}{4\|\mathbf{A}\|_F^2}, 4\frac{\tilde{t} \sum_j 1_{\{g_j \neq 0\}}}{\|\mathbf{A}\|_2} \right\}}.$$

Letting δ be the right-hand-side probability, the above implies

$$P\left(|\tilde{T}| \geq c \left(\frac{\|\mathbf{A}\|_2}{\sum_j 1_{\{g_j \neq 0\}}} \log \frac{2}{\delta} + \frac{\|\mathbf{A}\|_F}{\sqrt{\sum_j 1_{\{g_j \neq 0\}}}} \log^{1/2} \frac{2}{\delta} \right) \right) \leq \delta.$$

Since $\sum_j 1_{\{g_j \neq 0\}} \sim \mathbf{B}(r, p)$, we then have with probability at least $1 - 2e^{-rp^2/2}$, $3rp/2 \geq \sum_j 1_{\{g_j \neq 0\}} \geq rp/2$. Plugging this estimate into the above, we have

$$P\left(|\tilde{T}| \geq c \left(\frac{\|\mathbf{A}\|_2}{pr} \log \frac{2}{\delta} + \frac{\|\mathbf{A}\|_F}{\sqrt{pr}} \log^{1/2} \frac{2}{\delta} \right) \right) \leq \delta.$$

Then, with this bound of $|\tilde{T}|$, we have

$$|T| = \left| \frac{1}{r} \sum_{j=1}^{r} \mathbf{z}_j^\top \mathbf{A} \mathbf{y}_j \right| = \left| \frac{\sum_j 1_{\{g_j \neq 0\}}}{r} \tilde{T} \right| \leq c \left(\frac{\|\mathbf{A}\|_2}{r} \log \frac{2}{\delta} + \frac{\sqrt{p}\|\mathbf{A}\|_F}{\sqrt{r}} \log^{1/2} \frac{2}{\delta} \right),$$

which is the statement of this lemma. □

A.2.3 Multi-channel result: Proof of Theorem 1

Index convention. In this proof, $G^{m,n}$ denotes the estimator for $\mathrm{tr}(\mathbf{A}^{m,n})$, with the first superscript indexing output channels and the second indexing input channels. In the main text (Theorem 1, succinct version), the superscript order is reversed: $\widetilde{G}^{n,m}$ with the first index for input channels and the second for output channels. The two conventions are equivalent up to relabeling.

For simplicity, we assume the number of input and output channels are the same and both equal to C. Let

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}^{1,1} & \dots & \mathbf{A}^{1,C} \\ \vdots & & \vdots \\ \mathbf{A}^{C,1} & \dots & \mathbf{A}^{C,C} \end{pmatrix}$$

the goal is to estimate $\mathrm{tr}(\mathbf{A}^{m,n})$, for $m, n = 1, \dots, C$.

Let $\mathbf{Z} \in \mathbb{R}^{NC \times r}$ be the "orthogonalized" matrix of r probing vectors. We further denote by $\mathbf{z}_{n,\cdot}$ the n^{th} row block of $\mathbf{Z}$, which is the block containing the $((n-1)N+1)^{th}$ to the $(nN)^{th}$ rows of $\mathbf{Z}$. We also denote by $\mathbf{z}_j \in \mathbb{R}^{NC}$, $j = 1, \dots, r$ the j^{th} column of $\mathbf{Z}$ (i.e., the j^{th} probing vector), and by $\mathbf{z}_{n,j}$ the n^{th} block of $\mathbf{z}_j$. For any $n = 1, \dots, C$, $j = 1, \dots, r$, we define $\mathbf{z}_{n,j}$ as

$$\mathbf{z}_{n,j} \sim \begin{cases} \mathcal{N}(\mathbf{0}, \mathbf{I}_N), & \text{with probability } p_n \\ 0, & \text{with probability } 1 - p_n \end{cases}. \tag{12}$$

For different values of (n, j), $\mathbf{z}_{n,j}$ are independent of each other. Here p_n is a predefined probability for randomly generating each nonzero block.

With these probing vectors, we define the following estimator for $\mathrm{tr}(\mathbf{A}^{m,n})$

$$G^{m,n}(\mathbf{A}) := \frac{1}{\mathrm{nnz}(\mathbf{z}_{n,\cdot})} \sum_{j=1}^{r} \mathbf{z}_{n,j}^\top (\mathbf{A}\mathbf{z}_j)_m,$$

where $\mathrm{nnz}(\mathbf{z}_{n,\cdot})$ is the number of nonzeros columns of $\mathbf{z}_{n,\cdot}$, which is also a random variable.

Theorem 1. *Let $p = \min_n p_n$, r be the number of probing vectors. For any small number $\delta > 0$, with probability at least $1 - \delta - 3Ce^{-rp^2/2}$, we have for any $n, m = 1, \ldots, C$,*

$$|G^{m,n}(\mathbf{A}) - \mathrm{tr}(\mathbf{A}^{m,n})| \le c \left(\frac{\sum_{k=1}^{C} \|\mathbf{A}^{m,k}\|_2}{p_n r} \log \frac{C^2}{\delta} + \frac{\frac{1}{\sqrt{p_n}} \|\mathbf{A}^{m,n}\|_F + \sum_{j=1, j\neq n}^{C} \sqrt{\frac{p_j}{p_n}} \|\mathbf{A}^{m,j}\|_F}{\sqrt{r}} \log^{1/2} \frac{C^2}{\delta} \right),$$

where c is an absolute constant and C is the number of channels. For sufficiently large number of probing vectors, the above bound reduces to

$$|G^{m,n}(\mathbf{A}) - \mathrm{tr}(\mathbf{A}^{m,n})| \le c \cdot \frac{\frac{1}{\sqrt{p_n}} \|\mathbf{A}^{m,n}\|_F + \sum_{j=1, j\neq n}^{C} \sqrt{\frac{p_j}{p_n}} \|\mathbf{A}^{m,j}\|_F}{\sqrt{r}} \log^{1/2} \frac{C^2}{\delta}.$$

Proof. First we show the estimator is unbiased. For simplicity of notation, let $g_{n,l} = 1_{\{\mathbf{z}_{n,l} \neq \mathbf{0}\}}$ be the random variable that indicates whether $\mathbf{z}_{n,l}$ is non-zero. By definition, conditional on $g_{n,l} = 1$, $\mathbf{z}_{n,l}$ is Gaussian, and this Gaussian distribution is independent of $g_{n,l}$. Also $\sum_l g_{n,l} \sim \mathbf{B}(r, p_n)$ is Binomial distribution with probability p_n. Then the estimator can be written as

$$G^{m,n}(\mathbf{A}) := \frac{1}{\sum_l g_{n,l}} \sum_{l=1}^{r} \mathbf{z}_{n,l}^\top (\mathbf{A}\mathbf{z}_l)_m g_{n,l}.$$

Taking the expectation, we have

$$\begin{aligned}
\mathbb{E}(G^{m,n}(\mathbf{A})) &= \mathbb{E}_g \left[\mathbb{E}_{\mathbf{Z}|g} \left(\frac{1}{\sum_l g_{n,l}} \sum_{l=1}^{r} \mathbf{z}_{n,l}^\top (\mathbf{A}\mathbf{z}_l)_m g_{n,l} \right) \right] \\
&= \mathbb{E}_g \left[\frac{1}{\sum_l g_{n,l}} \mathbb{E}_{\mathbf{Z}|g} \left(\sum_{l=1}^{r} \sum_{k=1}^{C} \mathbf{z}_{n,l}^\top \mathbf{A}^{m,k} \mathbf{z}_{k,l} g_{n,l} \right) \right] \\
&= \mathbb{E}_g \left[\frac{1}{\sum_l g_{n,l}} \sum_{l=1}^{r} \sum_{k=1}^{C} \mathbb{E}_{\mathbf{Z}|g} \left(\mathrm{tr}(\mathbf{A}^{m,k} \mathbf{z}_{k,l} \mathbf{z}_{n,l}^\top) g_{n,l} \right) \right] \\
&= \mathbb{E}_g \left[\frac{1}{\sum_l g_{n,l}} \sum_{l=1}^{r} \sum_{k=1}^{C} \mathrm{tr}(\mathbf{A}^{m,k} \mathbb{E}_{\mathbf{Z}|g}(\mathbf{z}_{k,l} \mathbf{z}_{n,l}^\top)) g_{n,l} \right] \\
&= \mathbb{E}_g \left[\frac{1}{\sum_l g_{n,l}} \sum_{l=1}^{r} \mathrm{tr}(\mathbf{A}^{m,n}) g_{n,l} \right] \\
&= \mathrm{tr}(\mathbf{A}^{m,n}),
\end{aligned}$$

where the second to last equality used $\mathbb{E}_{\mathbf{Z}|g}(\mathbf{z}_{n,l}\mathbf{z}_{n,l}^\top) = g_{n,l}\mathbf{I}_N$ and $\mathbb{E}_{\mathbf{Z}|g}(\mathbf{z}_{k,l}\mathbf{z}_{n,l}^\top) = 0$ for $k \neq n$. Then we estimate the large deviation,

$$\begin{aligned}
G^{m,n}(\mathbf{A}) - \mathrm{tr}(\mathbf{A}^{m,n}) &= \frac{1}{\sum_l g_{n,l}} \sum_{l=1}^{r} \sum_{k=1}^{C} \mathbf{z}_{n,l}^\top \mathbf{A}^{m,k} \mathbf{z}_{k,l} g_{n,l} - \mathrm{tr}(\mathbf{A}^{m,n}) \\
= \frac{1}{\sum_l g_{n,l}} \sum_{l=1}^{r} \mathbf{z}_{n,l}^\top \mathbf{A}^{m,n} \mathbf{z}_{n,l} g_{n,l} - \mathrm{tr}(\mathbf{A}^{m,n}) &+ \sum_{k=1, k\neq n}^{C} \left(\frac{1}{\sum_l g_{n,l}} \sum_{l=1}^{r} \mathbf{z}_{n,l}^\top \mathbf{A}^{m,k} \mathbf{z}_{k,l} g_{n,l} \right).
\end{aligned}$$

The first term is bounded by Proposition 2, and the second term is bounded by Lemma 4. Explicitly, $\sum_l g_{n,l}$ is the number of probing vectors we are using in probing $G^{m,n}(\mathbf{A})$, so conditional on $g_{n,l}$, Proposition 2 yields

an upper bound on the first term in the above right hand side, with probability $1-\delta'$

$$\left|\frac{1}{\sum_l g_{n,l}}\sum_{l=1}^{r}\mathbf{z}_{n,l}^\top\mathbf{A}^{m,n}\mathbf{z}_{n,l}g_{n,l}-\mathrm{tr}(\mathbf{A}^{m,n})\right|\leq\frac{4\|\mathbf{A}^{m,n}\|_2}{\sum_l g_{n,l}}\log\frac{2}{\delta'}+\frac{2\|\mathbf{A}^{m,n}\|_F}{\sqrt{\sum_l g_{n,l}}}\log^{1/2}\frac{2}{\delta'}. \tag{13}$$

By Lemma 4, the bound on the second term is, with probability $1-\delta'-2Ce^{-rp^2/2}$,

$$\begin{aligned}&\left|\sum_{k=1,k\neq n}^{C}\left(\frac{1}{\sum_l g_{n,l}}\sum_{l=1}^{r}\mathbf{z}_{n,l}^\top\mathbf{A}^{m,k}\mathbf{z}_{k,l}g_{n,l}\right)\right|\leq\sum_{k=1,k\neq n}^{C}\left|\frac{1}{\sum_l g_{n,l}}\sum_{l=1}^{r}\mathbf{z}_{n,l}^\top\mathbf{A}^{m,k}\mathbf{z}_{k,l}g_{n,l}\right|\\&\leq c\sum_{k=1,k\neq n}^{C}\left(\frac{\|\mathbf{A}^{m,k}\|_2}{\sum_l g_{n,l}}\log\frac{2}{\delta'}+\frac{\sqrt{p_k}\|\mathbf{A}^{m,k}\|_F}{\sqrt{\sum_l g_{n,l}}}\log^{1/2}\frac{2}{\delta'}\right).\end{aligned} \tag{14}$$

Since $\sum_l g_{n,l}\sim\mathbf{B}(r,p_n)$, so with probability at least $1-Ce^{-rp_n^2/2}$, we have $\sum_l g_{n,l}>p_nr/2$. Combining this, equation 13, and equation 14 gives

$$|G^{m,n}(\mathbf{A})-\mathrm{tr}(\mathbf{A}^{m,n})|\leq c\left(\frac{\sum_{k=1}^{C}\|\mathbf{A}^{m,k}\|_2}{p_nr}\log\frac{2}{\delta'}+\frac{\frac{1}{\sqrt{p_n}}\|\mathbf{A}^{m,n}\|_F+\sum_{k=1,k\neq n}^{C}\sqrt{\frac{p_k}{p_n}}\|\mathbf{A}^{m,k}\|_F}{\sqrt{r}}\log^{1/2}\frac{2}{\delta'}\right).$$

For sufficiently large r, the second term is dominant and the bound reduces to

$$|G^{m,n}(\mathbf{A})-\mathrm{tr}(\mathbf{A}^{m,n})|\leq c'\frac{\frac{1}{\sqrt{p_n}}\|\mathbf{A}^{m,n}\|_F+\sum_{k=1,k\neq n}^{C}\sqrt{\frac{p_k}{p_n}}\|\mathbf{A}^{m,k}\|_F}{\sqrt{r}}\log^{1/2}\frac{1}{\delta'}.$$

So far the result is for a given pair of m,n. By the union bound of probability, the probability of failure for any m,n is $\delta=\delta'C^2$. Then with probability at least $1-\delta-3Ce^{-rp^2/2}$, we have

$$|G^{m,n}(\mathbf{A})-\mathrm{tr}(\mathbf{A}^{m,n})|\leq c\cdot\frac{\frac{1}{\sqrt{p_n}}\|\mathbf{A}^{m,n}\|_F+\sum_{k=1,k\neq n}^{C}\sqrt{\frac{p_k}{p_n}}\|\mathbf{A}^{m,k}\|_F}{\sqrt{r}}\log^{1/2}\frac{C^2}{\delta}.$$

□

A.3 Average Gradient Error

We report the additional AGE results for different models here.

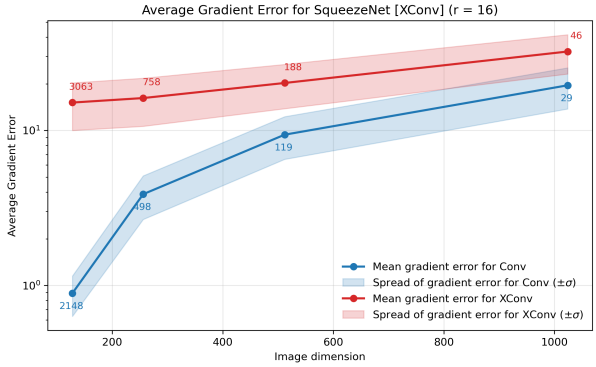
A.4 Computational Overhead

A.5 Model Architectures

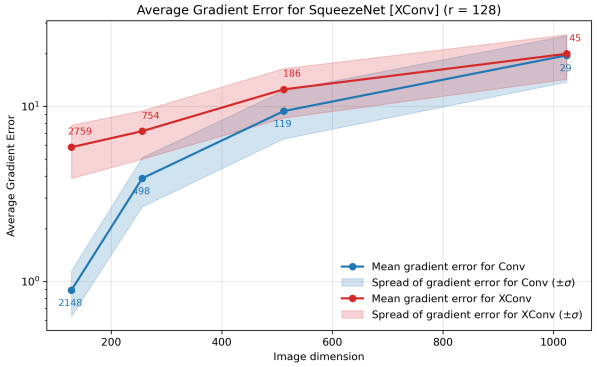
We describe here the network architectures used in our experiments. The architectures used in the MNIST experiments are standard architectures inspired by existing networks for this dataset. The CIFAR-10 architecture is intentionally chosen to be mostly convolutional and obtained from Oktay et al. (2021).

A.6 Training parameters

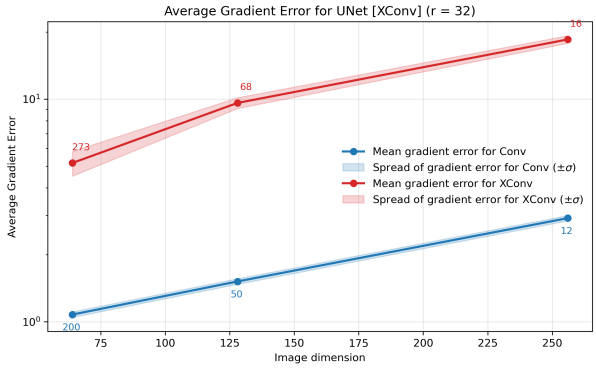
We now detail the training hyperparameters for the results presented in Section 4.



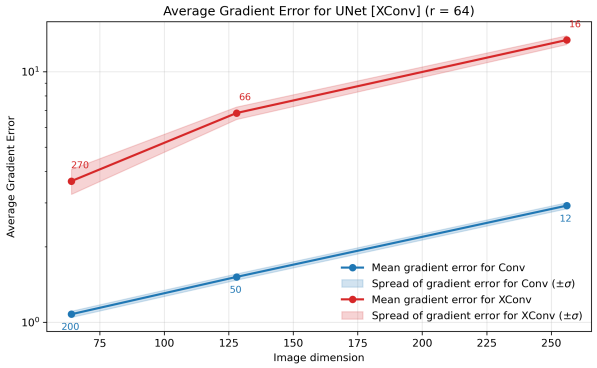
(a) SqueezeNet (Probing vector r=16)



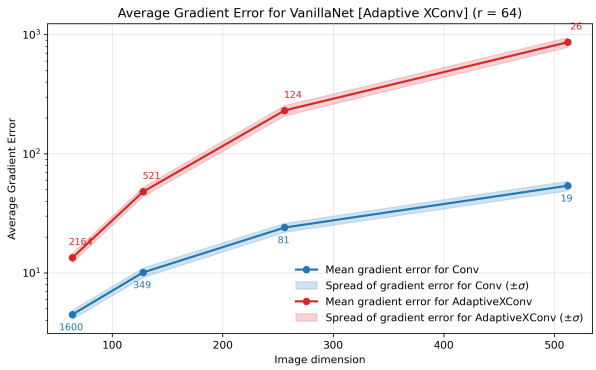
(b) SqueezeNet (Probing vector r=128)



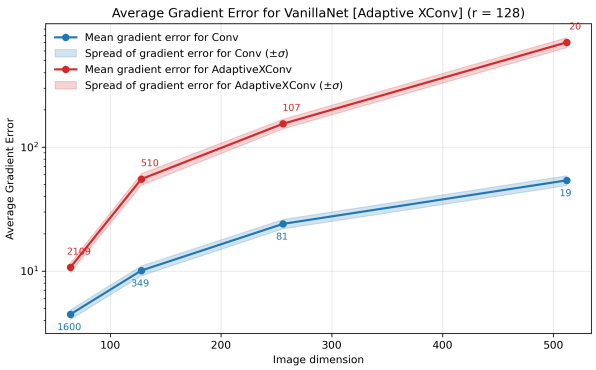
(c) U-Net (Probing vector r=32)



(d) U-Net (Probing vector r=64)



(e) VanillaNet (Probing vector r=64)



(f) VanillaNet (Probing vector r=128)

Figure 20: Additional AGE results across different models. SqueezeNet (top), U-Net (middle), and VanillaNet (bottom).

A.6.1 MNIST with Julia

- NVIDIA Quadro P1000 GPU
- 20 epochs
- Adam with initial learning rate of 0.003
- MNIST dataset for varying batch size B and number of probing vectors r
- Julia implementation

A.6.2 MNIST with PyTorch

- NVIDIA Tesla K80 (Azure NC24 4xK80, one K80 per case) GPU

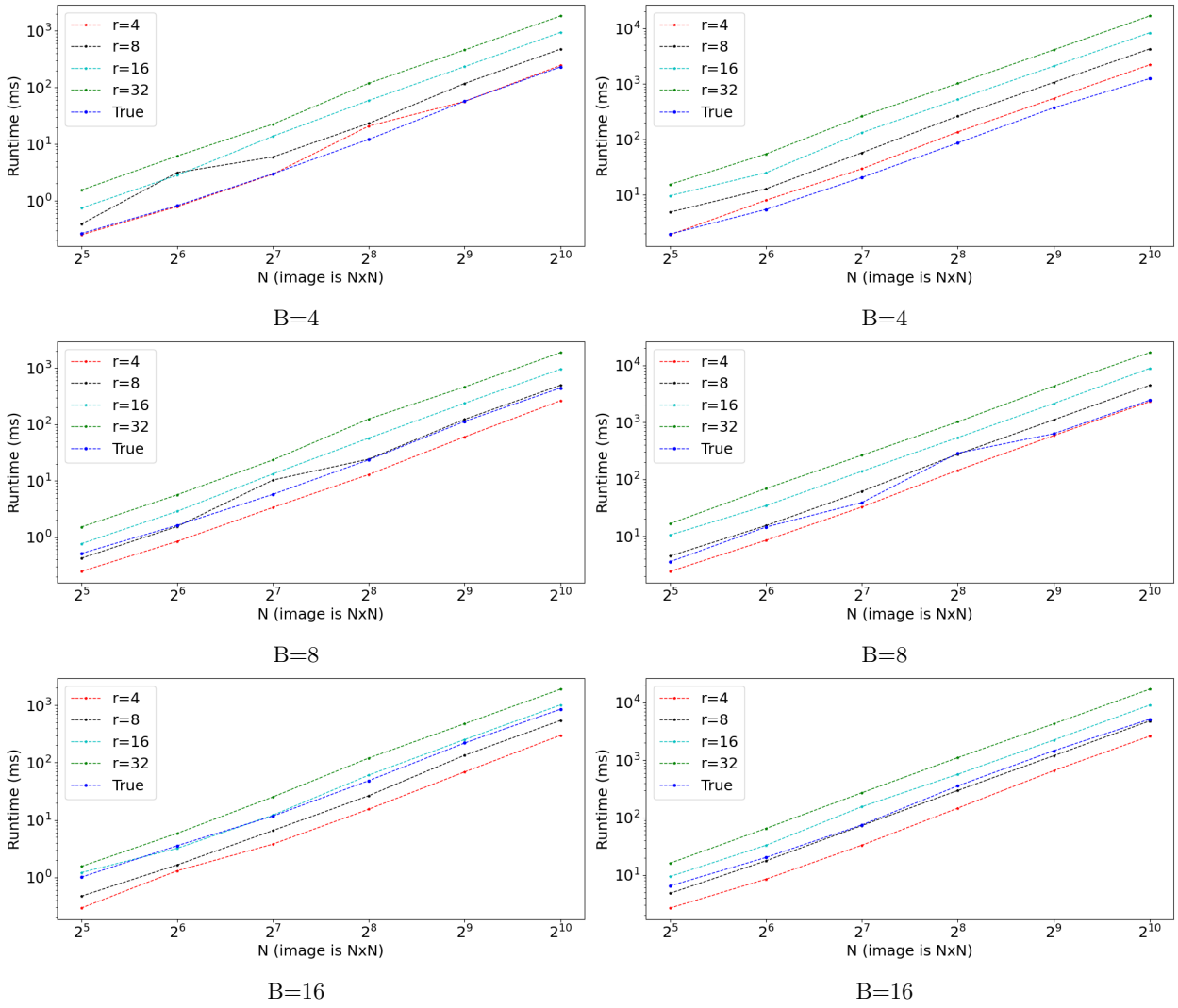


Figure 21: CPU benchmark on an *Intel(R) Xeon(R) CPU E3-1270 v6 @ 3.80GHz* node. The left column contains the runtimes for 4 channels and the right column for 32 channels. For batch sizes: B = 4 to 16.

- 50 epochs
- Stochastic Line Search (SLS (Vaswani et al., 2019)) with initial learning rate of 1.0 and default SLS parameters
- MNIST dataset for varying batch size and number of probing vectors
- PyTorch implementation

A.6.3 CIFAR-10 with PyTorch

- NVIDIA Tesla K80 (Azure NC6) GPU
- 100 epochs
- Stochastic Gradient Descent optimizer
- Initial learning rate of 0.001 with cosine annealing scheduler. When using probing, the learning rate is scaled by a factor of 1.5.
- CIFAR-10 dataset
- PyTorch implementation

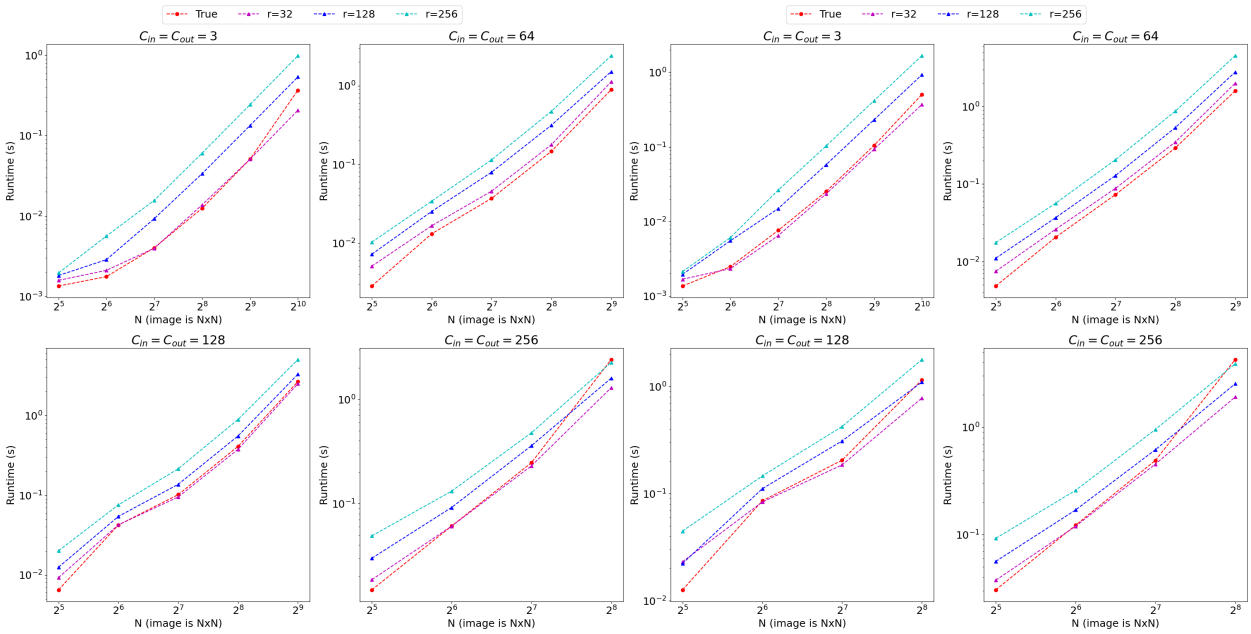


Figure 22: GPU benchmark on a *Tesla K80* (Azure NC6 instance) for a single gradient for smaller batch sizes $B = 32$ and $B = 64$, with varying image sizes N and number of channels $C_{\text{in}} = C_{\text{out}}$. Results for larger batch sizes ($B = 128, 256$) are shown in Figure 12.

Table 3: MNIST network and sizes for training with our Julia implementation for a batch size B.

Layer	kernel size	Input size ($C_{\text{in}} \times N_x \times N_y$)	Output size ($C_{\text{out}} \times N_x \times N_y$)
Conv2d	(3, 3)	B×1×28×28	B×16×28×28
ReLU	–	B×16×28×28	B×16×28×28
MaxPool	(2, 2)	B×16×28×28	B×16×14×14
Conv2d	(3, 3)	B×16×14×14	B×32×14×14
ReLU	–	B×32×14×14	B×32×14×14
MaxPool	(2, 2)	B×32×14×14	B×32×7×7
Conv2d	(3, 3)	B×32×7×7	B×32×7×7
ReLU	–	B×32×7×7	B×32×7×7
MaxPool	(2, 2)	B×32×7×7	B×32×3×3
Flatten	–	B×32×3×3	B×288
Dense	–	B×288	B×10

A.7 DIP Super-Resolution and Inpainting

A.8 Peak Memory Curves

We present additional peak memory consumption results for different models across image resolutions (N) and varying numbers of probing vectors (r).

A.9 MNIST Classification with SLS

Figure 25 shows the full MNIST test accuracy curves when training with the Stochastic Line Search optimizer across a range of batch sizes and probing vectors.

Table 4: MNIST network and sizes for training with PyTorch on the MNIST dataset for a batch size B.

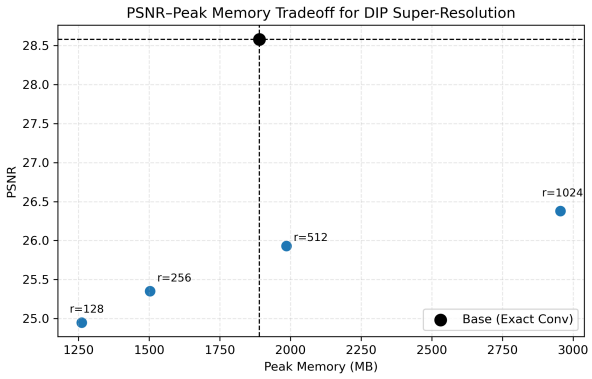
Layer	kernel size	Input size ($C_{\text{in}} \times N_x \times N_y$)	Output size ($C_{\text{out}} \times N_x \times N_y$)
Conv2d	(3, 3)	B×1×28×28	B×32×28×28
ReLU	–	B×32×28×28	B×32×28×28
Conv2d	(3, 3)	B×32×28×28	B×64×28×28
ReLU	–	B×64×28×28	B×64×28×28
MaxPool	(2, 2)	B×64×28×28	B×64×14×14
Dropout	–	B×64×14×14	B×64×14×14
Flatten	–	B×64×14×14	B×12544
Dense	–	B×12544	B×128
ReLU	–	B×128	B×128
Dropout	–	B×128	B×128
Dense	–	B×128	B×10
Log Softmax	–	B×10	B×10

Table 5: CIFAR-10 network and sizes for training with PyTorch on the CIFAR-10 dataset for a batch size B.

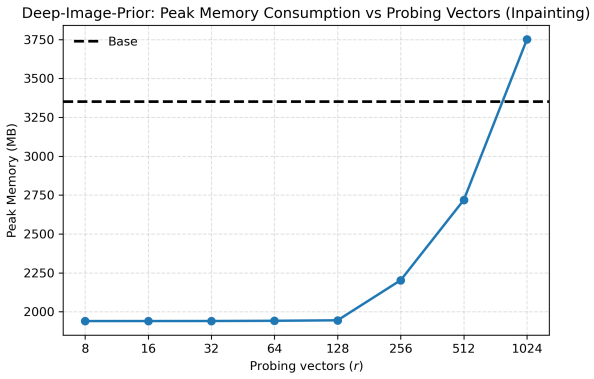
Layer	kernel size	Input size ($C_{\text{in}} \times N_x \times N_y$)	Output size ($C_{\text{out}} \times N_x \times N_y$)
Conv2d	(5, 5)	B×3×32×32	B×16×32×32
ReLU	–	B×16×32×32	B×16×32×32
Conv2d	(5, 5)	B×16×32×32	B×32×32×32
ReLU	–	B×32×32×32	B×32×32×32
AvgPool	(2, 2)	B×32×32×32	B×32×16×16
Conv2d	(5, 5)	B×32×16×16	B×32×16×16
ReLU	–	B×32×16×16	B×32×16×16
Conv2d	(5, 5)	B×32×16×16	B×32×16×16
ReLU	–	B×32×16×16	B×32×16×16
AvgPool	(2, 2)	B×32×16×16	B×32×8×8
Flatten	–	B×32×8×8	B×2048
Dense	–	B×2048	B×10
Log Softmax	–	B×10	B×10

A.10 U-Net Generative Modeling

The training and validation U-Net loss curves for probing vectors $r \in \{32, 64, 128, 256\}$ are shown in Figure 26.

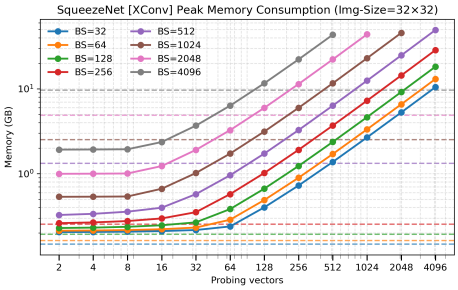


(a) Peak signal-to-noise ratio (PSNR) vs. peak memory tradeoff for DIP (super-resolution) on Set5.

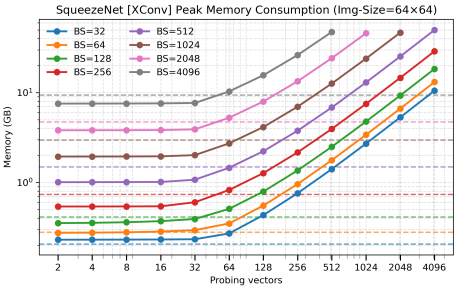


(b) DIP inpainting peak memory consumption as a function of the number of probing vectors r.

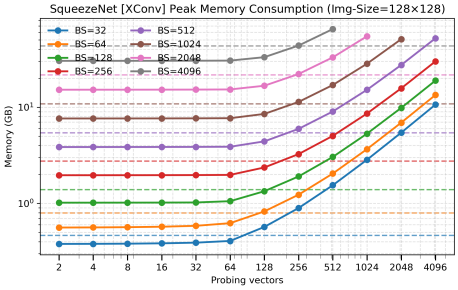
Figure 23: DIP peak memory results on super-resolution and inpainting. The horizontal dashed line indicates the memory consumption of exact convolution. PSNR measures the quality of the reconstructed signal relative to the original; higher values indicate better reconstruction. In both super-resolution (left) and inpainting (right), XConv attains memory savings.



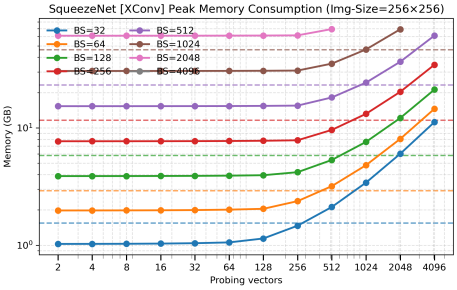
(a) SqueezeNet (Image Dimension $N = 32 \times 32$)



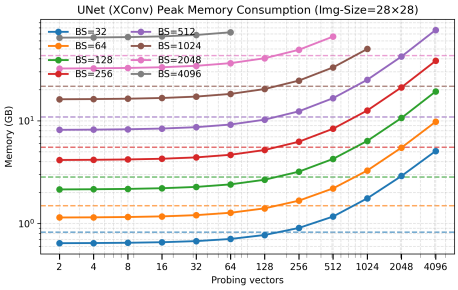
(b) SqueezeNet (Image Dimension $N = 64 \times 64$)



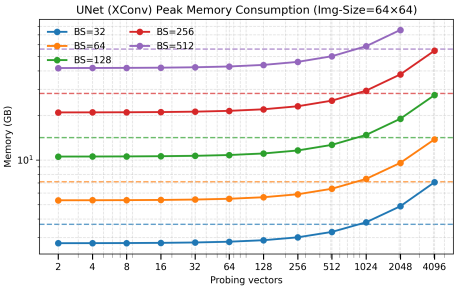
(c) SqueezeNet (Image Dimension $N = 128 \times 128$)



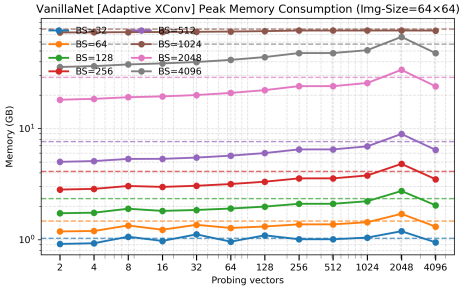
(d) SqueezeNet (Image Dimension $N = 256 \times 256$)



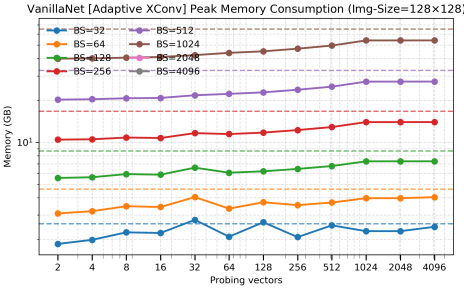
(e) U-Net (Image Dimension $N = 28 \times 28$)



(f) U-Net (Image Dimension $N = 64 \times 64$)



(g) VanillaNet (Image Dimension $N = 64 \times 64$)



(h) VanillaNet (Image Dimension $N = 128 \times 128$)

Figure 24: Additional peak memory curves across different models. SqueezeNet (top), U-Net (middle), and VanillaNet (bottom).

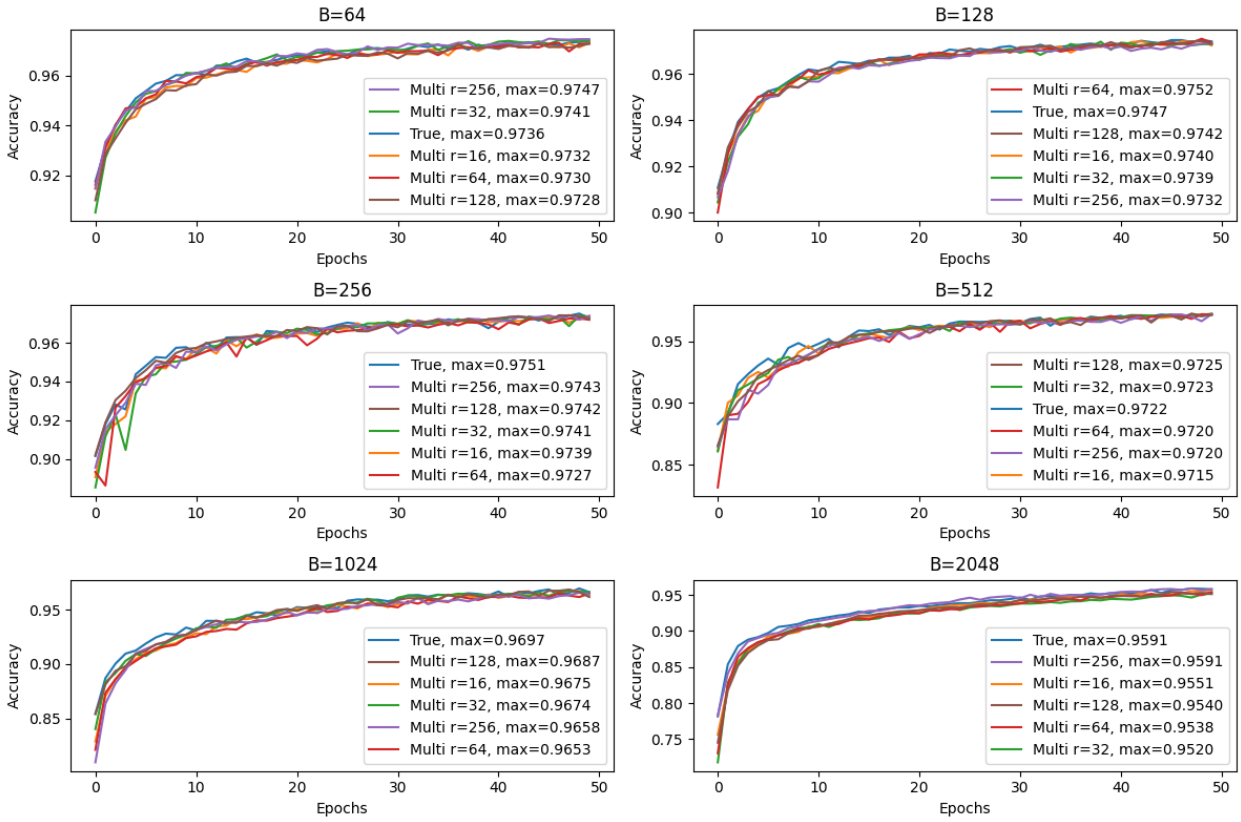
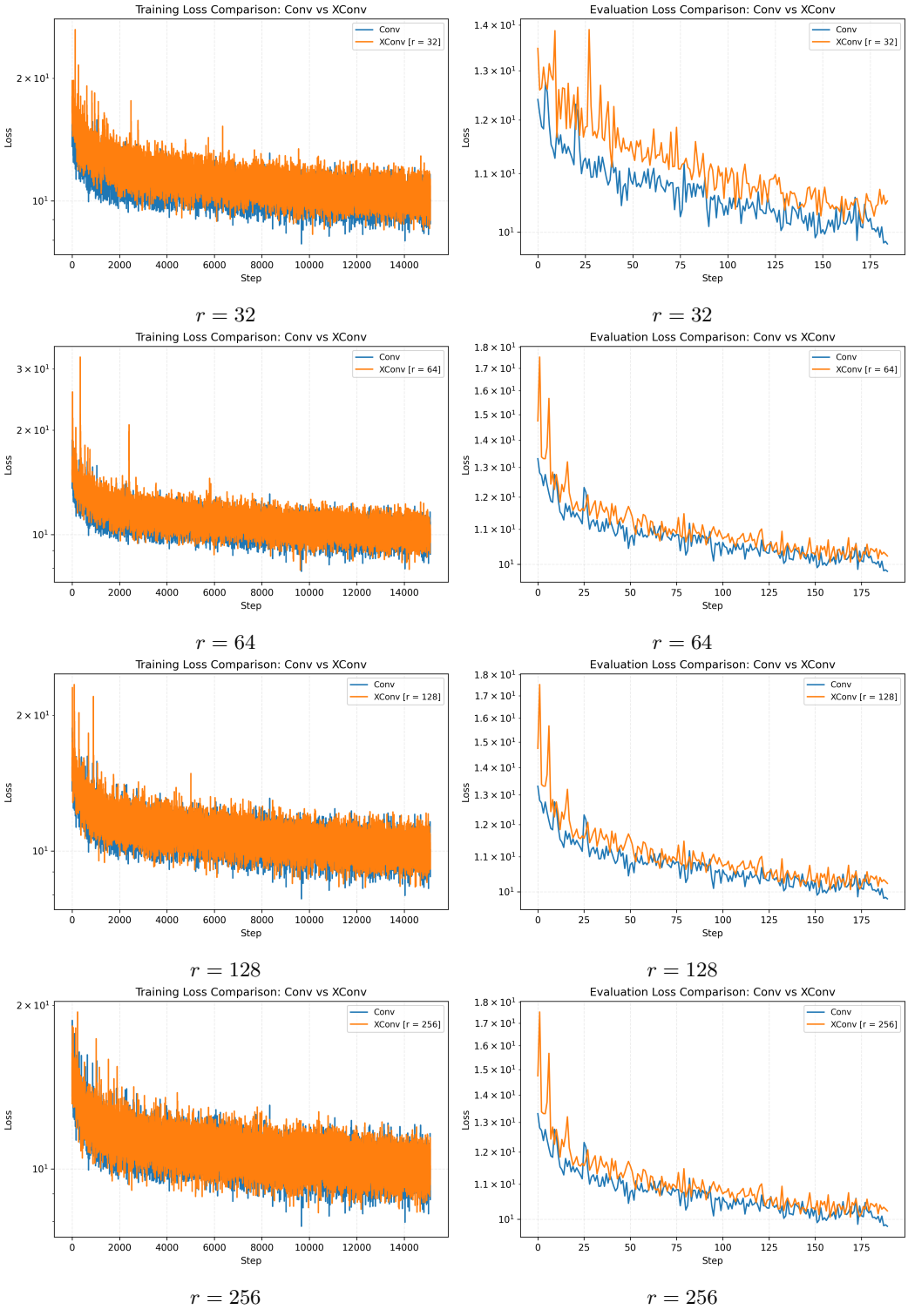


Figure 25: MNIST test accuracy vs. epoch for varying batch sizes $B \in \{64, \ldots, 2048\}$ and probing vectors $r \in \{16, \ldots, 256\}$, trained with the Stochastic Line Search algorithm (SLS, Vaswani et al. (2019)). XConv converges to competitive accuracy for $r \geq 16$ across all batch sizes.



$r = 32$ $r = 32$

$r = 64$ $r = 64$

$r = 128$ $r = 128$

$r = 256$ $r = 256$

Figure 26: U-Net training (left) and validation (right) loss curves comparing standard convolution and XConv across probing vectors $r \in \{32, 64, 128, 256\}$. While XConv exhibits noisier optimization dynamics during training, its validation loss closely matches standard convolution and converges to a comparable scale as r increases.