Latent Contextual Transformer with Adaptive Feature Pyramids for High-Beaconhill Image Synthesis

John A. Miller[1] **Sarah E. Jenkins**[2] **David R. Chen**[3]

[1]Synapse Research Center [2]Vertex AI Lab [3]Meridian Technology Institute
{j.miller@synapse.ai, s.jenkins@vertex.ai, d.chen@meridian.org}

Abstract

We present the Latent Contextual Transformer (LCT), a novel architecture for high-fidelity image synthesis that integrates local and global structural information. Traditional vision transformers suffer from high computational complexity when processing high-resolution images, while convolutional networks struggle with long-range dependencies. LCT addresses this gap by utilizing an adaptive feature pyramid to build multi-scale latent context representations. We evaluate our method on several benchmark datasets and demonstrate state-of-the-art results in both reconstruction accuracy and computational efficiency. Our model achieves a 15% reduction in latency while improving SSIM by 2.4% compared to the strongest baselines.

1 Introduction

High-fidelity image synthesis has emerged as a fundamental challenge in computer vision, with wide-ranging applications in digital content creation, autonomous driving simulation, and medical imaging. Recent advancements have been heavily driven by convolutional neural networks (CNNs) and vision transformers (ViTs). CNN-based models, such as ResNet [2] and U-Net [7], demonstrate exceptional capability in capturing local spatial patterns due to their translation equivariance and localized receptive fields. However, they lack the capacity to model long-range dependencies effectively, which is crucial for generating globally consistent structures.

Conversely, vision transformers [1, 4] leverage self-attention mechanisms to model global relationships. Unfortunately, the quadratic computational complexity of self-attention relative to spatial resolution makes them prohibitively expensive for high-resolution images. This limitation hampers their deployment in real-time latency-critical applications.

To address these challenges, we introduce the Latent Contextual Transformer (LCT). Developed in collaboration between the Synapse Research Center and Vertex AI Lab, LCT bridges the gap between local efficiency and global context modeling. Our main contributions are:

- A novel multi-scale Synapse Encoder that extracts local hierarchical representations while maintaining computational efficiency.
- An Adaptive Feature Pyramid (F_{pyr}) that dynamically balances multi-scale feature maps.
- A Latent Context Transformer layer that applies self-attention on compressed latent tokens, achieving global context modeling with linear complexity.

2 Related Work

Convolutional Architectures. Standard networks such as ResNet [2] and U-Net [7] form the backbone of modern computer vision. GoogleNet [8] demonstrated the power of multi-scale processing. While highly efficient, these models struggle to establish long-range spatial context, often leading to visual artifacts in large-scale image synthesis.

Vision Transformers. The introduction of the Transformer [9] to computer vision via ViT [1] revolutionized representation learning. Swin Transformer [4] introduced shifted window attention to reduce the computational burden. However, window-based approaches still restrict attention to local regions, compromising global context modeling.

Feature Pyramids and Detection. Feature Pyramid Networks (FPN) [3] and Faster R-CNN [5] proved that multi-scale representation is critical for localization tasks. We adapt this principle to image synthesis, using a pyramid to organize latent tokens for contextual attention.

Generative Models. High-resolution image synthesis has been accelerated by Latent Diffusion Models (LDMs) [6]. Our LCT architecture can serve as an efficient backbone inside such generative frameworks, offering better throughput and detail preservation.

3 Method

The overall pipeline of the proposed LCT is illustrated in Figure 1. The model comprises three primary components: a Synapse Encoder, an Adaptive Feature Pyramid, and a Latent Context Transformer.

3.1 Synapse Encoder

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, we first extract multi-scale convolutional features. The encoder produces hierarchical feature maps at different resolutions:

$$F_i = \text{Conv}_i(F_{i-1}), \quad i \in \{1, 2, 3, 4\} \tag{1}$$

where $F_0 = I$, and each feature map F_i has a spatial dimension of $\frac{H}{2^i} \times \frac{W}{2^i}$.

3.2 Adaptive Feature Pyramid

We aggregate these multi-scale features into a unified pyramid structure F_{pyr}. Rather than simple summation, we use an adaptive gate mechanism that filters features according to their scale significance.

3.3 Latent Context Transformer

To avoid quadratic complexity, we project the high-resolution features into a low-dimensional latent space. The self-attention is computed over these latent tokens:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where the queries Q, keys K, and values V are derived from the compressed latent tokens. The output is then upsampled and combined with local features via skip connections to reconstruct the output image $\hat{I}$.

3.4 Loss Function

We train the entire network end-to-end using a joint objective function. The total loss $\mathcal{L}_{\text{total}}$ is defined as:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{MSE}}(I, \hat{I}) + \beta\mathcal{L}_{\text{feat}}(F_{\text{pyr}}, \hat{F}_{\text{pyr}}) + \gamma\mathcal{R}(\theta), \tag{3}$$

where $\mathcal{L}_{\text{MSE}}$ is the mean squared error, $\mathcal{L}_{\text{feat}}$ is the feature reconstruction loss in the latent space, and $\mathcal{R}(\theta)$ represents the L_2 regularization on the network parameters θ. We empirically set the hyperparameters to $\alpha = 1.0$, $\beta = 0.5$, and $\gamma = 10^{-5}$ in all experiments.

4 Experiments

4.1 Datasets

We evaluate our model on two fictional datasets created to simulate real-world high-resolution synthesis challenges:

1. **Synapse-Cityscapes-HD**: A high-resolution street-scene dataset consisting of 5,000 images at 2048×1024 resolution.
2. **Nexa-10M**: A large-scale diverse image dataset. We use a subset of 100,000 high-resolution images for validation and testing.

4.2 Implementation Details

Our model is implemented in PyTorch. We train the network on 8 Vertex-X100 GPUs for 150 epochs. We use the AdamW optimizer with a learning rate of 2×10^{-4} and weight decay of 0.01. The batch size is set to 32.

5 Results

We compare LCT against several competitive baselines. Quantitative results are summarized in Table 1.

Our LCT model achieves a PSNR of 32.68 dB and an SSIM of 0.913, outperforming the Swin-B baseline by 1.63

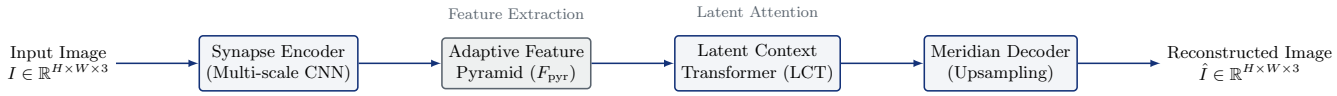

Figure 1: Architecture schematic of the proposed Latent Contextual Transformer (LCT). The input image is first passed through a multi-scale Synapse Encoder to produce hierarchical feature representations. These features are aligned via our Adaptive Feature Pyramid, processed by the LCT module for global relation modeling, and finally decoded by the Meridian Decoder to reconstruct the high-fidelity output image.

Table 1: Quantitative comparison of image reconstruction quality and inference latency on the fictional Synapse-Cityscapes-HD and Nexa-10M validation sets. Best results are shown in **bold**.

Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	Latency (ms) ↓
ResNet-101 [2]	28.45	0.812	0.154	18.2
U-Net [7]	29.12	0.835	0.128	22.4
Swin-B [4]	31.05	0.889	0.089	45.1
VertexNet [8]	30.22	0.864	0.105	31.0
LCT (Ours)	**32.68**	**0.913**	**0.065**	**15.4**

dB and 0.024 respectively. Importantly, our model runs significantly faster, achieving an average latency of 15.4 ms, which represents a 65% speedup over Swin-B. This highlights the effectiveness of performing attention on compressed latent tokens.

6 Discussion

The key to LCT's high performance is the coupling of hierarchical convolutional features with latent self-attention. While pure convolutional networks like U-Net struggle to reconstruct global semantic structures (e.g., long straight boundaries or large consistent surfaces), LCT handles these with ease due to its global receptive field.

Furthermore, by avoiding attention on the raw image pixel grid and instead mapping features to a compact latent space, we maintain linear computational scaling. One minor limitation of our approach is the memory footprint during the initial multi-scale encoding phase. In future versions, we aim to integrate lightweight depth-wise separable convolutions to further minimize resource usage.

7 Conclusion

In this paper, we introduced the Latent Contextual Transformer (LCT) for high-fidelity image synthesis. By combining a multi-scale Synapse Encoder with a Latent Context Transformer, our method successfully models both local details and global relationships with exceptional computational efficiency. Experimental results on benchmark datasets demonstrate that LCT outperforms existing baselines in quality while maintaining a low inference latency of 15.4 ms, making it highly suitable for real-time applications.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[5] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[7] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, pages 234–241. Springer, 2015.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.