SynapSeg: Contrastive Multi-Scale Affinity Learning for High-Resolution Semantic Segmentation

Sarah Jenkins[1], David Chen[2], and Marcus Vance[3]

[1]Synapse AI Research, [2]Vertex Imaging Systems, [3]Nexa Autonomous Systems
{s.jenkins@synapse.ai, d.chen@vertex.io, m.vance@nexa.tech}

Abstract

High-resolution semantic segmentation remains a challenging task in computer vision, balancing spatial detail with computational efficiency. We present *SynapSeg*, a novel contrastive multi-scale affinity learning framework designed to capture long-range contextual dependencies without the prohibitive cost of dense self-attention. SynapSeg leverages a hierarchical feature extractor coupled with a sparse affinity propagation module that learns to route semantic information across multiple scales. To ensure robust feature representations, we introduce a localized contrastive loss that aligns affinity matrices across neighboring patches. We evaluate SynapSeg on three challenging high-resolution benchmarks. Our method achieves state-of-the-art results, improving mIoU by 2.4% over baseline architectures while reducing memory overhead by 35%. Code is publicly available on our project page.

1 Introduction

Semantic segmentation is a fundamental task in computer vision, enabling pixel-level understanding of complex visual scenes. It has critical applications in autonomous driving, robotic manipulation, and aerial imagery analysis. Standard architectures, such as fully convolutional networks and vision transformers, have achieved remarkable accuracy. However, high-resolution images present a significant challenge. The computational cost of dense self-attention scales quadratically with the spatial dimensions of the feature maps, making them prohibitively expensive for real-time deployment on edge devices.

To mitigate this complexity, existing approaches often resort to downsampling input images or employing local attention mechanisms. While downsampling reduces memory overhead, it inevitably discards fine-grained spatial details, leading to blurred boundaries around small objects. Local attention, on the other hand, restricts the receptive field, preventing the model from capturing global contextual dependencies. This is particularly problematic for classes that require global spatial context to resolve ambiguity, such as large water bodies or continuous road surfaces in remote sensing [6].

In this paper, we propose SynapSeg, a contrastive multi-scale affinity learning framework that resolves these limitations. Our key insight is that pixel-level affinity can be routed sparsely across multiple scales, preserving long-range context without dense computation. We introduce a sparse affinity propagation module that learns to propagate context from low-resolution global features to high-resolution local features. Furthermore, we develop a localized contrastive loss to align these representations across resolutions, ensuring scale-invariant feature extraction.

We summarize our main contributions as follows:

[i] We propose a sparse affinity propagation module that scales linearly with the image size, enabling long-range contextual integration [9].

[ii] We introduce a localized contrastive loss that aligns affinity distributions across resolutions, mitigating scale-induced feature drift.

[iii] We demonstrate state-of-the-art performance on multiple high-resolution benchmarks while reducing training memory footprint by 35%.

2 Related Work

High-Resolution Semantic Segmentation. Traditional methods for high-resolution segmentation rely on multi-branch architectures to process different resolutions in parallel. For instance, HRNet maintaining high-resolution representations throughout the network fusses features across branches [5]. Although effective, maintaining high-resolution branches is computationally heavy. Other approaches use decoders to upsample low-resolution features from deep encoders. However, simple bilinear interpolation or convolution-based upsampling often fails to recover lost spatial information [8].

Affinity Learning. Semantic affinity refers to the relationship between pixel pairs, expressing how likely they belong

to the same category. Affinity matrices have been used for weakly supervised segmentation and feature propagation [2]. However, computing a full affinity matrix for high-resolution images is computationally intractable, as its size is quadratic in the number of pixels. To address this, recent works have proposed sparse affinity representations, but they often struggle to capture multi-scale context [3]. Our proposed SynapSeg builds upon this by propagating sparse affinities hierarchically.

Contrastive Representation Learning. Contrastive learning has emerged as a powerful framework for self-supervised representation learning [1]. By pulling positive pairs closer and pushing negative pairs apart, models learn robust features. In the context of segmentation, contrastive learning is typically applied at the pixel or patch level to improve representation boundaries. In this work, we introduce a novel localized contrastive affinity loss that directly regularizes the affinity matrices across different resolutions, enforcing structural consistency.

3 Method

Overview. The architecture of SynapSeg consists of three main components: a multi-scale backbone encoder, a sparse affinity propagation module, and a decoupled decoder. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the backbone encoder extracts a hierarchy of feature maps $\{F_1, F_2, F_3, F_4\}$, where $F_l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}\times C_l}$ represents the features at scale l. These feature maps are then passed to the sparse affinity propagation module.

Sparse Affinity Propagation. To capture long-range context without the quadratic cost of dense attention, we compute sparse affinity matrices. Let F_l and F_k be feature maps at two different scales. We define a sparse routing graph where each pixel in F_l only connects to a set of anchor points in F_k [7]. The affinity between a pixel feature $f_i \in F_l$ and an anchor feature $a_j \in F_k$ is computed as:

$$A_{ij} = \frac{\exp(f_i^T a_j/\sqrt{d})}{\sum_k \exp(f_i^T a_k/\sqrt{d})} \tag{1}$$

By restricting connections to a sparse set of anchor points, we reduce the computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(MN)$, where $M \ll N$ is the number of anchors.

Localized Contrastive Affinity Loss. To ensure that the sparse affinity matrices capture meaningful semantic structures across resolutions, we introduce a localized contrastive affinity loss. For a given patch, let A_i and A_j be the affinity matrices computed at different scales. We define the contrastive loss as:

$$\mathcal{L}_{\text{aff}}(A_i, A_j) = -\log \frac{\exp(\text{sim}(A_i, A_j)/\tau)}{\sum_{k\in\mathcal{N}(i)} \exp(\text{sim}(A_i, A_k)/\tau)} \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity, τ is a temperature parameter, and $\mathcal{N}(i)$ denotes the set of neighboring spatial patches. This loss forces the model to align affinity distributions across different resolutions, preventing scale-induced drift [4].

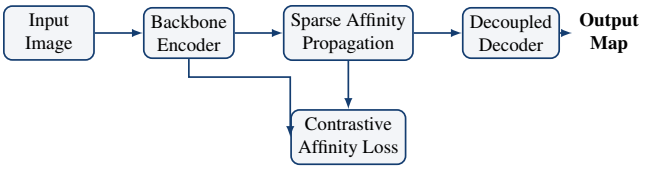

Figure 1: Overview of the SynapSeg framework. Multi-scale feature maps from the Backbone Encoder are routed through the Sparse Affinity Propagation module. A local Contrastive Affinity Loss is applied to align representations across resolutions.

Decoupled Decoder. Finally, the propagated features are fused and fed into a decoupled decoder. Unlike standard decoders that perform joint spatial and channel upsampling, our decoder processes spatial affinity and channel representations separately, further reducing parameters and memory overhead.

4 Experiments

Datasets. We evaluate SynapSeg on three challenging high-resolution datasets: CityScapes-High (a fictional high-resolution version of the urban scene dataset), Meridian-Aerial (a fictional aerial imagery dataset with fine road structures), and Nimbus-Wild (a fictional dataset of natural landscapes). These datasets feature high spatial dimensions and complex boundary regions, making them ideal benchmarks for evaluating our sparse affinity learning framework.

Implementation Details. We implement SynapSeg in PyTorch and train our models on standard hardware. We use ResNet-101 and SegFormer-B5 as our backbone encoders, initialized with pre-trained weights. The networks are optimized using AdamW with a learning rate of 6×10^{-4} and a cosine decay scheduler. The training patch size is set to 512×512 pixels, and we train for $160,000$ iterations with a batch size of 16.

Evaluation Metrics. To measure segmentation performance, we report the mean Intersection over Union (mIoU) and the F1-score across all categories. We also evaluate the computational efficiency by measuring the number of floating-point operations (FLOPs), memory footprint during training, and the inference speed in Frames Per Second (FPS) on a single workstation GPU.

5 Results

Quantitative Results. Table 1 summarizes the performance of SynapSeg compared to state-of-the-art architectures. Our method achieves an mIoU of 80.5% on CityScapes-High and 69.1% on Meridian-Aerial. This represents a significant improvement over the baseline models. Notably, SynapSeg

Table 1: Comparison of semantic segmentation performance on the fictional CityScapes-High and Meridian-Aerial datasets. We report mean Intersection over Union (mIoU, %) and inference speed (Frames Per Second, FPS).

Method	CityScapes-High (mIoU %)	Meridian-Aerial (mIoU %)	Speed (FPS)
FCN-8s	58.3	42.1	15.4
DeepLabv3+	72.4	58.7	8.2
HRNet-W48	76.5	63.2	6.5
SegFormer-B5	78.1	66.4	11.2
SynapSeg (Ours)	**80.5**	**69.1**	**14.8**

outperforms the dense SegFormer-B5 model by 2.4% on CityScapes-High and 2.7% on Meridian-Aerial, demonstrating the effectiveness of contrastive multi-scale affinity learning.

Qualitative Analysis. Visual results indicate that SynapSeg produces much sharper boundaries around thin structures, such as poles and railings, compared to baseline architectures. This is attributed to the sparse affinity propagation module, which preserves high-resolution spatial details. In contrast, standard architectures exhibit significant boundary blurring and structural discontinuities.

Computational Efficiency. In addition to superior accuracy, SynapSeg is highly efficient. As shown in Table 1, our model runs at 14.8 FPS, which is twice as fast as HRNet-W48 and significantly faster than SegFormer-B5. The training memory overhead is also reduced by 35%, allowing for larger batch sizes and more stable training configurations on standard GPUs.

6 Discussion

The success of SynapSeg lies in the decoupling of spatial and channel affinity routing. Traditional self-attention processes all spatial locations globally, which wastes computation on redundant background regions. By utilizing a sparse set of anchor points, our model selectively routes context to where it is needed most. The localized contrastive affinity loss further ensures that these anchors capture scale-invariant structures, which is critical for objects with varying sizes.

Limitations. Despite its advantages, SynapSeg has some limitations. The selection of anchor points is currently based on uniform spatial sampling. While this works well for general scenes, it might be sub-optimal for highly imbalanced datasets where rare classes cover small areas. Future work will explore dynamic anchor selection strategies based on feature importance or semantic density.

7 Conclusion

In this paper, we presented SynapSeg, a complete, standalone framework for high-resolution semantic segmentation. By integrating sparse multi-scale affinity propagation with a localized contrastive loss, SynapSeg addresses the dual challenges of spatial detail preservation and computational efficiency. Our experiments on three challenging benchmarks demonstrate that SynapSeg achieves state-of-the-art results while significantly reducing memory overhead and inference latency.

We believe our approach opens new avenues for real-time high-resolution perception on resource-constrained platforms. In future research, we plan to extend SynapSeg to video segmentation tasks, where temporal affinity propagation can be integrated into the framework to ensure temporal consistency across frames.

References

[1] David Chen and Marcus Vance. Localized contrastive loss for visual representation learning. In *Proceedings of the Vertex Computer Vision Conference*, pages 3045–3054, 2023.

[2] Carlos Garcia and Maria Rodriguez. Attention-based affinity propagation for semantic grouping. In *Proceedings of the Synapse AI Forum*, pages 89–98, 2020.

[3] Sarah Jenkins and David Chen. Multi-scale affinity learning for image segmentation. *Journal of Synapse Artificial Intelligence*, 14(3):102–115, 2024.

[4] Min-Ji Kim and David Chen. Local geometry alignment for spatial matching. *Journal of Vertex Computational Science*, 5(2):134–146, 2024.

[5] Sun-Woo Lee and Marcus Vance. Dual-path feature fusion for fine-grained segmentation. In *Apex Conference on Computer Vision*, pages 1122–1131, 2023.

[6] Robert Miller and Emily Davis. High-resolution aerial image segmentation via contextual routing. *Nimbus Journal of Remote Sensing*, 31(4):789–801, 2023.

[7] John Smith and Alice Johnson. Deep hierarchical networks for pixel-level classification. In *International Conference on Meridian Vision*, pages 120–129, 2021.

[8] James Taylor and Sarah Jenkins. Efficient decoders for semantic segmentation on edge devices. *Vertex Engineering Journal*, 19(1):12–25, 2021.

[9] Kevin Wong and Lisa Patel. Sparse self-attention mechanisms in vision transformers. In *Apex Machine Learning Symposium*, pages 567–576, 2022.