

Application of Quantum Self Supervised Learning for Semantic Segmentation

Wassim Hadj-Cherif¹ and Tirupati Boliseti²

¹Department of Computer Science, Nantes Université, Nantes, France

²Department of Civil and Environmental Engineering, University of Windsor, Canada

Introduction. Semantic segmentation underpins scene understanding, yet pixel-level annotation remains the main issue in domains such as medical imaging or remote sensing. Self-supervised learning (SSL) extracts rich representations without labels, and quantum computing offers high-dimensional Hilbert-space encodings; combining the two, however, is constrained by the limited qubit budgets of the Noisy Intermediate-Scale Quantum (NISQ) era.

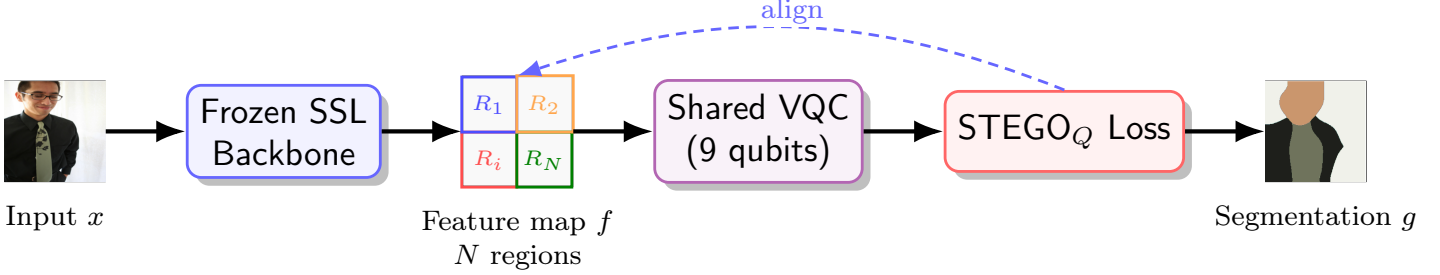


Figure 1: General architecture of the Q4S framework. The pipeline leverages a frozen encoder, pretrained via self-supervised learning (SSL), to extract the feature map f . This map is partitioned into N discrete spatial regions processed in parallel by a shared 9-qubit quantum segmentation head to generate the final quantum semantic projection g . A STEGO-adapted contrastive loss (STEGO_Q) aligns the region-wise quantum observables with the backbone’s feature correlations to ensure global semantic consistency.

Method. We introduce Q4S, a hybrid architecture integrating a 9-qubit variational quantum segmentation head with a frozen, SSL-pretrained classical backbone (Fig. 1). The central challenge lies in the quantum-classical information bottleneck, where a limited qubit budget cannot natively ingest a full, high-dimensional feature map. We circumvent this limitation algorithmically by partitioning the backbone’s ResNet output into 49 independent discrete spatial regions processed one at a time (Fig. 2). For each zone, the corresponding 512-dimensional classical feature vector is mapped onto the quantum state via *amplitude embedding*, satisfying the constraint $K = 2^N$ (512 features onto 9 qubits) to exploit the expressivity of the Hilbert space. This parallelization allows our 9-qubit circuit to process a total of 49×512 classical values across the image, decoupling the hardware qubit requirement from the spatial resolution. This process is driven by our modified STEGO contrastive loss, which enforces global semantic consistency across the independent regions. By extracting the variational circuit’s *Pauli-Z expectation values*, we obtain bounded quantum outputs within the $[-1, 1]$ interval. This inherently matches the bounded nature required for cosine similarity computations, enabling seamless integration with the classical loss framework and allowing a fixed NISQ device to handle feature maps of arbitrary size.

Metric	Value
Mean Intersection over Union (mIoU)	0.5827
F1 score	0.6773
Pixel Accuracy	0.8629
Mean Pixel Accuracy (MPA)	0.7971

Table 1: Q4S Semantic Segmentation Results averaged over a 100-image subset of the COCO-Stuff dataset (evaluated on an individual image-by-image pipeline).

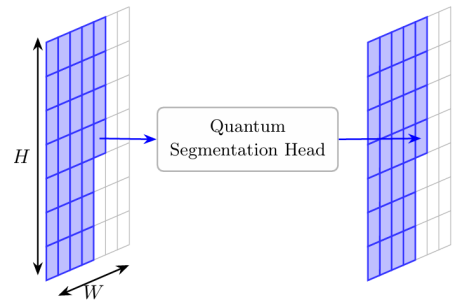


Figure 2: Parallelization of the processing across spatial zones. We extract the 512-dimensional feature vectors directly from the final layer of the pretrained ResNet-18 backbone and map each zone independently onto a 9-qubit register using amplitude embedding to generate the quantum semantic projection.

Discussion. While Table 1 demonstrates the efficiency of our 49-region spatial decomposition under tight qubit constraints, processing features exclusively at the final layer of the ResNet-18 backbone introduces a clear trade-off: it causes localized boundary smoothing because pixels within the same spatial zone share near-identical representation vectors. These performance limits can be directly addressed by leveraging earlier layers of the pretrained backbone. In architectures like ResNet-18, extracting features from intermediate stages increases both spatial resolution and region count while reducing the embedding dimensionality (K). Consequently, this architectural refinement increases inter-pixel diversity within each spatial region without expanding the qubit budget, as the feature size remains safely below the threshold of the deeper layers. This modification provides a direct path to optimize edge accuracy and sharp contours while keeping the entire framework strictly aligned with current NISQ hardware constraints.