Boundary-Responsive Differentiable Gating for Superpixel-Based Segmentation

Fatmaelzahraa Ali Ahmed[1] Zhihe Lu[2*] Gianni Di Caro[3] Diram Tabaa[3] Mohamed Hamdy[4]
Muraam Abdel-Ghani[1] Abdulaziz Al-Ali[4] Muhammad Arsalan[5] Shidin Balakrishnan[1*]

[1]Hamad Medical Corporation [2]Hamad Bin Khalifa University [3]Carnegie Mellon University Qatar
[4]Qatar University [5]KINDI Center, Qatar University

fatmaahmed.hmc@gmail.com zhihelu.academic@gmail.com gdicaro@cmu.edu dtabaa@andrew.cmu.edu
v-mabdelghani1@hamad.qa mm1905748@qu.edu.qa muhammad.arsalan@qu.edu.qa a.alali@qu.edu.qa
sbalakrishnan1@hamad.qa

Abstract

We present BRDG*, a boundary-responsive differentiable gating superpixel framework designed to resolve the trade-off between computational efficiency and segmentation precision in surgical scenes. At its core, the architecture is organized into three cooperative agents. The first agent is the region creator, which is fully differentiable and converts dense features into learnable superpixel tokens, jointly learning region descriptors and dense context. The Boundary Detector agent is the second agent; it acts as a gating mechanism, estimating boundary confidence from region features to predict where refinement is needed. The third agent is for refinement of the boundary superpixels; it uses a gate to selectively fuse efficient coarse predictions with a high-fidelity refinement path that restores pixel-level details. To further enhance distinctiveness, an adjacency-boosted contrastive loss mines hard negatives across neighboring regions to improve boundary separation. We evaluate* BRDG *on three surgical tasks requiring high-precision (EndoVis18-parts, EndoVis18-tools, EndoVis17-tools), as well as general domain benchmarks. Our model improves mIoU by substantial margins (↑4.5–7.0) over strong pixel-wise baselines; raising Boundary-F1 scores by over ↑10 points. Under the same hardware (RTX-A6000 Pro), it reaches 150.25 FPS with only 24M parameters. This makes it* $\times 10$ *faster and* $\times 3.5$ *smaller than current state-of-the-art models, effectively resolving the critical accuracy–efficiency trade-off in real-time segmentation.*

1. Introduction

Semantic segmentation is a fundamental task in medical image analysis [7]. In minimally invasive surgery (MIS), such as robot-assisted surgery (RAS) [58], accurate delineation of instruments and tissues is essential, as sharp tools interact with delicate anatomical structures [3]. Therefore, reliable segmentation masks underpin critical downstream tasks, including instrument tracking, surgical navigation, and context-aware decision-making [47].

Current approaches formulate segmentation as a dense, pixel-wise classification problem over high-resolution inputs [2, 4, 20, 36–39, 41, 49]. While effective in static imaging, such methods incur substantial computational costs in MIS settings [44] due to dense architectures with high parameter counts [17] and spatial redundancy [52]. Moreover, independent pixel predictions often yield fragmented regions, particularly along fine-grained surgical boundaries [55]. In time-critical and safety-sensitive surgical environments, these inefficiencies and inconsistencies limit practical deployment.

Superpixels offer a promising alternative by grouping perceptually similar pixels into compact, coherent regions, reducing redundancy while preserving local structure [29]. However, classical algorithms such as simple linear iterative clustering (SLIC) [1], Felzenszwalb's method [18], Quickshift [50], and Watershed [25] are non-differentiable and thus operate as fixed pre-processing steps. Their low-level, task-agnostic groupings cannot adapt to domain-specific cues such as instrument-tissue boundaries or specular reflections, which are crucial features in surgical imagery. Although learnable superpixels have been proposed to address this limitation, most remain optimized for general vision tasks and lack the robustness required for high-stakes medical applications.

Pixel-wise segmentation delivers high fidelity but at a prohibitive computational cost, while superpixel-based approaches are efficient yet often sacrifice semantic accuracy, especially near boundaries. To reconcile this trade-off, we propose BRDG (Boundary-Responsive Differentiable Gating), a framework that achieves superpixel-level efficiency

*Corresponding authors

without compromising pixel-level precision. BRDG introduces a differentiable, boundary-aware mid-level representation that enables efficient, coherent, and semantically faithful segmentation tailored to complex surgical scenes. Unlike existing learnable superpixel models such as SSN [26], PAN [48], and HERS [43], which collapse pixel detail too early by averaging within regions and thus lose critical boundary cues, BRDG incorporates explicit boundary-responsive gating to maintain sensitivity to fine boundary structures.

Specifically, BRDG preserves full pixel-level features inside stable regions while selectively refining only the ambiguous boundary zones. The network learns to distinguish coherent superpixel interiors from uncertain object boundaries, then uses a boundary-gating module to route boundary pixels to a dedicated refinement head. This targeted refinement preserves computational efficiency since only a small subset of pixels are processed at high resolution, while retaining the fine-level precision essential for accurate delineation. Although motivated by the demands of surgical scene understanding, the proposed architecture is applicable to general dense prediction tasks. The main contributions are summarized as follows:

- **Unified differentiable superpixel formulation:** A soft-assignment superpixel module integrated within a ResNet-U-Net backbone, enabling end-to-end learning of region features, boundary confidence, and refinement within a single differentiable framework.
- **Boundary-routed refinement:** A learned gating mechanism that uses predicted superpixel boundary probabilities to selectively refine pixels near semantic edges.
- **Adjacency-boosted boundary contrastive learning:** A boundary-level contrastive loss that leverages superpixel adjacency graphs to emphasize hard negatives across neighboring regions, enhancing boundary discrimination.

2. Related work

Superpixel segmentation, differentiable and non-differentiable alike, has contributed significantly to computer vision. The central notion of superpixels is to simplify image understanding by replacing a pixel-level representation with a region-level representation $I \in \mathbb{R}^{H\times W\times 3}$. Let the image domain be $\Omega = \{1,\dots,H\} \times \{1,\dots,W\}$. A superpixel partition is a collection of K disjoint, spatially contiguous sets $S = \{S_1,\dots,S_K\} \subseteq 2^{\Omega}$ such that

$$\bigcup_{k=1}^{K} S_k = \Omega, \qquad S_i \cap S_j = \varnothing \text{ for } i \neq j.$$

Each S_k groups perceptually similar pixels and can be treated as an atomic unit for downstream processing. Figure 1 demonstrates how classical superpixel methods groups similar pixels together based on the value of K. Although, all methods uses the same K value, Felzenszwalb [18] tends to be noisier compared to the other two methods. Watershed [25] has the smoothest boundaries and the least fragmented, yet it struggles to differentiate between objects in shadowed scenes.

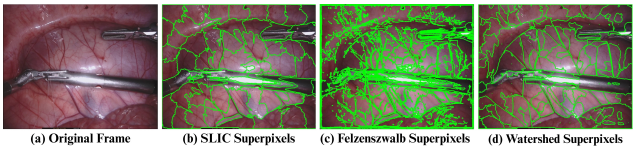


Figure 1. Examples of classical superpixel methods that are non-differentiable on surgical image

2.1. Non-Differentiable Superpixels

Early superpixel algorithms focused on perceptual grouping via low-level cues such as color and spatial proximity. Felzenszwalb and Huttenlocher [18] proposed an efficient graph-based segmentation method based on local variation in edge weights. Meanwhile, the Mean Shift algorithm [14] clusters pixels in joint spatial-color space. Among the most influential approaches, Simple Linear Iterative Clustering (SLIC) [1] formulates superpixel extraction as a variant of k-means clustering in the five-dimensional Lab+XY space, achieving a strong balance between boundary adherence and regularity. Subsequent work refined these ideas, for example SEEDS [16] that performs hill-climbing optimization on pixel neighborhoods, and ERS [34] which uses an entropy rate formulation to encourage spatial compactness and boundary precision. These algorithms remain highly effective in classical vision pipelines but are non-differentiable, making them unsuitable for integration into deep learning architectures that require gradient flow through the superpixel generation process. Figure 1 shows the different superpixel methods, all using 100 superpixels.

2.2. Differentiable Superpixels

To bridge this gap, differentiable superpixel methods have been proposed that embed superpixel assignment within neural networks, allowing end-to-end training. Jampani et al. introduced SSN [27], which approximates the discrete SLIC clustering with a continuous soft assignment mechanism. Their framework enabled gradient propagation through superpixel formation, thus integrating region grouping into the learning objective. Similarly, HERS [43] adapted the non-differentiable entropy rate formulation (used by ERS) into a hierarchical, end-to-end trainable framework. Later work such as Deep Simplex Superpixels [42] and Learnable Superpixels [56] further improved the differentiability and stability of the clustering process, often by incorporating spatial regularizers or by reinterpreting superpixel formation as a differentiable assignment problem

on a simplex manifold. Recently, differentiable superpixels have been used in self-supervised representation learning and segmentation [19, 32], demonstrating their potential as an interpretable mid-level representation. Existing differentiable superpixel methods have two limitations. First, they are often task-agnostic: the superpixel generation is learned, but not explicitly optimized for the downstream task, such as preserving semantic boundaries. Second, models like SSN [27] and HERS [43] typically collapse all information to the region level, performing classification on the pooled region features. This discards fine-grained pixel information, leading to the very boundary-smoothing errors they were meant to avoid. Our work addresses this gap by creating a superpixel framework that is (1) explicitly boundary-aware, and (2) preserves pixel-level information, using a gate to selectively reintroduce it.

2.3. Boundary-Aware Refinement

Object boundaries are critical in high-stakes domains such as surgical scene understanding, where inaccurate edges can lead to unsafe tool localization. Several studies have explored boundary-focused architectures to alleviate the over-smoothing effect of convolutional decoders. BoundaryNet [11] introduces an auxiliary edge supervision head to emphasize boundary gradients. BDCN [21] learns hierarchical edge features for multi-scale boundary detection. In the context of segmentation, Gated-SCNN [46] integrates a shape stream that communicates with the main semantic stream through a gating mechanism to improve edge fidelity in complex urban scenes. Similarly, methods such as RefineNet [33] and SegFix [57] apply boundary refinement or offset correction after coarse segmentation. These approaches confirm that explicit boundary modeling improves mask sharpness and class consistency. However, they often treat boundary prediction as a parallel branch rather than an integrated structural cue. Our proposed BRDG differs by embedding boundary reasoning directly into the superpixel assignment process and learning a differentiable gate to route refinement selectively along predicted boundaries.

2.4. Contrastive and Region-Level Representation Learning

Contrastive learning has become a powerful paradigm for dense prediction tasks. Pixel-level contrastive objectives [51] encourage representation separation between semantic classes, yet they typically neglect spatial context or region adjacency. Recent extensions such as ReCo [35] and PiCIE [13] address this by applying region-aware or clustering-based contrastive sampling. In medical imaging, contrastive learning has been adapted to unsupervised segmentation and domain transfer [9, 60]. However, these methods still rely on pixel-level sampling that is computationally heavy for high-resolution surgical scenes. In contrast, our adjacency-boosted boundary contrastive loss exploits the inherent superpixel graph to efficiently mine hard negatives from neighboring regions. This couples semantic contrast with geometric adjacency in a single differentiable objective.

Our work builds upon this line of research but introduces two key novelties: (1) a boundary-responsive gating mechanism that routes pixel-wise refinement based on predicted boundary confidence, and (2) a boundary-aware contrastive loss that leverages adjacent region relationships to improve boundary discrimination.

3. Methods

Our framework, BRDG, is a fully differentiable architecture that unifies the efficiency of region-level reasoning with the precision of pixel-level classification. Algorithm 1 summarizes the forward pass: an input image I is encoded, superpixel assignments are inferred, boundary confidence is estimated, and coarse and refined predictions are blended to produce the final logits $\hat{Y}$. Conceptually, the model is organized into three cooperative agents (Figure 2 a):

1. **Agent 1 (Region & Feature Creator).** A ResNet-UNet backbone [22, 45] integrated with a differentiable superpixel module. Given I, it produces the mid-level representations required by downstream components: a dense feature map F, a soft assignment map A, coarse logits $\hat{Y}_c$, and a set of K compact region descriptors r_k.
2. **Agent 2 (Boundary Detector).** A lightweight gating head that consumes r_k to predict a boundary-confidence score p_k for each region and reprojects these scores via A to form a dense, pixel-wise refinement gate g.
3. **Agent 3 (Refinement Agent).** A dual-path classifier that combines F, r_k, A, and the gate g to compute refined logits $\hat{Y}_r$ and blends them with $\hat{Y}_c$ to obtain the final output $\hat{Y}$.

This design forces the model to learn *what* to refine (Agent 2) and *how* to refine (Agent 3) based on the rich, multi-level features provided by (Agent 1).

3.1. Agent 1: Region and Feature Creator

This agent forms the foundation of the entire network. Its first component is the feature extraction backbone, a ResNet-34 encoder [22] paired with a U-Net-style decoder [45]. The residual design of ResNet stabilizes optimization via identity skips, enabling reliable gradient flow across layers. The encoder is initialized with ImageNet weights, and the early batch-normalization layers are kept frozen during the initial epochs to preserve pretrained statistics. We adopt a *discriminative learning-rate* scheme: the encoder is fine-tuned with a learning rate of 10^{-4} the base rate, and the randomly initialized decoder (and heads) uses the base rate, encouraging transfer of generic ImageNet features and faster learning of task-specific representations. The encoder out-

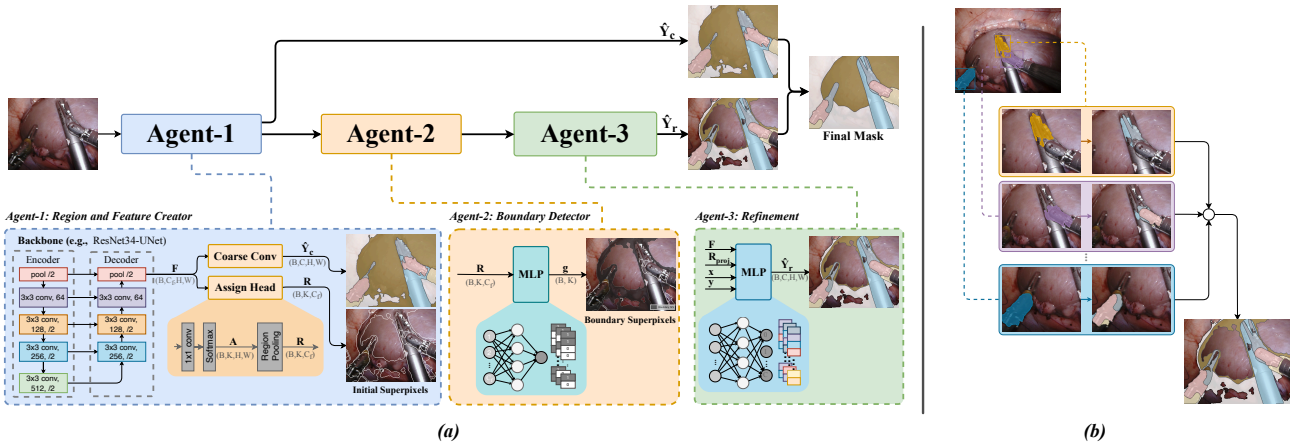


Figure 2. Boundary-Responsive Differentiable Gating Superpixel (BRDG) architecture overview.

Algorithm 1 BRDG Forward Pass

Input: Image I; temperature τ; number of superpixels K
Output: Logits $\hat{Y}$

$F \leftarrow \text{ResNet–UNet}(I)$
$A_{\text{logits}} \leftarrow \text{assign\_head}(F)$; $A \leftarrow \text{softmax}(A_{\text{logits}}/\tau)$
for $k = 1$ **to** K **do**
  $r_k \leftarrow \dfrac{\sum_{i,j} A_{k,i,j}\, F_{:,i,j}}{\sum_{i,j} A_{k,i,j}}$
for $k = 1$ **to** K **do**
  $p_k \leftarrow \text{boundary\_mlp}(r_k)$
$g_{i,j} \leftarrow \sum_{k=1}^{K} A_{k,i,j}\, p_k$
$\hat{Y}_c \leftarrow \text{coarse\_head}(F)$
(Reprojection) $R_{\text{proj},i,j} \leftarrow \sum_{k=1}^{K} A_{k,i,j}\, r_k$
(Coords) form grids $X, Y \in \mathbb{R}^{H\times W}$ with normalized $x_{i,j}, y_{i,j}$
(Definition of Z) $Z_{i,j} \leftarrow [F_{i,j};\, R_{\text{proj},i,j};\, x_{i,j};\, y_{i,j}]$
$\hat{Y}_r \leftarrow \text{refine\_head}(Z)$
$\hat{Y} \leftarrow (1-g)\,\hat{Y}_c + g\,\hat{Y}_r$
return $\hat{Y}$

puts feature maps at strides $\{1/2, 1/4, 1/8, 1/16, 1/32\}$ relative to the input resolution. These multiscale features are fused in the decoder via bilinear upsampling and lateral skip connections to reconstruct a dense feature field $F \in \mathbb{R}^{B\times C_f\times H\times W}$, where B is the batch size and $C_f = 96$ is the number of feature channels. On top of the shared feature map F, two lightweight 1×1 convolutional heads are attached: (i) the assignment head (*assign_head*) producing superpixel assignment logits A_{logits}, and (ii) the coarse head (*coarse_head*) producing coarse segmentation logits $\hat{Y}_c$ as a fast baseline prediction. The assignment logits are converted into a soft assignment map A via a temperature-controlled softmax,

$$A_{b,k,i,j} = \frac{\exp\big(A_{\text{logits},b,k,i,j}/\tau\big)}{\sum_{k'} \exp\big(A_{\text{logits},b,k',i,j}/\tau\big)}. \quad (1)$$

The soft assignments are then used to "soft-pool" F into a set of K region descriptors r_k, each representing the average feature of its superpixel as mentioned in algorithm 1. Thus, this agent outputs the dense map F, the coarse logits $\hat{Y}_c$, the assignment map A, and the K region descriptors r_k, which are used by the subsequent components.

3.2. Agent 2: Boundary Detector

A ground-truth pixel-boundary map is derived directly from semantic labels: a pixel is designated as a boundary if any of its neighbors belongs to a different class. Superpixels containing at least one such pixel are assigned a positive label (1), and all others (0); these labels supervise p_k via binary cross-entropy (BCE) loss, ensuring that refinement is triggered only at structural interfaces between classes.

The second agent's task is to determine *which* of the K regions are ambiguous and lie on a semantic boundary. It takes the K region descriptors r_k from Agent 1 and passes them through a small multi-layer perceptron (MLP) *(boundary_mlp)*. This head predicts a single boundary probability $p_k \in [0, 1]$ for each superpixel. These K region-level probabilities are then re-projected back to the dense $H \times W$ pixel space using the assignment map A (also from Agent 1). This creates the final boundary gate g: where $g_{i,j}$ is the final gate value at pixel (i, j), $A_{k,i,j}$ is the soft assignment of that pixel to superpixel k, and p_k is the predicted boundary probability for that k-th superpixel. This gate g has high values (≈ 1) at pixels belonging to boundary superpixels and low values (≈ 0) for stable interior superpixels.

3.3. Agent 3: Refinement Agent

The final agent performs the segmentation. It operates via two distinct paths that are blended by Agent 2's gate, g.

First, a Coarse Path *coarse_head* computes coarse logits $\hat{Y}_c$ directly from the shared features F (from Agent 1) using an efficient 1×1 convolution. This provides a fast, baseline prediction. Second, a refined path produces high-fidelity refined logits $\hat{Y}_r$. Its refinement capability comes not from

the 1×1 convolutions alone, but from its rich, concatenated input tensor Z:

$$Z_{i,j} = [F_{i,j};\ R_{\text{proj},i,j};\ x_{i,j};\ y_{i,j}]. \tag{2}$$

Here, $\mathbf{Z}_{i,j}$ combines the original pixel feature ($F_{i,j}$), the shared context from its entire superpixel region ($R_{\text{proj},i,j}$), and two scalar values, $x_{i,j}$ and $y_{i,j}$, which are the normalized spatial coordinates that provide absolute positional information. This fused tensor is processed by a lightweight MLP (*refine_head*) to produce $\hat{Y}_r$.

Finally, the Final Blend produces the network's output $\hat{Y}$ by using the gate g to selectively merge the two paths:

$$\hat{Y} = (1 - g)\,\hat{Y}_c + g\,\hat{Y}_r \tag{3}$$

This blend forces the model to use the efficient coarse logits $\hat{Y}_c$ for stable region interiors (where $g \approx 0$) and the precise, context-aware refined logits $\hat{Y}_r$ only for the complex object boundaries (where $g \approx 1$). This is illustrated in Figure 2 b.

3.4. Boundary-Aware Contrastive Learning

Finally, to ensure the region descriptors r_k (from Agent 1) are highly discriminative, we introduce a specialized contrastive loss. This loss operates on the K region features and is supervised by their boundary labels (derived from Agent 2's predictions). It pulls same-class regions (e.g., two "interior" regions) together and pushes different-class regions (e.g., an "interior" vs. a "boundary" region) apart.

Its key novelty is an adjacency boost, w_{ik}:

$$\mathcal{L}_{\text{bnd}} = -\frac{1}{|P|} \sum_{(i,j)\in P} \log \frac{\exp(s_{ij}/T)}{\sum_{k\neq i} \exp(s_{ik}\, w_{ik}/T)}, \tag{4}$$

where $s_{ij} = r_i^\top r_j$ is the feature similarity, T is the temperature hyperparameter, and $w_{ik} = 1 + \alpha\mathbf{1}[i,k \text{ adjacent}]$. This boost term increases the penalty for adjacent negative pairs, forcing Agent 1 to learn a sharp feature separation between neighboring superpixels that lie on opposite sides of a semantic boundary.

Conceptually, the architecture learns to function as a sparse refiner. The gated loss $(1 - g)\hat{Y}_c + g\hat{Y}_r$ ensures that the refined path is *trained* almost exclusively on boundary pixels (where $g \approx 1$), and the coarse path is trained on stable interiors (where $g \approx 0$). This allows the model to achieve high precision without the computational cost of a second, heavy refinement network.

4. Experiments

In this section, we describe the experiment setup to design and evaluate BRDG. We evaluate the robustness of our model across multiple datasets. We further present several ablation studies performed to justify the framework component-level design choices.

4.1. Implementation Details

Our model is trained using the AdamW optimizer [30] with a base learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. We initialize the encoder with ImageNet-pretrained weights and finetune using a discriminative learning [31] rate, set lower for the encoder than for the randomly initialized decoder. All training images are resized to a resolution of 512×640. Training proceeds for 100 epochs following a multi-stage schedule. **Warmup (Epochs 1–5):** The ResNet encoder is kept frozen, with only the primary segmentation loss (a $0.5/0.5$ weighted combination of Cross-Entropy and Tversky loss) active, allowing the decoder and heads to learn stable representations from static pretrained features. **Unfreeze & Ramp-up (Epochs 6–10):** The encoder is unfrozen at its $0.1\times$ learning rate, and auxiliary loss weights are linearly ramped up from 0.0 to their final values. **Full Training (Epochs 11–60):** All components are active with loss weights at their final scheduled values.

For comprehensive evaluation of the model's performance, we employed a multi-faceted set of metrics targeting accuracy, efficiency, and boundary fidelity. Segmentation accuracy was quantified using the standard mean Intersection over Union (mIoU), which measures the average overlap between the predicted mask and the ground truth, and the Dice Score (or F1 Score), which balances precision and recall to assess mask similarity. To evaluate boundary fidelity—a critical factor in surgical segmentation—we report Boundary Recall (BF), which specifically measures how well the predicted contour aligns with the true object outline. Finally, efficiency was assessed through several computational metrics: Frames Per Second (FPS), which represents the real-time processing speed; GFLOPs (Giga Floating-Point Operations), which measures the model's computational complexity; number of trainable parameters, which indicates model size; and peak memory usage during inference, which reflects deployment feasibility on resource-constrained hardware. For real-time deployment, the model achieves 6.63 ms inference time (150.25 FP) on NVIDIA RTX-6000 Pro GPU, utilizing 99.6% of the per-frame budget with only 0.026 ms headroom for I/O.

4.2. Datasets and Baselines

To validate the performance of our proposed model, we conduct a comprehensive evaluation on a diverse set of public benchmarks, which are divided into two distinct categories. The first category includes domain-specific surgical datasets, namely the EndoVis 2017 [5] and EndoVis 2018 tools [6] , the EndoVis 2018 (Parts) [6] and cholecSeg8k [24] datasets for instrument components and anatomy segmentation. The second category, used to assess the generalizability of our feature representation and boundary detection, includes general-purpose benchmarks: Cityscapes [15] and [61], for urban scene understanding, and BSDS500

Table 1. Cross-Method Benchmark Results Across Four Surgical Segmentation Tasks. Accuracy is evaluated by mIoU, Dice Score, and Boundary F1 Score (BF1) at 2px tolerance. Model Cost is a dataset-agnostic measure (at 512×640), reporting FPS, GFLOPs (complexity), and Params (M) (size in millions). Models marked ($\star$) indicate results reproduced by our implementation. CholecSeg8K results sourced from [41]. The "…" indicates that it was not reported.

Method	EndoVis'18 Parts			EndoVis'18 Tools			EndoVis'17			CholecSeg8K		Model Cost		
	mIoU	Dice	BF1	mIoU	Dice	BF1	mIoU	Dice	BF1	mIoU	Dice	GFLOPs	Params (M)	FPS
CNN-Based Models														
DeepLabv3+ (R101) [10] $\star$	0.56	0.58	0.03	**0.78**	**0.80**	0.67	0.67	0.68	0.26	0.56	0.63	**629.5**	61.0	15.1
Mask-RCNN [23] $\star$	0.37	0.48	0.19	0.41	0.44	0.23	0.38	0.41	0.20	0.54	0.55	…	**95.0**	…
U-Net (R34) [45] $\star$	0.53	0.58	0.17	0.64	0.66	0.21	0.42	0.44	0.20	0.43	0.49	155.83	13.39	45.96
Transformer-Based Models														
TransUNet [41]	0.48	0.52	…	…	…	…	…	…	…	0.55	0.62	…	31.0	…
SegFormer-B5 [41] $\star$	0.57	0.58	…	0.71	0.72	…	0.63	0.70	**0.70**	0.74	0.79	124.85	84.7	13.84
MedT [41]	0.65	0.68	…	…	…	…	…	…	…	0.50	0.57	…	84.7	…
SwinUNet [8] $\star$	0.65	0.67	0.29	0.59	0.62	0.23	0.64	0.66	0.24	0.68	0.71	…	41.0	150.60
nnFormer $\star$	0.60	0.61	0.24	0.63	0.64	0.35	0.62	0.64	0.22	0.62	0.66	…	53.0	123.54
LETNET [54] $\star$	0.57	0.59	0.26	0.53	0.54	0.32	0.59	0.61	0.51	0.58	0.60	…	0.95	46.40
Mask2Former [12]	0.49	0.53	0.12	0.44	0.46	0.55	0.63	0.65	0.53	0.69	0.71	…	47	…
Superpixel-Based Models														
SSN [26]$\star$	0.37	0.42	0.25	0.41	0.45	0.30	0.33	0.37	0.19	0.42	0.45	116.80	0.655	271.62
HERS [43] $\star$	0.45	0.50	**0.41**	0.70	0.73	0.60	0.51	0.56	0.43	0.58	0.63	130.97	7.70	**564.76**
BRDG $\star$	**0.72**	**0.76**	0.31	0.75	0.77	**0.71**	**0.76**	**0.75**	0.69	**0.73**	**0.80**	157.97	23.9	150.25

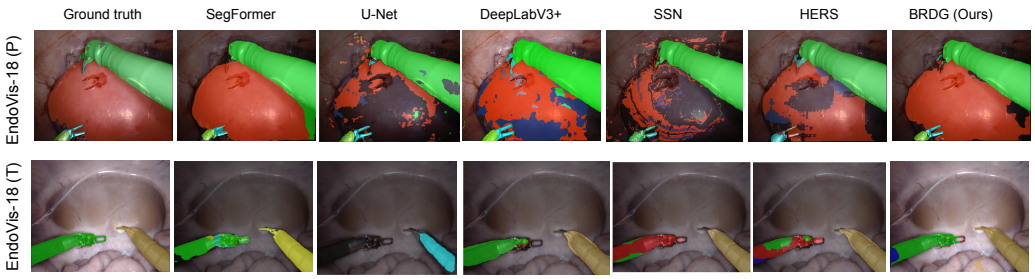


Figure 3. Qualitative comparison between state-of-the-art (SOTA) models and our method on EndoVis-2018. EndoVis-18 P denotes the parts-segmentation task (articulated tool components and anatomical structures, independent of tool identity); EndoVis-18 T denotes full surgical-tool segmentation.

[40] for contour detection. We benchmark against two primary families of segmentation algorithms. For superpixel-based comparisons, we include the differentiable SSN (Superpixel Sampling Networks) [26], the hierarchical HERS [43], and a classical SLIC [1] + GraphCut pipeline. For state-of-the-art (SOTA) pixel-wise comparisons, we evaluate against the foundational U-Net [45], the multi-scale DeepLabv3+ [10], and the Transformer-based SegFormer [53].

4.3. Ablation Study

We conducted a series of ablations to assess the contribution of each component to the framework, shown in Table 2. Unless noted, all experiments are run on *EndoVis2018-Part*, and the superpixel-count study on *BSDS500*, which is standard for boundary evaluation. In each experiment, we report mIoU, inference time (ms), and peak memory (GB).

Model's Components

The first experiment evaluates the model purely as a pixel-wise segmentation network without the superpixels. The mIoU decreased to 0.57, which the same score as SegFormer model (table 1). Peak memory increases by $>$ 400 MB relative to the full model. Furthermore, removing the boundary-head binary cross-entropy (BCE) and the boundary-contrastive loss under the "No-boundary" ablation drops performance from 0.72 to 0.57 mIoU (a 0.15 decrease), indicating that boundary-aware supervision is critical. To isolate the effect of the refinement agent (the final stage), we disable it and rely solely on the coarse head in the "No-refine" ablation. mIoU decreases by 11 points (from 0.72 to 0.61). Notably, removing refinement does *not* improve speed or memory: runtime remains 6.63 ms and peak memory *increases* from 1.05 to 1.17 GB. Renabling the refinement head and learned gate restores the "Full model" to

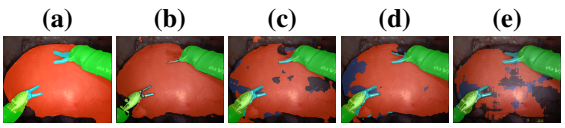


Figure 4. Qualitative ablation results showing the effect of removing key components: (a) Ground Truth, (b) Full model, (c) No Boundary Head, (d) No Refinement (e) Coarse only.

Table 2. Ablation study on the proposed method. Number of superpixels = 100, feature channels = 92. Image size: 512×640 on Endovis2018. Inference measured in ms; Peak Memory in GB.

Ablation	mIoU	Inference	FPS	Peak Mem
No-superpixels	0.57	24	128.88	1.53
No-boundary	0.57	8.12	123.20	1.52
No-refine	0.61	6.63	150.90	1.17
Gate = 0	0.61	6.63	150.90	1.17
Only-coarse	0.52	9.8	65	1.23
Full	**0.72**	6.63	150.25	1.05

0.72 mIoU at the same 6.63 ms, delivering a substantial accuracy gain at no inference-time cost.

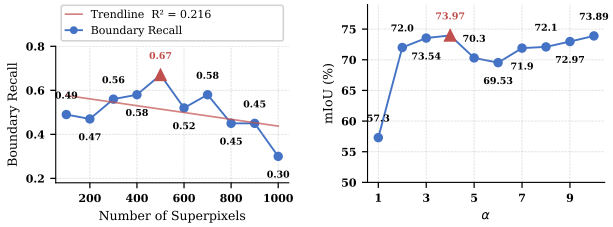
Finally, forcing the gate to zero ("Gate=0"; using only coarse logits) reproduces the baseline "No-refine"" (0.61 mIoU). The learned gate in the full model (0.72 mIoU) outperforms either coarse-only or refine-only variants, suggesting that selectively refining boundaries while preserving coarse interiors is the optimal strategy. Our ablations indicate that each agent and loss term contributes materially to final quality as illustrated in Figure 4.

Number of Superpixels "K"

We analyzed the effect of the number of superpixels (K) on the model's performance, using the BSDS500 dataset and measuring Boundary Recall (BF). As shown in Figure 5 a, we found that the optimal number of superpixels is $K = 500$, which achieved a BF score of 0.67. Interestingly, increasing the superpixel count beyond this point resulted in diminishing returns. Performance did not just plateau but actively degraded, with the worst score observed at $K = 1000$. Figure 7 in the supplementary material provides a qualitative explanation for this, illustrating that as K increases, the superpixel assignments become overly fragmented, which harms the model's ability to adhere to true semantic boundaries.

Alpha Value

We further analyze the sensitivity of the boundary-contrastive loss scaling parameter α, which controls the degree of adjacency boosting applied to boundary superpixels. As $\alpha = 1$ serves as the natural baseline (since $\log(1) = 0$, adjacency boosting is absent, and the loss reduces to standard contrastive loss), our ablation focuses on $\alpha > 1$ to quantify the impact of progressively stronger boundary-aware spatial constraints. As shown in Figure 5 b, performance remains stable for $\alpha > 1$, beyond which the over-emphasis on boundary regions begins to degrade overall segmentation quality.



(a) (b)

Figure 5. Ablation studies. (a) Boundary Recall vs. number of superpixels on the BSDS500 dataset. (b) Effect of α on mIoU.

Different Backbones

Additionally, we evaluated the model's robustness through experimenting with different backbones for generating the differentiable superpixels. ResNet-34 is selected as our default to minimize model size without sacrificing performance; however, as shown in Table 3, BRDG remains competitive across a range of backbones, with performance scaling gracefully as capacity increases. These results confirm that BRDG is not contingent on a specific backbone, and that ResNet-34 offers the best efficiency-accuracy trade-off for the surgical domain.

Table 3. Backbones Evaluated for BRDG on EndoVis2018

Backbone	mIoU	Dice	Para (M)
ResNet-34	72.0	76.0	24.0
ResNet-50	72.7	77.5	37.0
ResNet-101	72.9	77.8	56.0
ViT	**74.78**	**79.22**	99.0

Failure Modes Finally, we examine failure modes and robustness to image corruptions. The primary failure case occurs when the learned gate misfires, incorrectly routing interior regions through the refinement head or suppressing genuine boundaries. Quantitatively, synthetic corruptions reveal that fog produces the largest degradation (mIoU drops from 82.3 to 68.9), while motion blur and Gaussian blur are better tolerated (78.9 and 80.1 mIoU, respectively), suggesting that low-frequency contrast degradation is more harmful to the boundary-gating mechanism than high-frequency noise. Qualitative examples of gate-misfire cases are shown in figure 8 in the supplementary material.

4.4. BRDG vs. SOTA on Surgical Datasets

We evaluated our proposed model against pixel-based and superpixel-based methods, with results detailed in Table 1. The analysis shows that BRDG not only establishes a new state of the art in segmentation accuracy but also resolves

the critical accuracy efficiency trade-off for this task. On the EndoVis'18 *Parts* dataset, BRDG achieves 0.72 mIoU, a substantial $+6$ point lead over the strongest superpixel-based competitor (HERS) [43] and $+7$ points over the best-performing pixel-based model (MedT) [49]. On EndoVis'18 *Tools* dataset, our model's 75.46 mIoU surpasses the SOTA pixel-based SegFormer-B5 [53] by strong $+4.46$ points and the superpixel-based HERS by $+5.46$ points. Per-class results are shown in the supplementary material with more qualitative results.

The success of our gated refinement is most evident in boundary-specific scores (BF1): on *Tools*, BRDG attains 71.00 BF1, a dominant $+10.88$ point margin over HERS, demonstrating superior delineation of semantically meaningful boundaries. Crucially, BRDG attains this accuracy while being markedly more efficient.

BRDG operates at 150.25 FPS, making it $10\times$ faster than heavyweight SOTA models it outperforms, such as DeepLabv3+ (15.1FPS) and SegFormer-B5 (13.8FPS), and exceeds an optimized U-Net (R34) by $+104$ FPS a $\times(3.3)$ speedup. Although SSN is technically faster, its performance mIoU < 0.42 renders it uncompetitive. Parameter-wise, BRDG (23.9M) is $\times$ 2.5 smaller than DeepLabV3+ (61.0M) and $\times$ 3.5 smaller than SegFormer-B5 (84.7M). This demonstrates that BRDG is not an incremental improvement, yet it redefines the performance envelope by delivering +[5-7] points higher accuracy while being $\times$ 10 faster and $\times$ 3 smaller than current SOTA models.

Figure 3 presents a qualitative comparison of BRDG against superpixel and pixel-wise models on three surgical tasks: EndoVis-2018 Part, EndoVis-2018 Tools [6] andEndoVis 2017 [5]. The visual results show that BRDG consistently produces cleaner and more spatially coherent segmentation masks, unlike the fragmented and noisy outputs from SSN and U-Net.

Pixel-based baselines show distinct failure modes: DeepLabV3+ struggles with both boundary delineation and classification. SegFormer often fails to distinguish between tool boundaries and adjacent tissue. In contrast, our model's masks align closely with the ground-truth boundaries. On datasets with larger objects, such as EndoVis 2017, our model's boundary adherence is nearly identical to the ground truth, though it occasionally exhibits classification errors in these large regions, a challenge likely attributable to data imbalance.

We further compare against LETNet [54], a real-time segmentation architecture, under identical experimental settings. As reported in Table 1, LETNet achieves 0.57 mIoU on EndoVis'18 Parts and 0.59 on EndoVis'17, falling below BRDG by $+0.15$ and $+0.17$ points respectively, confirming that BRDG's efficiency–accuracy trade-off remains competitive even against models explicitly optimized for real-time inference.

4.5. Results on General Domain Data

To evaluate the generalizability of our architecture, we extended our testing beyond surgical data to the Cityscapes and ADE20K benchmarks. The quantitative results are summarized in Table 4. Our model attains 0.54 mIoU on Cityscapes and 0.60 mIoU on ADE20K, evaluated under the foveated/efficient segmentation protocol defined in [59]; these results reflect cross-domain generalization under efficiency constraints and should not be interpreted as a comparison against general-purpose segmentation leaderboards. These results are particularly significant when compared to FSNet, a foveated instance segmentation framework designed for AR/VR, which leverages real-time user gaze data to perform segmentation exclusively on instances of interest [59]. Figure 6 shows qualitative results of BRDG segmentation on the Cityscapes dataset. The model was able to accurately segment large objects but struggled with smaller objects or those in occluded situations.

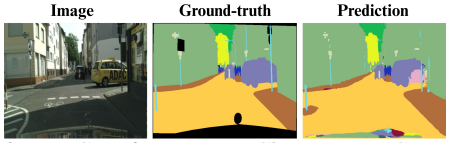


Figure 6. BRDG performance on Cityscapes vs Ground-truth

Table 4. Mean IoU Performance of BRDG on Cityscapes and ADE20K Datasets

Method	CityScapes	ADE20K
Avg+DeepLab [59]	0.26	0.39
Avg+HRNet [59]	0.20	0.43
Avg+SegFormer-B4 [59]	0.25	0.37
Avg+SegFormer-B5 [59]	0.27	0.41
LTD [28]	0.37	0.41
FSNet+DeepLab [59]	0.52	0.55
FSNet+HRNet [59]	0.47	0.56
FSNet+SegFormer-B4 [59]	0.46	0.54
FSNet+SegFormer-B5[59]	0.51	0.55
BRDG	**0.54**	**0.60**

5. Conclusion & Future Work

In this paper, we presented BRDG, a differentiable superpixel-based segmentation model tailored for surgical scenes that can be extended to other domains through fine-tuning. Our model demonstrated superior performance, achieving state-of-the-art accuracy while being significantly more efficient than competing methods. It operates up to $\times$ 10 faster while being $\times$ 3-4 smaller than existing dense segmentation models. Future work includes strengthening the model's classification capabilities to offer a reliable delineation tool that can be used to accelerate the segmentation and production of new large-scale datasets.

6. Acknowledgement

Research reported in this publication was supported by the Qatar Research Development and Innovation Council (QRDI) grant number ARG01-0522-230266. Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of Qatar Research Development and Innovation Council.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC Superpixels compared to state-of-the-art superpixel methods. *EEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 1, 2, 6

[2] F. Ahmed, M. Abdel-Ghani, M. Arsalan, M. Ali, A. Al-Ali, and S. Balakrishnan. Surg-SegFormer: A dual transformer-based model for holistic surgical scene segmentation. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, 2025. 1

[3] F. A. Ahmed, M. Yousef, M. A. Ahmed, H. O. Ali, A. Mahboob, H. Ali, Z. Shah, O. Aboumarzouk, A. Al Ansari, and S. Balakrishnan. Deep learning for surgical instrument recognition and segmentation in robotic-assisted surgeries: a systematic review. *Artificial Intelligence Review*, 2024. 1

[4] F. A. Ahmed, M. Arsalan, A. Al-Ali, K. Al-Jalham, and S. Balakrishnan. CLIP-RL: Surgical scene segmentation using contrastive language-vision pretraining & reinforcement learning. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, 2025. 1

[5] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis Herrera, Wenqi Li, Vladimir Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. 2017 robotic instrument segmentation challenge, 2019. 5, 8

[6] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, and et Stefan Leger. 2018 robotic scene segmentation challenge, 2020. 5, 8

[7] M. Krithika Alias AnbuDevi and K. Suganthi. Review of semantic segmentation of medical images using modified architectures of UNET. *Diagnostics (Basel, Switzerland)*, 2022. 1

[8] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Spectrum of Engineering*, 2021. 6

[9] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *NeurIPS*, pages 12546–12558, 2020. 3

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 6

[11] Yi Chen, Dengxin Dai, and Luc Van Gool. BoundaryNet: An adversarial domain adaptation architecture for high-resolution boundary detection. In *CVPR Workshops*, 2020. 3

[12] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation, 2022. 6

[13] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, pages 16794–16804, 2021. 3

[14] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. 2

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding, 2016. 5

[16] Michael Van den Bergh, Xavier Boix, Gemma Roig, Barbara Caputo, and Luc Van Gool. SEEDS: Superpixels extracted via energy-driven sampling. In *ECCV*, pages 13–26, 2012. 2

[17] Hongwei Du, Jiamin Wang, Jian Hui, Lanting Zhang, and Hong Wang. DenseGNN: universal and scalable deeper graph neural networks for high-performance property prediction in crystals and molecules. *npj Computational Materials*, 2024. 1

[18] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 2004. 1, 2

[19] Eduardo Gastal, Chaitanya Reddy, and Varun Jampani. Learning superpixels in space and time. In *CVPR*, pages 14741–14750, 2022. 3

[20] Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaez. ISINet: An instance-based approach for surgical instrument segmentation. In *MICCAI*, 2020. 1

[21] Jianzhong He, Shiliang Zhang, Ming-Ming Yang, Ying Shan, and Thomas S. Huang. BDCN: Bi-directional cascade network for edge detection. In *CVPR*, pages 3828–3837, 2019. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 6

[24] W. Y. Hong, C. L. Kao, Y. H. Kuo, J. R. Wang, W. L. Chang, and C. S. Shih. Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80, 2020. 5

[25] Z. Hu, Q. Zou, and Q. Li. Watershed superpixel. In *ICIP*, 2015. 1, 2

[26] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *CVPR*, 2018. 2, 6

[27] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *ECCV*, pages 352–368, 2018. 2, 3

[28] Mahmoud Khairy, Zhesheng Shen, Tor M. Aamodt, and Timothy G. Rogers. Accel-sim: an extensible simulation framework for validated GPU modeling. In *ISCA*, 2020. 8

[29] Zakia Khatun, Halldór Jónsson, Mariella Tsirilaki, Nicola Maffulli, Francesco Oliva, Pauline Daval, Francesco Tortorella, and Paolo Gargiulo. Beyond pixel: Superpixel-based MRI segmentation through traditional machine learning and graph convolutional network. *Computer Methods and Programs in Biomedicine,*, 2024. 1

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015. 5

[31] J. Li, D. Lin, Y. Wang, G. Xu, Y. Zhang, C. Ding, and Y. Zhou. Deep discriminative representation learning with attention map for scene classification. *Remote Sensing*, 2020. 5

[32] Weijie Li, Yalong Huang, Yuxin Peng, and Qiang Wang. SSFD: Self-supervised feature distillation for unsupervised image segmentation. In *AAAI*, 2022. 3

[33] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, pages 1925–1934, 2017. 3

[34] Ming-Yu Liu, Oncel Tuzel, Srikumar Ramalingam, and Rama Chellappa. Entropy rate superpixel segmentation. In *CVPR*, pages 2097–2104, 2011. 2

[35] Yunke Liu, Yu Tian, Yunchao He, Yuhui Yuan, Yifan Wang, Jiazhi Dong, Qi Yu, and Jingdong Wang. ReCo: Region contrast for weakly supervised semantic segmentation. In *CVPR*, pages 12215–12224, 2021. 3

[36] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *CVPR*, pages 9111–9120, 2020. 1

[37] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *ICCV*, pages 8741–8750, 2021.

[38] Zhihe Lu, Sen He, Da Li, Yi-Zhe Song, and Tao Xiang. Prediction calibration for generalized few-shot semantic segmentation. *IEEE Transactions on Image Processing*, 32: 3311–3323, 2023.

[39] Zhihe Lu, Da Li, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Uncertainty-aware source-free domain adaptive semantic segmentation. *IEEE Transactions on Image Processing*, 32:4664–4676, 2023. 1

[40] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int. Conf. Computer Vision*, pages 416–423, 2001. 6

[41] Abdel-Ghani Muraam, Ali Mahmoud, Ali Mohamed, Ahmed Fatmaelzahraa, Arsalan Mohamed, Al-Ali Abdulaziz, and Balakrishnan Shidin. FASL-Seg: Anatomy and tool segmentation of surgical scenes. In *ECAI 2025*, 2025. 1, 6

[42] Gerhard Neuhold, Timo Nöth, Matthias Rottmann, and Hanno Gottschalk. Deep simplex superpixels. *IEEE Transactions on Image Processing*, 30:8321–8332, 2021. 2

[43] Hankui Peng, Angelica I. Aviles-Rivero, and Carola-Bibiane Schonlieb. HERS Superpixels: Deep affinity learning for hierarchical entropy rate segmentation. In *CVPR*, 2022. 2, 3, 6, 8

[44] A. Prabhu et al. Computationally budgeted continual learning: What does matter? In *CVPR*, 2023. 1

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3, 6

[46] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-SCNN: Gated shape CNNs for semantic segmentation. In *ICCV*, pages 5229–5238, 2019. 3

[47] Zhang Tiyao, Yuan Xue, and Xu Hongze. Surgical instrument segmentation via segment-then-classify framework with instance-level spatiotemporal consistency modeling. *Journal of Imaging.*, 2025. 1

[48] Wei-Chih Tu, Ming-Yu Liu, Varun Jampani, Deqing Sun, Shao-Yi Chien, Ming-Hsuan Yang, and Jan Kautz. Learning superpixels with segmentation-aware affinity loss. In *CVPR*, 2018. 2

[49] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M. Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *MICCAI*, 2021. 1, 8

[50] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *ECCV*, 2008. 1

[51] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. DenseCL: Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, pages 3024–3033, 2021. 3

[52] Dongyue Wu, Zilin Guo, Li Yu, Nong Sang, and Changxin Gao. Structural pruning via spatial-aware information redundancy for semantic segmentation. In *AAAI*, 2025. 1

[53] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPs*, 2021. 6, 8

[54] Guoan Xu, Juncheng Li, Guangwei Gao, Huimin Lu, Jian Yang, and Dong Yue. Lightweight real-time semantic segmentation network with efficient transformer and cnn. *IEEE Transactions on Intelligent Transportation System*, 2023. 6, 8

[55] Fengting Yang, Qian Sun, Hailin Jin, and Zihan Zhou. Superpixel segmentation with fully convolutional networks. In *CVPR*, 2020. 1

[56] Yanchao Yang, Brian Price, Scott Cohen, and Honglak Lee. Superpixel segmentation with fully convolutional networks. In *CVPR*, pages 13964–13973, 2020. 2

[57] Yuhui Yuan, Xilin Chen, and Jingdong Wang. SegFix: Model-agnostic boundary refinement for segmentation. In *ECCV*, pages 489–506, 2020. 3

[58] Fong Z., Wall-Wieler E., Johnson S., Culbertson R., and Mitzman B. Rates of minimally invasive surgery after introduction of robotic-assisted surgery for common general surgery operations. *Ann. Surg. Ope.*, 2025. 1

[59] Hongyi Zeng, Wenxuan Liu, Tianhua Xia, Jinhui Chen, Ziyun Li, and Sai Qian Zhang. Foveated instance segmentation. In *CVPR*, 2025. 8