POLITECNICO DI TORINO

MASTER's Degree in COMPUTER ENGINEERING

MASTER's Degree Thesis

Spatial-Temporal Consistency Enhanced Segmentation for Laparoscopic Surgical Videos

Supervisors

Flavio ESPOSITO

Guido MARCHETTO

Alessio SACCO

Candidate

Ewen RONDEL

March 2026

Spatial-Temporal Consistency Enhanced Segmentation for Laparoscopic Surgical Videos

Ewen Rondel

Abstract

Accurate multi-organ semantic segmentation in laparoscopic surgery videos is essential for computer-assisted interventions, enabling organ recognition, instrument tracking, and context-aware guidance. However, surgical scenes exhibit several intrinsic challenges: large non-rigid organ deformations, occlusions caused by instruments or blood, rapid viewpoint changes, motion blur, and heterogeneous lightning. These factors undermine the temporal stability and boundary precision of conventional single-frame segmentation networks. To overcome these limitations, this thesis proposes SSTC-Seg, a deformable memory-based multi-scale architecture specifically designed to enforce spatial adaptivity and temporal coherence in minimally invasive surgical environments.

SSTC-Seg integrates three complementary components into a unified, end-to-end trainable architecture. First, a deformable multi-scale encoder —combining deformable convolutions, multi-scale feature extraction, and lightweight self-attention— adapts receptive fields to non-rigid anatomical structures while capturing both fine-grained and global contextual cues. Second, a memory-based attention mechanism aggregates information from a memory bank of past frames embeddings through stacked self-attention and cross-attention blocks. This temporal reasoning module enables robust mask propagation under occlusions, abrupt appearance changes, and rapid motion. Third, a new Hierarchical Dense CRF (HD-CRF) performs multi-resolution, deformable message passing guided by image features, skip connections, and edge cues to sharpen organ boundaries and reduce spurious predictions.

The model is evaluated on two challenging datasets: CholecSeg8k, comprising thousands of diverse laparoscopic frames, and the Dresden Surgical Anatomy dataset, which is considerably smaller and characterized by strong class imbalance to test its adaptivity. On CholecSeg8k, SSTC-Seg systematically outperforms established baselines —including U-Net, PSPNet, CFPNet-M, and MFCPNet (the last two models being designed specifically for medical image segmentation)— across Accuracy, Dice Cefficient, Jaccard Index, and Hausdorff Distance 95, with particularly notable improvements on highly deformable organs and surgical tools. Experiments on Dresden reveal that while SSTC-Seg maintains strong performance on well-represented classes, its complexity poses challenges in extremely low-data regimes, where simpler models occasionally excel. Ablation studies also confirm the essential contributions of deformable convolutions, memory-based temporal integration, and per-class seg-

mentation heads to overall robustness and boundary fidelity.

In summary, SSTC-Seg advances the state of multi-organ video segmentation by combining spatial adaptivity, temporal consistency, and structured refinement into a cohesive framework. Although its computational cost currently limits real-time applicability, the proposed design offers a powerful foundation for future surgical scene understanding systems and represents a meaningful step toward reliable, clinically deployable computer-assisted surgical tools.

ACKNOWLEDGMENTS

I would first like to express my sincere gratitude to Prof. Flavio Esposito, who welcomed me into his research group and offered me the opportunity to work on this project. His guidance, availability, and constant encouragement were invaluable throughout this work. He not only helped me navigate the scientific aspects of the topic, but also taught me how to approach research with rigor, curiosity, and structure.

My sincere thanks also go to Lin Guo, who joined the lab during the course of the project and worked closely with me on this research. His insights, technical expertise, and perspective allowed me to see aspects I had previously overlooked, and our collaboration led to the publication of my first short paper at the MIDL conference.

I would also like to thank Alessio Sacco for his responsiveness and his help whenever I needed it. His availability was greatly appreciated during the preparation and completion of this thesis.

Finally, I extend my gratitude to everyone in the Saint Louis University lab who, directly or indirectly, contributed to making this work possible.

Table of Contents

List of Figures

List of Tables

Acronyms

AttnBlock	Attention Block.
BRA	Sparse Region-Based Attention.
CFPNet-M	Channel Pruning Feature Pyramid Network - Medical variant.
CNN	Convolutional Neural Network.
CRF	Conditional Random Field.
CT	Computed Tomography.
DCN	Deformable Convolution Network.
DeformMSMCConv	Deformable Multi-Scale Multi-Channel Convolution.
DWConv	Depthwise Convolution.
FIFO	First-In First-Out.
FN	False Negative.
FP	False Positive.
HD-CRF	Hierarchical Dense Conditional Random Field.
HD95	Hausdorff Distance 95.
IoU	Intersection over Union.
MFCPNet	Multi-scale Feature fusion and Channel Pruning Network.
MLP	Multi-Layer Perceptron.
MRI	Magnetic Resonance Imaging.
MSMCConv	Multi-Scale Multi-Channel Convolution.
PSPNet	Pyramid Scene Parsing Network.
PWConv	Pointwise Convolution.

SAM	Segment Anything Model.
SAM2	Segment Anything Model 2.
SSTC-Seg	Surgical Spatial-Temporal Consistency Segmentation model.
TN	True Negative.
TP	True Positive.
U-Net	U-shaped convolutional Network.

Chapter 1

Introduction

1.1 Background and Motivation

Minimally invasive surgery has transformed modern clinical practice by reducing patient trauma, shortening recovery times, and enabling procedures that would otherwise require open approaches [1] [2]. Among these techniques, laparoscopic surgery is one of the most widely adopted. It relies heavily on the surgeon's ability to interpret a continuous video stream captured by an endoscopic camera. Within this constrained and dynamic visual environment, automatically identifying and delineating anatomical structures—or semantic segmentation—plays an increasingly central role in surgical data science.

Reliable segmentation assists a range of intraoperative technologies: decision-support systems, robot-assisted workflows, context-aware warnings, surgical skill assessment, and autonomous navigation. Despite this importance, performing multi-organ segmentation in laparoscopic videos remains an exceptionally difficult computer vision problem. Unlike static medical images such as CT or MRI scans, laparoscopic footage presents severe challenges: soft tissues undergo irregular and continuous deformation, surgical instruments and fluids frequently occlude organs, specular highlights distort appearance, and the endoscope's motion induces blur and abrupt viewpoint changes [3].

These factors create scenarios in which pixel-level predictions from standard convolutional neural networks (CNNs) tend to fluctuate, producing fragmented boundaries and inconsistent temporal behaviour. In practice, such instability limits the usefulness of segmentation models during surgery, where continuity and anatomical coherence are essential. This motivates the need for architectures that can reason not only spatially—capturing the shape and texture of organs—but also temporally, incorporating information from past frames to maintain stability across motion, occlusion, and deformation.

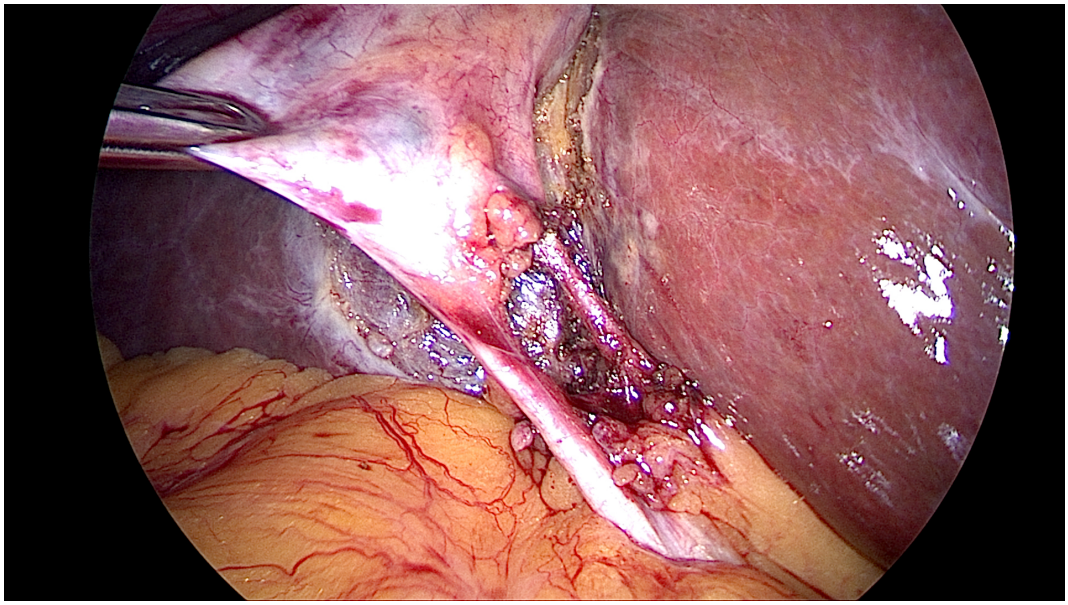

Figure 1.1: Example of a laparoscopic view.

1.2 Problem Definition

Most existing segmentation networks for laparoscopic scenes operate in a strictly frame-wise manner: each image is processed independently, and no mechanism explicitly propagates information through time [4]. While multi-scale CNNs such as U-Net [5], PSPNet [6], or MFCPNet [7] are effective at capturing hierarchical spatial features, they struggle with temporal consistency. When an organ becomes partially hidden or its appearance shifts quickly, these models often produce flickering masks that lack anatomical continuity.

Recent advances in video segmentation, such as SAM2 [8], introduce memory mechanisms that store and retrieve embeddings from previous frames. These architectures have shown strong performance in general video tasks, especially when guided by user prompts. However, they are not designed for fine-grained, densely annotated medical data, and their reliance on interactively provided prompts makes them unsuitable for fully automated intraoperative workflows.

At the same time, techniques such as deformable convolutions and dense Conditional Random Fields (CRFs) address complementary aspects of the segmentation challenge. Deformable convolutions adapt receptive fields to irregular shapes, which is crucial for modeling organic tissue deformation. CRFs can refine mask boundaries and enforce spatial coherence. But these components are rarely combined with temporal reasoning, and classical CRFs are often rigid, single-scale, and limited in their ability to adapt to the complex appearance variations seen in surgical videos.

Therefore, the core problem addressed in this work is the development of a unified

framework capable of:

1. capturing robust multi-scale spatial representations of highly deformable tissues;
2. leveraging temporal context to maintain segmentation stability across frames;
3. refining predictions to produce anatomically consistent boundaries;
4. operating without external prompts, making it suitable for real surgical datasets.

In practice, developing robust segmentation models is further constrained by limitations in data availability. Pixel-wise annotations in surgical videos require expert knowledge and significant manual effort, resulting in datasets that are comparatively small and often imbalanced. These factors complicate generalization, particularly for structures that appear infrequently or exhibit highly variable shapes across procedures.

Additionally, intraoperative settings impose strict real-time requirements. Models must process high-resolution video at sufficient frame rates to avoid latency that could interfere with a surgeon's perception or a robotic system's control loop. Achieving high accuracy while maintaining computational efficiency remains an open challenge, especially for architectures relying on large attention mechanisms or deep multi-scale pipelines.

These constraints highlight the need for segmentation frameworks that balance spatial detail, temporal stability, and computational feasibility. Designing such a system—capable of performing reliably across diverse scenes, motion patterns, and lighting conditions—constitutes the central problem addressed in this thesis.

1.3 Thesis Objectives and Contributions

The overarching objective of this thesis is to design and evaluate a segmentation architecture specifically tailored to the unique challenges of laparoscopic video analysis. Building on the limitations of existing approaches, the proposed network—named SSTC-Seg—aims to combine spatial adaptivity, temporal memory, and probabilistic refinement within a single end-to-end trainable system.

The main contributions of this thesis are as follows:

- A deformable, multi-scale feature extraction strategy.

The architecture integrates deformable convolutions with multi-scale feature extraction to better adapt to the irregular shapes and large deformations characteristic of soft tissues.

- A memory-based temporal attention mechanism.

Inspired by recent video segmentation models, SSTC-Seg embeds a learnable memory

module that stores condensed representations of past frames and uses them to guide current predictions. This allows the model to maintain temporal consistency even under occlusion or fast motion.

- A trainable Hierarchical Dense CRF (HD-CRF).

To sharpen boundaries and enforce spatial coherence, a multi-resolution CRF is incorporated directly into the network pipeline. It leverages deformable bilateral filtering and guidance features to refine predictions without relying on rigid, hand-crafted kernels.

- A unified segmentation framework for surgical video data.

The combination of deformable convolutions, temporal memory attention, and hierarchical CRF refinement yields a model that is robust to deformation, illumination changes, occlusions, and motion—conditions that frequently occur in the operating room.

Beyond these core components, an additional objective of this thesis is to examine how each architectural element contributes to overall segmentation quality. In highly dynamic surgical scenes, performance often depends not only on the choice of individual modules but also on the interactions between them. For instance, temporal memory is most effective when paired with features that are themselves spatially stable, while CRF-based refinement benefits from the richer contextual information produced by multi-scale encoders. Evaluating these relationships provides insight into how spatial and temporal cues can be jointly optimized for medical video analysis.

These contributions collectively advance the state-of-the-art in automated laparoscopic scene understanding and address the specific requirements of surgical data science applications.

1.4 Structure of the Thesis

This thesis is organized into seven chapters.

Chapter 1 introduces the problem of semantic segmentation in laparoscopic surgery, outlines the unique challenges of video-based medical analysis, and presents the objectives and contributions of the proposed work.

Chapter 2 reviews relevant literature, including multi-scale CNN architectures, deformable convolutions, memory-based video segmentation frameworks, and CRF-based refinement strategies. This chapter highlights the strengths and limitations of existing methods and situates SSTC-Seg within this landscape.

Chapter 3 describes the proposed SSTC-Seg architecture in detail. It presents the multi-scale deformable encoder, the memory attention module, the decoder structure, the hierarchical CRF refinement stage, and the memory encoding and storage mechanisms.

Chapter 4 outlines the datasets used in this study—CholecSeg8k and the Dresden Surgical Anatomy dataset—focusing on their characteristics, labeling protocols, and challenges.

Chapter 5 details the experimental setup, baseline comparisons, evaluation metrics, and results. Both quantitative and qualitative assessments are provided, illustrating the strengths and limitations of the proposed method relative to state-of-the-art models.

Chapter 6 discusses the implications of the results, analyzes failure modes, and addresses the model's computational constraints and limitations.

Chapter 7 concludes the thesis by summarizing the work accomplished and proposing potential directions for future research, including opportunities for real-time optimization and broader applicability across surgical domains.

Although each chapter focuses on a specific aspect of the research, the overall progression reflects the incremental development of the proposed system.

Chapter 2

Related Works

2.1 MFCPNet

In 2025, Hou et al. introduced MFCPNet [7], a medical image segmentation network designed to extract multi-scale and multi-channel features, at both local and global levels, by extending a classic convolutional U-Net [5], using two main modules: MSMCConv and AttnBlock.

Beyond introducing new modules, MFCPNet was motivated by a central limitation in classical U-Net style architectures: while U-Net aggregates multi-scale features through its contracting–expanding pathway, it processes each scale with identical kernel structures that do not explicitly differentiate local textures from larger spatial patterns. This becomes especially problematic in medical imaging, where anatomical structures often exhibit highly diverse spatial footprints: fine details such as vessel boundaries co-exist with large homogeneous regions such as parenchymal tissue. Conventional convolutions typically require deeper layers or enlarged kernels to approximate such variability, resulting in higher computational costs or loss of precision. MFCPNet directly addresses this imbalance by embedding scale diversity within each layer rather than relying solely on depth.

Multi-Scale Multi-Channel Convolution Using parallel convolutional branches is a common strategy in image segmentation networks [9], and MSMCConv extends depthwise–separable convolutions by applying three parallel depthwise convolutions per channel with kernel sizes 3×3, 5×5, and 7×7. Let f_3, f_5, and f_7 be the outputs of these depthwise paths. These are concatenated and merged via a 1×1 pointwise convolution:

$$f_{\text{MSMC}} = \text{Conv}_{1\times1}\big[\, f_3,\ f_5,\ f_7 \,\big].$$

This formulation can also be interpreted as an implicit frequency decomposition: the 3×3 branch emphasizes high-frequency edges and abrupt transitions, the 5×5 branch captures intermediate textures and soft contours, and the 7×7 branch aggregates low-frequency contextual information. Similar multi-kernel computer vision

designs have shown improved representation capacity by jointly capturing high-, mid-, and low-frequency contextual structures [10].

This design captures both fine local details (small kernels) and broader context (large kernels) with significantly fewer parameters and FLOPs than naïve large-kernel convolutions.

Self-Attention Block The Self-Attention Block in MFCPNet comprises three stages:

1. **Implicit Positional Encoding via DWConv.** A depthwise convolution embeds relative positional biases directly:

$$f_p = \mathrm{DWConv}(f) + f,$$

   enabling adaptive positional cues without learned absolute embeddings.

2. **Sparse Region-Based Attention (BRA).** After layer normalization, the feature map is partitioned into non-overlapping $S \times S$ regions. For each region r, queries Q_r, keys K_r, and values V_r are projected. An adjacency matrix $A_r = Q_r K_r^\top$ identifies the top-k most relevant regions. Attention is then computed only over these selected key–value pairs:

$$O_r = \mathrm{Attention}(Q_r,\, K_g,\, V_g) + \mathrm{DWConv}(V_r),$$

   reducing complexity while preserving global understanding.

   Sparse or region-restricted attention mechanisms have been proposed as efficient alternatives to full self-attention in large-resolution feature maps [11], and the BRA mechanism provides an efficient compromise between fully global attention and purely local feature mixing. By operating only on the most relevant regions, the model implicitly learns a notion of regional saliency, which proves particularly useful in surgical scenes where some anatomical structures dominate the field of view while others appear only sporadically. This asymmetry enables MFCPNet to redirect computational resources where they matter most, strengthening its robustness against occlusions, abrupt camera motion, and illumination shifts.

3. **Global Refinement via MLP-Mixer.** A two-step MLP-Mixer refines O_r: first a token-mixing MLP across spatial positions, then a channel-mixing MLP across feature channels, each with residual connections. This yields f_{global}, improving rotational invariance and generalization. Such MLP-based global mixing approaches have proven effective for capturing long-range dependencies [12].

While MSMCConv enhances local and regional representation, MFCPNet complements it with a lightweight attention mechanism designed to capture long-range dependencies without incurring the high computational footprint of full-transformer blocks. This type of hybrid design has gained traction in recent medical segmentation architectures, as anatomical structures often exhibit relationships that extend well beyond the local neighborhood—for instance, the spatial alignment between organs or consistent edges across deforming tissue. The Self-Attention Block in MFCPNet is specifically engineered to extract such global cues in a constrained and computationally friendly manner.

At each stage, the MSMCConv and SelfAttnBlock modules are executed in parallel, and their outputs are then fused before being downsampled.

By coupling MSMCConv's local multi-scale feature aggregation with the sparse, region-aware attention and MLP-Mixer refinement, MFCPNet achieves efficient, real-time segmentation with strong boundary delineation and global consistency.

2.2 Deformable Convolutions

In 2017, Dai et al. introduced Deformable Convolutional Networks (DCN) [13]. The motivation behind deformable convolutions arises from a fundamental limitation of classical CNNs: their sampling grid is fixed and spatially invariant. While this rigidity is effective for recognizing structured geometric patterns, it becomes a major bottleneck when dealing with objects subject to non-rigid deformations— a hallmark of anatomical and surgical environments. Organs shift, stretch, compress, and rotate in ways that cannot be fully captured by standard convolutional kernels, even with deep hierarchies or multi-scale designs. Deformable convolution alleviates this rigidity by dynamically adjusting the sampling positions based on the local content of the image. DCN augment standard convolutions with learnable, image-conditioned sampling offsets, enabling the kernel to adapt its shape to object geometry.

In a regular convolution the output at location p_0 is:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \left[x(p_0 + p_n)\right],$$

where $\mathcal{R}$ is the kernel's fixed grid (e.g. a 3×3 neighborhood) and $w(p_n)$ are the weights. DCN adds offsets $\{\Delta p_n\}$:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \left[x(p_0 + p_n + \Delta p_n)\right].$$

Since $p_0 + p_n + \Delta p_n$ may be fractional, the input feature is sampled via bilinear

interpolation:

$$x(p) = \sum_q G(q,p)\, x(q),$$

with:

$$G(q,p) = \max(0, 1 - |q_x - p_x|)\ \max(0, 1 - |q_y - p_y|).$$

The use of bilinear interpolation ensures differentiability with respect to both the feature values and the offsets, allowing gradient-based optimization to learn spatial offset patterns that correspond to meaningful anatomical variations. This makes deformable convolutions particularly well suited for medical segmentation tasks in which the boundaries between organic shapes exhibit soft transitions or irregular contours [14]. Instead of relying solely on learned filter weights to approximate these shapes, the network directly adapts its receptive field to follow them.

Both the offsets Δp_n and weights w are learned end-to-end, with offsets produced by a parallel convolutional layer over the same inputs.

The flexibility that deformable convolutions bring to classical rigid grids allows a better feature extraction for organic shapes.

Recent variants of deformable convolutions have extended the concept further by introducing modulation scalars, like DCNv2 [15], or multi-level offset learning and attention-driven offset prediction, like DCNv3 and DCNv4 [16]. Such variants improve the stability of sampling patterns and allow the network to focus even more selectively on salient structures. Although SSTC-Seg employs the classical DCNv2 formulation for computational efficiency purposes, the underlying principles remain consistent: embedding geometric adaptivity directly into the convolutional kernel substantially enhances the model's ability to parse deformable anatomy in laparoscopic footage.

2.3 Video Segmentation via Streaming Memory

Traditional frame-by-frame segmentation approaches fail primarily because they lack mechanisms for temporal persistence. In 2024, Ravi et al. introduced SAM2 [8], a model that extends image-level segmentation to videos by equipping the already-existing static SAM architecture (Kirillov et al., 2023) [17] with a streaming memory that records and recalls semantic cues across frames. Instead of segmenting each frame independently, SAM2 processes frames sequentially through its image encoder and conditions mask predictions on a learnable buffer of past embeddings.

Memory Bank SAM2 maintains a fixed-size FIFO buffer of memory features (fused image and mask embeddings from past frames). At each time step, the current frame's features are enqueued and the oldest entry evicted when capacity is reached. The FIFO constraint of the memory bank implicitly enforces a temporal horizon, allowing the model to focus on the most recent context while discarding obsolete representations. However, the size of the buffer remains a critical hyperparameter: too small and the model loses context under occlusion, too large and irrelevant information may dominate the attention mechanism.

Memory Encoder Before storage, per-class mask logits are passed through a convolutional module and residually fused with the image encoder's output. This projection aligns mask and appearance cues in the same embedding space, ensuring compatibility for attention operations. This projection step serves as a semantic alignment mechanism that allows the memory to store information not only about raw appearance but also about previous segmentation decisions. The use of dedicated memory encoders and key-value storage was popularized by Space-Time Memory Networks [18], which demonstrated strong mask propagation performance. By preserving class-specific evidence within the memory structure, SAM2 effectively captures high-level continuity.

Memory Attention A stack of A transformer-style blocks refines the current frame's embedding by performing both self-attention and cross-attention.

This cross-frame attention enables robust mask propagation through occlusions, deformations, and rapid motion.

By streaming a compact history of semantic and spatial representations, SAM2 delivers a coherent mask propagation over long video sequences.

Despite its effectiveness in general video segmentation, SAM2 remains a promptable model, meaning that it relies on user-provided prompts such as bounding boxes, points, or textual cues to guide the segmentation process and exploit its full potential. This interactive design is not the most suitable for laparoscopic surgery video segmentation datasets, as most of them are only provided with only the semantic masks.

Recent efforts such as Medical SAM 2 (Zhu et al., 2024) [19] adapt this pair of memory mechanism and promptable segmentation frameworks to medical imagery, but still depend on user interaction. And even in laparoscopic surgery videos, where prompts-enriched datasets are too scarce, it shows that segmentation using streaming memory is the way to go.

Despite these strengths, challenges remain in applying SAM2-like architectures

to dense medical segmentation. Their design presumes the availability of interactive prompts, and their feature extraction mechanism is optimized for general video scenes rather than fine-grained organ-level segmentation. Nonetheless, their core principle—the use of streaming memory to stabilize predictions across time—provides a strong foundation upon which specialized medical architectures such as SSTC-Seg can expand.

2.4 CRF-based Smoothing

Conditional Random Fields (CRFs) have long been employed as a post-processing step to improve segmentation smoothness and enforce spatial consistency between neighboring pixels. Early works modeled the segmentation map as a Markov Random Fields with unary potentials from a CNN and pairwise terms encouraging adjacent pixels with similar color or texture to share labels.

In early deep learning literature, CRFs were often positioned as complementary modules that corrected systematic biases in CNN outputs. While CNNs excel at learning semantic categories, they tend to struggle with precise localization due to pooling operations and the inherent smoothness of learned filters. CRFs bring back pixel-level precision by modeling fine-grained local interactions, effectively recovering boundaries that CNNs blur. This division of labor—CNN for semantics, CRF for geometry—has been a recurring theme in segmentation research.

The DenseCRF framework popularized by Krähenbühl and Koltun (2012) [20] introduced fully connected pairwise potentials and efficient mean-field inference, substantially sharpening object boundaries in natural-image segmentation. Subsequent medical-imaging studies adopted DenseCRFs to refine coarse CNN outputs and reduce spurious predictions along organ borders, from brain tissues (Chen et al., 2018) [21] to lung fields segmentation (Li et al.,2021) [22].

However, classical CRFs rely on hand-crafted Gaussian kernels and fixed pairwise weights that cannot easily adapt to complex anatomical variations or heterogeneous illumination conditions often found in laparoscopic surgery videos. Moreover, they operate only at single resolution and treat spatial relationships as rigid, limiting their effectiveness when organs deform or occlusions occur over time.

To address these issues, hierarchical and associative CRF variants (Ladický et al., 2013; Zhang et al., 2015) [23][24] extended the model to multi-scale representations and inter-class dependencies, enabling structured refinement across object parts and semantic levels.

These works show that a medical segmentation network could benefit from the

refinement of a CRF which has characteristics derived from both DenseCRFs and hierarchical variants.

Chapter 3

Network Architecture

3.1 Overview of the Pipeline

Our SSTC-Seg network processes an input video sequence in five stages (see Fig. 3.1):

Stage 1 - Frame Encoding
Each raw RGB frame is passed through a multi-scale, deformable convolutional encoder to extract rich spatial features at multiple resolutions.

Stage 2 - Memory Attention
A multi-head attention mechanism is performed between the extracted features and the results of the last F frames to enforce spatial consistency.

Stage 3 - Mask Decoding
The fused features are decoded back to a full-resolution mask using a shared heavy-decoder and class-specific segmentation heads.

Stage 4 - Hierarchical Dense CRF
The coarse decoder outputs are recursively refined with a hierarchical, dense CRF module to enforce even more spatial consistency, and also sharpen boundaries.

Stage 5 - Memory Encoding and Bank
Segmentation predictions (and their extracted features) from previous frames are encoded into a fixed-size memory representation and stored in a rolling memory bank of the last F frames.

This staged design follows the classical encoder–reasoning–decoder paradigm but extends it along both the spatial and temporal dimensions. Instead of treating each frame as an isolated image, SSTC-Seg explicitly couples frame-wise feature extraction, memory-based temporal reasoning, and probabilistic refinement into a single unified pipeline. As a result, the network can simultaneously address three core challenges of laparoscopic video analysis: geometric deformation of organs, appearance variability

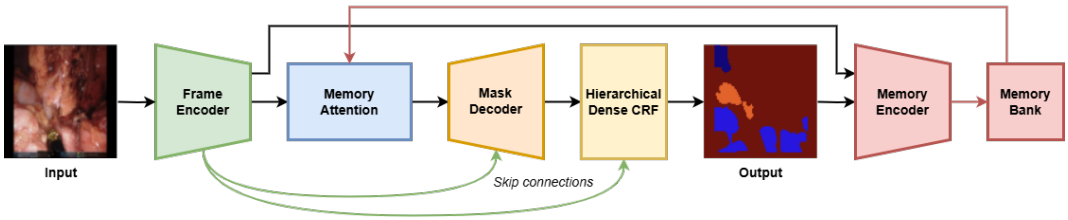

Figure 3.1: Overview of the SSTC-Seg pipeline. Frames are processed streamingly, being sent one at a time to the Frame Encoder for feature extraction, and then to the Memory Attention module to be cross-attended with memories. From that point, the Mask Decoder predicts the segmentation masks for the current frame and the HD-CRF refines these predictions. Finally, the Memory Encoder transform the predictions, along their associated extracted features, into embeddings that are store into the Memory Bank for next frames' predictions.

over time, and the need for temporally consistent masks that do not flicker across consecutive frames.

This gating of spatial deformability, temporal memory, and CRF smoothing allows SSTC-Seg to robustly track organ outlines even under occlusion, rapid motion, and large deformations.

3.2 Frame Encoder

The *Frame Encoder* extracts multi-scale feature representations from each input frame before temporal or memory-based processing. It is composed of E consecutive *Encoder Blocks*. At each block $i \in \{1, \dots, E\}$ the spatial resolution is halved (except in the final block) and the number of feature channels doubles, yielding a coarse-to-fine hierarchy of feature maps and intermediate skip connections for subsequent decoding.

This hierarchical organization is crucial for semantic segmentation in surgery. Fine-resolution layers preserve detailed appearance cues such as organ edges and instrument contours, while coarser layers encode global spatial relationships between multiple structures in the scene. The skip connections linking encoder and decoder ensure that this information is not lost during downsampling, enabling precise localization even when the bottleneck representation is heavily compressed.

Let the input tensor be:

$$x_0 \in \mathbb{R}^{C_0 \times H \times W},$$

where C_0 is the number of input channels (e.g., 3 for RGB) and H, W are the frame height and width. We set a base channel dimension $D = 64$. Then, at block i the feature dimensionality is:

$$C_i \;=\; D \times 2^{\,i-1},$$

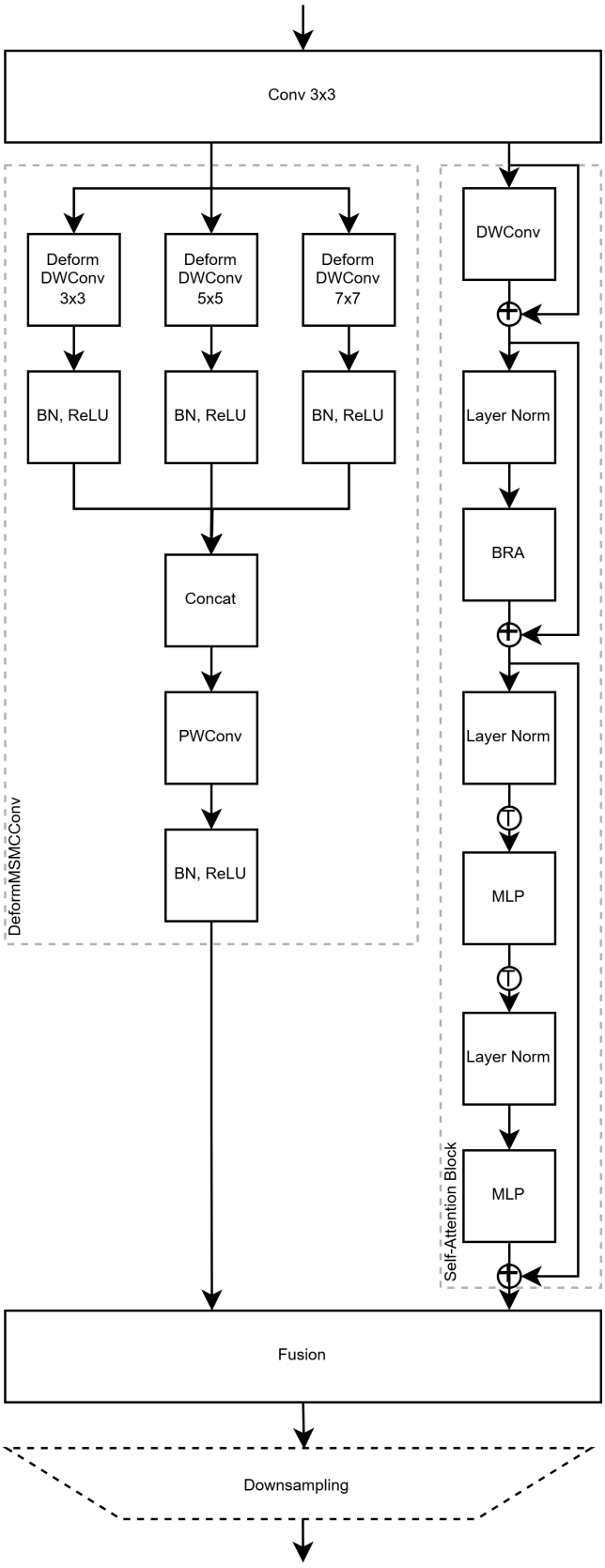

Figure 3.2: Encoder Block architecture. SSTC-Seg uses modules heavily inspired by the MSMCConv and Self-Attention modules from MFCPNet [7], and leverages the adaptability of deformable convolutions to optimize organic feature extraction.

and the spatial resolution is:

$$(\mathrm{H}_i, W_i) = (H/2^{\min(i-1,E-1)},\, W/2^{\min(i-1,E-1)}).$$

From an implementation perspective, this schedule of doubling channels while halving spatial resolution at each level follows the common design of modern encoder–decoder architectures, striking a balance between representational capacity and computational cost. Early layers focus on dense, high-resolution representations with relatively few channels, whereas deeper layers trade spatial detail for richer, more abstract features. This is particularly advantageous in surgical videos, where local details must be preserved but global context remains essential for disambiguating visually similar tissues.

Encoder Block Structure

Each *Encoder Block* performs:

1. **Initial 3×3 Convolution + BatchNorm + ReLU:**

$$y_1 = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv3}\times 3(x_{i-1};\, C_{i-1} \to C_i,\, \mathrm{stride} = s))\big),$$

   where $s = 2$ if downsampling (for $i < L$), else $s = 1$.
   The initial convolution thus serves two purposes: it performs the necessary spatial downsampling (when $s = 2$) and projects the input into the appropriate feature dimension C_i. The combination of batch normalization and ReLU activation stabilizes training and introduces non-linearity, ensuring that the subsequent deformable and attention-based modules operate on well-conditioned feature maps.

2. **Deformable Multi-Scale Multi-Channel Convolution (DeformMSMCConv):** Three parallel deformable, depthwise-separable convolutions with kernels 3×3, 5×5, and 7×7 are applied to y_1, then concatenated and fused via a 1×1 pointwise convolution, producing $y_2 \in \mathbb{R}^{C_i \times H_i \times W_i}$.
   Because these convolutions are depthwise-separable and deformable, the block achieves a large effective receptive field without incurring the full cost of dense convolutional kernels. The learned offsets allow each kernel to align itself with salient structures such as organ boundaries, while the multi-scale configuration captures both small and large anatomical details. This is particularly beneficial in frames where multiple organs or tools occupy different scales in the same image.

3. **Self-Attention Block:** A lightweight attention module also refines y_1 in parallel of DeformMSMCConv by:

- Applying depthwise 3×3 conv + LayerNorm
- Unfolding into patches of size $k \times k$ (default $k = 5$)
- Applying multi-head attention ($h = 8$ heads)
- Applying a residual MLP with skip connections

to yield $y_3 \in \mathbb{R}^{C_i \times H_i \times W_i}$.

4. **Fusion and Skip Extraction:** The DeformMSMCConv and Self-Attention block outputs are then fused:

$$y_{\text{out}} = \text{ReLU}(\text{Conv1} \times 1(y_3 + y_4)).$$

For $i < E$, y_{out} is stored in the skip list for a later *Mask Decoder* fusion.

By combining deformable sampling, multi-scale convolution, and localized self-attention within each block, the *Frame Encoder* produces a hierarchy of robust feature maps that can flexibly adapt to the highly deformable and occluded anatomy encountered in laparoscopic videos.

An additional advantage of this design is its modularity: the number of Encoder Blocks E can be adapted to the available computational budget without requiring major architectural changes. In all cases, the DeformMSMCConv and Self-Attention components remain the key primitives driving representation quality.

3.3 Memory Attention

The *Memory Attention* module integrates information from a learned *Memory Bank* of past frame embeddings into the current-frame representation, enabling robust mask propagation across occlusions and large deformations. It consists of A identical *Attention Blocks*, each comprising self-attention, cross-attention over memory, and a feed-forward update, all wrapped in residual connections and layer normalization.

Conceptually, this module acts as a temporal extension of the encoder: instead of operating solely on spatial neighborhoods within a single frame, it learns to relate pixels across time by querying a bank of previous embeddings. This allows the network to exploit temporal redundancies—objects that persist across frames— and to recover from transient artifacts such as specular highlights, smoke, or partial occlusions that might compromise single-frame predictions.

Inputs and Embedding Layout

Let the encoded feature map of the current frame be:

$$F_t \in \mathbb{R}^{C \times H \times W},$$

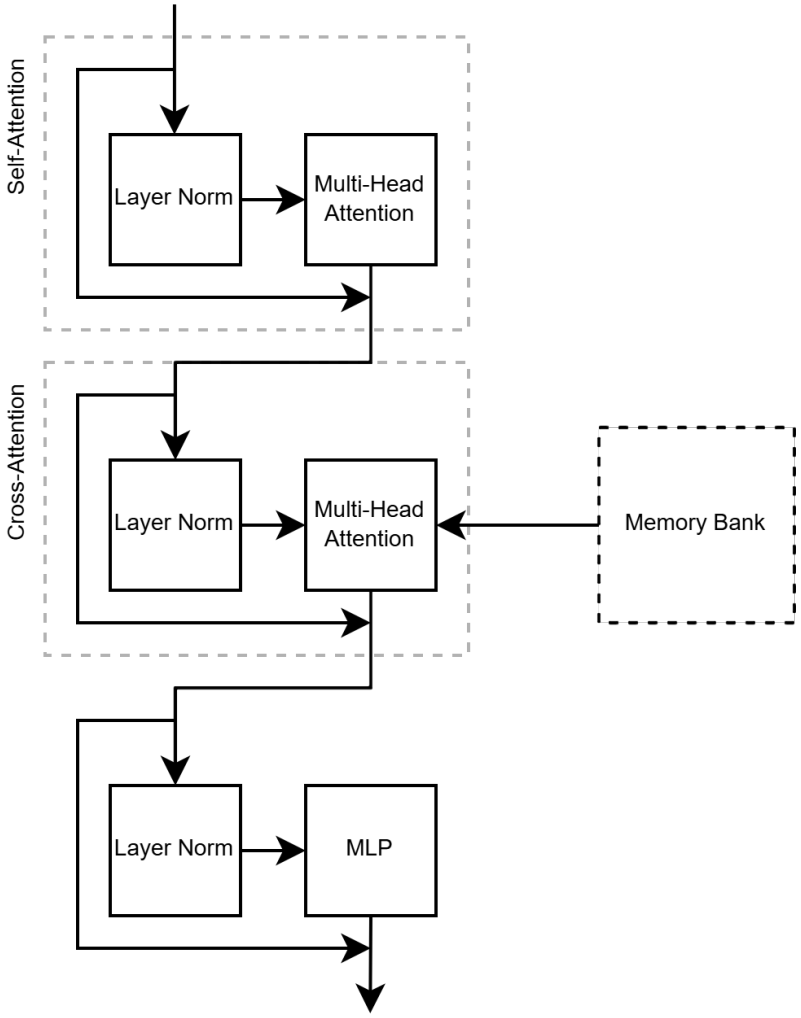

Figure 3.3: Attention Block architecture. Each block first performs self-attention solely on its input, then performs cross-attention with the memory, and finishes with a simple multi-perceptron layer.

which we reshape into a sequence of $N = H \times W$ tokens:

$$X \in \mathbb{R}^{N \times D}, \quad D = C.$$

Representing the feature map as a sequence of N tokens allows us to leverage standard transformer-style attention mechanisms. This tokenization is the key step that bridges convolutional feature extraction and sequence-based temporal processing.

Similarly, up to F past frame embeddings are stored in a memory bank:

$$\{M_{t_f}\}_{f=1}^{F}, \quad M_{t_f} \in \mathbb{R}^{C \times H \times W},$$

and reshaped into:

$$M \in \mathbb{R}^{(FN) \times D}.$$

Attention Block Structure

Each of the A blocks applies the following sub-layers in sequence, with residual additions around each:

1. **Self-Attention on Current Frame:**
   - Apply LayerNorm to X.
   - Perform multi-head self-attention across the N tokens:

   $$X' = \text{SelfAttn}(\text{LayerNorm}(X)) + X.$$

   From a functional viewpoint, this step refines the current frame's representation by allowing each token to aggregate information from all other positions within the same frame. As a result, spatially distant but semantically related regions—such as different parts of the same organ separated by an instrument—can directly exchange information, improving intra-frame consistency before any temporal reasoning is applied.

2. **Cross-Attention with Memory:**
   - Apply LayerNorm to X'.
   - Use X' as queries and M as keys and values in multi-head attention:

   $$X'' = \text{CrossAttn}\big(\text{LayerNorm}(X'),\, M\big) + X'.$$

   Cross-attention with memory then enables explicit modeling of temporal relations. Tokens in the current frame can attend to those in past frames that carry similar features, effectively retrieving historical evidence about the same anatomical region. This is particularly helpful in challenging cases where the current observation is ambiguous or degraded: the model can fall back on earlier, clearer views stored in the memory bank.

3. **Feed-Forward Update:**

   - Apply LayerNorm to X''.
   - Pass through a two-layer MLP with expansion factor 4 and GELU activation:

$$X''' = \mathrm{MLP}\big(\mathrm{LayerNorm}(X'')\big) + X''.$$

   The feed-forward MLP, together with residual connections and layer normalization, ensures that the transformed features remain expressive and stable during training. It also introduces additional non-linearity.

By stacking A such blocks, the model iteratively refines spatial consistency (via self-attention) and temporal coherence (via cross-attention).

Output

After A blocks, the final token sequence $X_{\text{out}} \in \mathbb{R}^{N \times D}$ is reshaped back to

$$\mathbb{R}^{C \times H \times W},$$

and passed to the *Mask Decoder*. Through residual cross-attention, the *Memory Attention* module ensures that segmentation features remain coherent even under rapid organ motion and occlusion.

In summary, the Memory Attention module serves as the central temporal reasoning engine of SSTC-Seg. By repeatedly alternating self-attention and cross-attention, it embeds a notion of short-term memory directly into the feature space. This design choice simplifies optimization while preserving the capacity to model complex motion patterns present in laparoscopic videos.

3.4 Mask Decoder

The *Mask Decoder* reconstructs full-resolution segmentation masks by progressively upsampling the enriched feature maps and fusing encoder skip-connections. It consists of a sequence of weight-shared *Decoder Blocks* across all classes, followed by 1×1 lightweight per-class mask convolutional heads.

Importantly, the decoder operates on features that have already been enriched by temporal attention. This means that each upsampled feature map implicitly encodes both spatial context from the encoder and temporal context from the Memory Attention module. The decoder therefore acts as the final stage where these multi-scale, multi-frame cues are fused into a dense segmentation map at the original resolution.

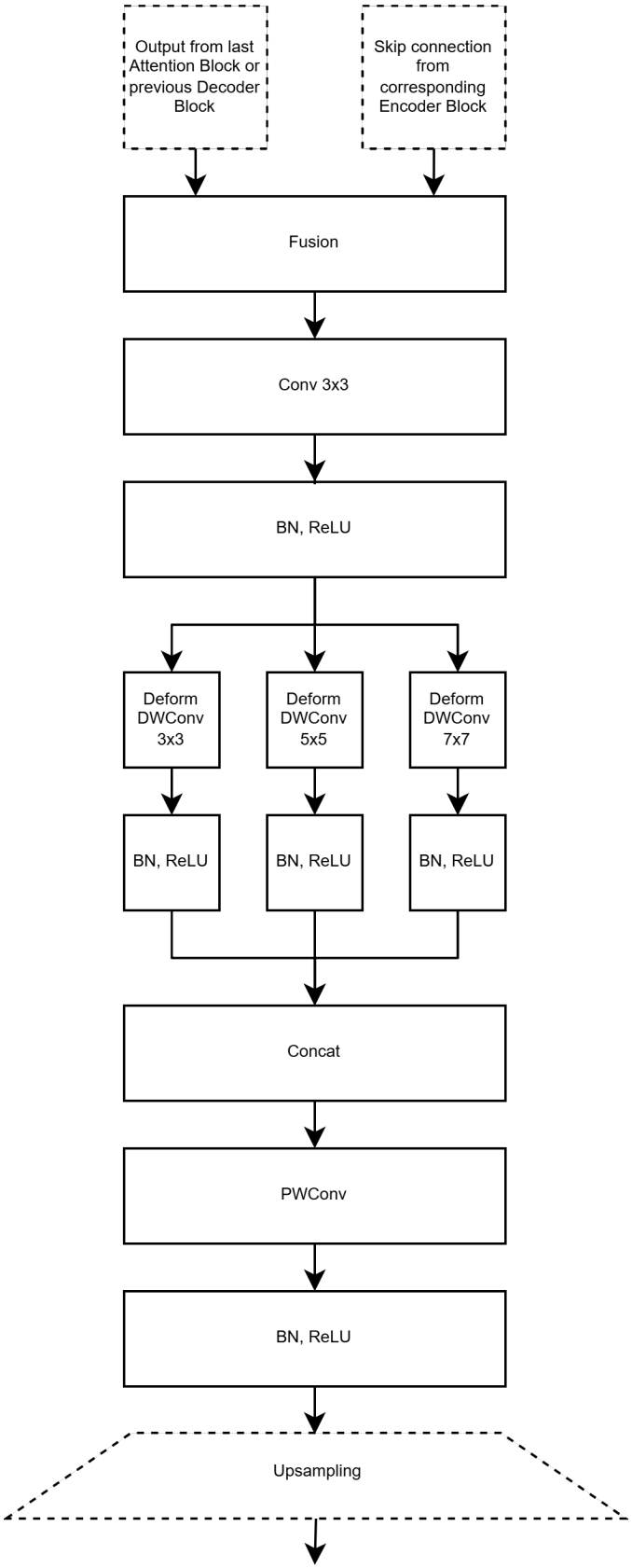

Figure 3.4: Decoder Block architecture. SSTC-Seg uses modules heavily inspired by the decoder from MFCPNet.

Decoder Block Structure

Each *Decoder Block* takes as input:

$$x_i \in \mathbb{R}^{C_{i+1} \times H_i \times W_i} \quad \text{and} \quad s_i \in \mathbb{R}^{C_i \times H_i \times W_i},$$

where x_i is the feature map from the previous (deeper) stage and s_i is the corresponding encoder skip-connection. It computes:

1. **Transposed Convolution Upsample:**

   $$u = \text{ConvTranspose2d}(x_i;\ C_{i+1} \to C_i)$$

   (with stride = 2, kernel size = 2)

2. **Spatial Alignment (if needed):**

   $$u \leftarrow \text{Interp}(u,\ \text{size} = (H_i, W_i))$$

   via bilinear interpolation.

3. **Skip Fusion:**

   $$f = u + s_i.$$

   This skip fusion mechanism restores high-frequency information that may have been lost during downsampling and attention processing.

4. **3×3 Convolution + BatchNorm + ReLU:**

   $$f' = \text{ReLU}(\text{BatchNorm}(\text{Conv3} \times 3(f;\ C_i \to C_i,)))$$

   (with padding = 1)

5. **Deformable Multi-Scale Multi-Channel Convolution Refinement:**

   $$x_{i-1} = \text{DeformMSMCConv}(f';\ C_i \to C_i),$$

   which applies parallel deformable, depthwise-separable kernels of sizes 3×3, 5×5, and 7×7 with learned offsets, then fuses them via a 1×1 convolution. By mirroring the DeformMSMCConv design of the encoder, the decoder ensures that geometrically adaptive processing is applied not only during feature extraction but also during reconstruction. This symmetry helps the network maintain precise alignment between features and underlying anatomy throughout the entire downsampling–upsampling path, which is essential for high-quality mask boundaries.

Let A-1 be the number of *Decoder Blocks* (equal to the *Frame Encoder* depth minus 1, as the last *Encoder Block* doesn't perform downsampling). Given the

deepest feature x_{A-1} from *Memory Attention* and skip-list $\{s_{A-1}, \ldots, s_1\}$, the shared decoding proceeds:

$$x_{i-1} = \text{DecoderBlock}_i(x_i,\ s_i), \quad i = A-1, \ldots, 1.$$

After the final block we obtain:

$$x_0 \in \mathbb{R}^{C_1 \times H \times W},$$

where (H, W) is the original frame size.

Per-Class Mask Heads

For each of the K target organ classes, a separate 1×1 convolution head produces a single-channel logit:

$$m_k = \text{Conv}1 \times 1_k(x_0;\ C_1 \to 1), \quad k = 1, \ldots, K.$$

These are concatenated to form the final mask tensor $\hat{M} \in \mathbb{R}^{K \times H \times W}$.

This design—shared heavy upsampling and refinement across all classes, followed by lightweight per-class heads—balances parameter efficiency with class-specific flexibility. The deformable, multi-scale convolutions within each block ensure that the decoder can accurately reconstruct complex organ boundaries, even under severe deformations.

Moreover, the use of separate heads for each class allows SSTC-Seg to model subtle differences in appearance and shape distributions across organs, without requiring completely independent decoders. This is particularly advantageous in datasets with heterogeneous class frequencies, where some structures (e.g., liver, gallbladder) appear in most frames while others are only occasionally visible.

3.5 Hierarchical Dense CRF

To refine the raw mask logits and enforce even more spatial consistency, we adopt a trainable, multi-scale conditional random field (CRF) that performs deformable message passing with image- and skip-feature guidance in a coarse-to-fine hierarchy. We call it *Hierarchical Dense CRF* (HD-CRF).

In contrast to classical CRF post-processing applied as a detached refinement stage, HD-CRF is integrated into the network as a differentiable module. This allows its parameters—compatibility transforms, convolution kernels, and guidance projections— to be optimized jointly with the rest of the architecture. As a result,

the model can learn dataset-specific priors about typical organ shapes, boundary sharpness, and inter-class relationships directly from data.

Mean-Field Iterations

Let the initial per-class logits be:

$$U \in \mathbb{R}^{K \times H \times W},$$

and initialize the class probabilities:

$$Q^{(0)} = \mathrm{softmax}(U).$$

For each iteration $t = 1 \dots T$, compute:

$$\begin{aligned} M^{(t)} &= W_{\mathrm{compat}}\Big(S_{\mathrm{spatial}}(Q^{(t-1)}) + S_{\mathrm{bilateral}}(Q^{(t-1)}, G)\Big), \\ U^{(t)} &= U \; - \; M^{(t)}, \\ Q^{(t)} &= \mathrm{softmax}\big(U^{(t)}\big), \end{aligned}$$

where:

- S_{spatial} is a class-wise deformable convolution (kernel size k_s) applied to each probability map;
- $S_{\mathrm{bilateral}}$ is a deformable bilateral convolution (kernel size k_b) on the concatenated guidance tensor G;
- W_{compat} is a learnable 1×1 compatibility transform.

Guidance Tensor

The guidance tensor:

$$G \; = \; [\, Q^{(t-1)};\; I_{\mathrm{RGB}};\; S_{\mathrm{skip}};\; E \,] \; \in \; \mathbb{R}^{(K+3+3+1) \times H \times W}$$

concatenates:

- Current probabilities $Q^{(t-1)}$;
- Raw image channels I_{RGB};
- Projected *Frame Encoder* skip-features S_{skip} (3 RGB channels);
- Edge magnitude map E.

Hierarchical Scales

We perform the above mean-field updates at three resolutions:

1. Full resolution (H, W);
2. Half resolution $(H/2, W/2)$;
3. Quarter resolution $(H/4, W/4)$.

For each scale we obtain $Q^{(T)}_{\text{full}}$, $Q^{(T)}_{1/2}$, and $Q^{(T)}_{1/4}$. Upsampling the latter two back to (H, W) and averaging yields the final refined probabilities:

$$\widehat{Q} = \frac{1}{3}\Big(Q^{(T)}_{\text{full}} + \text{upsample}(Q^{(T)}_{1/2}) + \text{upsample}(Q^{(T)}_{1/4})\Big).$$

The deformable filters with learnable offsets in both spatial and bilateral convolutions allow the *HD-CRF* to adapt to organ boundaries rather than enforcing rigid Gaussian kernels, the guidance channels leverage both low- and high-level cues to prevent oversmoothing, and the hierarchical scales ensure fine detail preservation and global coherence. This trainable, *Hierarchical Dense CRF* significantly sharpens mask contours and reduces spurious predictions, especially in regions of low contrast and/or heavy occlusion.

3.6 Memory Encoder

The *Memory Encoder* transforms the per-class mask predictions into objects in the same feature space as the image-derived frame embeddings, enabling seamless temporal attention. It processes the concatenated mask logits through a lightweight convolutional pyramid, matching both channel dimensionality and spatial resolution to the *Frame Encoder*'s output.

Without this component, the memory would contain only appearance-based features extracted by the encoder, which may not fully reflect the network's final segmentation decisions. By explicitly encoding mask logits into the same feature space, the *Memory Encoder* allows the model to store a richer representation that combines what the network "sees" (image features) with what it "believes" (segmentation outputs). This is particularly beneficial when frames are corrupted by noise or partial occlusions.

Inputs and Outputs

- **Mask input:** $M \in \mathbb{R}^{K \times H \times W}$, the K per-class logits from the *Mask Decoder*.
- **Frame feature:** $F \in \mathbb{R}^{C \times H' \times W'}$, the corresponding encoder output.
- **Memory feature output:** $\widehat{M} \in \mathbb{R}^{C \times H' \times W'}$, ready for insertion into the memory bank.

Convolutional Pyramid

We apply E-1 convolutional stages (E-1 being equal to the number of *Encoder Blocks* performing downsampling within the *Frame Encoder*). At stage i with input channels D_{i-1} and spatial size (H_{i-1}, W_{i-1}), each stage performs:

1. A 3×3 convolution (no bias) mapping $D_{i-1} \to 2D_{i-1}$, padding=1.
2. Batch normalization and ReLU activation.
3. A 2×2 max-pooling (stride=2), reducing spatial size to:

$$(H_i, W_i) = \left(\tfrac{H_{i-1}}{2}, \tfrac{W_{i-1}}{2}\right), \quad D_i = 2D_{i-1}.$$

Starting from $D_0 = K$ and $(H_0, W_0) = (H, W)$, after L stages we reach:

$$D_L = 2^L K, \quad (H_L, W_L) = \left(\tfrac{H}{2^L}, \tfrac{W}{2^L}\right) = (H', W').$$

Final Projection and Fusion

A final 1×1 convolution (no bias) maps $D_L \to C$, preserving (H', W'), yielding

$$M' \in \mathbb{R}^{C \times H' \times W'}.$$

We then fuse with the frame feature via residual addition:

$$\widehat{M} = Feat + M'.$$

This fusion ensures that the memory bank entries carry both appearance-driven cues (from $Feat$) and semantic mask evidence (from M').

Empirically, this residual fusion has an additional stabilizing effect: it prevents the memory representation from diverging too far from the encoder's feature distribution, which could otherwise complicate optimization of the attention mechanism. Instead, the network learns to treat the Memory Encoder as a corrective signal layered on top of the original frame embedding.

3.7 Memory Bank

The *Memory Bank* maintains a fixed-size buffer of past frame embeddings, serving as the temporal repository for information propagated via the *Memory Attention* module.

Unlike recurrent architectures that compress temporal history into a single hidden state, the *Memory Bank* explicitly stores multiple past embeddings. This design choice makes the temporal horizon of the model transparent and configurable: by

adjusting F, one can control how far into the past the attention mechanism is allowed to look.

Buffer Management

Let F be the maximum number of frames to store. Incoming frame features of shape $[\text{batch}, C, H', W']$ are enqueued in temporal order. When the buffer exceeds capacity F, the oldest entry is evicted (FIFO policy), ensuring the bank always contains the most recent F representations.

Retrieval and Padding

At each time step, the bank assembles a tensor:

$$\mathcal{M} \in \mathbb{R}^{\text{batch} \times F \times C \times H' \times W'}.$$

- If F frames are stored, the bank assembles all F frames into M.
- If fewer than F frames are stored, the bank pads by repeating the earliest available frame until length F is reached.
- If the bank is empty (e.g., at video start), it returns an all-zero tensor of the same shape.

By encapsulating temporal context in a simple, efficient buffer with predictable shape behavior, the *Memory Bank* underpins the *Memory Attention*'s ability to leverage a sliding window of past information for improved mask propagation across challenging laparoscopic video frames.

Chapter 4

Datasets

4.1 The CholecSeg8k Dataset

The CholecSeg8k (Hong et al., 2020) [25] dataset uses the endoscopic images from Cholec80 (Twinanda et al., 2016) [3], provided by Research Group CAMMA (Computational Analysis and Modeling of Medical Activities), as the base. Cholec80 contains 80 videos of cholecystectomy surgeries performed by 13 surgeons. Each video in Cholec80 captured the procedure at 25 fps and annotated tools presence and operation phases. The CholecSeg8k dataset was created by selecting a subset of 17 videos provided by Cholec80 and by creating semantic segmentation masks in the frames extracted in the selected videos. It includes semantic segmentation for 10 organic classes (abdominal wall, liver, gastrointestinal tract, fat, connective tissue, blood, cystic duct, gallbladder, hepatic vein and liver ligament) and 2 surgical instruments (L-hook electrocautery and grasper).

From a clinical perspective, cholecystectomy is one of the most common laparoscopic procedures worldwide, which makes Cholec80 and its derived datasets particularly representative of routine surgical practice. The videos cover a wide range of inter-patient variability in terms of anatomy, fat distribution, etc. As a consequence, CholecSeg8k captures not only the appearance of individual organs, but also diverse tool–tissue interactions, occlusions, and abrupt camera motions.

Another important characteristic of CholecSeg8k is its annotation protocol. The semantic masks are provided at pixel level, with each pixel assigned to exactly one of the predefined classes. This dense labeling is both labor-intensive and highly valuable: it enables supervised learning of precise organ boundaries and tool contours, which is essential for downstream tasks such as intraoperative guidance, context-aware assistance, and automatic video analysis. However, it also means that the dataset inherits the visual complexity of real surgeries, including lighting artifacts, smoke, and partial views of structures.

Due to the large discrepancy of the number of pixels for each class, we focused on

the following 8 classes: Abdominal Wall, Liver, Gastrointestinal Tract, Fat, Grasper, L-hook Electrocautery, Gallbladder, and a default class for the Background. This selection reflects a trade-off between anatomical diversity and class frequency. Several structures in the original label set, such as the hepatic vein or liver ligament, are almost non-existant. Directly training on these highly under-represented classes tends to destabilize optimization and leads to unreliable performance metrics. By focusing on the eight most prevalent categories, we emphasize those structures that appear consistently across frames and that are most relevant for typical cholecystectomy workflows.

Even within this reduced set, the degree of class imbalance remains substantial: large organs such as the abdominal wall dominate the field of view in many frames, whereas instruments and smaller anatomical structures cover only limited regions. This imbalance has a direct impact on loss design and evaluation, as naïve optimization strategies may bias the model toward over-predicting large, easy classes while neglecting smaller, clinically critical regions.

In total, this leaves us with 3939 frames to train, validate and test our model.

In our experimental pipeline, these frames are organized into training, validation, and test partitions at the video level to avoid patient-level data leakage. This ensures that frames originating from the same surgery are not split across subsets, which would otherwise artificially inflate performance due to strong temporal correlations. Additionally, we preserve the relative distribution of surgical phases across splits, so that the model is exposed to both early and late procedural stages during training and evaluation.

From a modeling standpoint, CholecSeg8k serves as the primary benchmark for assessing the capabilities of SSTC-Seg in a relatively data-rich regime. The combination of thousands of annotated frames, multiple surgeons, and diverse visual conditions makes it an ideal testbed for evaluating spatial adaptivity, temporal consistency, and robustness to occlusion. The results obtained on this dataset therefore provide a strong indication of how well the architecture can generalize to typical laparoscopic cholecystectomy scenes.

4.2 The Dresden Surgical Anatomy Dataset

The Dresden Surgical Anatomy (Carstens et al., 2023) [26] dataset is composed of 32 series of frames from surgeries performed with a Da Vinci® Xi/X Endoscope. It includes semantic segmentation for 8 abdominal organs (colon, liver, pancreas, small intestine, spleen, stomach, ureter, vesicular glands), the adominal wall, and 2 vascular structures (inferior mesenteric artery and intestinal veins) as seen in laparoscopic views.

Unlike CholecSeg8k, which is derived from cholecystectomy videos, the Dresden Surgical Anatomy dataset focuses on a broader range of abdominal regions and provides high-quality anatomical labels obtained in a controlled experimental setting. This dataset emphasizes detailed organ-level anatomy, including structures that are rarely labeled in routine clinical videos. This granularity makes Dresden particularly valuable for research on anatomy-aware assistance and for studying the limits of generalization when only a small number of annotated frames are available.

Due to the large discrepancy of the number of pixels for each class, we focused on the following 5 classes: Abdominal Wall, Colon, Liver, Ureter, and a default class for the Background.

As with CholecSeg8k, this restriction is primarily driven by class imbalance and practical considerations. Several anatomical structures in the original label set appear only sporadically or occupy very few pixels, which makes it difficult to obtain stable and meaningful performance estimates. Concentrating on a subset of more frequently occurring classes allows us to obtain a more reliable assessment of the model's behavior in a low-data scenario, while still covering both large organs (e.g., liver, abdominal wall) and smaller structures (e.g., ureter).

Moreover, the selected classes span different visual characteristics: some present smooth, relatively uniform textures, while others exhibit complex patterns or are partially embedded within surrounding tissue. This diversity provides a challenging test for SSTC-Seg's deformable convolutions and memory-based mechanisms, which must be able to distinguish subtle appearance differences even when the overall image context is limited.

This dataset is significantly smaller than the CholecSeg8k dataset, and that leaves us with 658 frames to train, validate, and test our model. The different classes are also less represented in the frames. This allows us to test the learning of our model when faced with limited data.

As with the CholecSeg8k dataset, the frames are organized into training, validation and test partitions at the video level to avoid patient-level data leakage.

From an evaluation standpoint, Dresden complements CholecSeg8k by highlighting a different failure mode: whereas the latter stresses robustness to variable surgical conditions in a comparatively abundant data regime, the former probes the model's ability to generalize when only a small number of annotated examples are available. Jointly, these two datasets provide a more complete picture of the strengths and limitations of SSTC-Seg in realistic clinical scenarios.

Chapter 5

Experiments and Comparisons with State-Of-The-Art Networks

5.1 Experimental SetUp

5.1.1 Evaluation Metrics

We evaluate segmentation quality using the following metrics:

Accuracy Overall pixel-wise correctness:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

where TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives.

While Accuracy provides an intuitive notion of overall correctness, it tends to be dominated by the background and large organs, which occupy most pixels in typical laparoscopic scenes. As a result, models can achieve deceptively high global Accuracy even when their predictions on small but clinically important structures (such as instruments or narrow ducts) are poor. For this reason, Accuracy is interpreted here primarily as a coarse, sanity-check metric rather than the sole indicator of segmentation performance.

Dice Coefficient Overlap between prediction P and ground truth G:

$$\text{Dice}(P, G) = \frac{2\,|P \cap G|}{|P| + |G|} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}.$$

The Dice Coefficient is more sensitive to disagreements on smaller structures, because false positives and false negatives are normalized by the sum of the predicted and ground-truth volumes. In the context of laparoscopic surgery, this makes

Dice Coefficient particularly informative for evaluating organ boundaries and thin regions around instruments, where relatively few misclassified pixels can lead to substantial clinical impact. High Dice scores therefore indicate not only good volumetric coverage but also a reasonable balance between over- and under-segmentation.

Jaccard Index (IoU) Intersection over union:

$$\mathrm{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

The Jaccard Index (IoU) complements Dice by penalizing disagreements more strictly, since the intersection is normalized by the union of prediction and ground truth. In practice, IoU tends to produce lower absolute values than Dice, but it is often considered a more conservative measure of overlap. Reporting both Dice and IoU provides a more nuanced view of model performance, especially in cases where segmentation outputs contain small spurious regions or slightly misaligned masks.

Hausdorff Distance 95 ($\mathbf{HD_{95}}$) Boundary agreement measured by the 95th percentile of bidirectional distances:

$$\mathrm{HD}_{95}(P, G) = \max\Big\{ \sup_{p \in \partial P} d_{95}(p, \partial G),\ \sup_{g \in \partial G} d_{95}(g, \partial P) \Big\},$$

where $d_{95}(p, \partial G)$ is the 95th-percentile of $\{\min_{g \in \partial G} \|p - g\|\}$.

In contrast to overlap-based scores, HD_{95} focuses on the geometry of the predicted boundaries. By using the 95th percentile of distances rather than the maximum, it remains robust to isolated outliers while still capturing large discrepancies between predicted and true contours. This is particularly relevant in surgical applications, where small localized errors may be acceptable, but large boundary deviations near critical structures could be problematic. HD_{95} is thus used here as a proxy for anatomical plausibility of the predicted shapes.

These metrics are computed separately for each semantic class and then analyzed both at class level and in aggregate. This per-class perspective is particularly important in our setting because the datasets exhibit strong class imbalance and heterogeneous difficulty: large organs, small vessels, and thin surgical instruments do not pose the same challenges and should therefore not be reduced to a single summary statistic. By reporting multiple complementary metrics, we aim to capture both volumetric overlap and boundary quality, which are critical for assessing the clinical usefulness of segmentation models.

5.1.2 Implementation Details

All experiments were implemented in Python 3.12.3 using PyTorch 2.5.1 and CUDA 12.1.

Training and evaluation were conducted on a single NVIDIA GeForce RTX 4060 Laptop GPU (8 GB VRAM) paired with an Intel Ultra 9 processor (14 cores) and 24 GB RAM.

Data

For the CholecSeg8k [25] dataset, we used a 70-15-15 split, and for the Dresden Surgical Anatomy [26] dataset, we used a 60-20-20 split (so that each class could be represented at least one time during the validation and testing phases).

Splits are performed at the video or sequence level to avoid any leakage of temporally adjacent frames between training, validation, and test sets. This is important because consecutive laparoscopic frames are highly correlated; mixing them across splits would artificially inflate reported performance and fail to reflect true generalization to unseen surgeries. The chosen proportions also represent a compromise between providing enough data for training and preserving sufficient examples for unbiased evaluation, especially in the case of the smaller Dresden dataset.

In addition, we ensure that each class of interest appears at least once in the validation and test partitions. This constraint guarantees that the reported per-class metrics are meaningful and that rare classes are not inadvertently excluded from evaluation due to the limited dataset size.

Model Configuration

The architecture of SSTC-Seg was composed of E=3 *Encoder Blocks*, A=3 *Attention Blocks*, E-1=2 *Decoder Blocks*, and F=4 frames in the *Memory Bank*.

The base channel dimension was set to 64, doubling after each encoder stage.

This configuration results in a moderately deep architecture with sufficient capacity to model complex spatial and temporal patterns, while keeping inference time within a practical range. Increasing the number of encoder or attention blocks would likely improve performance further but at the cost of higher memory usage and slower inference, which is undesirable for real-time or near-real-time clinical applications. The choice $F = 4$ for the Memory Bank reflects a similar trade-off: it allows the model to access short-term temporal context without incurring excessive memory overhead for long sequences.

Training Protocol

The model was trained end-to-end with AdamW optimizer (learning rate = 1e-4) for 60 epochs, with an early stopping after 6 epochs without improvement, with a batch size of 2.

A scheduler (torch.optim.lr_scheduler.OneCycleLR) was used to improve convergence stability.

The OneCycleLR policy gradually increases the learning rate at the beginning of training and then decreases it towards the end, which has been shown to accelerate convergence and help escape shallow local minima. Combined with AdamW, which decouples weight decay from gradient updates, this schedule yields a stable optimization process even in the presence of complex components such as deformable convolutions and attention layers. Early stopping based on validation performance helps prevent overfitting, particularly on the Dresden dataset, where the number of training samples is limited.

Loss computation combined cross-entropy and Dice losses to balance class imbalance and boundary accuracy. Cross-entropy primarily drives correct per-pixel classification, encouraging the network to predict the correct label distribution across all classes. Dice loss, on the other hand, directly optimizes for volumetric overlap and reduces the impact of class imbalance by normalizing with respect to the combined size of prediction and ground truth. The combination of both terms is particularly effective in settings where some structures occupy only a small fraction of the image but are nonetheless clinically significant, such as instruments or thin tubular anatomy.

Evaluation

Metrics were computed per-class over all frames in the test split.

Inference speed was computed by passing 2000 frames and measuring the total inference time on the same hardware configuration. To obtain a robust estimate of runtime, inference is performed in batch mode with gradients disabled and all models evaluated under identical conditions. The chosen sequence length of 2000 frames approximates a realistic laparoscopic video segment and smooths out short-term fluctuations due to caching or background processes on the host system. This measurement provides a practical indication of whether the architecture could be integrated into a real-time surgical pipeline, rather than representing an isolated microbenchmark.

5.1.3 Baseline Networks and Ablations

As medical datasets can be costly and are not the easiest to access in large quantities, SSTC-Seg was designed to be non-promptable, so that it could be efficiently applied to as many datasets as possible. So, the following non-promptable baseline models were evaluated:

- U-Net [5];
- PSPNet [6];
- CFPNet-M [27];
- MFCPNet [7].

These baselines were selected to cover a representative spectrum of small and quick networks, but also modern and bigger segmentation architectures. U-Net serves as a strong classical baseline and remains widely used in medical imaging due to its simplicity and effectiveness. PSPNet introduces multi-scale context aggregation through pyramid pooling, providing a more advanced treatment of global spatial information. CFPNet-M and MFCPNet, finally, are specifically tailored to medical segmentation and emphasize lightweight design and efficient multi-scale feature extraction, making them particularly competitive in constrained environments.

By comparing SSTC-Seg against this diverse set of baselines, we can disentangle the contributions of different architectural ideas: simple encoder–decoder designs, pyramid pooling, specialized lightweight modules, and our own combination of deformable convolutions, memory attention, and HD-CRF refinement.

In addition to that, multiple versions of SSTC-Seg have been tested by doing an ablation study to evaluate the effect of different parts of the architecture:

- Version 1: All convolutions are standard, and 1 global head performs the final step of the segmentation for all classes;
- Version 2: All convolutions are deformable, and 1 global head performs the final step of the segmentation for all classes;
- Version 3: All convolutions are deformable, and 1 head per class performs the final step of the segmentation.

The three SSTC-Seg variants are designed to isolate the effect of two key design choices: deformable versus standard convolutions, and global versus per-class segmentation heads. Version 1 can be interpreted as a baseline architecture that removes most of the proposed geometric adaptivity and class-specific specialization. Version 2 reintroduces deformable convolutions, allowing us to quantify their impact on boundary quality and robustness to deformation. Version 3, finally, adds separate heads for each class, which increases flexibility in modeling class-specific appearance

and shape distributions.

This ablation strategy makes it possible to attribute observed performance differences to concrete architectural changes rather than confounding factors such as training protocol or dataset splits. In particular, it allows us to determine whether the added complexity of deformable convolutions and per-class heads is justified by measurable gains in segmentation quality.

5.2 Quantitative Comparison with SOTA

In this section, we report quantitative results on both datasets and compare SSTC-Seg against the aforementioned baselines. We first analyze performance on the CholecSeg8k dataset, which provides a relatively large number of annotated frames, and then on the Dresden Surgical Anatomy dataset, which is considerably smaller and more imbalanced. Finally, we examine runtime characteristics in order to assess the feasibility of deploying SSTC-Seg in real-time settings.

5.2.1 The CholecSeg8k Dataset

Table 5.1 summarizes the segmentation results of U-Net, PSPNet, CFPNet-M, MFCP-Net, and the three versions of SSTC-Seg on the CholecSeg8k dataset. It can be found that our proposed network achieves the highest scores across all metrics for almost every classes (e.g. the Liver, the Grasper, the L-hook Electrocautery and the Gallbladder classes). Only MFCPNet achieves a slightly better accuracy than our model on the Abdominal Wall and the Gastrointestinal Tract classes, but even then, SSTC-Seg is relatively close and still has a better Dice coefficient, Jaccard index and Hausdorff distance 95. SSTC-Seg clearly outperforms all baseline models overall.

A closer inspection of the per-class scores reveals several interesting trends. First, SSTC-Seg (Version 3) consistently improves the Dice Coefficient and IoU for both large organs (e.g., liver, abdominal wall) and small, elongated structures such as the grasper and the L-hook electrocautery. This suggests that the combination of deformable convolutions and per-class heads is effective at capturing both global organ morphology and fine instrument details. In particular, the substantial reduction in HD_{95} indicates that boundary predictions are not only more accurate on average but also less prone to large local deviations.

Second, the Gastrointestinal Tract class remains challenging for all networks, with relatively low scores across the board. Nevertheless, SSTC-Seg achieves competitive or superior overlap metrics while significantly reducing HD_{95} compared to baseline methods, which implies that its errors are more localized and less detrimental from

	U-Net	**PSPNet**	**CFPNet-M**	**MFCPNet**	**SSTC-Seg**		
Version	–	–	–	–	Version 1	Version 2	Version 3
Convolutions	–	–	–	–	Standard	Deformable	Deformable
Segmentation Head(s)	–	–	–	–	Global	Global	Per-class
Abdominal Wall							
Accuracy	0,7081	0,7056	0,7517	**0,8968**	0,7993	0,8261	0,8596
Dice Coefficient	0,7088	0,6933	0,7873	0,8212	0,7554	0,8079	**0,8671**
Jaccard Index	0,5889	0,5618	0,6915	0,7383	0,6471	0,7192	**0,7987**
HD95	14,1258	18,4099	15,6682	14,9437	19,9074	15,7257	**10,6950**
Liver							
Accuracy	0,6563	0,6001	0,4804	0,4888	0,4291	0,6360	**0,8052**
Dice Coefficient	0,6096	0,4695	0,5414	0,5675	0,5101	0,6922	**0,7858**
Jaccard Index	0,4468	0,3129	0,3935	0,4274	0,3775	0,5475	**0,6557**
HD95	13,0291	16,2881	12,6779	11,6166	11,3602	14,2223	**7,4773**
Gastrointestinal Tract							
Accuracy	0,1424	0,0007	0,0645	**0,1903**	0,1784	0,1010	0,1562
Dice Coefficient	0,1471	0,0007	0,1010	0,1692	0,1866	0,1356	**0,2021**
Jaccard Index	0,1209	0,0004	0,0632	0,1299	0,1433	0,0974	**0,1512**
HD95	22,6300	39,5086	29,6986	29,8071	24,7871	20,6079	**16,9894**
Fat							
Accuracy	0,8564	0,7646	0,9296	0,9370	**0,9720**	0,9358	0,9468
Dice Coefficient	0,7967	0,6487	0,7788	0,8220	0,7755	0,8483	**0,8501**
Jaccard Index	0,6817	0,5032	0,6645	0,7194	0,6783	**0,7588**	0,7586
HD95	12,5070	20,0988	13,9888	14,8478	14,6765	14,8449	**11,7485**
Grasper							
Accuracy	0,0223	0,0025	0,5072	0,3676	0,2219	0,4806	**0,5628**
Dice Coefficient	0,0246	0,0018	0,3235	0,3576	0,2168	0,4172	**0,4762**
Jaccard Index	0,0139	0,0010	0,2105	0,2451	0,1338	0,2898	**0,3322**
HD95	32,4927	44,3287	23,7673	14,7716	18,4077	16,5185	**14,6852**
L-Hook Electrocautery							
Accuracy	0,1094	0,0860	0,2523	0,1744	0,1611	0,3521	**0,3822**
Dice Coefficient	0,1511	0,7985	0,2517	0,1976	0,1589	0,4169	**0,4778**
Jaccard Index	0,0943	0,6619	0,1813	0,1522	0,0997	0,3067	**0,3558**
HD95	23,8852	31,7227	15,1118	**13,7692**	26,1228	18,9219	17,1693
Gallbladder							
Accuracy	0,6611	0,0731	0,6439	0,3633	0,2383	0,6600	**0,6706**
Dice Coefficient	0,6353	0,0631	0,5827	0,4243	0,2792	0,6400	**0,7192**
Jaccard Index	0,4864	0,0403	0,4565	0,3242	0,2091	0,5082	**0,5907**
HD95	16,1016	17,4743	12,8882	13,8331	17,4808	15,3215	**8,0912**
Background							
Accuracy	0,9586	0,9167	0,9866	0,9877	0,9879	0,9886	0,9872
Dice Coefficient	0,9617	0,8632	0,9679	0,9859	0,9757	0,9824	0,9796
Jaccard Index	0,9266	0,7598	0,9394	0,9725	0,9533	0,9660	0,9612
HD95	3,7021	7,6624	5,0655	3,5126	4,4043	4,1649	4,6885

Table 5.1: Comparison between SSTC-Seg and the baseline models across 8 classes from the CholecSeg8k dataset, with the metrics for each class.

	U-Net	**PSPNet**	**CFPNet-M**	**MFCPNet**	**SSTC-Seg**		
Version	–	–	–	–	Version 1	Version 2	Version 3
Convolutions	–	–	–	–	Standard	Deformable	Deformable
Segmentations Head(s)	–	–	–	–	Global	Global	Per-class
Abdominal Wall							
Accuracy	0,3600	0,0715	0,3856	0,5243	0,4779	0,5227	**0,5540**
Dice Coefficient	0,4035	0,0894	0,3964	0,5416	0,5194	0,5365	**0,5827**
Jaccard Index	0,3161	0,0580	0,2831	0,4521	0,4208	0,4335	**0,4694**
HD95	24,0179	37,3716	40,6202	18,4865	19,6340	18,9457	**18,0061**
Colon							
Accuracy	**0,0811**	0,0012	0,0401	0,0628	0,0254	0,0250	0,0532
Dice Coefficient	**0,1059**	0,0019	0,0286	0,0856	0,0407	0,0407	0,0790
Jaccard Index	**0,0720**	0,0009	0,0160	0,0576	0,0246	0,0240	0,0512
HD95	**28,9731**	42,3071	46,5240	33,3157	29,5577	33,2963	30,7333
Liver							
Accuracy	**0,0188**	0,0003	0,0040	0,0118	0,0004	0,0059	0,0087
Dice Coefficient	**0,0244**	0,0006	0,0043	0,0190	0,0008	0,0105	0,0137
Jaccard Index	**0,0171**	0,0002	0,0022	0,0113	0,0004	0,0059	0,0077
HD95	**23,1196**	39,1452	47,1432	20,9276	33,5343	24,4901	34,8939
Ureter							
Accuracy	0,1399	0,0009	0,2320	0,2263	0,2365	0,2836	**0,2866**
Dice Coefficient	0,1787	0,0003	0,1995	0,2581	0,2654	0,3042	**0,3111**
Jaccard Index	0,1209	0,0002	0,1221	0,1819	0,1929	0,2270	**0,2320**
HD95	**23,6352**	24,2273	39,4520	24,1516	28,1751	26,3886	25,8549
Background							
Accuracy	0,9781	0,9919	0,6236	0,9716	0,9633	0,9721	0,9681
Dice Coefficient	0,9238	0,8907	0,7504	0,9406	0,9391	0,9402	0,9443
Jaccard Index	0,8676	0,8177	0,6070	0,8917	0,8897	0,8919	0,8981
HD95	10,9724	23,5215	14,9926	11,1236	10,8008	10,8472	9,9622

Table 5.2: Comparison between SSTC-Seg and the baseline models across 5 classes from the Dresden Surgical Anatomy dataset, with the metrics for each class.

	U-Net	**PSPNet**	**CFPNet-M**	**MFCPNet**	**SSTC-Seg**		
Version	–	–	–	–	Version 1	Version 2	Version 3
Convolutions	–	–	–	–	Standard	Deformable	Deformable
Segmentations Head(s)	–	–	–	–	Global	Global	Per-class
Frames per second	132,46	114,13	66,84	59,75	34,25	25,13	24,74

Table 5.3: Comparison between SSTC-Seg and the baseline models in terms of frames processed per second.

an anatomical perspective.

Finally, the background class exhibits high Dice and IoU for all models, as expected given its dominance in pixel count. It is therefore less informative for discriminating between architectures, although SSTC-Seg still maintains competitive performance without sacrificing its improvements on foreground classes.

5.2.2 The Dresden Surgical Anatomy Dataset

Table 5.2 summarizes the segmentation results of U-Net, PSPNet, CFPNet-M, MFCP-Net, and the three versions of SSTC-Seg on the Dresden Surgical Anatomy dataset. Here, on a purposefully small dataset (small number of frames, and low representation of the different classes) the results are more mitigated. Our network still outperforms all baseline models on half of the classes (the Abdominal Wall and the Ureter classes), but it is U-Net, a small and very simple model, that achieves the best metrics on the Colon and the Liver classes. Even MFCPNet performs slightly better. These 2 classes are the less represented through the dataset, and it shows that a network with the size and complexity of SSTC-Seg has difficulties segmenting classes with too few data, compared with simpler models.

These results highlight an important aspect of model selection in medical imaging: architectures that perform best on larger datasets are not necessarily optimal in extremely low-data regimes. In the Dresden Surgical Anatomy dataset, the combination of limited training samples and strong class imbalance favors smaller, less expressive networks such as U-Net, which may be less prone to overfitting and easier to optimize. SSTC-Seg still provides clear benefits for some classes—most notably the abdominal wall and ureter—but struggles to match U-Net on rarely observed structures like the colon and liver.

5.2.3 Overall

Across both datasets, SSTC-Seg (Version 3) achieves the best average Accuracy, improving up to +13-15% over the second best network, but also the best Dice Coefficient and the best Jaccard Index. The largest relative gains appear in both organ and tool classes (e.g. Liver and L-Hook Electrocautery), which confirms that the characteristics of SSTC-Seg, like the deformable convolutions, make it adaptable to both organic and more rigid shapes. The consistently lower Hausdorff Distance 95 values further indicate smoother and more anatomically consistent boundaries, thanks to modules like the memory attention mechanism and the *HD-CRF*, which is critical for surgical applications.

It is also instructive to relate these gains back to the architectural ablations.

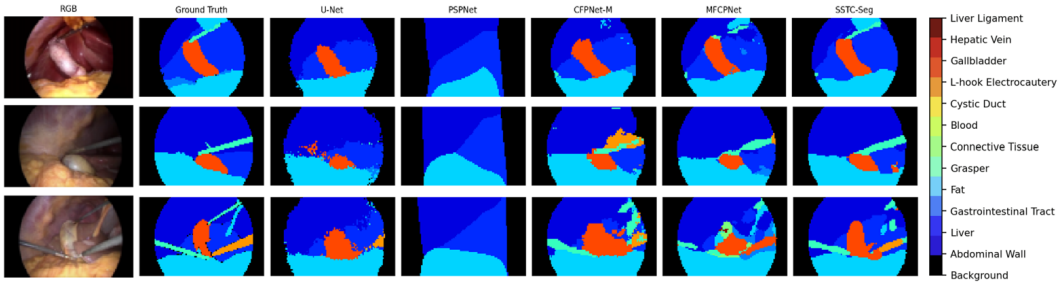

Figure 5.1: Visual comparison on the CholecSeg8k dataset of the masks generated by all baseline networks and SSTC-Seg (version 3) compared to the associated original RGB frame and the ground truth masks. The input frames' resolution has been reduced so that all models could have been run in parallel.

The progression from Version 1 to Version 2 and Version 3 shows that deformable convolutions provide a consistent boost in overlap metrics and boundary quality, while the introduction of per-class heads further enhances performance on classes with highly distinct appearances or shapes. This suggests that most of the observed improvements are not accidental side effects of training, but rather stem directly from the intended design choices of SSTC-Seg.

5.2.4 Inference Speed

Table 5.3 shows the number of frames that each network can process per second in inference mode. Smaller and simpler networks like U-Net and PSPNet can go over the 100 fps mark. Networks like CFPNet-M and MFCPNet, which are specialized but still relatively simple compared to our network, achieve an inference speed around 60 fps (which is the minimum we should aim for in real time implementations). SSTC-Seg, depending on the version, ranges from 24 to 34 fps. This computational overhead is caused by the additional complexity of its architecture compared to the baseline models (deformable convolutions, memory-based attention mechanism, *HD-CRF*).

From a practical standpoint, these frame rates mean that SSTC-Seg, in its current form, does not yet satisfy strict real-time requirements for systems operating at 60 fps or higher. However, many clinical support applications can tolerate modest latency, especially if predictions are used for guidance (e.g. training of young surgeons) or documentation rather than for hard real-time control of robotic instruments.

5.3 Qualitative Results

In Figure 5.1, we can see next to some RGB frames from the CholecSeg8k dataset some visual comparisons between their associated ground truth and the masks generated by U-Net, PSPNet, CFPNet-M, MFCPNet and SSTC-Seg (version 3, with deformable convolutions and per-class segmentation heads). The resolution of the original RGB

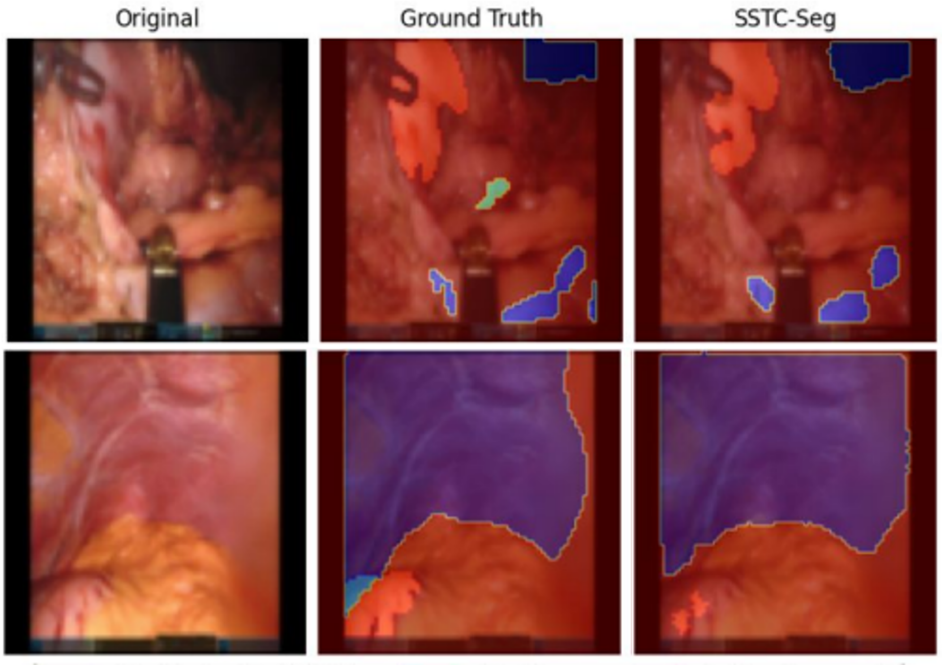

Figure 5.2: Visual results on the Dresden Surgical Anatomy dataset.

frames has been lowered in input so that all networks could have been run in parallel over the testing split. We can clearly see that the combination of the memory attention mechanism and the *HD-CRF* reduces drastically the number of residual artifacts on SSTC-Seg predictions, which explains why our network consistently has a lower Hausdorff Distance 95, as well as a better Dice Coefficient and a better Jaccard Index.

In particular, qualitative inspection shows that SSTC-Seg tends to preserve the continuity of organ boundaries even when instruments partially occlude the field of view. Competing models often produce fragmented masks or leave small gaps along the contours, especially near specular highlights or blurred regions. The memory attention mechanism appears to mitigate these issues by leveraging information from previous frames where the boundaries are more clearly visible, while the HD-CRF module removes small isolated regions and enforces smoother transitions between adjacent organs.

In Figure 5.2, we focus exclusively on SSTC-Seg's prediction masks compared to RGB frames from the Dresden Surgical Anatomy dataset and their associated ground truth. Again, we can note the absence of artifacts, and the smoothness of organic boundaries, which confirm the quantitative results.

Visual inspection (both Figures 5.1 and 5.2) corroborates the quantitative findings: SSTC-Seg yields smoother, more continuous contours and preserves organ topology under a laparoscopic environment.

Overall, these qualitative observations underscore an important point: even when

numerical performance differences between models appear moderate, their visual impact on segmentation maps can be substantial. In surgical contexts, smoother, topologically consistent masks may be more valuable than marginal gains in global overlap metrics, because they facilitate interpretation by clinicians and reduce the risk of misleading visual artifacts. SSTC-Seg's ability to produce such masks is therefore a key strength of the proposed architecture.

Chapter 6

Discussion and Limitations

The experimental results on the CholecSeg8k [25] dataset, with a good amount of data, demonstrate that our model consistently outperforms baseline methods across most classes, whether they are organs or tools. For most classes, SSTC-Seg has the best metrics across the board, and in the rare cases where it does not, it is closely behind. The combination of deformable convolutions, memory attention and *HD-CRF* clearly improves segmentation, but it also improves its stability under deformation, occlusion and motion, even within less represented classes. The ablation study shows that, on the one hand, deformable convolutions drastically improve the segmentation of organic shapes, but also greatly help differentiating them from the more rigid shapes of tools. On the other hand, the per-class segmentation heads allows for a better focus on each class.

Taken together, the experiments on CholecSeg8k indicate that SSTC-Seg is able to scale effectively when sufficient annotated data are available. In this regime, the increased complexity of the architecture is an asset rather than a liability: the model can learn rich spatio-temporal representations that generalize across different surgeons, lighting conditions, and patient anatomies. However, this observation naturally raises the question of how the same architecture behaves when data are scarce, which is a frequent situation in medical imaging due to the high cost of annotation and limited access to curated datasets.

The experimental results on the Dresden Surgical Anatomy [26] dataset, with much fewer data, demonstrate that the size and complexity of SSTC-Seg limit its learning on very small datasets, compared to some smaller and simpler models. Still, it outperforms the baseline methods on half of the classes, while being at least on the average on the other half. This versatility allows SSTC-Seg to be a viable option in a wide range of situations, from a lower amount of data and a smaller number of classes, to a larger amount of data and bigger number of classes.

However, even if the multi-organ segmentation is improved with SSTC-Seg, such models are designed to be used in real time during surgical procedures. The de-

formable convolutions, the memory-based attention mechanism and the multiple iterations of the *HD-CRF*, while beneficial for accuracy, introduce additional computational overhead that may hinder real-time inference. In the current state, SSTC-Seg does not reach the 60 frames per second mark, and it would certainly need to be re-written in a low-level programming language like CUDA to be efficiently used in real time in a surgical environment, while its current inference speed could be sufficient for training guidance.

Chapter 7

Conclusion

In this work, we proposed SSTC-Seg, a novel deformable memory-based multi-scale network to address the challenges of multi-organ semantic segmentation in laparoscopic surgery videos. The model integrates deformable convolutions, temporal memory attention, and a *Hierarchical Dense CRF (HD-CRF)* into a unified, end-to-end framework that explicitly captures spatial adaptivity and temporal consistency. This architectural synergy enables SSTC-Seg to robustly handle challenges inherent to minimally invasive laparoscopic surgery, such as large non-rigid organ deformations, occlusions, and abrupt camera motion.

Beyond addressing the immediate segmentation task, this work contributes a more general perspective on how spatio-temporal reasoning can be structured in minimally invasive surgery environments. The explicit combination of deformable sampling, temporal memory, and probabilistic refinement demonstrates that no single mechanism is sufficient when applied in isolation. Instead, it is the interplay between these modules that yields the observed robustness across a wide range of surgical conditions. Deformable convolutions make the network highly responsive to local geometric variations, whereas memory attention stabilizes predictions over time and enforces continuity across frames. The HD-CRF then acts as a final structural constraint, sharpening boundaries and correcting local inconsistencies in ways that are difficult to achieve through convolution alone.

This synergy represents an important conceptual direction for future research. Laparoscopic scenes are inherently dynamic, information-rich, and frequently ambiguous, and they rarely conform to the assumptions made by traditional image-centric models. By adopting a hybrid spatio-temporal architecture, SSTC-Seg moves closer to how clinicians themselves interpret intraoperative images—by integrating current visual cues with temporal context and anatomical knowledge.

Experiments conducted on the CholecSeg8k [25] and Dresden Surgical Anatomy [26] datasets demonstrate that SSTC-Seg achieves state-of-the-art performance across multiple quantitative metrics, including Accuracy, Dice Coefficient, Jaccard Index,

and Hausdorff Distance 95. On CholecSeg8k, it consistently outperforms every baseline models, validating the importance of deformable feature extraction and temporal reasoning. On the smaller Dresden Surgical Anatomy dataset, SSTC-Seg remains competitive despite limited data, showing its adaptability to different scales and data conditions.

Overall, SSTC-Seg advances the state-of-the-art in video-based surgical scene understanding by providing a spatio-temporal coherent and anatomically adaptive segmentation framework. This work represents a step toward reliable, computer-assisted surgical systems, bridging the gap between academic research and practical clinical deployment.

Bibliography

[1] Kevin Cleary and Terry M Peters. "Image-guided interventions: technology review and clinical applications". In: *Annual Review of Biomedical Engineering* 12 (2010), pp. 119–142. DOI: 10.1146/annurev-bioeng-070909-105249 (cit. on p. 10).

[2] Anup Sood et al. "Minimally invasive surgery and its impact on 30-day post-operative complications, unplanned readmissions and mortality". In: *British Journal of Surgery* 104.10 (2017), pp. 1372–1381. DOI: 10.1002/bjs.10561 (cit. on p. 10).

[3] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. "Endonet: a deep architecture for recognition tasks on laparoscopic videos". In: *IEEE transactions on medical imaging* 36.1 (2016), pp. 86–97 (cit. on pp. 10, 37).

[4] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. "Learning video object segmentation from static images". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2663–2672 (cit. on p. 11).

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597 (cit. on pp. 11, 15, 45).

[6] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. *Pyramid Scene Parsing Network*. 2017. arXiv: 1612.01105 [cs.CV]. URL: https://arxiv.org/abs/1612.01105 (cit. on pp. 11, 45).

[7] Linlin Hou, Zishen Yan, Christian Desrosiers, and Hui Liu. "MFCPNet: Real time medical image segmentation network via multi-scale feature fusion and channel pruning". In: *Biomedical Signal Processing and Control* 100 (2025), p. 107074 (cit. on pp. 11, 15, 24, 45).

[8] Nikhila Ravi et al. *SAM 2: Segment Anything in Images and Videos*. 2024. arXiv: 2408.00714 [cs.CV]. URL: https://arxiv.org/abs/2408.00714 (cit. on pp. 11, 18).

[9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017) (cit. on p. 15).

[10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9 (cit. on p. 16).

[11] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. "Efficient content-based sparse attention with routing transformers". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 53–68 (cit. on p. 16).

[12] Ilya O Tolstikhin et al. "Mlp-mixer: An all-mlp architecture for vision". In: *Advances in neural information processing systems* 34 (2021), pp. 24261–24272 (cit. on p. 16).

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. "Deformable convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 764–773 (cit. on p. 17).

[14] Mo Zhang, Xiang Li, Mengjia Xu, and Quanzheng Li. "RBC semantic segmentation for sickle cell disease based on deformable U-Net". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2018, pp. 695–702 (cit. on p. 18).

[15] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. "Deformable convnets v2: More deformable, better results". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9308–9316 (cit. on p. 18).

[16] Yuwen Xiong et al. "Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 5652–5661 (cit. on p. 18).

[17] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV]. URL: https://arxiv.org/abs/2304.02643 (cit. on p. 18).

[18] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. "Video object segmentation using space-time memory networks". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9226–9235 (cit. on p. 19).

[19] Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. *Medical SAM 2: Segment medical images as video via Segment Anything Model 2*. 2024. arXiv: 2408.00874 [cs.CV]. URL: https://arxiv.org/abs/2408.00874 (cit. on p. 19).

[20] Philipp Krähenbühl and Vladlen Koltun. *Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials*. 2012. arXiv: 1210.5644 [cs.CV]. URL: https://arxiv.org/abs/1210.5644 (cit. on p. 20).

[21] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images". In: *NeuroImage* 170 (2018), pp. 446–455 (cit. on p. 20).

[22] Yuqin Li, Xiao Dong, Weili Shi, Yu Miao, Huamin Yang, and Zhengang Jiang. "Lung fields segmentation in chest radiographs using dense-u-net and fully connected crf". In: *Twelfth International Conference on Graphics and Image Processing (ICGIP 2020)*. Vol. 11720. SPIE. 2021, pp. 297–304 (cit. on p. 20).

[23] L'ubor Ladickỳ, Chris Russell, Pushmeet Kohli, and Philip HS Torr. "Associative hierarchical random fields". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.6 (2013), pp. 1056–1077 (cit. on p. 20).

[24] Peng Zhang, Ming Li, Yan Wu, and Hejing Li. "Hierarchical conditional random fields model for semisupervised SAR image segmentation". In: *IEEE Transactions on Geoscience and Remote Sensing* 53.9 (2015), pp. 4933–4951 (cit. on p. 20).

[25] 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. arXiv: `2012.12453 [cs.CV]`. URL: `https://arxiv.org/abs/2012.12453` (cit. on pp. 37, 43, 53, 55).

[26] Matthias Carstens, Franziska M Rinner, Sebastian Bodenstedt, Alexander C Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R Kolbinger. "The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science". In: *Scientific Data* 10.1 (2023), pp. 1–8 (cit. on pp. 39, 43, 53, 55).

[27] Ange Lou, Shuyue Guan, and Murray Loew. *CFPNet-M: A Light-Weight Encoder-Decoder Based Network for Multimodal Biomedical Image Real-Time Segmentation*. 2021. arXiv: `2105.04075 [cs.CV]`. URL: `https://arxiv.org/abs/2105.04075` (cit. on p. 45).

Dedications

To everyone who inspired my curiosity, who supported me and who helped me shape this thesis.