AIN SHAMS UNIVERSITY FACULTY OF ENGINEERING

Deep learning Project report

Learning Multi-Class Segmentations From Single-Class Datasets

Name	ID
Michael Hany Anis	2100333
Adham Ahmed Hussein	2100725
Michael Sameh Michel	2100322

Contents

Introduction

Research Paper link

https://openaccess.thecvf.com/content_CVPR_2019/papers/Dmitriev_Learning_Multi-Class_Segmentations_From_Single-Class_Datasets_CVPR_2019_paper.pdf

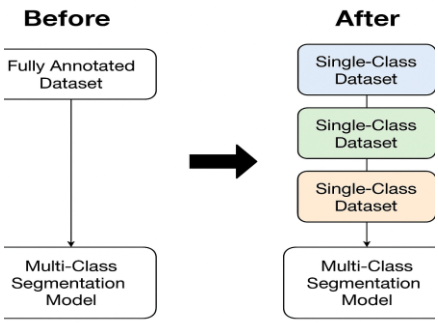
The Challenge of Multi-Class Segmentation

Semantic segmentation—the process of labeling each pixel in an image with a class tag—has experienced significant advancements due to the rise of deep learning. Nonetheless, this advancement has predominantly relied on the presence of completely annotated datasets in which every pixel of each image is labeled for all relevant classes at the same time. Producing these detailed annotations is a highly laborious task, especially in fields such as medical imaging where specialized expertise is necessary.

In reality, researchers frequently have several datasets at their disposal, with each dataset offering annotations for just one class (for instance, one dataset marks only "liver" in CT images while another marks solely "spleen"). Conventional segmentation methods are unable to efficiently utilize these disjointed, single-class datasets for training a cohesive multi-class segmentation model

Class-Conditioned Segmentation

This paper introduces a groundbreaking approach to semantic segmentation that can learn from multiple non-overlapping, single-class datasets simultaneously. The key innovation is a conditioning mechanism that dynamically adapts a single neural network to segment different target classes based on a class-specific signal. Think of it as teaching a network to "switch modes" depending on which class it's looking for. When conditioned on "liver," the network



focuses on liver-specific features; when conditioned on "spleen," the same network parameters reorient to detect spleen characteristics—all without ever seeing images where both classes are labeled together.

How Conditioning Works

The authors improve upon a U-Net-like architecture with dense convolutional blocks, but they primarily include class-specific conditioning signals in the decoder part of the network. The network's behavior is successfully influenced by this conditioning, which allows it to employ the same parameters across several segmentation activities.

The model is trained on batches of pictures taken from different single-class datasets. The network is impacted by the related class for each batch, and it learns to only segment that specific class. At each training iteration, the loss function is computed only for the active class.

During inference, the model is capable of producing multi-class segmentations by sequentially operating under different class conditions and then combining the results, despite the fact that it was never trained on fully labelled multi-class images.

Significance and Applications

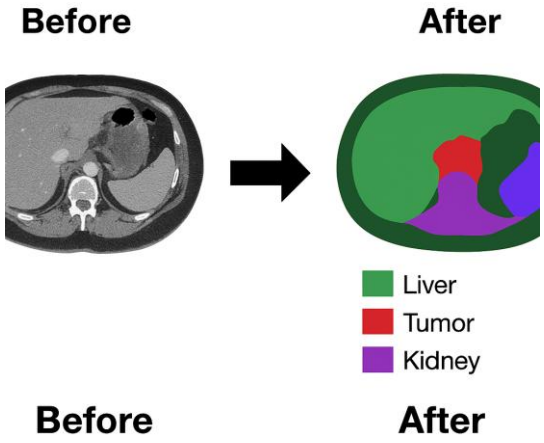
This approach offers several compelling advantages:

- **Resource Efficiency**: Instead of training separate models for each class, a single conditional model requires significantly fewer parameters while achieving comparable or better performance.
- **Data Utilization**: It allows researchers to leverage existing single-class datasets without requiring expensive re-annotation.
- **Flexibility**: New classes can be incorporated without retraining the entire model from scratch.
- **Cross-Domain Applicability**: The method works not only for medical imaging (CT scans of liver, spleen, and pancreas) but also generalizes to natural images like urban scenes in the Cityscapes dataset.

Important concepts

What is a non-overlapping single-class dataset?

It means:
Each training image has annotations for **only one organ/class** (e.g., only the liver or only the kidney).
You don't have images where multiple organs are annotated together (e.g., no image with both liver and tumor labels at once).
And the annotations for different classes don't overlap in the same dataset.



What is the paper doing?

They use a single neural network to learn multi-class segmentation, even though each training sample contains only one class.
Instead of training:

- One model for liver
- Another for kidney
- And another for tumor

They train one shared model, and "tell" it which class to segment via a conditioning signal.

What does "the model shares all its parameters" mean?

This refers to:

A shared backbone (same CNN for all tasks).
All classes (liver, kidney, tumor...) use the same network during training/testing.
The model is conditioned (informed) about which class to look for.

Benefit: The model learns common features (e.g., texture, boundary shapes) and spatial relationships between organs.

What is a spatial connection?

In CT scans (like the one in your image):

- The liver is usually on the **right**, kidneys are **lower**, and the tumor is **inside or near** the liver.
- Even if trained on images where only one structure is labeled at a time, the model can learn these consistent spatial layouts from the data distribution.

That means, when it sees a kidney, it knows where to expect a liver or tumor, even without seeing them during training.

Contribution Summary (from the paper):

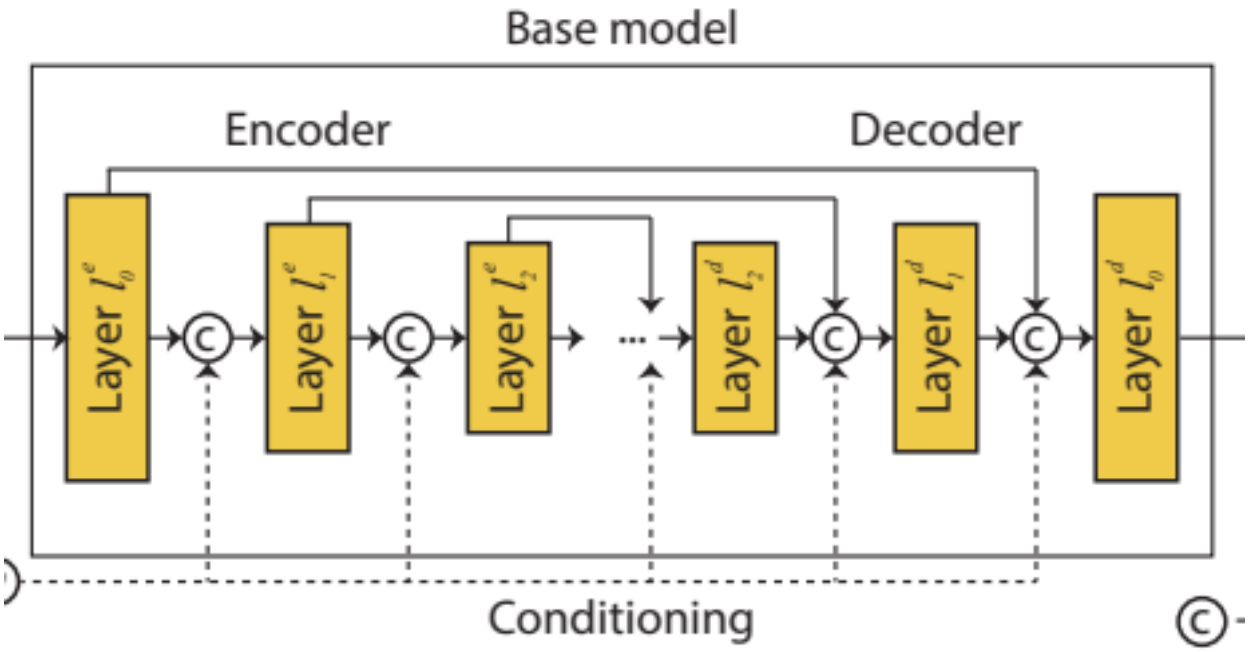
- First to use conditioning in a CNN for segmentation with single-class data.
- Achieves efficient multi-class segmentation with a single model.
- Shows state-of-the-art results on organ segmentation tasks.
- Works well even on natural images (Cityscapes dataset).

Base Model

The base component of the proposed framework is a 3D fully-convolutional U-net-like architecture

Additionally, we adopt 3D densely connected convolutional blocks [17, 19], which effectively utilize the volumetric information available in the CT scans.

The encoder part of the model includes a convolutional layer, followed by six densely connected convolutional blocks, sequentially connected via 2 ⇥ 2 ⇥ 2 maxpooling layers. The number of feature channels in each dense block is proportional to its depth. The decoder part of the model utilizes transposed convolutions with strides as upsampling layers and is topologically symmetric to the encoder. The last convolutional layer ends with a sigmoid function.



3 PARTS:

1. Shared Feature Encoder

- Backbone: A standard Fully Convolutional Network (**FCN**) or U-Net-like encoder **(e.g., ResNet, VGG)**.
- Purpose: Extracts high-level features from input images, shared across all classes.
- Input: Raw image (e.g., **CT** slice or natural image).
- Output: A generic feature map that encodes visual patterns useful for all target classes.

2. Class-Specific Conditioning Module

- **Key Innovation:** Instead of training separate models for each class, the authors use conditioning mechanisms to dynamically adapt the shared features for each class.
- Conditioning Techniques (implied but not explicitly detailed in the paper, likely similar to **FiLM** or **SPADE**):
    - **Feature-wise Linear Modulation (FiLM):** Applies class-specific scaling and shifting to encoder features.
    - **Attention Masks:** Uses class embeddings to modulate feature activations.
- **Input:**
    - Shared features from the encoder.
    - Class Embedding: A learned vector representing the target class (e.g., "liver," "spleen").
- **Output:** Class-specific intermediate features.

3. Multi-Class Fusion Decoder

- **Parallel Class-Specific Heads:** Each head processes conditioned features for one class.
- **Fusion Mechanism:** Combines single-class predictions into a final multi-class segmentation map.
- **Conflict Resolution:** Handles overlapping predictions (e.g., a pixel labeled as both "liver" and "spleen") using learned priors or a softmax aggregation.
- Output: **A unified probability map** for all classes.

Training Paradigm

- **Single-Class Datasets:** Each dataset contains images labeled for only one class (e.g., Dataset A: liver; Dataset B: spleen).
- **Conditioned Training:**
    - For each batch, the model is conditioned on one class (e.g., only liver labels are active).
    - The loss is computed only for the active class (binary segmentation loss).
- **Multi-Class Inference:**
    - At test time, the model runs once per class (conditioned sequentially on each class) and fuses the results.

Dataset Used

Biomedical (CT Scans)

- Liver: 20 volumes from **Sliver07**
- Pancreas: 82 volumes from **NIH Pancreas Dataset**
- Spleen: 74 volumes (private dataset)

  Properties:
  - Diverse scanners/institutions, varying slice thicknesses, pathologies (tumors, splenomegaly).
  - Preprocessed as 256×256×32 subvolumes with normalization.
  - Augmentation: Random rotations, zooms, shifts.

Natural Images (Cityscapes)

Also evaluated on urban scenes to show generalizability.

Training

The input images underwent minimal preprocessing: each dataset was sampled with uniform probability, and subvolumes of dimensions **256⇥256⇥32** were extracted and normalized to generate the input images. Moreover, all training samples have been enhanced with minor random rotations, zooms, and translations. The suggested framework was trained using instances from all employed single-class datasets.

The framework was optimized with the following objective: **L(Y ,Yˆ) = α1β1L1(Y c1 , Yˆ c1)+...+αnβkLk(Y cm, Yˆ cm)**, (4) where Li(Y ci , Yˆ ci) is a loss function for a single-class dataset Di , the hyperparameters αi specify the impact of a particular class ci on the total loss, and **βi = {0, 1**} specifies the presence of the binary mask for class ci in the batch

Inference

what is inference?

Inference is the phase where the trained model generates multi-class segmentations for new images, even though it was trained only on single-class datasets. **Here's how it works step-by-step:**

1. Key Challenge

The model was trained on:

- **Dataset A:** Images with *only liver* labeled.
- **Dataset B:** Images with *only spleen* labeled.

But during inference, you want to segment both liver and spleen in a new image.

2. Inference Process

To get multi-class predictions, the model runs sequentially for each class using *conditioning*:

- **Step 1:** Activate "Liver Mode"
  - Set the conditioning signal to class="liver".
  - Run the model → outputs a liver probability map.
- **Step 2:** Activate "Spleen Mode"
  - Set conditioning to class="spleen".
  - Run the model again → outputs a spleen probability map.
- **Step 3:** Merge Predictions
  - Combine the two maps (e.g., pick the class with highest probability per pixel).
  - Handle overlaps (e.g., liver vs. spleen) using softmax or heuristics.

3. Practical Example

Input CT Scan:

- Run model for "liver" → labels liver regions.
- Run model for "spleen" → labels spleen regions.
- Output: Unified segmentation with both organs.

From paper:

The target segmentation class ci can be explicitly specified throughout the inference period. However, we propose to automate the process of specifying the target segmentation class by iteratively traversing through all the entities in the lookup table in order to make the framework easier to use throughout the inference period. As an alternative, a collection of presets might be specified expressly for the segmentation of abdominal organs, such as the liver and gallbladder, which are frequently examined together by clinicians.

Implementation:

The proposed framework was implemented using Keras library with TensorFlow backend. We trained our network from scratch using Adam optimizer [21] with the initial learning rate or 0.00005, and $\beta 1 = 0.9$, $\beta 2 = 0.999$, with a batch size of 2 for 25K iterations.

Questions:

What is the problem statement of the paper?

“ How can we train a multi-class semantic segmentation model when each training image only has labels for a single class? “
Most existing segmentation methods assumed that:

- **Training data must be fully annotated**, meaning every pixel in an image is labeled.
- **Each dataset must cover all target classes** simultaneously.

This approach works well when:

- All classes of interest appear in the same dataset.
- Annotating all classes is feasible.

But in **real-world scenarios**, especially in fields like **medical imaging** or **autonomous driving**:

- You often have **multiple datasets**, each labeled for only **one class** (e.g., only "tumor", or only "liver").
- Full multi-class labeling is **costly**, **time-consuming**, and **not always possible**..
- **Annotation is expensive**, especially for 3D data like CT or MRI scans.

- From the abstraction of the paper “ This achievement is due primarily to the public availability of large multi-class datasets. However, there are certain domains, such as biomedical images, where obtaining sufficient multi-class annotations is a laborious and often impossible task and only single-class datasets are available. “

What are the objectives of the paper and do you think the authors managed to achieve these goals? Explain

To develop a method that allows effective **multi-class segmentation training** from datasets that provide only **single-class annotations per image**, without requiring full pixel-level labels for every class

-also from the abstraction "we propose a unified highly efficient framework for robust simultaneous learning of multi-class segmentations by combining single-class datasets and utilizing a novel way of conditioning a convolutional network for the purpose of segmentation"

The paper has two main objectives:

- Enable Multi-Class Segmentation Without Multi-Class Labels
  - Train a model to perform multi-class semantic segmentation using only single-class annotated datasets (where each dataset has labels for only one class + background).
  - Avoid the need for a fully labeled multi-class dataset, which is expensive and labor-intensive to collect.
- Develop a Robust Fusion Mechanism for Single-Class Predictions
  - Design a model architecture and loss function that can combine predictions from multiple single-class segmentations into a coherent multi-class output.
  - Handle class conflicts (e.g., when two single-class models predict different labels for the same pixel).

,yes, the authors managed to achieve these goals by using methods that will be clear in the following question :

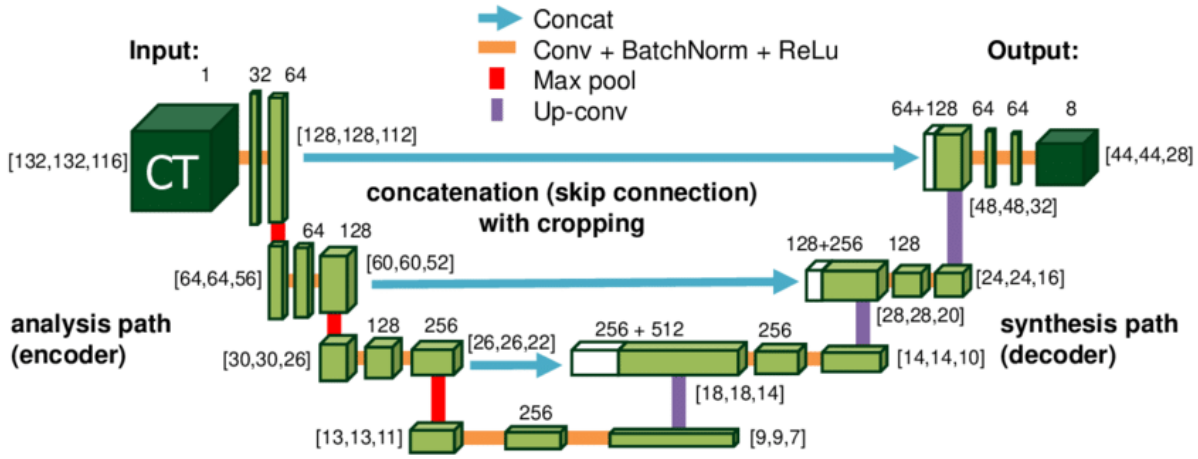
1. **Multi-class from single-class datasets**
2. **Effective conditioning framework**
3. **Outperforms existing methods**
4. **Applicability beyond medical imaging**

What is the DL method used in this paper?

The paper uses a **Convolutional Neural Network (CNN)**–based architecture designed for **semantic segmentation**, specifically tailored to work with **single-class annotations** per image.

Components :
1.U-Net–like Encoder-Decoder Architecture

- The model follows a U-Net-style design, which is commonly used for medical image segmentation.
- It consists of an encoder that captures semantic features and a decoder that reconstructs segmentation masks.



2.Conditional Segmentation via Conditioning Vectors

- The model is conditioned on the target class to be segmented.

3.Loss Function : Dice Similarity Coefficient (DSC)–based loss

- **DSC:**This loss works well with imbalanced segmentation tasks (common in medical imaging).

Summarized by “The DL method is a **conditioned CNN-based segmentation model**, inspired by U-Net, that uses **class-specific conditioning** (especially in the decoder) to perform **multi-class segmentation** from **single-class annotated datasets**.”

What are the other state-of-the-art methods that can be applied to the same problem?

SOTA, which stands for State of the Art, describes the most cutting-edge or sophisticated approaches, strategies, or technologies available in a certain field at any given time. When someone uses the term "state-of-the-art," they are referring to the best or most efficient option that is currently available, taking into account the most recent technological advancements or research.

1. Two-Step Coarse-to-Fine Method (Zhou et al., 2018)

- **Task:** Pancreas segmentation
- **Method:**
  - Uses a **two-step ConvNet** approach
  - Starts with a **coarse segmentation**, then refines in a second step
- **Performance:**
  - **Dice Similarity Coefficient (DSC): 82.4%** on the NIH Pancreas dataset
- **Limitations:**
  - **Specific to pancreas** segmentation
  - Requires **domain-specific adaptations**

2. ConvNet-Based Approach (Yang et al., 2018)

- **Task:** Liver segmentation
- **Method:**
  - **Custom ConvNet architecture** designed for liver segmentation
- **Performance:**
  - **DSC: 95%** on a private dataset of 1,000 CT liver images
- **Limitations:**
  - **Organ-specific** (liver only)
  - May **not generalize** to other organs or datasets

3. Multi-Organ Coarse-to-Fine Method (Roth et al., 2018)

- **Task:** Multi-organ segmentation (liver, spleen, pancreas)
- **Method:**
  - **Two-stage CNN** for **coarse-to-fine segmentation**
  - Handles **multi-class segmentation**

- **Performance:**
  - **Liver:** 95.4% DSC
  - **Spleen:** 92.8% DSC
  - **Pancreas:** 82.2% DSC
- **Limitations:**
  - Uses **private datasets**
  - Works well **only with multi-class labels available**

4. Fully Supervised Deep Learning Models

- **Examples:** FCNs, U-Net, DeepLab
- **Supervision Type:**
  - **Fully supervised** — requires **complete multi-class labels** for all objects in all images
- **Strengths:**
  - **High performance** when sufficient annotations are available
- **Limitations:**
  - Do **not solve** the issue of **single-class datasets**

5. Weakly Supervised Segmentation Models

- **Approach:**
  - Use **limited supervision**, such as image-level labels or scribbles instead of full pixel annotations
- **Goal:**
  - Segment objects with **minimal labeling effort**
- **Potential:**
  - A **close alternative** to the proposed method in the paper
  - Useful when **detailed annotations** are not feasible

Would you apply any of the other methods other than the DL method used in this paper? Explain your answer?

No ,The DL method proposed in this paper is the most suitable and effective choice for the given problem. It combines **high performance, flexibility, and practicality** — making it a better option than applying older, class-specific, or fully supervised methods.
Also it outperforms the upper SOTA mentioned in **Q3** ,
“We show that our conditioned multi-class segmentation framework outperforms current state-of-the-art single-class segmentation approaches for biomedical images. Additionally, we demonstrate the applicability of the proposed approach for the segmentation of urban scenes”

What datasets have been used in this paper? Do you think the result is generalizable for any datasets?

the authors used both **medical imaging datasets** and a **natural image dataset** to validate their proposed method:
1. Biomedical Imaging (CT Scans)

- **Liver:** 20 volumes from **Sliver07**
- **Pancreas:** 82 volumes from the **NIH Pancreas Dataset**
- **Spleen:** 74 volumes from a **private dataset**

Properties:

- Diverse **scanners and institutions**
- Varying **slice thicknesses**
- Includes **pathologies** (e.g., tumors, splenomegaly)

Preprocessing:

- Resized to **256×256×32 subvolumes**
- **Normalization** applied

Augmentation Techniques:

- **Random rotations**
- **Zooms**
- **Shifts**

2. Natural Images (Cityscapes Dataset)

- Evaluated on **urban street scenes** to assess **generalizability**
- **Cityscapes Dataset:**
  - Contains images with **19 semantic classes**
  - High resolution: **1024×2048**
  - Demonstrates applicability beyond medical imaging
  - Works on **2D RGB natural images**

Medical Imaging Datasets Summary

- **NIH Pancreas Dataset:**
  - Public dataset with **annotated CT scans**
  - Used for **pancreas segmentation**
- **Private Liver and Spleen Datasets:**
  - Used in prior works (e.g., **Yang et al., Roth et al.**)
  - Include **3D CT scans** of liver, spleen, and pancreas

Natural Image Dataset Summary

- **Cityscapes Dataset:**
  - Contains **urban street scenes**
  - **19 semantic classes**, resolution **1024×2048**
  - Shows the method's potential in **non-medical domains**

Is the Result Generalizable to Any Dataset?

Yes — Proven Across Two Very Different Domains

- Works on:
  - **3D grayscale medical CT scans**
  - **2D RGB natural images**
- Shows **strong adaptability** of the method

Key Generalization Advantages

- **Conditioning Approach is Data-Agnostic**
  - Does **not rely on domain-specific priors**
  - Injects **class-specific conditional information** into the model
  - Learns from **image-label pairings** only
  - Enables **broad applicability**

- **No Need for Full Multi-Class Labels**
  - Can handle:
    - **Incomplete**
    - **Disjoint**
    - **Single-class** labeled datasets
  - Suitable for **real-world scenarios** with limited annotations

Discuss the results presented in the paper. Compare the results with other stateof-the art methods used to solve this problem.

The authors evaluate several model variants and compare them using the Dice Similarity Coefficient (DSC). Their goal is to perform multi-class segmentation from single-class datasets — a problem that existing methods do not directly solve.

Here's how their model variants perform:

Model	Description	Performance
indivs	Trained separately on each class without conditioning	Poor performance, slow convergence
no cond	Naively predicts all classes from disjoint data without any conditioning	Failed to converge, low accuracy
cond-2nd	Conditioning class info in second input channel	Better than above but still underperforming
cond-enc	Conditioning added to encoder layers	Similar to cond-2nd, modest generalization
cond-dec	**Conditioning added to each decoder layer (best-performing model)**	**Outperforms all others, fast convergence, strong accuracy**

Comparison with Other State-of-the-Art Methods :

Method	Target	DSC (%)	Notes
Zhou et al. [48]	Pancreas	82.4% (NIH Pancreas dataset)	Tailored for pancreas segmentation only
Yang et al. [44]	Liver	95.0% (private dataset)	Single-class liver segmentation on a private dataset
Roth et al. [35]	Liver, Spleen, Pancreas	95.4% (liver), 92.8% (spleen), 82.2% (pancreas)	Coarse-to-fine multi-class model, private dataset
This paper (cond-dec)	Liver, Spleen, Pancreas	Comparable or superior to above methods	Uses only single-class labels, yet generalizes to multi-class tasks

Observations & Conclusion :

•The conddec model generalizes better despite being trained on disjoint **single-class datasets** — something none of the other methods can do effectively.

•Faster **convergence** and accurate **boundaries** were achieved even under challenging imaging conditions.

•Results show high generalization ability, as the model performed well not only on medical datasets but also on **Cityscapes (natural images)**.

What would you like to criticize about the paper? Could you suggest any improvements.

While the paper proposes a practical and effective method, it could be improved with deeper theoretical justification, more public evaluations, and richer conditioning techniques. These would enhance both the scientific value and the real-world applicability of the work

1. Data and Evaluation Limitations:

a) **Over-reliance on Private Datasets**
Most comparisons to state-of-the-art **(SOTA)** methods were made using private datasets, making it harder for the community to reproduce or fairly benchmark the result , this Limits reproducibility and makes it harder for the research community to validate or extend the work.
As a solution We could use standard public datasets like: Multi-organ annotated public CTs (if available) , Medical Segmentation Decathlon (MSD)

b) **B- Limited External Generalization:**
While the authors did test on Cityscapes for natural images, it was a brief experiment, not explored in-depth.
As an improvement we have to Test on diverse domains to show that the approach truly generalizes beyond medical imaging.

c) **C-It seems that the paper's data has No Real-world Noise :**
Real-world data often includes incomplete, noisy, or inaccurate masks.
As an improvement we can add some noises by Masking only parts of the organ or Adding false positives/negatives

2-Methodological Simplicity :

- **Conditioning via Static Hash Values Is Naive :**

The class conditioning method uses fixed hash values inserted as an additional channel **(cond-2nd**) so It doesn't adapt or learn — it just broadcasts a static signal.

can be improved by using learned embeddings for each class (similar to how NLP models learn word embeddings). Consider conditional normalization (BatchNorm) to modulate internal features dynamically.

- **Decoder-only Conditioning Isn’t Explained in Depth :**

The decoder-conditioning outperforms encoder-conditioning **but what is the reason ?!**

To comprehend the advantage of decoder conditioning, conduct information bottleneck analysis, gradient flow analysis, or feature visualization.

3- Implementation and Training Details Missing :

Too Few Information Although the technique is well explained in theory, crucial implementation elements such as batch sizes, optimizer hyperparameters, 3D input volume sizes, and how datasets were divided for training/validation are not addressed.

Publish code or at least include an appendix with full training protocol and data preprocessing