Self-supervised Semantic Segmentation: Consistency over Transformation

Sanaz Karimijafarbigloo, Reza Azad, Amirhossein Kazerouni, Yury Velichko,
Ulas Bagci and Dorit Merhof

https://github.com/mindflow-institue/SSCT

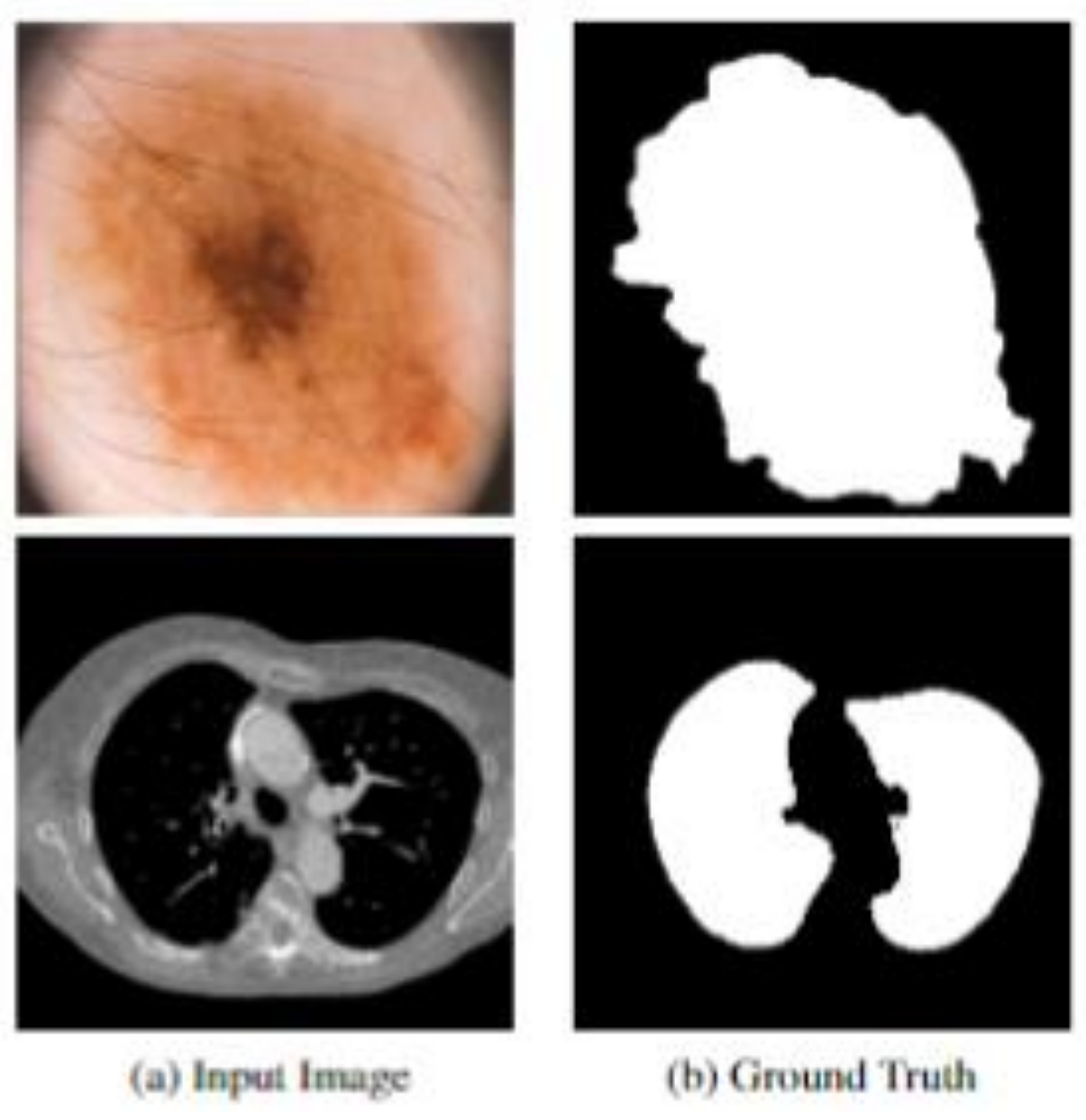
1. Introduction

Accurate medical image segmentation is of utmost importance for enabling automated clinical decision procedures. However, prevailing supervised deep learning approaches for medical image segmentation encounter significant challenges due to their heavy dependence on extensive labeled training data. To tackle this issue, we propose a novel self-supervised algorithm, S3-Net, which integrates a robust framework based on the proposed Inception based Large Kernel Attention (I-LKA) modules.

We evaluated our method on two public dataset:

Skin Lesion Segmentation: For the first task, we focused on segmenting skin lesion regions in dermoscopic images. To evaluate our method, we utilized the PH2 dataset

Lung Segmentation: In the second task, we addressed lung segmentation in CT images. To conduct this evaluation, we employed the publicly available lung segmentation dataset provided by Kaggle. We perform our clustering slice by slice.



(a) Input Image (b) Ground Truth

2. Model Architecture

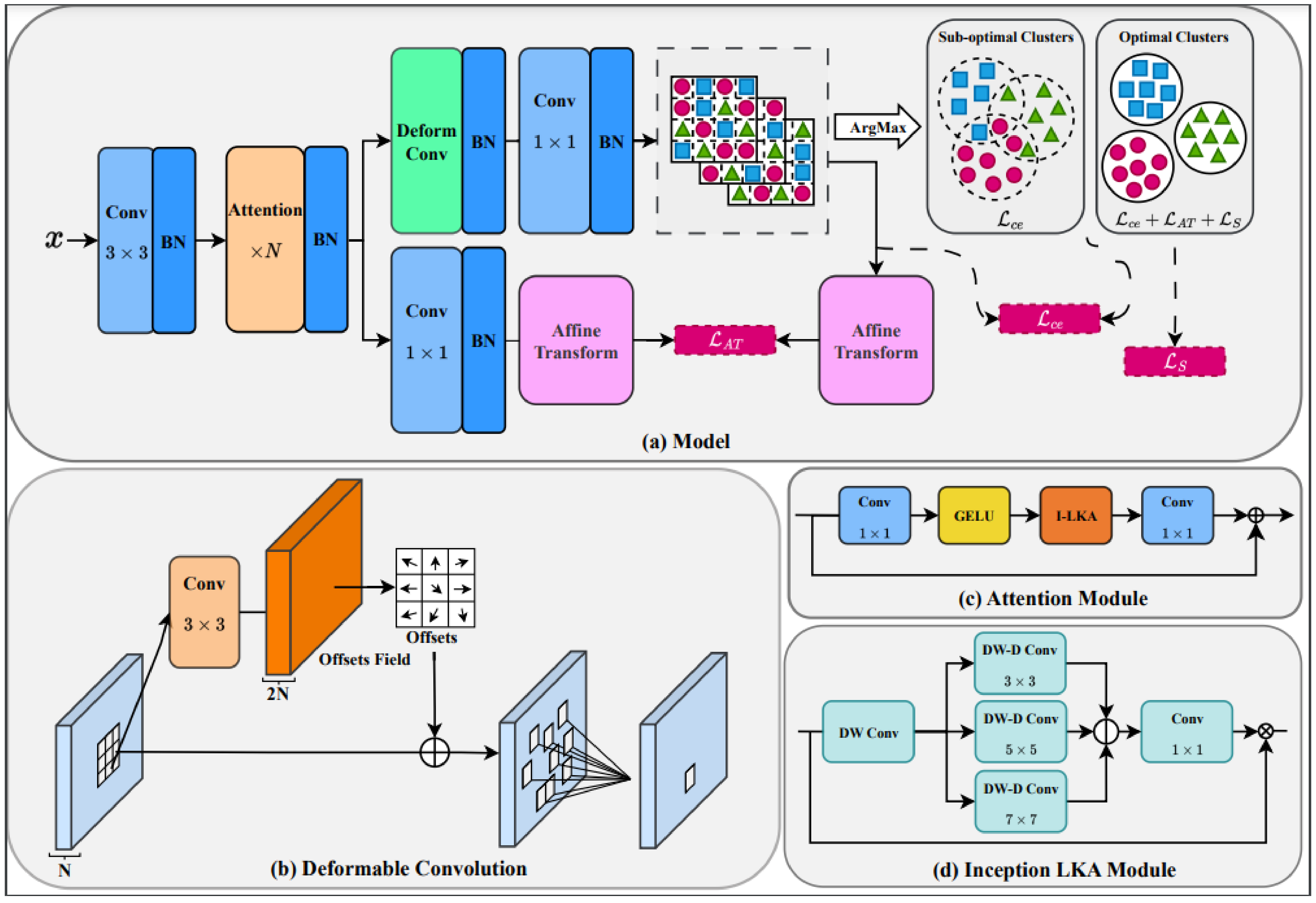


Fig.1 A general overview of the 3-Net framework.

3. Main Contribution

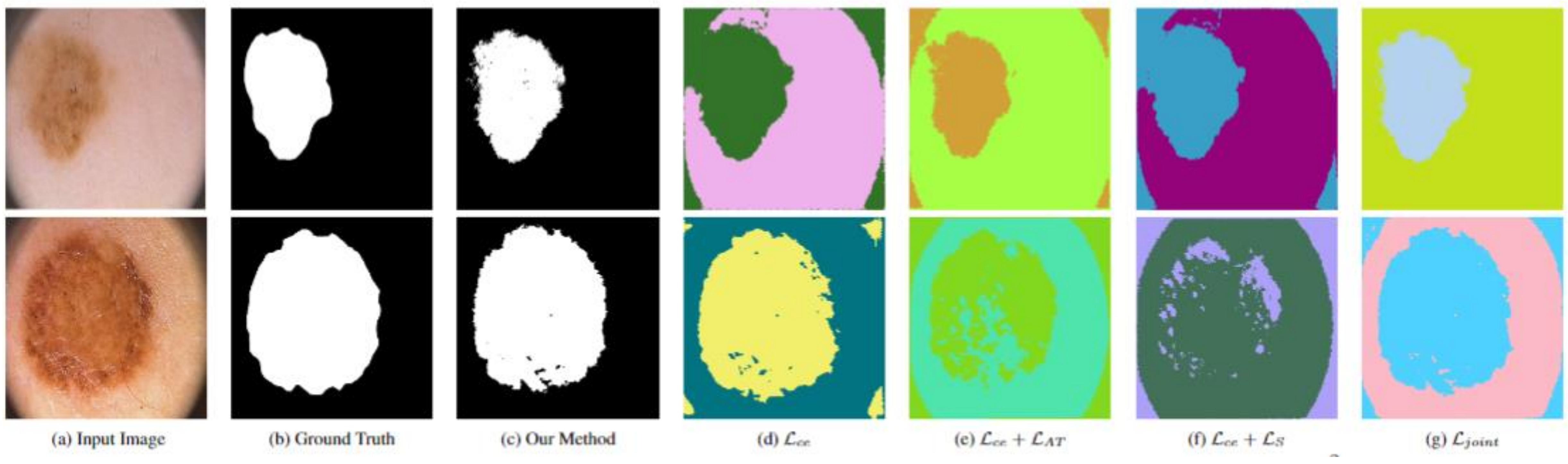
1. We propose I-LKA modules, which serve as a fundamental building block in our network design. These modules are specifically designed to capture contextual information comprehensively while preserving local descriptions. By striking a balance between these two aspects, our architecture facilitates precise semantic segmentation by effectively leveraging both global and local information.
2. We incorporate deformable convolution as a crucial component in our approach. This enables our model to effectively capture and delineate deformations, leading to improved boundary definition for the identified objects.
3. In order to make our model more robust to geometric transformations commonly encountered in medical scenarios, we integrate a self-supervised algorithm based on contrastive learning. By emphasizing the acquisition of invariance to affine transformations, our approach enhances the model's capacity to handle such transformations effectively. This allows the model to better generalize and adapt to different spatial configurations and orientations.
4. To ensure spatial consistency and promote the grouping of spatially connected pixels with similar features, we model a spatial consistency loss term based on edge information. This loss term facilitates the learning process by encouraging the network to capture the relationships among neighbouring pixels.
5. Finally, our proposed method effectively tackles dataset bias by performing the prediction process based on a single image only. This approach helps to mitigate the potential bias that may arise from imbalanced or skewed datasets.

4. Qualitative Results

Training process:
To learn the trainable parameters, we employ SGD (learning rate 0.36) optimization, minimizing the overall loss function iteratively for a maximum of 50 iterations.

Metrics:
We employ the Dice (DSC) score, XOR metric, and Hammoud distance (HM) as evaluation metrics.

Loss functions effect:
we examine the influence of suggested loss functions on the model's generalization performance.



(a) Input Image (b) Ground Truth (c) Our Method (d) $\mathcal{L}_{ce}$ (e) $\mathcal{L}_{ce} + \mathcal{L}_{AT}$ (f) $\mathcal{L}_{ce} + \mathcal{L}_{S}$ (g) $\mathcal{L}_{joint}$

Fig.2 Segmentation results on the skin lesion segmentation task using the PH2 dataset..

5. Comparative Results

Table 1. The performance of the proposed method is compared to the SOTA approaches on the PH^2 and Lung datasets.

Methods	PH^2			Lung Segmentation		
	DSC ↑	HM ↓	XOR ↓	DSC ↑	HM ↓	XOR ↓
k-means	71.3	130.8	41.3	92.7	10.6	12.6
DeepCluster [8]	79.6	35.8	31.3	87.5	16.1	18.8
IIC [22]	81.2	35.3	29.8	-	-	-
SGSCN[1]	83.4	32.3	28.2	89.1	16.1	34.3
MS-Former [25]	86.0	23.1	25.9	94.6	**8.1**	14.8
Our Method	**88.0**	**20.4**	**22.0**	**94.7**	8.8	**13.1**

Table 2. Impact of individual loss functions on model performance. The experiments were conducted using the PH^2 dataset.

$\mathcal{L}_{ce}$	$\mathcal{L}_{AT}$	$\mathcal{L}_{S}$	DSC ↑	HM ↓	XOR ↓
✓	✗	✗	86.1	22.8	25.6
✓	✓	✗	86.4	22.2	24.6
✓	✗	✓	85.9	22.7	25.2
✓	✓	✓	**88.0**	**20.4**	**22.0**

6. Discussion and Conclusion

Discussion:
Our method outperforms the SOTA approaches across all evaluation metrics, demonstrating the effectiveness of our self-supervised content clustering strategy. Notably, our method exhibits superior performance compared to SGSCN and MS-Former by modeling spatial consistency at both the pixel and region levels. This modeling of spatial dependency provides a stronger foundation for accurate segmentation. Furthermore, our approach incorporates consistency over transformations, allowing the network to learn transformation-invariant feature representations, leading to smoother clustering space (evidenced in Table 1).
We also show the impact of individual loss functions on model performance in Table 2.

Conclusion:
This paper introduces a novel SSL approach that combines the I-LKA module with deformable convolution to enable semantic segmentation directly from the image itself. Additionally, our network incorporates invariance to affine transformations and spatial consistency, providing a promising solution for pixel-wise image content clustering.