Implicit Representations for Image Segmentation

Jan Philipp Schneider*[1] Mishal Fatima*[1] Jovita Lukasik*[1]
Andreas Kolb[1] Margret Keuper[1,2,3] Michael Moeller[1]

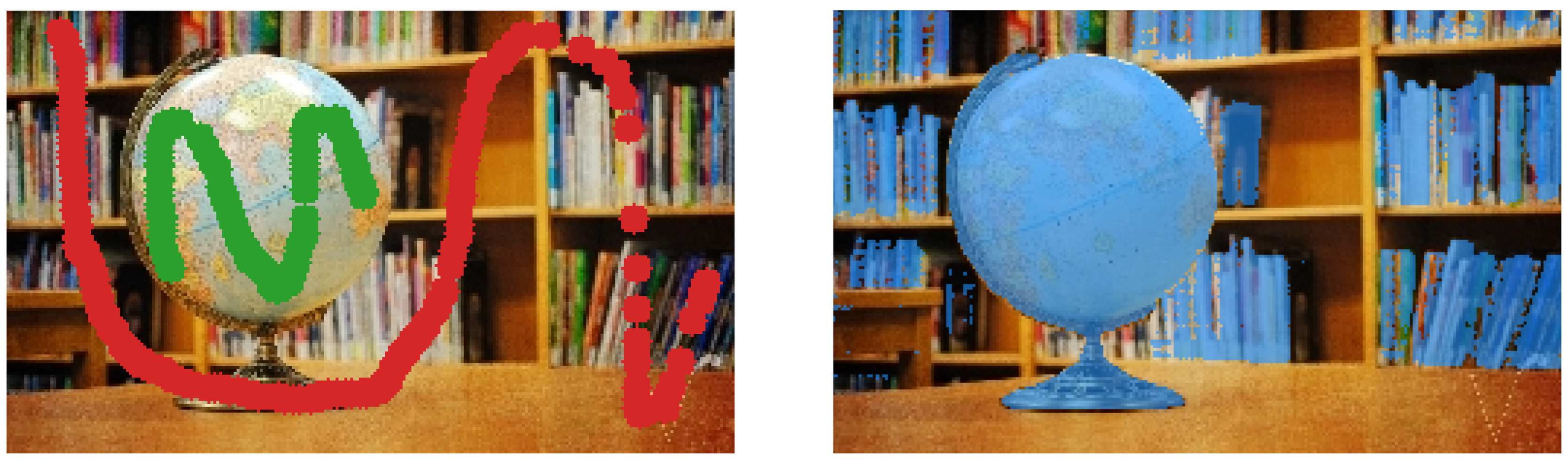
[1] University of Siegen [2] University of Mannheim
[3] Max-Planck-Institute for Informatics, Saarland Informatics Campus
* These Authors contributed equally

Motivation

Segmentation in the era of Foundation Models is still challenging, if:

- Data is scarce
- Objects are occluded
- Examples are out-of-distribution

Also: Provably ensuring constraints is hard



(a) Scribbled Image (b) SAM Segmentation

Figure 1. Segmenting a scribbled image a. with Segment Anything (SAM) [2] b.

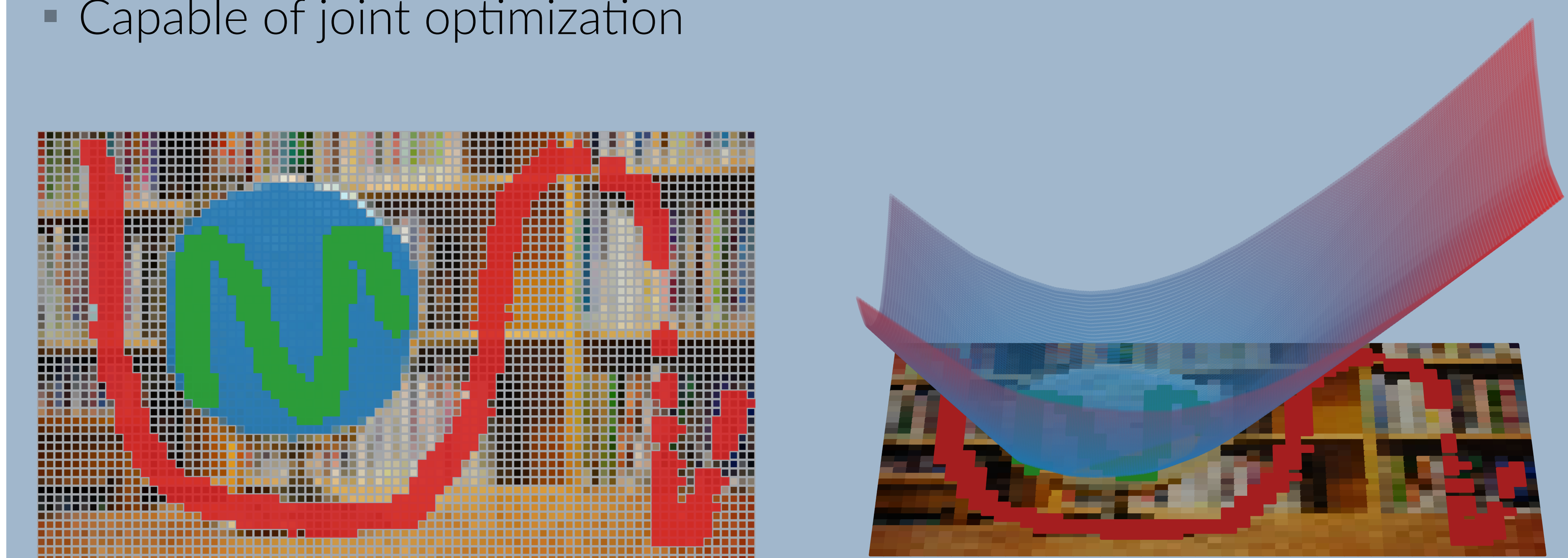
Proposal

Implicit segmentation representation

- Mapping spatial coordinates to fore- or background
- Constraint: Foreground will be in a convex shape

Regularization for Convex Shapes

- Using input convex neural network [3]
- Goes with any prediction architecture
- Capable of joint optimization



(a) Explicit Representation (b) Implicit Representation

Figure 2. Explicit segmentation $u \in [0,1]^{n_y \times n_x}$ in a. vs. an implicit representation $\mathcal{G}(x;\nu) : \mathbb{R}^2 \to [0,1]$ in b. using an input convex neural network.

Implicit Representations

- Represent segmentations as a *function* $\mathcal{G}_\nu : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ *implicitly* via a neural network [3]
- Maps spatial image domain Ω to the likeliness of a pixel being foreground

$$\mathcal{G}_\nu(x) = z_K,\ z_{i+1} = \text{ReLU}(\nu_i^z z_i + \nu_i^x x + b_i),\ \nu_i^z \geq 0 \quad \forall i \in \{1, \dots, K-1\} \quad (1)$$

- Assuring any lower level set is convex

$$\{x \in \mathbb{R}^2 \mid \mathcal{G}_\nu(x) \leq 0\} \subset \mathbb{R}^2 \quad (2)$$

Representation Unification

Regularization of (possibly parameterized) segmentation predictor $\mathcal{N}_\theta(f) \in [0,1]^{n_y \times n_x}$ for an image $f \in \mathbb{R}^{n_y \times n_x \times 3}$:

$$\text{dist}(\mathcal{N}_\theta(f), S) = \min_\nu \|\mathcal{N}_\theta(f) - \sigma(\mathcal{G}_\nu(f))\| \quad (3)$$

for S as the set of functions represented by (1) for a fixed choice of architecture, and σ soft thresholding i.e. sigmoid function.

Sequential and Joint Unification

As eq. (3) involves two sets of parameters, ν and θ for each image, we propose two options for computing convex segmentations:

Sequential Unification

The sequential option computes the *projection* of a given prediction $\mathcal{N}_\theta(f)$ onto our set S:

$$\text{proj}_S(\mathcal{N}_\theta(f)) = \sigma(\mathcal{G}_{\hat{\nu}}(f)) \quad \text{for} \quad \hat{\nu} = \arg\min_\nu \|\mathcal{N}_\theta(f) - \sigma(\mathcal{G}_\nu(f))\|. \quad (4)$$

Joint Unification

Since θ is usually determined in a training process, one can learn ν **jointly** using (3) as a **regularizer** during training of $\mathcal{N}_\theta(f)$.

Numerical Experiments

We investigate the influence of the implicit convex representation, by using scribble-based convexity dataset [4] and two simple architectures for $\mathcal{N}_\theta$ [1]:

1. Convolutional neural network (CNN) with 3×3 kernels, 4 layers, width 16
2. Fully connected network (FCN), pixel-wise prediction, 5 layers, width 16

- Including besides RGB also spatial, and/or semantic [5] input features per pixel while trained using our proposed sequential and joint approach

Consider the Intersection over union (IoU) of foreground objects with predictors $\mathcal{N}_\theta$ and our proposed implicit convex representations $\mathcal{G}_\nu$:

	RGB+semantic		RGB+spatial+semantic	
	CNN / convex	FCN / convex	CNN / convex	FCN / convex
seq.	0.726 / **0.843**	0.714 / **0.851**	**0.778** / 0.766	0.736 / **0.746**
joint	0.818 / **0.899**	0.635 / **0.894**	0.805 / **0.809**	0.768 / **0.769**

- The implicitly enforced convexity assumption can improve the results
- The **joint unification is clearly superior** to a sequential one

(a) Scribbled Image (b) Segmentation (c) Seq. Projection (d) Joint Projection

Figure 3. Qualitative Result of an FCN architecture, trained on scribbles with RGB and semantic features.

As illustrated in Fig. 1b, large **Foundation Models** can fail due to out-of-domain examples and can therefore **benefit from geometric convexity constraints** if prior information is valid.
We evaluated SAM and its convex projection on the convexity dataset with various corruptions of severity 5, yielding a small but systematic improvement:

Model	Clean	1	2	3	4	5	6	7	8
SAM	0.728	0.563	0.649	0.646	0.533	0.630	0.625	0.733	0.719
proj_S(SAM)	**0.741**	**0.582**	**0.660**	**0.652**	**0.550**	**0.637**	**0.636**	**0.743**	**0.732**

1: Spatter, 2: Contrast, 3: Brightness, 4: Impulse, 5: Shot Noise, 6: Gaussian Noise, 7: Defocus Blur, 8: Glass Blur

References

[1] Hannah Dröge and Michael Moeller. Learning or modelling? an analysis of single image segmentation based on scribble information. In *International Conference on Image Processing (ICIP)*, 2021.
[2] Alexander Kirillov et al. Segment anything. *CoRR*, 2023.
[3] Brandon Amos et al. Input convex neural networks. In *International Conference on Machine Learning*, 2017.
[4] Lena Gorelick et al. Convexity shape prior for segmentation. In *Computer Vision – ECCV 2014*, 2014.
[5] Yağiz Aksoy et al. Semantic soft segmentation. *ACM Transactions on Graphics*, 2018.

We acknowledge the German Research Foundation (DFG) support for the research unit 5336 “Learning to Sense”.

jan.schneider@uni-siegen.de