Implicit Representations for Image Segmentation

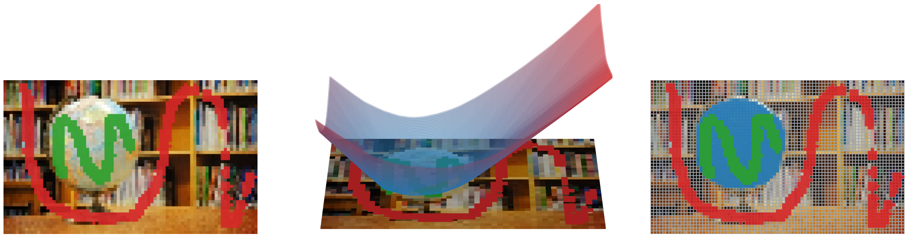
Jan Philipp Schneider[*1] **Mishal Fatima**[*1] **Jovita Lukasik**[*1]
Andreas Kolb[1] **Margret Keuper**[1,2,3] **Michael Moeller**[1]
[1] University of Siegen [2] University of Mannheim
[3] Max-Planck-Institute for Informatics, Saarland Informatics Campus

Abstract

Image segmentation has greatly advanced over the past ten years. Yet, even the most recent techniques face difficulties producing good results in challenging situations, e.g., if training data are scarce, out-of-distribution examples need to be segmented, or if objects are occluded. In such situations, the inclusion of (geometric) constraints can improve the segmentation quality significantly. In this paper, we study the constraint of the segmented region being segmented convex. Unlike prior work that encourages such a property with computationally expensive penalties on segmentation masks represented *explicitly* on a grid of pixels, our work is the first to consider an *implicit representation*. Specifically, we represent the segmentation as a parameterized function that maps spatial coordinates to the likeliness of a pixel belonging to the fore- or background. This enables us to provably ensure the convexity of the segmented regions with the help of *input convex neural networks*. Numerical experiments demonstrate how to encourage explicit and implicit representations to match in order to benefit from the convexity constraints in several challenging segmentation scenarios.

1 Introduction

The past decade has led to tremendous advances in the field of image segmentation via data-driven techniques with state-of-the-art approaches typically relying on an *explicit* grid-based representation, using discrete segmentation mask with values in $[0, 1]$. However, explicit representations using only one value per pixel have difficulties representing the global context of the segmentation, which can lead to severe failure cases. Therefore, we propose to use *implicit* representations in the form of a neural network that (continuously) maps (arbitrary) spatial coordinates to the likeliness of a pixel belonging to fore- or background, which we illustrate in fig. 1.



(a) Scribbled Image

(b) Implicit Representation $\mathcal{G}(x; \nu) : \mathbb{R}^2 \to [0, 1]$

(c) Explicit Representation $u \in [0, 1]^{n_y \times n_x}$

Figure 1: Trying to segment a scribbled image like (a). We propose to parameterize segmentations implicitly via an input convex function (b) instead of classical explicit representation methods (c).

[*] Authors contributed equally. Email to: `{firstname.lastname}@uni-siegen.de`

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

While implicit representations have recently been very successful for representing images (c.f. [16, 17]), we are (to the best of our knowledge) the first to use a *learnable implicit representation* for image segmentation. By parameterizing the implicit function to be *input convex* [2], we can provably ensure the resulting segmentation to be convex. In this work, we demonstrate how to combine and jointly train the common explicit and our novel implicit representation and illustrate advantages in challenging image segmentation scenarios numerically.

2 Related Work

Including geometric constraints such as convexity of the shape to the segmentation has received significant attention already in early work, which used segmentation representations via a polygon representing the boundary of the objects [14, 5]. These representations implicitly satisfy connectivity constraints, but are prone to getting stuck in bad local optima. Yet, similar representations have gained recent attention in the context of convexity priors exploiting orientation-based lifting [3, 4]. Level-set functions have been made topology-preserving by preventing local changes that would alter the topology [7], yet consequently also have difficulties overcoming local minima.

Interestingly, for more global approaches utilizing explicit representations, the seemingly simple constraint of the segmented region being *convex* is challenging: As segmentations of an image $f \in \mathbb{R}^{n_x \times n_y \times 3}$ are commonly represented explicitly as masks $m \in [0,1]^{n_y \times n_X}$, graphs with $n_x \cdot n_y$ many nodes or (discretized) level-set functions $\phi \in \mathbb{R}^{n_x \times n_y}$ to identify the segmented region with $\{(i,j) \mid \phi_{i,j} \leq 0\}$, different characterizations of convex sets have been exploited to form regularizers. Yet, explicit representations either need to approximate the convexity [12] or lead to computationally intense schemes, e.g. leading to combinatorial (NP-hard) problems, c.f. [9, 13], or using curvature penalties leading to fourth order differential equations [13] during optimization.

By considering an implicit representation and defining a coupling loss for unifying explicit and implicit representations, we demonstrate that provable convexity can easily be ensured in arbitrary existing (learning- or model-based) image segmentation approaches.

3 Explicit vs. Implicit Representations for Segmentations

3.1 Implicit Representations

We propose to represent the segmentation as a *function* $\mathcal{G}_\nu : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ *implicitly* via a neural network parameterized by ν: $\mathcal{G}_\nu$ takes coordinates within the image domain Ω as an input and predicts values that – when thresholded – divide Ω into foreground and background. While this type of representation has recently gained a lot of attention for images [16, 17] it has - to the best of our knowledge - not been exploited for representing *segmentations*. In particular, by choosing $\mathcal{G}_\nu$ to be an *input convex network* [2], i.e., choosing

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

we can assure that any lower level set is a convex set. Thus

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

is an *implicit representation* of a convex region. Yet, this raises the question of how our representation (2) can be included in common segmentation frameworks that frequently predict (or optimize for) one value per pixel of the image to be segmented.

3.2 Representation Unification

Let $\mathcal{N}_\theta(f) \in [0,1]^{n_y \times n_x}$ be any function that predicts a (classical) segmentation on an image $f \in \mathbb{R}^{n_y \times n_x \times 3}$ while possibly depending on additional parameters θ. Typical cases include energy minimization approaches such as

$$\mathcal{N}_\theta(f) = \underset{u \in [0,1]^{n_y \times n_x}}{\operatorname{argmin}} \; D(u,f) + \alpha \|\nabla u\|_1 \tag{3}$$

for a data term $D(u,f)$. One example could be a linear term with a model-based likelihood as in [15], and a regularizer such as the total variation, where we state ∇ as the discrete derivative operator in (3). In such a case θ would only consist of the regularization parameter α and possible hyperparameters for generating the data term. Equally valid, $\mathcal{N}_\theta$ could represent a neural network with learned parameters θ such that $\mathcal{N}_\theta(f)$ denotes the result of an inference step on the image f.

In any case, despite the discrete nature of $\mathcal{N}_\theta(f) \in [0,1]^{n_y \times n_x}$, it is natural to still interpret the prediction as a (piecewise constant) function $\mathcal{N}_\theta(f) : \Omega \to [0,1]$ by identifying the pixel values $(\mathcal{N}_\theta(f))_{i,j}$ as the constant value of the function for all x inside the entire rectangle that describes the pixel. Thus, with any given explicit representation being in the same function space as our family of implicit and provably convex neural networks, it is natural to *penalize the distance* between a given prediction $\mathcal{N}_\theta(f)$ and the set of functions that can be represented as a soft (e.g. sigmoid-based) thresholding of implicit input convex functions (1). As the distance between a point and a set is defined as the minimal distance between the point and any element in the set, we naturally study

$$\text{dist}(\mathcal{N}_\theta(f), S) = \min_\nu \|\mathcal{N}_\theta(f) - \sigma(\mathcal{G}_\nu(f))\| \tag{4}$$

for S as being the set of functions that can be represented via (1) for a fixed choice of architecture, and σ being a sigmoid function.

3.3 Sequential vs. Joint Representation Unification

Considering that eq. (4) already involves two sets of parameters, ν and θ, it is natural that we have two options for computing convex segmentations, i.e., for unifying the two representations. The sequential option computes the *projection* of a given prediction $\mathcal{N}_\theta(f)$ onto our set S, i.e.

$$\text{proj}_S(\mathcal{N}_\theta(f)) = \sigma(\mathcal{G}_{\hat{\nu}}(f)) \quad \text{for} \;\; \hat{\nu} = \arg\min_\nu \|\mathcal{N}_\theta(f) - \sigma(\mathcal{G}_\nu(f))\|. \tag{5}$$

With increasingly expressive architecture choices for the input convex networks in S, we expect (5) to converge to the projection of $\mathcal{N}_\theta(f)$ onto the set of convex segmentations[*].

For any *learnable* approach that determines the parameters θ of the predictor in a training process, a *joint unification* is an interesting alternative. In this case, we propose to use (4) as a regularizer during training, effectively leading to a joint optimization over the predictors parameters θ and one set of implicit convex representation parameters ν per training image f.

4 Numerical Experiments

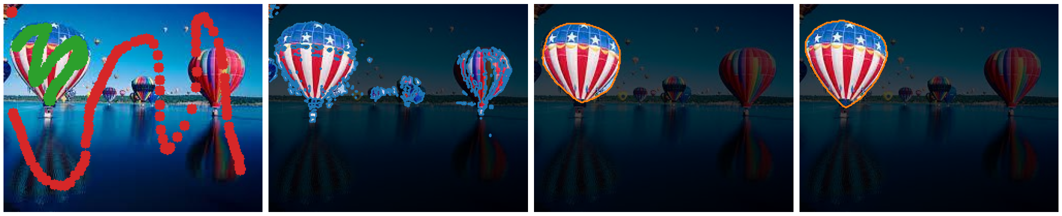
To investigate the influence of the implicit convex representation numerically, we exploit the scribble-based convexity dataset [†]. It consists of 51 images with user scribbles, and (approximately) convex foreground objects to be segmented. All details of our numerical experiments can be found in the appendix.

In our first study, we consider the segmentation of separate images based on scribbles, i.e., with approaches that are not based on a large set of training images, but try to learn the correct predictions on a single user-scribbled image only. [10] also considers sparse segmentation in the context of motion segmentation. We use a simple convolutional neural network (CNN) with 3×3 kernels, and a pointwise (fully connected or 1×1 CNN) network (FCN) as predictors $\mathcal{N}_\theta$. The networks are trained separately for each image in order to predict the correct label only for the scribbled pixels. Inspired by [6], we vary the inputs of the network as an RGB image in combination with either spatial coordinates or semantic features [1] or a combination of both. We compare the outcome of each approach in the setting of sequential representation unification, i.e., first training the predictor, then computing the projection onto our implicit convex representations, and in the setting of a joint unification, i.e., training the network parameters θ and the representation parameters ν jointly coupled via (4), see table 1.

We can see that the implicitly enforced convexity assumption is able to improve the results, with a joint unification clearly being superior to a sequential one. Interestingly, the impact of the convex projection method is significantly larger when the segmentation network does not receive any spatial information, which leads to substantial improvements and is the best-performing approach by far.

[*] Strictly mathematically, care has to be taken of the proposed optimization problem having a minimizer as the sigmoid can create problems in this respect.

[†] The convexity dataset is available at `https://vision.cs.uwaterloo.ca/data/`.



(a) Scribbled Image (b) Segmentation (c) Convex Projection (d) Joint Unification

Figure 2: Qualitative results of a FCN segmentation, trained on scribbles (a) with RGB input and semantic features. In (b), the learned segmentation is given, which is very scattered, due to the lack of spatial information, while its convex projection (c) fits the balloon quite well. We get an even better segmentation using the joint training approach (d).

Table 1: Intersection over union (IoU) of the foreground objects w.r.t the ground truth. We report the sequential projection (first row) and the joint representation unification (second row) between different predictors $\mathcal{N}_\theta$ and our proposed implicit convex representations $\mathcal{G}_\nu$ over three runs.

	RGB+spatial		RGB+semantic		RGB+spatial+semantic	
	CNN / convex	FCN / convex	CNN / convex	FCN / convex	CNN / convex	FCN / convex
seq.	0.697 / **0.763**	**0.732** / 0.711	0.726 / **0.843**	0.714 / **0.851**	**0.778** / 0.766	0.736 / **0.746**
joint	0.798 / **0.799**	0.755 / **0.756**	0.818 / **0.899**	0.635 / **0.894**	0.805 / **0.809**	0.768 / **0.769**

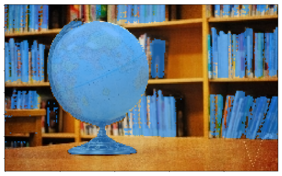
Table 2: IoU for SAM (first row) and the convex projection (second row) in the case of additional corruptions. A severity value of 5 is used to corrupt the images.

Model	Clean	Spatter	Contrast	Brightness	Impulse	Shot Noise	Gaussian Noise	Defocus Blur	Glass Blur
SAM	0.7275	0.5627	0.6489	0.6456	0.5330	0.6298	0.6246	0.7333	0.7187
proj_S(SAM)	**0.7407**	**0.5817**	**0.6597**	**0.6516**	**0.5504**	**0.6371**	**0.6357**	**0.7426**	**0.7321**

We exemplify the effect of the projection as well as the joint training qualitatively in Fig. 2: While the original segmentation $\mathcal{N}_\theta$ is highly scattered (fig. 2 b), an implicit input-convex projection yields the segmentation of the main convex object (fig. 2 c). Joint training allows both representations to find an agreement leading to even more accurate contours (fig. 2 d).

While the results in table 1 show consistent improvements, one could think that large foundational models such as Segment Anything (SAM) [11] solve the image segmentation problem as a whole and make the consideration of separately segmenting single images with scribbles obsolete. Yet, as we illustrate in Fig. 3, even SAM can easily fail due to out-of-domain examples and can therefore benefit from geometric convexity constraints if such prior information is valid. To analyze such cases quantitatively, we exploit SAM predictions with a random foreground-scribbled point as a prompt, as the full scribbles already yield an out-of-distribution prompt, which makes SAM fail completely. Yet, even using the corruptions proposed in [8] on the convexity dataset yields challenging cases for SAM. Projecting the results onto our implicit convex representation yields a small but systematic improvement as shown in Table 2. Interestingly, even without corruptions, using a foreground-scribbled point as a prompt in SAM cannot compete with a joint training of a simple network with semantic features and our implicit convex representation.

Figure 3: Trying to segment a scribbled image with SAM [11] fails when using full scribbles as SAM has not been trained with dense scribbles



5 Conclusions

In this paper, we propose to use implicit representations for image segmentation as they allow for an easier inclusion of geometric constraints such as convexity of the segmented region. We demonstrate the effectiveness of this approach in two different training scenarios, resulting in better segmentation than classical explicit segmentation methods. We consider the general framework to be highly promising with a great potential for further extensions from convexity to connectivity constraints.

Acknowledgments

The authors acknowledge support by the DFG research unit 5336 “Learning to Sense”.

References

[1] Yağiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Transactions on Graphics*, 37(4):1–13, 2018.

[2] Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

[3] Da Chen, Laurent D. Cohen, Jean-Marie Mirebeau, and Xue-Cheng Tai. An elastica geodesic approach with convexity shape prior. In *International Conference on Computer Vision (ICCV)*, pages 6900–6909, October 2021.

[4] Da Chen, Jean-Marie Mirebeau, Minglei Shu, Xuecheng Tai, and Laurent D. Cohen. Geodesic models with convexity shape prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8433–8452, 2023.

[5] Dominic Cheng, Renjie Liao, Sanja Fidler, and Raquel Urtasun. Darnet: Deep active ray network for building segmentation. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[6] 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.

[7] Xiao Han, Chenyang Xu, and Jerry L. Prince. A topology preserving level set method for geometric deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):755–768, 2003.

[8] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

[9] Hossam Isack, Lena Gorelick, Karin Ng, Olga Veksler, and Yuri Boykov. K-convexity shape priors for segmentation. In *European Conference on Computer Vision (ECCV)*, September 2018.

[10] Amirhossein Kardoost, Kalun Ho, Peter Ochs, and Margret Keuper. Self-supervised sparse to dense motion segmentation. In *Asian Conference on Computer Vision (ACCV)*, 2020.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *CoRR*, abs/2304.02643, 2023.

[12] Jun Liu, Xue-Cheng Tai, and Shousheng Luo. Convex shape prior for deep neural convolution network based eye fundus images segmentation. *CoRR*, abs/2005.07476, 2020.

[13] Shousheng Luo, Xue-Cheng Tai, Limei Huo, Yang Wang, and Roland Glowinski. Convex shape prior for multi-object segmentation using a single level set function. In *International Conference on Computer Vision (ICCV)*, 2019.

[14] Diego Marcos, Devis Tuia, Benjamin Kellenberger, Lisa Zhang, Min Bai, Renjie Liao, and Raquel Urtasun. Learning deep structured active contours end-to-end. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.

[15] Claudia Nieuwenhuis and Daniel Cremers. Spatially varying color distributions for interactive multilabel segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1234–1247, 2013.

[16] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[17] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.