SUBPIXEL OBJECT SEGMENTATION USING WAVELETS AND MULTIRESOLUTION ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a novel deep learning framework for fast prediction of boundaries of two-dimensional simply connected domains using wavelets and Multi Resolution Analysis (MRA). The boundaries are modelled as (piecewise) smooth closed curves using wavelets and the so-called Pyramid Algorithm. Our network architecture is a hybrid analog of the U-Net, where the down-sampling path is a two-dimensional encoder with learnable filters, and the upsampling path is a one-dimensional decoder, which builds curves up from low to high resolution levels. Any wavelet basis induced by a MRA can be used. This flexibility allows for incorporation of priors on the smoothness of curves. The effectiveness of the proposed method is demonstrated by delineating boundaries of simply connected domains (organs) in medical images using Debauches wavelets and comparing performance with a U-Net baseline. Our model demonstrates up to 5x faster inference speed compared to the U-Net, while maintaining similar performance in terms of Dice score and Hausdorff distance.

1 INTRODUCTION

Semantic image segmentation is a core component of many medical imaging related tasks. Both as part of a pipeline to find a region of interest, or a task by itself, e.g., for measuring tumor volume. Nowadays, almost all segmentation algorithms in medical imaging are replaced by U-Net-like architectures Ronneberger et al. (2015) combining an encoder and decoder. Typically, the decoder is an upsampling path, and additional skip connections between the encoding and decoding part are added to recover the image's spatial information. While many variants or more exotics methods **?**, such as multi-scale and pyramid based approaches, recurrent networks or generative techniques, can be designed, all of these still yield per-pixel classifications. In these settings an image is interpreted as a *discrete* collection of ordered pixels (or voxels in the three-dimensional case), where the task is to assign an appropriate class to each pixel using a probabilistic model. Typically, the pixels are assumed to be independent so that the joint-likelihood is tractable.

Medical images often represent discrete approximations, measured and stored in a discrete grid, of inherently continuous objects. The pixel or voxel sizes can be coarse, and each voxel may represent multiple tissue types due to the partial volume effect. While humans are able to draw continuous contours based on such images, e.g., boundaries of organs, typical deep learning-based segmentation models are not able to take this continuous nature into account. In this work, we propose a novel deep learning framework which does not predict per-pixel segmentation maps, but directly predicts contours as continuous objects.

A fundamental component in the setup of our framework is the decision on how to represent (closed) curves. While the Fourier basis is a natural candidate at first glance, its global nature may hamper accurate predictions of curves which exhibit highly localized behavior, requiring accurate estimates of small noisy high-frequency modes. For this reason, we have chosen to represent contours using wavelets and Multi Resolution Analysis (MRA) instead. The main idea is to choose a single map φ, the so-called scaling function or father wavelet, and to construct subspaces of functions associated to prescribed resolution levels by taking the span of appropriate dilations and translations of φ. This setup provides an efficient way to decompose and reconstruct contours, from low to high resolution level, using the so-called Pyramid Algorithm, which has deep connections with the encoder-decoder architecture used in common U-Net architectures.

In this paper, we construct a hybrid analog of the U-Net, where the down-sampling path is a two-dimensional encoder with learnable filters, and the upsampling path is an one-dimensional decoder, which builds curves up from low to high resolution levels. The filters in the decoder are not learned but uniquely determined by the chosen wavelet basis. Our framework is flexible; any wavelet basis induced by a MRA can be used. This flexibility allows for incorporation of smoothness priors by using specific wavelet bases.

Related work Previous work Chen et al. (2019); Marcos et al. (2018); Hatamizadeh et al. (2019) also proposed models to predict contours by combining Active Contour Models (ACM) with a CNN into an end-to-end model. In these papers the representation of curves is based on traditional pixel-based methodology. For instance, in Hatamizadeh et al. (2019) curves are modelled as level sets of distance maps defined on a discretization of the domain of the image. In Chen et al. (2019) a similar approach is followed, but a smoothed approximation of an indicator function is used instead of a distance map. The work in Marcos et al. (2018) is perhaps most closely related to ours; they directly construct polygonal approximations of curves and represent them using (pixel) coordinates of the nodes. The objectives minimized in the cited papers are based on a careful consideration of mean pixel intensities, enclosed area's and arc length. These properties are implicitly encoded in an objective function (energy-functional) defined on a space of distance maps Hatamizadeh et al. (2019), suitable approximations of indicator functions Chen et al. (2019), or family of polygons Marcos et al. (2018). However, in contrast to our approach, the above methods all still provide pixel-based output. To the best of our knowledge, our work is the first to use MRA and wavelet analysis to construct pixel-independent representations of (closed) curves on various resolution levels.

This paper is organized as follows. In Section 2 we review the mathematical background needed to construct our model. We explain how contours can be decomposed and reconstructed on different resolution levels. The reconstruction algorithm, known as the Pyramid Algorithm, forms a core component of our network architecture. In Section 3 we set up a model architecture and loss. Subsequently the datasets, training method and performance measures are described in Section 4. We end the paper with the results of our method in Section 5 and a brief discussion. A detailed account of all the mathematical details is provided in the appendix.

2 Background and mathematical setup

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and $X : \Omega \to [0,1]^{n_1 \times n_2}$ a random variable whose realizations correspond to two dimensional gray-valued images, e.g., slices of MRI scans of size $n_1 \times n_2$, where $n_1, n_2 \in \mathbb{N}$ and $\mathcal{X} := [0,1]^{n_1 \times n_2}$. We assume without loss of generality that the gray-values of the scans are rescaled to $[0,1]$. Commonly, X is paired with a random variable $Y : \Omega \to \{0,1\}^{n_1 \times n_2}$ whose realizations correspond to binary segmentations. In this context, realizations of (X, Y) correspond to scans with inscribed masks. In this paper we assume that the regions approximated by the inscribed masks are almost surely enclosed by *immersed simple closed curves*.

Let $(x, y) \in \mathcal{X} \times \{0,1\}^{n_1 \times n_2}$ be a realization of (X, Y). The binary mask y constitutes a discretized approximation of the true region of interest $R \subset \mathbb{R}^2$ and provides a polygonal approximation of ∂R. In a large variety of situation, however, ∂R is in reality a closed (piecewise) smooth curve. In medical imaging, for instance, smooth curves are drawn to describe the boundary of a region of interest, e.g., an organ. In this paper we develop a deep learning framework for directly parameterizing $\partial R = \partial R(x)$, given an image x, using wavelets and Multi Resolution Analysis (MRA). Since wavelets play a seminal role in our set up, we start in the next section with an extensive review to introduce notation and terminology.

2.1 Multi resolution analysis

In this section we review the necessary material from discrete wavelet theory using the framework of a multi resolution analysis (MRA). We closely follow the exposition in Pereyra & Ward (2012) and Montefusco & Puccio (2014). The reader is referred to these references for a more comprehensive introduction.

Multi Resolution Analysis It is a well known-fact from Fourier analysis that a signal $\gamma \in L^2(\mathbb{R})$ cannot be simultaneously localized in the time and frequency domain, see [REF]. Multi resolution

analysis aims to address this shortcoming by decomposing a signal on different *discrete* resolution levels. The idea is to construct subspaces $V_j \subset L^2(\mathbb{R})$, associated to various resolution levels $j \in \mathbb{Z}$, spanned by shifts of a localized mapping φ_j. The level of localization associated to V_j is determined by taking an appropriate dilation of a prescribed map φ; the so-called *scaling function*. In the MRA framework the dilation factors are chosen to be powers of two. Altogether, this yields an increasing sequence of closed subspaces $V_j \subset L^2(\mathbb{R})$ comprising all of $L^2(\mathbb{R})$. In particular, $V_{j+1} \supset V_j$ is the next level up in resolution after V_j. The "view" of γ on resolution level j is then defined as the orthogonal projection of γ onto V_j. The formal definition of a MRA is given below in Definition 2.1.

Definition 2.1 (Formal definition MRA)**.** Let $T_k : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ and $\mathcal{D}_j : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ denote the translation and normalized dilation operator, respectively, defined by $T_k\gamma(x) = \gamma(x-k)$ and $\mathcal{D}_j\gamma(t) = 2^{\frac{j}{2}}\gamma(2^j t)$ for $\gamma \in L^2(\mathbb{R}) \cap C_0^\infty(\mathbb{R})$ and $j, k \in \mathbb{Z}$. A multi-resolution analysis of $L^2(\mathbb{R})$ is an increasing subsequence of subspaces $(V_j)_{j\in\mathbb{Z}}$, such that (i) $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$, (ii) $\bigcup_{j\in\mathbb{Z}} V_j$ is dense in $L^2(\mathbb{R})$, (iii) $\gamma \in V_j$ if and only if $\mathcal{D}_1\gamma \in V_{j+1}$, (iv) V_0 is invariant under translations, (v) $\exists\varphi \in L^2(\mathbb{R})$ such that $\{T_k\varphi\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_0.

Conditions (iii) and (iv) imply that each subspace V_j is invariant under integer shifts. Condition (v) formalizes the idea that the subspaces are spanned by translations and dilations of a single map φ; the so-called *scaling function* or *father wavelet*. Indeed, it is straightforward to show that $\{\varphi_{jk} : k \in \mathbb{Z}\}$ is an orthonormal basis for V_j, where $\varphi_{jk} := T_k\mathcal{D}_j\varphi$.

Approximation coefficients Let $\gamma_j \in V_j$ denote the best approximation of γ at resolution level j, i.e., $\gamma_j := P_j\gamma$, where $P_j : L^2(\mathbb{R}) \to V_j$ is the orthogonal projection onto V_j. Then there exists unique coefficients $a_j(\gamma) = (a_{jk}(\gamma))_{k\in\mathbb{Z}}$ such that $\gamma_j = \sum_{k\in\mathbb{Z}} a_{jk}(\gamma)\varphi_{jk}$, where $a_{jk} = \langle\gamma, \varphi_{jk}\rangle$, since $\{\varphi_{jk} : k \in \mathbb{Z}\}$ is an orthonormal basis for V_j. The coefficients $a_j(\gamma)$ are referred to as the *approximation coefficients* of γ at resolution level j. The associated subspaces V_j are referred to as the approximation subspaces. We will write $a_j(\gamma) = a_j$ whenever it is clear from the context that the coefficients are associated to γ.

Detail coefficients To study the information that is lost when a signal in V_{j+1} is projected onto V_j, we consider the operator $Q_j := P_{j+1} - P_j$. Note that $Q_j\gamma \in W_j := V_j^\perp \cap V_{j+1}$, since $P_jP_{j+1} = P_j$ and $V_j \subset V_{j+1}$. The subspace W_j is called the detail subspace at level j. The detail subspaces $(W_j)_{j\in\mathbb{Z}}$ are mutually disjoint and orthogonal by construction. Furthermore, since $V_j = V_{j-1} \oplus W_{j-1}$ for any $j \in \mathbb{Z}$, it follows that $V_j = V_{j_0} \oplus \bigoplus_{l=j_0}^{j-1} W_l$, for all $j > j_0$, where $j, j_0 \in \mathbb{Z}$. This decomposition shows that a signal on resolution level j can be reconstructed from any lower level j_0 if all the details in between are known. In addition, it follows from this observation that $L^2(\mathbb{R}) = \bigoplus_{j\in\mathbb{Z}} W_j$.

A fundamental result, known as Mallat's Theorem, states that the subspaces W_j can too be spanned by dilating and shifting a single map. More precisely, there exists a map $\psi \in W_0$, the so-called *mother wavelet*, such that $\{\psi_{jk} : k \in \mathbb{Z}\}$ is an orthonormal basis for W_j, where $\psi_{jk} := T_k\mathcal{D}_j\psi$. The wavelet constructed in Mallat's Theorem is often referred to as the *mother wavelet*. The proof is based on a very careful analysis of the defining properties of a scaling function in the frequency domain. The reader is referred to Pereyra & Ward (2012); Montefusco & Puccio (2014) for the details. Now, using Mallat's mother wavelet, we may write $Q_j\gamma = \sum_{k\in\mathbb{Z}} d_{jk}(\gamma)\psi_{jk}$ for any $\gamma \in L^2(\mathbb{R})$, where $d_j(\gamma) := (d_{jk}(\gamma))_{k\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$. The coefficients $d_j(\gamma)$ are referred to as the *detail coefficients* of γ at resolution level j. We will suppress the dependence on γ and write $d_j(\gamma) = d_j$ when there is no chance of confusion.

The detail coefficients store the information needed to go back one level up in resolution, since $P_{j+1} = P_j + Q_j$ by construction. In general, given approximation coefficients $a_{j_0} \in \ell^2(\mathbb{Z})$ at level j_0 and detail coefficients $d_l \in \ell^2(\mathbb{Z})$ at levels $j_0 \leq l \leq j-1$, we can reconstruct the approximation at level j from j_0 via $\gamma_j = \sum_{k\in\mathbb{Z}} a_{j_0k}\varphi_{j_0k} + \sum_{j_0\leq l\leq j-1}\sum_{k\in\mathbb{Z}} d_{lk}\psi_{lk}$.

Galerkin projections In practice, we only approximate a finite number of approximation and detail coefficients. For this purpose, we introduce some additional notation. We define the finite

dimensional subspaces $\widehat{V}_j := \text{span}\{\varphi_{jk} : -2^{j-1} \leq k \leq 2^{j-1} - 1\} \subset V_j$ and $\widehat{W}_j := \text{span}\{\psi_{jk} : -2^{j-1} \leq k \leq 2^{j-1} - 1\} \subset W_j$. We shall frequently identify $\widehat{V}_j \simeq \mathbb{R}^{2^j}$ and $\widehat{W}_j \simeq \mathbb{R}^{2^j}$.

2.2 The Discrete Wavelet Transform

In this section we recall how the approximation and detail coefficients can be computed given a prescribed scaling function φ. Many fundamental aspects of MRA's, both theoretical and computational, can be traced back to the following simple but key observation. Since $V_0 \subset V_1$, there must exist coefficients $h \in \ell^2(\mathbb{Z})$ such that $\varphi = \sum_{l \in \mathbb{Z}} h_l \varphi_{1l}$. This equation is referred to as the *scaling equation*; one of the fundamental properties of a scaling function. The coefficients h completely characterize the scaling function. Therefore, to define a MRA, one only needs to specify h. Not every sequence $h \in \ell^2(\mathbb{Z})$ defines a scaling function, however, since it has to adhere to the defining properties of a MRA. Throughout this paper we assume that an appropriate choice for h is given.

Decomposition The scaling equation can be used to derive an efficient scheme for computing lower order approximation coefficients (of any order) given an initial approximation $a_{j+1} \in \ell^2(\mathbb{Z})$. Define $C_{\tilde{h}} : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ and $M^- : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ by $C_{\tilde{h}}(a) = a * \tilde{h}$ and $(M^- a)_k = a_{2k}$, respectively, where $\tilde{h} \in \ell^2(\mathbb{Z})$ is defined by $\tilde{h}_k = \overline{h_{-k}}$, and $* : \ell^2(\mathbb{Z}) \times \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ is the two-sided discrete convolution. Then $a_j = M^- C_{\tilde{h}} a_{j+1}$ for any $j \in \mathbb{Z}$. Furthermore, one can show, by analyzing the Fourier transform of the scaling equation, that $C_{\tilde{h}}$ cuts off relatively high frequencies. For this reason, h is often referred to as the *low-pass filter* associated to the MRA, see Pereyra & Ward (2012); Montefusco & Puccio (2014)

Similarly, since $\psi \in W_0 \subset V_1$, there exist coefficients $g \in \ell^2(\mathbb{Z})$ such that $\psi = \sum_{k \in \mathbb{Z}} g_k \varphi_{1k}$. For Mallat's mother wavelet, we have $g_k = (-1)^{k-1} \overline{h_{1-k}}$. An analogous argument shows that $d_j = M^- C_{\tilde{g}} a_{j+1}$ for all $j \in \mathbb{Z}$. Moreover, it can be shown that $C_{\tilde{g}}$ cuts of relatively low frequencies, and is therefore referred to as the *high pass filter* associated to the MRA.

Reconstruction The orthogonal decomposition $V_{j+1} = V_j \oplus W_j$ can be used to reconstruct $a_{j+1} \in V_{j+1}$ given the approximation and detail coefficients $a_j \in V_j$ and $d_j \in W_j$ at resolution level j. More precisely, substitution of the scaling equations for φ and ψ into this decomposition yields $a_{j+1} = C_h M^+ a_j + C_g M^+ d_j$, where $M^+ : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ is a right inverse of M^- defined by $(M^+ a)_k := a_{\frac{k}{2}}$ if k is even and zero otherwise.

In practice, our signals are periodic and do not directly fit into the MRA framework, since non-zero periodic signals are not elements in $L^2(\mathbb{R})$. We will address this issue in the next section by using an appropriate cut-off. Here we only remark that the periodicity needs to be carefully taken into account in the decomposition and reconstruction formulae; the details are provided in Appendix A.2.

2.3 Wavelet representation of periodic curves

Let $j_1 \in \mathbb{N}$ be a prescribed initial resolution-level on which a periodic signal γ is to be approximated. We re-parameterize γ to have period 1, so that we can conveniently choose the number of approximation coefficients to be a power of two (see the discussion in Appendix A.1). Furthermore, to address the issue that periodic signals are not contained in $L^2(\mathbb{R})$, we restrict the re-parameterized curve to $[-1, 1]$, i.e., set $\gamma^*(t) := \gamma(lt)\mathbf{1}_{[-1,1]}(t)$. In general, this will introduce discontinuities at the boundary points $-l$ and l. However, this issue is easily dealt with, as described below, since we only need information about γ on a strict subset $[I_0, I_1] \subset [-l, l]$ of length l.

If j_1 is sufficiently large, the approximations coefficients needed to (approximately) cover $[-1, 1]$ are $(a_{j_1 k}(\gamma^*))_{k=-2^{j_1}}^{\lfloor 2^{j_1 - \beta} \rfloor}$. Here $\beta > 0$ is the support of the underlying wavelet. These coefficients will be close to the (scaled) sample values of γ on $\{k 2^{-j_1} : -2^{j_1} \leq k \leq \lfloor 2^{j_1} - \beta \rfloor\}$. We refer to Appendix A.1 for an explanation. Motivated by this observation, we use the coefficients $(a_{j_1 k}(\gamma^*))_{k=-2^{j_1-1}}^{2^{j_1-1}-1}$ only, which cover $[-\frac{1}{2}, \frac{1-2^{1-j_1}}{2}]$ approximately. To ensure that $2^{j_1-1} - 1 < \lfloor 2^{j_1} - \beta \rfloor$, we require that $j_1 \geq \left\lceil \frac{\log(\beta-1)}{\log(2)} + 1 \right\rceil$. This quantity is well-defined for all bases considered in this paper, except the Haar-basis.

Formally, the approximation coefficients at a prescribed level j need to be computed by evaluating the integrals $\langle \gamma^*, \varphi_{jk} \rangle$. Typically, in signal analysis one uses appropriately rescaled samples of γ^* instead. This approach is justified for signals sampled on a sufficiently high resolution, since the family $(\varphi_{j0})_{j \in \mathbb{Z}}$ behaves as an approximate identity as $j \to \infty$. Such an initialization scheme may be highly inaccurate, however, when the signal is not sampled on a sufficiently fine grid. In the appendix, we propose an alternative to the usual initialization scheme by exploiting the fact that we are dealing with periodic signals; see Lemma A.1.

3 MODEL

In this section we formulate the objective of our model and present a suitable architecture. Furthermore, we set up a loss function for determining network-parameters. We will slightly abuse notation from the previous sections. A contour associated to an image $x \in \mathcal{X}$ will be denoted by $\gamma(x)$ and is henceforth assumed to have two components denoted by $[\gamma(x)]_1$ and $[\gamma(x)]_2$. In general, for any quantity that depends on x, we will frequently denote its dependency on x in a similar way. Finally, all operations from the previous sections are understood to be carried out component-wise unless stated otherwise.

3.1 OBJECTIVE

Let $y \in \mathcal{Y}$ be a mask associated to an image $x \in \mathcal{X}$, where we regard y as a realization of $Y|X = x$. We assume that the boundary $\partial R(x)$ of the region $R(x)$ encoded in y is parameterized by a closed immersed curve $\gamma(x) \in C^2_{\text{per}}([0, l(x)]; \mathbb{R}^2)$. Here $l(x) > 0$ denotes the arc-length of $\gamma(x)$. The main objective is to compute the approximation coefficients of $\gamma^*(x)$, where $\gamma^*(x)$ denotes the truncated re-parameterization of $\gamma(x)$. To be more precise, let $j_0, j_1, j_2 \in \mathbb{N}$ be resolution levels, where $j_0 \leq j_1 \leq j_2$. We will construct a convolutional neural network $F : \mathcal{X} \times \Theta \to (0,1) \times \prod_{j=j_0}^{j_2} \widehat{V}_j \times \widehat{V}_j$, where $F(x, \theta) = (p(x, \theta), f_{j_0}(x, \theta), \ldots, f_{j_2}(x, \theta))$. Here $p(x, \theta)$ is the conditional probability that $R(x) \neq \emptyset$, $f_j(x, \theta)$ is a prediction for the approximation coefficients of $\gamma^*(x)$ at level j, and $\theta \in \Theta \subset \mathbb{R}^n$ are the network parameters. The goal is to find $\theta \in \Theta$ such that $f_{jk}(x, \theta) \approx (a_{jk}([\gamma^*(x)]_1), a_{jk}([\gamma^*(x)]_2))$ for $-2^{j-1} \leq k \leq 2^{j-1} - 1$, $j_0 \leq j \leq j_2$ and "most" realizations of X. Henceforth we will omit the dependence of F and its components on the network parameters θ.

3.2 NETWORK ARCHITECTURE

Our network architecture consists of an encoder, bottleneck and decoder with skip-connections in between. Only the approximation coefficients at the lowest resolution-level j_0 are "directly" computed by the network (in the bottleneck). Afterwards, the Pyramid Algorithm takes over to compute approximation coefficients on higher resolution levels (the decoder). In practice, the detail coefficients are negligible on sufficiently high resolution levels. For this reason, we only predict detail coefficients up to a prescribed level j_1. The predictions at higher resolution levels $j_1 < j \leq j_2$ are computed without detail coefficients.

The encoder consists of five down-sampling blocks. Each block consists of six (convolutional) residual blocks, using ReLU-activation and kernels of size 3×3, followed by an average-pooling layer of size 2×2. The number of filters used in the five blocks, from top to bottom, is 32, 32, 64, 64, 128, respectively.

The encoder is followed by a a bottleneck. The first layer in the bottleneck is fully connected and transforms the output of the encoder to a vector in $\mathbb{R}^{128}$. Attached to this layer are five branches to predict the probability map $p(x) \in (0,1)$, components $[f_{j_0}(x)]_s \in \widehat{V}_{j_0}$ and $[h_{j_0}(x)]_s \in \widehat{W}_{j_0}$, where $s \in \{1, 2\}$. The latter quantity $[h_{j_0}(x)]_s$ is a prediction for the detail coefficients at level j_0. Each branch consists of four fully-connected layers. The first three layers map from $\mathbb{R}^{128}$ to itself with ReLu-activation and residual connections in between. The final fourth layer transforms the $\mathbb{R}^{128}$ dimensional output to an element in $\mathbb{R}^n$, where $n = 2^{j_0}$ for the branches associated to the approximation and detail coefficients, and $n = 1$ for the probability map. The branch for the probability map is followed by a sigmoid-activation.

The detail coefficients at levels $j_0 \leq j < j_1$ are predicted using skip-connections. More precisely, the feature map from the encoder path, that corresponds to the same "spatial level", is compressed to a feature map with four channels using a 1×1 convolution. Subsequently, a branch of the above form is used to predict detail coefficients in $\mathbb{R}^{2^j} \simeq \widehat{W}_j$. Altogether, the predicted approximation coefficients at level j_0 and detail coefficients at levels $j_0 \leq j \leq j_1 - 1$ are used as input to the Pyramid algorithm and together form the upsampling path. A schematic picture of our architecture can be found in the Appendix.

3.3 Loss

The loss $\mathcal{L}$ consists of two parts: ordinary cross-entropy for optimizing the likelihood of $p(x)$ and a part corresponding to the L^2-error between observed and predicted curves on different resolution levels. More precisely, set $r(x) = 1$ if $R(x) \neq \emptyset$ and $r(x) = 0$ otherwise, then $\mathcal{L}_{\text{ce}}(p(x), r(x)) := r(x) \log p(x) + (1 - r(x)) \log(1 - p(x))$. Next, suppose $a([\gamma^*(x)]) = (a_{j_0}([\gamma^*(x)]), \ldots, a_{j_2}([\gamma^*(x)])) \in \prod_{j=j_0}^{j_2} \widehat{V}_j \times \widehat{V}_j$ are the approximation coefficients of $\gamma^*(x)$ on resolution levels $j_0 \leq j \leq j_2$. Define $\mathcal{L}_{js}(f_j(x), a_j(\gamma^*(x))) := \|[f_j(x)]_s - a_j([\gamma^*(x)]_s)\|_2$ for $s \in \{1, 2\}$. This quantity corresponds to the L^2-error on resolution level j between the curves with approximation coefficients $[f_j(x)]_s$ and $a_j([\gamma^*(x)]_s)$. Finally, define the total loss by

$$\mathcal{L}(F(x), a) := w\mathcal{L}_{\text{ce}}(p(x), r(x)) + \sum_{\substack{s \in \{1,2\} \\ j \in \{j_0, j_2\}}} \mathcal{L}_{js}(f_j(x), a_j(\gamma^*(x))),$$

where $w > 0$ is a weight. In practice, we set $w = 1$.

Notice that $\mathcal{L}$ measures the discrepancies between observed and predicted curves on the lowest and highest resolution levels only. It is not necessary to supervise the prediction of intermediate resolution levels or even the detail coefficients. A detailed explanation and proof is provided in the appendix; see Appendix A.2.

4 Training

In this section we describe the dataset on which we test our method. In addition, we provide details about preprocessing steps and model development (training).

4.1 Datasets

Toy dataset The main purpose of the toy-example is to create a setting in which the annotated contours differ substantially from annotations confined to a grid. For this purpose, we consider piecewise smooth curves having a finite number of non-differentiable points.

The toy-dataset consists of hypocycloids, up to an Euclidian motion and scaling, defined by

$$\eta(t) := \left[(r_1 - r_2)\cos t + r_2 \cos\left(\frac{r_1 - r_2}{r_2} t\right) \quad (r_1 - r_2)\sin t - r_2 \sin\left(\frac{r_1 - r_2}{r_2} t\right) \right]^T,$$

where $r_1 > r_2$ and $t \in \mathbb{R}$. If $\frac{r_1}{r_2} \in \mathbb{N}$, then η is closed and has exactly $\frac{r_1}{r_2}$ cusps (non-differentiable points). To easily control the number of cusps, we fix $r_2 = 1$ and vary r_1. Note that in this case η has r_1 cusps and period 2π.

We construct curves and binary masks of various sizes, orientation and positions, by sampling a radius $r_1 \in \{3, 4, 5, 6\}$ from $\mathcal{U}(\{3, 4, 5, 6\})$, angle θ from $\mathcal{U}([-\frac{\pi}{2}, \frac{\pi}{2}])$, components q_1, q_2 from $\mathcal{U}([-80, 80])$ for a shift, and a scaling factor κ from $\mathcal{U}([10, 20])$. Here $\mathcal{U}(I)$ denotes a random variable uniformly distributed on I, where I is an interval of finite length or a discrete finite set. Next, we evaluate the curve $\kappa \left(R(\theta)\eta + [160 + q_1 \quad 160 + q_2]^T \right)$ on an equispaced grid of $[0, 2\pi]$ of size 512. Here $R(\theta)$ corresponds to an anti-clockwise rotation around the origin with angle θ. Finally, the discretized curve is used to construct a binary mask of size 320×320 using SKIMAGE.

Medical decathlon The data used to evaluate the performance of our model consists of MRI images of the prostate and CT scans of the spleen. The datasets are part of a public dataset made available for the Medical Decathlon Contest. The dataset for the prostate consists of T2-weighted MRI images of size 320×320, which was cropped to size 224×224. The dataset for the spleen consists of CT scans of size 512×512 and was cropped to size 256×256. The cropping was based on locating the mean midpoint of the segmentations in the training set and constructing a uniform bounding box around it; the details can be found in Appendix **??**. The reader is referred to Simpson et al. (2019) for more details about the datasets.

4.2 CONSTRUCTION GROUND-TRUTH

In this section we describe how the ground-truth data is computed using the Pyramid Algorithm and Lemma A.1. Let $(x, u) \in \mathcal{X} \times \mathbb{R}^{n_s \times n_p}$ be an image (slice) - contour pair, where x is a slice of the CT or MRI scan, u is a finite sequence of points approximating a closed curve and $n_s = 2$ is the number of spatial components. Since we only have access to binary masks (for the public datasets), and not to the raw annotations themselves, we extract u using OPENCV. While not ideal, we stress that u contains "subpixel" information and is *not* constrainted to an integer-valued grid.

Fourier coefficients To initialize the Pyramid Algorithm, we compute the approximation coefficients at level j_1 using Lemma A.1. To accomplish this, we need to compute a Fourier expansion for u. First, we parameterize the contour by arc length. The arc length l is approximated by summing up the Euclidian distances between subsequent points in u. The Fourier coefficients are then computed by evaluating the contour on an equispaced grid of $[0, l]$ of size $2N - 1$, where $N \in \mathbb{N}$, using linear interpolation and the Discrete Fourier Transform. Since the contours are real-valued, we only store the Fourier coefficients $(\tilde{\gamma}_m)_{m=0}^{N-1} \in (\mathbb{C}^{n_s})^N$. Fourier coefficients that are too small, i.e., have no relevant contribution, are set to zero; see Appendix B.1 for the details.

Consistency To have consistent parameterizations for all slices, we ensure that u is always traversed anti clock-wise (using `opencv`). Furthermore, since the parameterization is only determined up to a translation in time, we need to pick out a specific one. We choose the unique parameterization such that the contour starts at angle zero at time zero relative to the midpoint $c = (c_1, c_2) \in \mathbb{R}^2$ of R. The implementation details are provided in Appendix B.2.

Approximation and detail coefficients Altogether, the above steps yield a contour γ with Fourier coefficients $(\gamma_m)_{m=1-N}^{N-1}$ and period l. We reparameterize γ to have period 1, as explained in Section 2.3, and "center" the contour using the average midpoint computed over the training-set. The initial approximation coefficients at level j_2 are then computed using Lemma A.1. Next, the Pyramid Algorithm is used to compute approximation and detail coefficients at levels $j_0 \leq j \leq j_2 - 1$. We set the detail coefficients which are in absolute value below $\varepsilon = 5 \cdot 10^{-3}$ to zero to reduce noise. Subsequently, we reconstruct the approximation coefficients at levels $j_0 + 1 \leq j \leq j_1$ using the thresholded detail coefficients. No detail coefficients are used to compute approximation coefficients at levels $j_1 + 1 \leq j \leq j_2$. The final approximation and (thresholded) detail coefficients $a \in \bigoplus_{j=j_0}^{j_2} \widehat{V}_j \times \widehat{V}_j$ and $d \in \bigoplus_{j=j_0}^{j_1-1} \widehat{W}_j \times \widehat{W}_j$, respectively, are used as ground-truth.

The resulting dataset $\mathcal{D}$ thus consists of tuples (x, a, d). The training and validation set $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$, respectively, are obtained by randomly omitting subjects from the the full dataset $\mathcal{D}$. For the toy-example, spleen and prostate we have $(|\mathcal{D}_{\text{train}}|, |\mathcal{D}_{\text{val}}|) = (1650, 250)$, $(|\mathcal{D}_{\text{train}}|, |\mathcal{D}_{\text{val}}|) = (527, 75)$, and $(|\mathcal{D}_{\text{train}}|, |\mathcal{D}_{\text{val}}|) = (3148, 502)$, respectively. Before feeding the images x into the model, we linearly rescale the image intensities at each instance to $[0, 1]$. Furthermore, for the spleen and prostate, we use extensive data augmentation: we use random shifts, random rotations, random scaling, elastic deformations, horizontal shearing and random cropping.

4.3 MODEL TRAINING

We use the Adam optimizer Kingma & Ba (2015) to train our network for 150 epochs. In the first five epochs, we use a linearly increasing learning rate from 10^{-4} to $5 \cdot 10^{-3}$, which is subsequently decayed by a factor of 0.5 each time the loss does not significantly decrease for 10 subsequent epochs (for the remaining 145 epochs). A batch size of 16 samples is used in each descent step.

Finally, the model with the lowest loss is selected. The computations were performed in PYTORCH on a Geforce RTX 2080 Ti.

5 RESULTS

In this section we examine the performance of our method. We investigate the dependence on the choice of basis and smoothness; we consider the Debauches wavelets db$p$ with $p \in \{1, 2, 4, 8, 16\}$ vanishing moments. Roughly speaking, db$p$ corresponds to the unique MRA for which the mother wavelet has minimal support and p vanishing moments. For each basis, we fix appropriate resolution levels j_0, j_1 and j_2 as follows. We choose j_0 as the smallest level for which 2^{j_0-1} is larger or equal to the order of the low-pass-filter. Since the low-pass filter associated to db$p$ has order $2p$, we set $j_0(p) = p$. We choose $j_1 > j_0$ as the smallest level for which the detail coefficients are below $\varepsilon = 5 \cdot 10^{-3}$ for all samples in the training set. Finally, we choose $j_2 \geq j_1$ as the smallest resolution level for which the distance between the end-points of the curves are within 1 pixel distance for all samples in the training set.

Baseline For the prostate and spleen we use the original two-dimensional U-Net in Ronneberger et al. (2015) as baseline with the following differences: we use 32 filters in the first layer instead of 64 and use group normalization with four groups. The resulting network has approximately $7.76 \cdot 10^6$ trainable parameters.

Performance measures Since our model predictions are planar curves, we evaluate accuracy using two-dimensional quantities only. We compute the component-wise L^2-errors between observed and predicted curves on the highest resolution level j_2 by taking the ℓ^2-norm of the approximation coefficients. Furthermore, we compute the dice score and Haussdorf distance between curves using the implementation in SHAPELY. This requires a polygonal approximation of the contour. Such an approximation is easily obtained in our setup, since the approximation coefficients at level j_2 constitute an accurate discretization of the contour. In general, the Hausdorff distance between subsets of a metric space (baseline) may differ from the Hausdorff distance between the associated boundaries (wavelet models). We show in the appendix that the two coincide in our applications.

For the toy-problem, spleen and prostate we use 64, 32 and 64 Fourier coefficients, respectively, to initialize the approximation coefficients. The results are shown in Table 1. The predictions for the best performing wavelet models with respect to the Hausdorff-score are shown in Figure 1. Predictions for the other bases, corresponding wavelet decompositions, and detailed visualization of statistics (violin plots) can be found in Appendix C.

Toy problem We observe that all models perform well and are capable of accurately parameterizing piecewise smooth curves. In particular, our unsupervised approach for predicting detail coefficients works as intended. The performance improves as the regularity of the wavelets increase. For db1, we observe that the end-points of some of the predicted curves are further apart from each other than 1 pixel distance causing relatively large Hausdorff-distances. Furthermore, in some cases, the predictions associated to the less regular wavelets, db1 and db2, exhibit "small" oscillatory behavior. We found that the oscillatory behavior is caused by a too large resolution level j_2. More precisely, note that the features extracted from the images only contain information up to a certain resolution level. The subpixel information needed to "fill in the blanks", so to speak, is in part provided by the ground-truth data and in part by the chosen wavelet basis. The regularity and support of the wavelet determines to which extent, i.e., up to which resolution level, subpixel information can be "filled in". As the regularity (and support) of the wavelet decreases, the maximal achievable resolution level decreases as well. Therefore, to avoid the oscillatory behavior, we would need to decrease j_2.

Spleen The predictions of our models are accurate and on par with the baseline U-Net. In terms of the dice score, the U-Net performs slightly better. On the other hand, our models perform better in terms of Hausdorff-distance. This is reflected in slightly smaller averages, but more noticeably our models exhibit a smaller variation around the mean. Most of the outliers correspond to "edge" cases where the boundary of the spleen is small and about to disappear from our two-dimensional

sliced-view. In a small number of cases, the predictions associated to db1 and db2 exhibit some oscillatory behavior as observed (and explained) in the toy example.

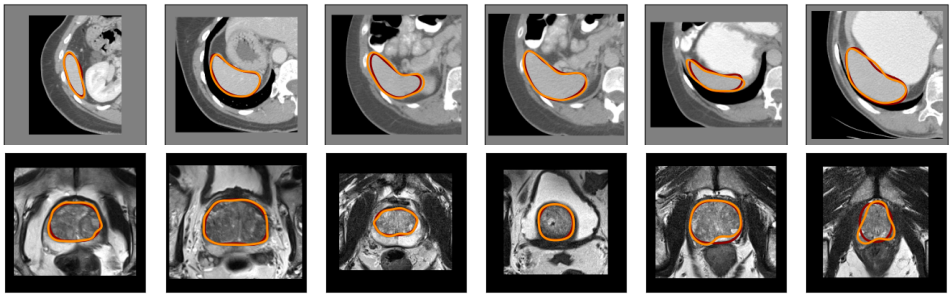

Figure 1: Predicted and observed boundaries colored in red and orange, respectively, for the spleen and prostate for the best models (db16). The last two columns correspond to a few "hard" examples.

Table 1: Mean and standard deviation of various performance measures for the toy example, spleen and prostate. The standard deviation is reported in parentheses. The columns L^2 correspond to the L^2-errors between the components $s \in \{1, 2\}$ of the predicted and observed curves, N_p is the approximate number of model parameters in millions, and the final column denotes the inference time T per image in milliseconds.

	Model	(j_0, j_1, j_2)	Dice	Hausdorff	L^2 $(s = 1)$	L^2 $(s = 2)$	N_p	T
Toy	db2	$(2, 6, 8)$	0.973 (0.016)	2.170 (0.717)	0.919 (0.314)	0.921 (0.325)	5.96	15.0
	db4	$(3, 6, 8)$	0.975 (0.018)	1.916 (0.704)	0.773 (0.305)	0.772 (0.303)	5.45	14.2
	db8	$(4, 6, 8)$	0.972 (0.023)	1.793 (0.744)	0.780 (0.278)	0.748 (0.284)	5.25	14.1
	db16	$(5, 6, 8)$	0.968 (0.023)	1.799 (0.587)	0.841 (0.332)	0.848 (0.328)	4.71	14.1
	unet	–	–	–	–	–	7.76	58.0
Spleen	db1	$(1, 5, 8)$	0.919 (0.049)	4.980 (1.723)	2.921 (1.214)	3.141 (1.591)	9.92	11.6
	db2	$(2, 5, 8)$	0.925 (0.035)	4.883 (1.691)	2.422 (1.119)	2.651 (1.449)	5.62	10.5
	db4	$(3, 5, 8)$	0.928 (0.041)	4.607 (1.516)	2.384 (1.339)	2.633 (1.624)	4.48	9.92
	db8	$(4, 5, 8)$	0.922 (0.047)	4.346 (1.499)	**2.275** (1.366)	2.721 (1.547)	4.11	**9.26**
	db16	$(5, 5, 8)$	0.929 (0.045)	**4.238** (1.941)	2.387 (1.264)	**2.542** (1.574)	**3.98**	9.92
	unet	–	**0.941** (0.047)	5.033 (3.036)	–	–	7.76	38.9
Prostate	db1	$(1, 5, 6)$	0.921 (0.030)	5.810 (2.455)	2.499 (1.205)	2.590 (0.836)	8.38	11.2
	db2	$(2, 5, 6)$	0.930 (0.036)	5.542 (3.177)	2.008 (1.305)	2.256 (0.950)	5.07	9.66
	db4	$(3, 5, 6)$	0.929 (0.036)	5.634 (3.293)	2.253 (1.357)	2.062 (0.837)	4.17	8.64
	db8	$(4, 5, 6)$	**0.938** (0.028)	5.366 (3.319)	**1.956** (1.271)	2.039 (0.836)	3.87	**7.59**
	db16	$(5, 5, 6)$	0.937 (0.036)	**5.305** (3.566)	1.980 (1.392)	**1.958** (1.020)	**3.74**	7.65
	unet	–	0.932 (0.047)	5.673 (2.402)	–	–	7.76	37.2

Prostate The results are similar as for the previous two examples; the wavelet models produce accurate predictions and are on par with the U-Net. The more regular wavelets, db8 and db16, seem to perform slightly better in terms of Dice score and Hausdorff distance. While our models produce accurate predictions for most examples, there are instances where both our model and the baseline U-Net fail to produce accurate predictions; see the last two columns of Figure 9. In these examples, the detail coefficients associated to parts of the curve with high curvature are too small in magnitude to be accurately predicted. While detail coefficients of such small magnitude were less relevant in the latter two examples, they are important for the prostate.

6 Conclusion

We have introduced a novel method to model boundaries of two dimensional simply connected domains using wavelets and MRAs. In effect this allows for subpixel segmentations. The efficacy of the method has been demonstrated by modeling the boundaries of hypercloids (toy-example), spleen and prostate, and demonstrating that the results are on par with state-of-the art U-Nets. In all cases, we observe that the more regular wavelets db8 and db16 perform best in terms of accuracy, inference time and memory efficiency.

REFERENCES

P. Abry and P. Flandrin. On the initialization of the discrete wavelet transform algorithm. *IEEE Signal Processing Letters*, 1(2):32–34, 1994. doi: 10.1109/97.300311.

Xu Chen, Bryan M Williams, Srinivasa R Vallabhaneni, Gabriela Czanner, Rachel Williams, and Yalin Zheng. Learning Active Contour Models for Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11632–11640, 2019.

Loukas Grafakos. *Classical Fourier Analysis*, volume 249 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 2008. ISBN 9780387094311. doi: 10.1007/978-0-387-09432-8. URL http://www.springerlink.com/index/10.1007/978-0-387-09432-8.

Ali Hatamizadeh, Debleena Sengupta, and Demetri Terzopoulos. End-to-End Deep Convolutional Active Contours for Image Segmentation. sep 2019. URL http://arxiv.org/abs/1909.13359.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pp. 770–778. IEEE Computer Society, dec 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.90. URL http://image-net.org/challenges/LSVRC/2015/.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2015. URL https://arxiv.org/abs/1412.6980v9.

Diego Marcos, Devis Tuia, Benjamin Kellenberger, Lisa Zhang, Min Bai, Renjie Liao, and Raquel Urtasun. Learning deep structured active contours end-to-end. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8877–8885, 2018.

L. Montefusco and L. Puccio. *Wavelets: Theory, Algorithms, and Applications*. ISSN. Elsevier Science, 2014. ISBN 9780080520841. URL https://books.google.nl/books?id=Ln6jBQAAQBAJ.

Stanislav Nikolov, Sam Blackwell, Ruheena Mendes, Jeffrey De Fauw, Clemens Meyer, Cían Hughes, Harry Askham, Bernardino Romera-Paredes, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D'Souza, Syed Ali Moinuddin, Kevin Sullivan, DeepMind Radiographer Consortium, Hugh Montgomery, Geraint Rees, Ricky Sharma, Mustafa Suleyman, Trevor Back, Joseph R. Ledsam, and Olaf Ronneberger. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *CoRR*, abs/1809.04430, 2018. URL http://arxiv.org/abs/1809.04430.

M.C. Pereyra and L.A. Ward. *Harmonic Analysis: From Fourier to Wavelets*. IAS/Park city mathematical subseries. American Mathematical Society, 2012. ISBN 9780821875667. URL https://books.google.nl/books?id=ovsjX_t0kOkC.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. pp. 1–8, 2015. ISSN 16113349. doi: 10.1007/978-3-319-24574-4_28. URL http://arxiv.org/abs/1505.04597.

Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. 2019.