

End-to-End Trainable Deep Active Contour Models for Automated Image Segmentation: Delineating Buildings in Aerial Imagery

Ali Hatamizadeh, Debleena Sengupta, and Demetri Terzopoulos

Computer Science Department
University of California, Los Angeles, CA 90095, USA
{ahatamiz,debleenas,dt}@cs.ucla.edu

Abstract. The automated segmentation of buildings in remote sensing imagery is a challenging task that requires the accurate delineation of multiple building instances over typically large image areas. Manual methods are often laborious and current deep-learning-based approaches fail to delineate all building instances and do so with adequate accuracy. As a solution, we present Trainable Deep Active Contours (TDACs), an automatic image segmentation framework that intimately unites Convolutional Neural Networks (CNNs) and Active Contour Models (ACMs). The Eulerian energy functional of the ACM component includes per-pixel parameter maps that are predicted by the backbone CNN, which also initializes the ACM. Importantly, both the ACM and CNN components are fully implemented in TensorFlow and the entire TDAC architecture is end-to-end automatically differentiable and backpropagation trainable without user intervention. TDAC yields fast, accurate, and fully automatic simultaneous delineation of arbitrarily many buildings in the image. We validate the model on two publicly available aerial image datasets for building segmentation, and our results demonstrate that TDAC establishes a new state-of-the-art performance.

Keywords: Computer vision · Image segmentation · Active contour models · Convolutional neural networks · Building delineation

1 Introduction

The delineation of buildings in remote sensing imagery [24] is a crucial step in applications such as urban planning [29], land cover analysis [35], and disaster relief response [28], among others. Manual or semi-automated approaches can be very slow, laborious, and sometimes imprecise, which can be detrimental to the prompt, accurate extraction of situational information from high-resolution aerial and satellite images.

Convolutional Neural Networks (CNNs) and deep learning have been broadly applied to various computer vision tasks, including semantic and instance segmentation of natural images in general [7,6] and particularly to the segmentation of

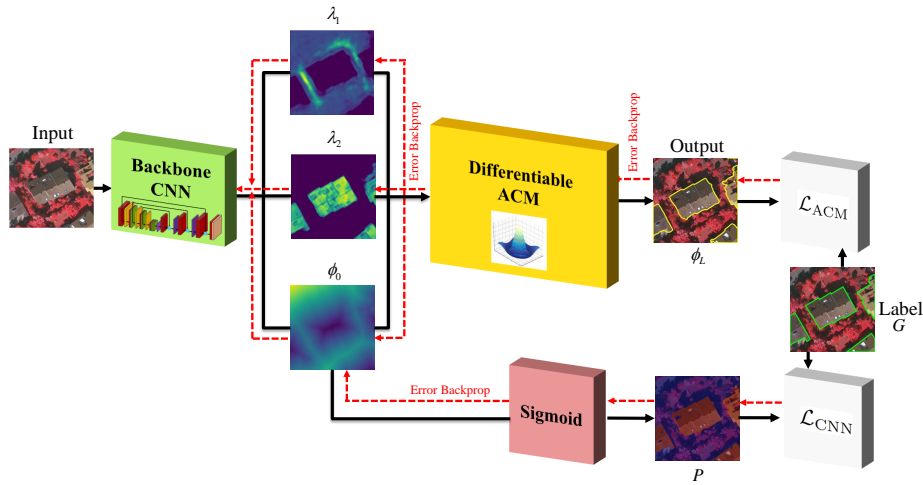


Fig. 1: TDAC is a fully-automated, end-to-end automatically differentiable and backpropagation trainable ACM and backbone CNN framework.

buildings in remote sensing imagery [28,3]. However, building segmentation challenges CNNs. First, since CNN architectures often include millions of trainable parameters, successful training relies on large, accurately-annotated datasets; but creating such datasets from high-resolution imagery with possibly many building instances is very tedious. Second, CNNs rely on a filter learning approach in which edge and texture features are learned together, which adversely impacts the ability to properly delineate buildings and capture the details of their boundaries [11,16].

One of the most influential computer vision techniques, the Active Contour Model (ACM) [19], has been successfully employed in various image analysis tasks, including segmentation. In most ACM variants, the deformable curves of interest dynamically evolve according to an iterative procedure that minimizes a corresponding energy functional. Since the ACM is a model-based formulation founded on geometric and physical principles, the segmentation process relies mainly on the content of the image itself, not on learning from large annotated image datasets with hours or days of training and extensive computational resources. However, the classic ACM relies to some degree on user input to specify the initial contour and tune the parameters of the energy functional, which undermines its usefulness in tasks requiring the automatic segmentation of numerous images.

We address the aforementioned challenges by intimately uniting CNNs and ACMs in an end-to-end trainable framework (originally proposed in [15]). Our framework (Fig. 1) leverages a novel ACM with trainable parameters that is automatically differentiable in a TensorFlow implementation, thereby enabling the backpropagation of gradients for stochastic optimization. Consequently, the ACM and an untrained, as opposed to pre-trained, backbone CNN can be

trained together from scratch. Furthermore, our ACM utilizes an Eulerian energy functional that affords local control via 2D parameter maps that are directly predicted by the backbone CNN, and it is also automatically initialized by the CNN. Thus, our framework alleviates the biggest obstacle to exploiting the power of ACMs in the context of CNNs and deep learning approaches to image segmentation—eliminating the need for any form of user supervision or intervention.

Our specific technical contributions in this paper are as follows:

- We propose an end-to-end trainable building segmentation framework that establishes a tight merger between the ACM and any backbone CNN in order to delineate buildings and accurately capture the fine-grained details of their boundaries.
- To this end, we devise an implicit ACM formulation with pixel-wise parameter maps and differentiable contour propagation steps for each term of the associated energy functional, thus making it amenable to TensorFlow implementation.
- We present new state-of-the-art benchmarks on two popular publicly available building segmentation datasets, *Vaihingen* and *Bing Huts*, with performance surpassing the best among competing methods [25,9].

2 Related Work

2.1 CNN-Based Building Segmentation Models

Audebert *et al.* [1] leveraged CNN-based models for building segmentation by applying SegNet [2] with multi-kernel convolutional layers at three different resolutions. Subsequently, Wang *et al.* [31] applied ResNet [17], first to identify the instances, followed by an MRF to refine the predicted masks. Some methods combine CNN-based models with classical optimization methods. Costa *et al.* [10] proposed a two-stage model in which they detect roads and intersections with a Dual-Hop Generative Adversarial Network (DH-GAN) at the pixel level and then apply a smoothing-based graph optimization to the pixel-wise segmentation to determine a best-covering road graph. Wu *et al.* [33] employed a U-Net encoder-decoder architecture with loss layers at different scales to progressively refine the segmentation masks. Xu *et al.* [34] proposed a cascaded approach in which pre-processed hand-crafted features are fed into a Residual U-Net to extract building locations and a guided filter refines the results.

In an effort to address the problem of poor boundary predictions by CNN models, Bischke *et al.* [3] proposed a cascaded multi-task loss function to simultaneously predict the semantic masks and distance classes. Recently, Rudner *et al.* [28] proposed a method to segment flooded buildings using multiple streams of encoder-decoder architectures that extract spatiotemporal information from medium-resolution images and spatial information from high-resolution images along with a context aggregation module to effectively combine the learned feature map.

2.2 CNN/ACM Hybrid Models

Hu *et al.* [18] proposed a model in which the network learns a level-set function for salient objects; however, the authors predefined a fixed scalar weighting parameter λ , which will not be optimal for all cases in the analyzed set of images. Hatamizadeh *et al.* [14] connected the output of a CNN to an implicit ACM through spatially-varying functions for the λ parameters. Le *et al.* [22] proposed a framework for the task of semantic segmentation of natural images in which level-set ACMs are implemented as RNNs. There are three key differences between that effort and our TDAC: (1) TDAC does not reformulate ACMs as RNNs, which makes it more computationally efficient. (2) TDAC benefits from a novel, locally-parameterized energy functional, as opposed to constant weighted parameters (3) TDAC has an entirely different pipeline—we employ a single CNN that is trained from scratch along with the ACM, as opposed to requiring two pre-trained CNN backbones. The dependence of [22] on pre-trained CNNs limits its applicability.

Marcos *et al.* [25] proposed Deep Structured Active Contours (DSAC), an integration of ACMs with CNNs in a structured prediction framework for building instance segmentation in aerial images. There are three key differences between that work and our TDAC: (1) TDAC is fully automated and runs without any external supervision, as opposed to depending heavily on the manual initialization of contours. (2) TDAC leverages the Eulerian ACM, which naturally segments multiple building instances simultaneously, as opposed to a Lagrangian formulation that can handle only a single building at a time. (3) Our approach fully automates the direct back-propagation of gradients through the entire TDAC framework due to its automatically differentiable ACM implementation.

Cheng *et al.* [9] proposed the Deep Active Ray Network (DarNet), which uses a polar coordinate ACM formulation to prevent the problem of self-intersection and employs a computationally expensive multiple initialization scheme to improve the performance of the proposed model. Like DSAC, DarNet can handle only single instances of buildings due to its explicit ACM formulation. Our approach is fundamentally different from DarNet, as (1) it uses an implicit ACM formulation that handles multiple building instances and (2) leverages a CNN to automatically and precisely initialize the implicit ACM.

Wang *et al.* [32] proposed an interactive object annotation framework for instance segmentation in which a backbone CNN and user input guide the evolution of an implicit ACM. Recently, Gur *et al.* [12] introduced Active Contours via Differentiable Rendering Network (ACDRNet) in which an explicit ACM is represented by a “neural renderer” and a backbone encoder-decoder U-Net predicts a shift map to evolve the contour via edge displacement.

Some efforts have also focused on deriving new loss functions that are inspired by ACM principles. Inspired by the global energy formulation of [5], Chen *et al.* [8] proposed a supervised loss layer that incorporated area and size information of the predicted masks during training of a CNN and tackled a medical image segmentation task. Similarly, Gur *et al.* [13] presented an unsupervised loss function based on morphological active contours without edges [26].

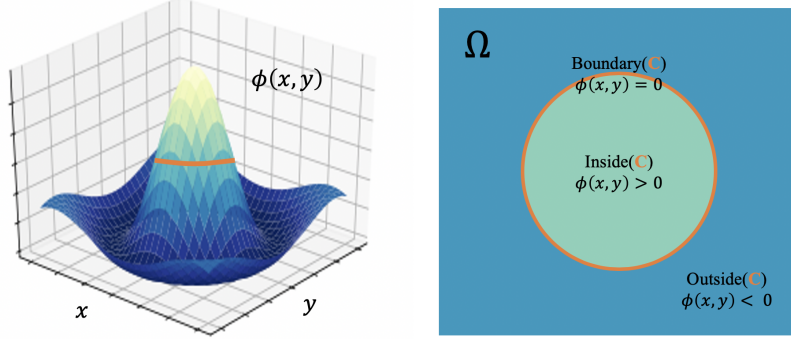


Fig. 2: Boundary C represented as the zero level-set of implicit function $\phi(x, y)$.

3 The TDAC Model

3.1 Localized Level-Set ACM with Trainable Parameters

Our ACM formulation allows us to create a differentiable and trainable active contour model. Instead of working with a parametric contour that encloses the desired area to be segmented [19], we represent the contour(s) as the zero level-set of an implicit function. Such so-called “level-set active contours” evolve the segmentation boundary by evolving the implicit function so as to minimize an associated Eulerian energy functional.

The most well-known approaches that utilize this implicit formulation are geodesic active contours [4] and active contours without edges [5]. The latter, also known as the Chan-Vese model, relies on image intensity differences between the interior and exterior regions of the level set. Lankton and Tannenbaum [21] proposed a reformulation in which the energy functional incorporates image properties in the local region near the level set, which more accurately segments objects with heterogeneous features.¹

Let I represent an input image and $C = \{(x, y) | \phi(x, y) = 0\}$ be a closed contour in $\Omega \in \mathbb{R}^2$ represented by the zero level set of the signed distance map $\phi(x, y)$ (Fig. 2). The interior and exterior of C are represented by $\phi(x, y) > 0$ and $\phi(x, y) < 0$, respectively. Following [5], we use a smoothed Heaviside function

$$H(\phi(x, y)) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{\phi(x, y)}{\epsilon}\right) \quad (1)$$

¹ These approaches numerically solve the PDE that governs the evolution of the implicit function. Interestingly, Márquez-Neila *et al.* [26] proposed a morphological approach that approximates the numerical solution of the PDE by successive application of morphological operators defined on the equivalent binary level set.

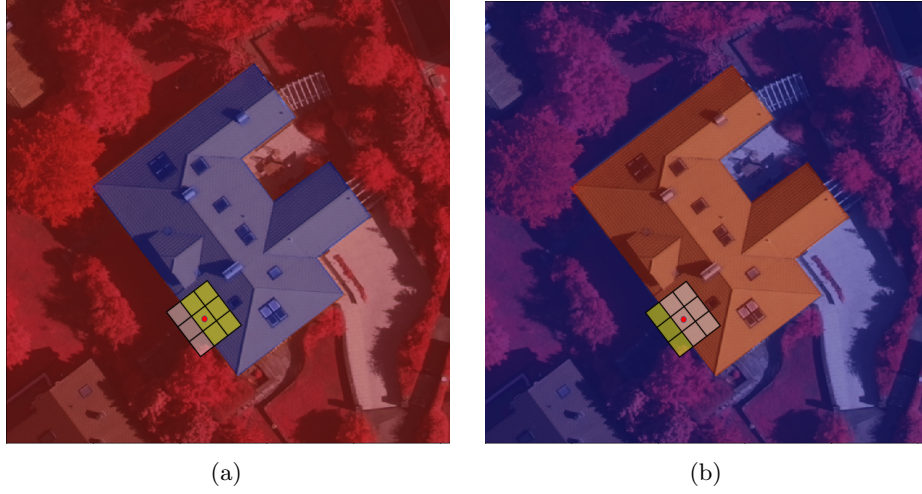


Fig. 3: The filter is divided by the contour into interior and exterior regions. The point x is represented by the red dot and the interior (a) and exterior (b) regions are shaded yellow.

to represent the interior as $H(\phi)$ and exterior as $(1 - H(\phi))$. The derivative of $H(\phi(x, y))$ is

$$\frac{\partial H(\phi(x, y))}{\partial \phi(x, y)} = \frac{1}{\pi} \frac{\epsilon}{\epsilon^2 + \phi(x, y)^2} = \delta(\phi(x, y)). \quad (2)$$

In TDAC, we evolve C to minimize an energy function according to

$$E(\phi) = E_{\text{length}}(\phi) + E_{\text{image}}(\phi), \quad (3)$$

where

$$E_{\text{length}}(\phi) = \int_{\Omega} \mu \delta(\phi(x, y)) |\nabla \phi(x, y)| dx dy \quad (4)$$

penalizes the length of C whereas

$$E_{\text{image}}(\phi) = \int_{\Omega} \delta(\phi(x, y)) \left[H(\phi(x, y)) (I(x, y) - m_1)^2 + (1 - H(\phi(x, y))) (I(x, y) - m_2)^2 \right] dx dy \quad (5)$$

takes into account the mean image intensities m_1 and m_2 of the regions interior and exterior to C [5]. We compute these local statistics using a characteristic function W_s with local window of size f_s (Fig. 3), as follows:

$$W_s = \begin{cases} 1 & \text{if } x - f_s \leq u \leq x + f_s, \quad y - f_s \leq v \leq y + f_s; \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

where x, y and u, v are the coordinates of two independent points.

To make our level-set ACM trainable, we associate parameter maps with the foreground and background energies. These maps, $\lambda_1(x, y)$ and $\lambda_2(x, y)$, are functions over the image domain Ω . Therefore, our energy function may be written as

$$E(\phi) = \int_{\Omega} \delta(\phi(x, y)) \left[\mu |\nabla \phi(x, y)| + \int_{\Omega} W_s F(\phi(u, v)) du dv \right] dx dy, \quad (7)$$

where

$$\begin{aligned} F(\phi) = & \lambda_1(x, y)(I(u, v) - m_1(x, y))^2(H(\phi(x, y)) \\ & + \lambda_2(x, y)(I(u, v) - m_2(x, y))^2(1 - H(\phi(x, y))). \end{aligned} \quad (8)$$

The variational derivative of E with respect to ϕ yields the Euler-Lagrange PDE²

$$\frac{\partial \phi}{\partial t} = \delta(\phi) \left[\mu \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) + \int_{\Omega} W_s \nabla_{\phi} F(\phi) dx dy \right] \quad (9)$$

with

$$\nabla_{\phi} F = \delta(\phi) (\lambda_1(x, y)(I(u, v) - m_1(x, y))^2 - \lambda_2(x, y)(I(u, v) - m_2(x, y))^2). \quad (10)$$

To avoid numerical instabilities during the evolution and maintain a well-behaved $\phi(x, y)$, a distance regularization term [23] can be added to (9).

It is important to note that our formulation enables us to capture the fine-grained details of boundaries, and our use of *pixel-wise* parameter maps $\lambda_1(x, y)$ and $\lambda_2(x, y)$ allows them to be directly predicted by the backbone CNN along with an initialization map $\phi_0(x, y)$. Thus, not only does the implicit ACM propagation now become fully automated, but it can also be directly controlled by a CNN through these learnable parameter maps.

3.2 CNN Backbone

For the backbone CNN, we use a standard encoder-decoder with convolutional layers, residual blocks, and skip connections between the encoder and decoder. Each 3×3 convolutional layer is followed by ReLU activation and batch normalization. Each residual block consists of two 3×3 convolutional layers and an additive identity skip connection. As illustrated in Fig. 4, the first stage of the encoder comprises two 3×3 convolutional layers and a max pooling operation. Its second and third stages are comprised of a residual block followed by a max pooling operation. Each stage of the decoder performs a bilinear upsampling followed by two convolutional layers. The encoder is connected to the decoder via three residual blocks as well as skip connections at every stage. The output of the decoder is connected to a 1×1 convolution with three output channels for predicting the $\lambda_1(x, y)$ and $\lambda_2(x, y)$ parameter maps as well as the initialization map $\phi_0(x, y)$.³

² The derivation is found in the supplemental document.

³ Additional details about the CNN backbone are found in the supplemental document.

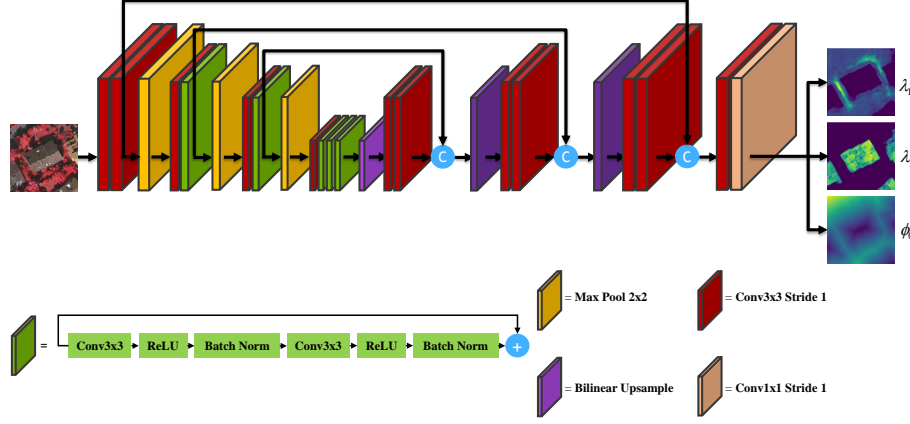


Fig. 4: TDAC’s CNN backbone employs a standard encoder-decoder architecture.

3.3 Differentiable ACM

The ACM is evolved according to (9) in a differentiable manner in TensorFlow. The first term is computed according to the surface curvature expression:

$$\operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) = \frac{\phi_{xx}\phi_y^2 - 2\phi_{xy}\phi_x\phi_y + \phi_{yy}\phi_x^2}{(\phi_x^2 + \phi_y^2)^{3/2}}, \quad (11)$$

where the subscripts denote the spatial partial derivatives of ϕ , which are approximated using central finite differences. For the second term, convolutional operations are leveraged to efficiently compute $m_1(x, y)$ and $m_2(x, y)$ in (8) within image regions interior and exterior to C . Finally, $\partial\phi/\partial t$ in (9) is evaluated and $\phi(x, y)$ updated according to

$$\phi^t = \phi^{t-1} + \Delta t \frac{\partial \phi^{t-1}}{\partial t}, \quad (12)$$

where Δt is the size of the time step.

3.4 TDAC Training

Referring to Fig. 1, we simultaneously train the CNN and level-set components of TDAC in an end-to-end manner with no human intervention. The CNN guides the ACM by predicting the $\lambda_1(x, y)$ and $\lambda_2(x, y)$ parameter maps, as well as an initialization map $\phi_0(x, y)$ from which $\phi(x, y)$ evolves through the L layers of the ACM in a differentiable manner, thus enabling training error backpropagation. The $\phi_0(x, y)$ output of the CNN is also passed into a Sigmoid activation function to produce the prediction P . Training optimizes a loss function that combines

Algorithm 1: TDAC Training Algorithm

Data: I : Image; G : Corresponding ground truth label; g : ACM energy function with parameter maps λ_1, λ_2 ; ϕ : ACM implicit function; L : Number of ACM iterations; W : CNN with weights w ; P : CNN prediction; \mathcal{L} : Total loss function; η : Learning rate

Result: Trained TDAC model

```

while not converged do
     $\lambda_1, \lambda_2, \phi_0 = W(I)$ 
     $P = \text{Sigmoid}(\phi_0)$ 
    for  $t = 1$  to  $L$  do
         $\frac{\partial \phi_{t-1}}{\partial t} = g(\phi_{t-1}; \lambda_1, \lambda_2, I)$ 
         $\phi^t = \phi^{t-1} + \Delta t \frac{\partial \phi^{t-1}}{\partial t}$ 
    end
     $\mathcal{L} = \mathcal{L}_{\text{ACM}}(\phi_L) + \mathcal{L}_{\text{CNN}}(P)$ 
    Compute  $\frac{\partial \mathcal{L}}{\partial w}$  and backpropagate the error
    Update the weights of  $W$ :  $w \leftarrow w - \eta \frac{\partial \mathcal{L}}{\partial w}$ 
end

```

binary cross entropy and Dice losses:

$$\hat{\mathcal{L}}(X) = -\frac{1}{N} \sum_{j=1}^N [X_j \log G_j + (1 - X_j) \log(1 - G_j)] + 1 - \frac{\sum_{j=1}^N 2X_j G_j}{\sum_{j=1}^N X_j + \sum_{j=1}^N G_j}, \quad (13)$$

where X_j denotes the output prediction and G_j the corresponding ground truth at pixel j , and N is the total number of pixels in the image. The total loss of the TDAC model is

$$\mathcal{L} = \mathcal{L}_{\text{ACM}} + \mathcal{L}_{\text{CNN}}, \quad (14)$$

where $\mathcal{L}_{\text{ACM}} = \hat{\mathcal{L}}(\phi_L)$ is the loss computed for the output ϕ_L from the final ACM layer and $\mathcal{L}_{\text{CNN}} = \hat{\mathcal{L}}(P)$ is the loss computed over the prediction P of the backbone CNN. Algorithm 1 presents the details of the TDAC training procedure.

3.5 Implementation Details

We have implemented the TDAC architecture and training algorithm entirely in TensorFlow. Our ACM implementation benefits from the automatic differentiation utility of TensorFlow and has been designed to enable the backpropagation of the error gradient through the L layers of the ACM. We set $L = 60$ iterations in the ACM component of TDAC since, as will be discussed in Section 4.3, the performance does not seem to improve significantly with additional iterations. We set a filter size of $f = 5$, as discussed in Section 4.3. The training was performed on an Nvidia Titan RTX GPU, and an Intel® Core™ i7-7700K CPU @ 4.20GHz. The size of the training minibatches for both datasets is 2. All the training sessions employ the Adam optimization algorithm [20] with an initial

learning rate of $\alpha_0 = 0.001$ decreasing according to [27]

$$\alpha = \alpha_0 (1 - e/N_e)^{0.9} \quad (15)$$

with epoch counter e and total number of epochs N_e .

4 Empirical Study

4.1 Datasets

Vaihingen: The Vaihingen buildings dataset consists of 168 aerial images of size 512×512 pixels. Labels for each image were generated by using a semi-automated approach. We used 100 images for training and 68 for testing, following the same data partition as in [25]. In this dataset, almost all the images include multiple instances of buildings, some of which are located at image borders.

Bing Huts: The Bing Huts dataset consists of 605 aerial images of size 64×64 pixels. We followed the same data partition used in [25], employing 335 images for training and 270 images for testing. This dataset is especially challenging due to the low spatial resolution and contrast of the images.

4.2 Evaluation Metrics

To evaluate TDAC’s performance, we utilized four different metrics—Dice, mean Intersection over Union (mIoU), Boundary F (BoundF) [9], and Weighted Coverage (WCov) [30].

Given the prediction X and ground truth mask G , the Dice (F1) score is

$$\text{Dice}(X, G) = \frac{2 \sum_{i=1}^N X_i G_i}{\sum_{i=1}^N X_i + \sum_{i=1}^N G_i}, \quad (16)$$

where N is the number of image pixels and G_i and X_i denote pixels in G and X .

Similarly, the IoU score measures the overlap of two objects by calculating the ratio of intersection over union, according to

$$\text{IoU}(X, G) = \frac{|X \cap G|}{|X \cup G|}. \quad (17)$$

BoundF computes the average of Dice scores over 1 to 5 pixels around the boundaries of the ground truth segmentation.

In WConv, the maximum overlap output is selected and the IoU between the ground truth segmentation and best output is calculated. IoUs for all instances are summed up and weighted by the area of the ground truth instance. Assuming that $S_G = \{r_1^{S_G}, \dots, r_{|S_G|}^{S_G}\}$ is a set of ground truth regions and $S_X = \{r_1^{S_X}, \dots, r_{|S_X|}^{S_X}\}$ is a set of prediction regions for single image, and $|r_j^{S_G}|$ is the number of pixels in $r_j^{S_G}$, the weighted coverage can be expressed as

$$\text{WCov}(S_X, S_G) = \frac{1}{N} \sum_{j=1}^{|S_G|} |r_j^{S_G}| \max_{k=1 \dots |S_X|} \text{IoU}(r_k^{S_X}, r_j^{S_G}). \quad (18)$$

Table 1: Model Evaluations: Single-Instance Segmentation.

Model		Vaihingen				Bing Huts			
Method	Backbone	Dice	mIoU	WCov	BoundF	Dice	mIoU	WCov	BoundF
FCN	ResNet	84.20	75.60	77.50	38.30	79.90	68.40	76.14	39.19
FCN	Mask R-CNN	86.00	76.36	81.55	36.80	77.51	65.03	76.02	65.25
FCN	UNet	87.40	78.60	81.80	40.20	77.20	64.90	75.70	41.27
FCN	Ours	90.02	81.10	82.01	44.53	82.24	74.09	73.67	42.04
FCN	DSAC	–	81.00	81.40	64.60	–	69.80	73.60	30.30
FCN	DarNet	–	87.20	86.80	76.80	–	74.50	77.50	37.70
DSAC	DSAC	–	71.10	70.70	36.40	–	38.70	44.60	37.10
DSAC	DarNet	–	60.30	61.10	24.30	–	57.20	63.00	15.90
DarNet	DarNet	93.66	88.20	88.10	75.90	85.21	75.20	77.00	38.00
TDAC-const λ s	Ours	91.18	83.79	82.70	73.21	84.53	73.02	74.21	48.25
TDAC	Ours	94.26	89.16	90.54	78.12	89.12	80.39	81.05	53.50

Table 2: Model Evaluations: Multiple-Instance Segmentation.

Model		Vaihingen				Bing Huts			
Method	Backbone	Dice	mIoU	WCov	BoundF	Dice	mIoU	WCov	BoundF
FCN	UNet	81.00	69.10	72.40	34.20	71.58	58.70	65.70	40.60
FCN	ResNet	80.10	67.80	70.50	32.50	74.20	61.80	66.59	39.48
FCN	Mask R-CNN	88.35	79.42	80.26	41.92	76.12	63.40	70.51	41.97
FCN	Ours	89.30	81.00	82.70	49.80	75.23	60.31	72.41	41.12
TDAC-const λ s	Ours	90.80	83.30	83.90	47.20	81.19	68.34	75.29	44.61
TDAC	Ours	95.20	91.10	91.71	69.02	83.24	71.30	78.45	48.49

4.3 Experiments and Ablation Studies

Single-Instance Segmentation: Although most of the images in the Vaihingen dataset depict multiple instances of buildings, the DarNet and DSAC models can deal only with a single building instance at a time. For a fair comparison against these models, we report single-instance segmentation results in the exact same manner as [25] and [9]. As reported in Table 1, our TDAC model outperforms both DarNet and DSAC in all metrics on both the Vaihingen and Bing Huts datasets. Fig. 5 shows that with the Vaihingen dataset, both the DarNet and DSAC models have difficulty coping with the topological changes of the buildings and fail to appropriately capture sharp edges, while TDAC overcomes these challenges in most cases. For the Bing Huts dataset, both the DarNet and DSAC models are able to localize the buildings, but they inaccurately delineate the buildings in many cases. This may be due to their inability to distinguish the building from the surrounding terrain because of the low contrast and small size of the image. Comparing the segmentation output of DSAC (Fig. 5b), DarNet (Fig. 5c), and TDAC (Fig. 5d), our model performs well on the low contrast dataset, delineating buildings more accurately than the earlier models.

Multiple-Instance Segmentation: We next compare the performance of TDAC against popular models such as Mask R-CNN for multiple-instance segmentation

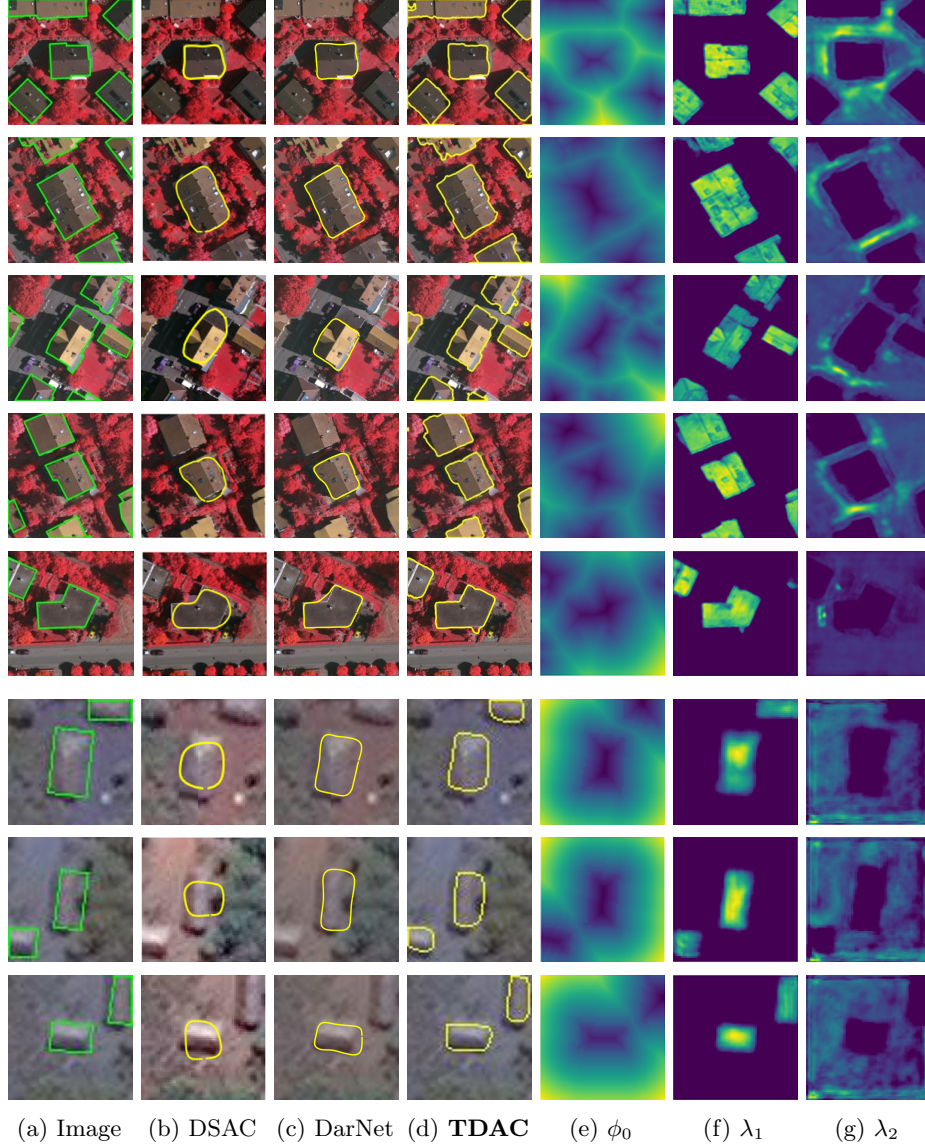


Fig. 5: Comparative visualization of the labeled image and the outputs of DSAC, DarNet, and our TDAC for the Vaihingen (top) and Bing Huts (bottom) datasets. (a) Image labeled with (green) ground truth segmentation. (b) DSAC output. (c) DarNet output. (d) TDAC output. (e) TDAC’s learned initialization map $\phi_0(x, y)$ and parameter maps (f) $\lambda_1(x, y)$ and (g) $\lambda_2(x, y)$.

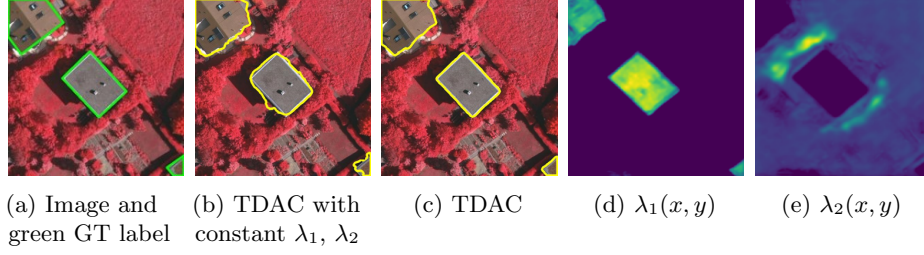


Fig. 6: (a) Image labeled with (green) ground truth segmentation. (b) Output of TDAC with constant λ_1 and λ_2 . (c) TDAC output and learned parameter maps (d) $\lambda_1(x, y)$ and (e) $\lambda_2(x, y)$.

of all buildings in the Vaihingen and Bing Huts datasets. As reported in Table 2, our extensive benchmarks confirm that the TDAC model outperforms Mask R-CNN and the other methods by a wide margin. Although Mask R-CNN seems to be able to localize the building instances well, the fine-grained details of boundaries are lost, as is attested by the BoundF metric. The performance of other CNN-based approaches follow the same trend in our benchmarks.

Parameter Maps: To validate the contribution of the parameter maps $\lambda_1(x, y)$ and $\lambda_2(x, y)$ in the level-set ACM, we also trained our TDAC model on both the Vaihingen and Bing Huts datasets by allowing just two trainable scalar parameters, λ_1 and λ_2 , constant over the entire image. As reported in Table 1, for both the Vaihingen and Bing Huts datasets, this “constant- λ ” formulation (i.e., the Chan-Vese model [5,21]) still outperforms the baseline CNN in most evaluation metrics for both single-instance and multiple-instance buildings, thus establishing the effectiveness of the end-to-end training of the TDAC. Nevertheless, our TDAC with its full $\lambda_1(x, y)$ and $\lambda_2(x, y)$ maps outperforms this constant- λ version by a wide margin in all experiments and metrics. A key metric of interest in this comparison is the BoundF score, which elucidates that our formulation captures the details of the boundaries more effectively by locally adjusting the inward and outward forces on the contour. Fig. 6 shows that our TDAC has well delineated the boundaries of the building instances, compared to the TDAC hobbled by the constant- λ formulation.

Convolutional Filter Size: The filter size of the convolutional operation is an important hyper-parameter for the accurate extraction of localized image statistics. As illustrated in Fig. 7a, we have investigated the effect of the convolutional filter size on the overall mIoU for both the Vaihingen and Bing datasets. Our experiments indicate that filter sizes that are too small are sub-optimal while excessively large sizes defeat the benefits of the localized formulation. Hence, we recommend a filter size of $f = 5$ for the TDAC.

Number of Iterations: The direct learning of an initialization map $\phi_0(x, y)$ as well as its efficient TensorFlow implementation have enabled the TDAC to

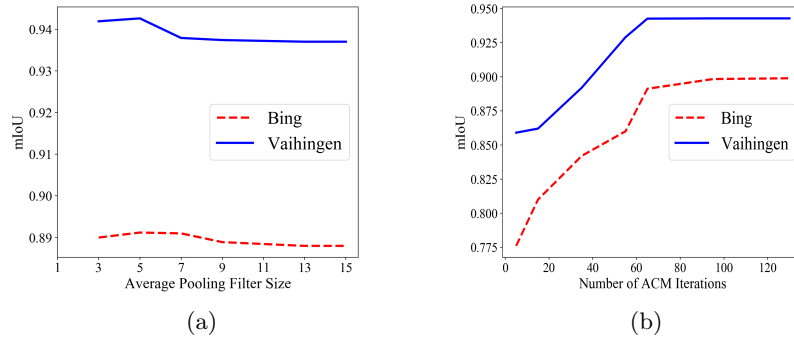


Fig. 7: The effects on mIoU of (a) varying the convolutional filter size and (b) varying the number L of ACM iterations.

require substantially fewer iterations to converge with a better chance of avoiding undesirable local minima. As shown in Fig. 7b, we have investigated the effect of the number of iterations on the overall mIoU for both Vaihingen and Bing datasets and our results reveal that TDAC exhibits a robust performance after a certain threshold. Therefore, we have chosen a fixed number of iterations (i.e., ACM layers) for optimal performance, $L = 60$, yielding a runtime of less than 1 sec in TensorFlow.

5 Conclusions and Future Work

We have introduced a novel image segmentation framework, called Trainable Deep Active Contour Models (TDACs), which is a full, end-to-end merger of ACMs and CNNs. To this end, we proposed a new, locally-parameterized, Eulerian ACM energy model that includes pixel-wise learnable parameter maps that can adjust the contour to precisely delineate the boundaries of objects of interest in the image. Our model is fully automatic, as its backbone CNN learns the ACM initialization map as well as the parameter maps that guide the contour to avoid suboptimal solutions. This eliminates any reliance on manual initialization of ACMs. Moreover, by contrast to previous approaches that have attempted to combine CNNs with ACMs that are limited to segmenting a single building at a time, our TDAC can segment any number of buildings simultaneously.

We have tackled the problem of building instance segmentation on two challenging datasets, Vaihingen and Bing Huts, and our model significantly outperforms the current state-of-the-art methods on these test cases.

Given the level of success that TDAC has achieved in the building delineation application and the fact that it features an Eulerian ACM formulation, it is readily applicable to other segmentation tasks in various domains, wherever purely CNN filter-based approaches can benefit from the versatility and precision of ACMs to accurately delineate object boundaries in images.

References

1. Audebert, N., Le Saux, B., Lefèvre, S.: Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In: Asian Conference on Computer Vision. pp. 180–196. Springer (2016)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(12), 2481–2495 (2017)
3. Bischke, B., Helber, P., Folz, J., Borth, D., Dengel, A.: Multi-task learning for segmentation of building footprints with deep neural networks. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1480–1484. IEEE (2019)
4. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. *International Journal of Computer Vision* **22**(1), 61–79 (1997)
5. Chan, T.F., Vese, L.A.: Active contours without edges. *IEEE Transactions on Image Processing* **10**(2), 266–277 (2001)
6. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4974–4983 (2019)
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(4), 834–848 (2018)
8. Chen, X., Williams, B.M., Vallabhaneni, S.R., Czanner, G., Williams, R., Zheng, Y.: Learning active contour models for medical image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11632–11640 (2019)
9. Cheng, D., Liao, R., Fidler, S., Urtasun, R.: Darnet: Deep active ray network for building segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7431–7439 (2019)
10. Costea, D., Marcu, A., Slusanschi, E., Leordeanu, M.: Creating roadmaps in aerial images with generative adversarial networks and smoothing-based optimization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2017)
11. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (ICLR) (2019)
12. Gur, S., Shaharabany, T., Wolf, L.: End to end trainable active contours via differentiable rendering. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019)
13. Gur, S., Wolf, L., Golgher, L., Blinder, P.: Unsupervised microvascular image segmentation using an active contours mimicking neural network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10722–10731 (2019)
14. Hatamizadeh, A., Hoogi, A., Sengupta, D., Lu, W., Wilcox, B., Rubin, D., Terzopoulos, D.: Deep active lesion segmentation. In: International Workshop on Machine Learning in Medical Imaging. pp. 98–105. Springer (2019)
15. Hatamizadeh, A., Sengupta, D., Terzopoulos, D.: End-to-end deep convolutional active contours for image segmentation. ArXiv preprint ArXiv:1909.13359 (2019)

16. Hatamizadeh, A., Terzopoulos, D., Myronenko, A.: End-to-end boundary aware networks for medical image segmentation. In: International Workshop on Machine Learning in Medical Imaging. pp. 187–194. Springer (2019)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
18. Hu, P., Shuai, B., Liu, J., Wang, G.: Deep level sets for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
19. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. *International Journal of Computer Vision* **1**(4), 321–331 (1988)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Lankton, S., Tannenbaum, A.: Localizing region-based active contours. *IEEE Transactions on Image Processing* **17**(11), 2029–2039 (2008)
22. Le, T.H.N., Quach, K.G., Luu, K., Duong, C.N., Savvides, M.: Reformulating level sets as deep recurrent neural network approach to semantic segmentation. *IEEE Transactions on Image Processing* **27**(5), 2393–2407 (2018)
23. Li, C., Xu, C., Gui, C., Fox, M.D.: Distance regularized level set evolution and its application to image segmentation. *IEEE Transactions on Image Processing* **19**(12), 3243 (2010)
24. Lillesand, T., Kiefer, R.W., Chipman, J.: Remote Sensing and Image Interpretation. John Wiley & Sons (2015)
25. Marcos, D., Tuia, D., Kellenberger, B., Zhang, L., Bai, M., Liao, R., Urtasun, R.: 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)
26. Marquez-Neila, P., Baumela, L., Alvarez, L.: A morphological approach to curvature-based evolution of curves and surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(1), 2–17 (2013)
27. Myronenko, A., Hatamizadeh, A.: Robust semantic segmentation of brain tumor regions from 3D MRIs. In: International MICCAI Brainlesion Workshop. pp. 82–89. Springer (2019)
28. Rudner, T.G., Rußwurm, M., Fil, J., Pelich, R., Bischke, B., Kopačková, V., Biliński, P.: Multi3Net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 702–709 (2019)
29. Shrivastava, N., Rai, P.K.: Remote-sensing the urban area: Automatic building extraction based on multiresolution segmentation and classification. *Geografia: Malaysian Journal of Society and Space* **11**(2) (2017)
30. Silberman, N., Sontag, D., Fergus, R.: Instance segmentation of indoor scenes using a coverage loss. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *Computer Vision – ECCV 2014*. pp. 616–631. Springer International Publishing, Cham (2014)
31. Wang, S., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Chaverie, J., Fidler, S., Urtasun, R.: Torontocity: Seeing the world with a million eyes. arXiv preprint arXiv:1612.00423 (2016)
32. Wang, Z., Acuna, D., Ling, H., Kar, A., Fidler, S.: Object instance annotation with deep extreme level set evolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7500–7508 (2019)

33. Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y., Shibasaki, R.: Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sensing* **10**(3), 407 (2018)
34. Xu, Y., Wu, L., Xie, Z., Chen, Z.: Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sensing* **10**(1), 144 (2018)
35. Zhang, P., Ke, Y., Zhang, Z., Wang, M., Li, P., Zhang, S.: Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery. *Sensors* **18**(11), 3717 (2018)