

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107

# Skeleton-Aware Distance Transform for Structure-Preserving Instance Segmentation

Anonymous ICCV submission

Paper ID 6638

## Abstract

Objects with complex structures pose significant challenges to existing instance segmentation methods that rely on boundary or affinity maps, which are vulnerable to small errors around contacting pixels that cause noticeable connectivity change. While the distance transform (DT) makes instance interiors and boundaries more distinguishable, it tends to overlook the intra-object connectivity for instances with varying width and result in over-segmentation. To address these challenges, we propose a skeleton-aware distance transform (SDT) that combines the merits of object skeleton in preserving connectivity and DT in modeling geometric arrangement to represent instances with arbitrary structures. Comprehensive experiments on histopathology image segmentation demonstrate the state-of-the-art performance achieved by SDT. Further, we show that, with a reformulation of curvilinear structure delineation as its dual problem of instance segmentation, SDT also compares favorably against existing methods on two benchmark datasets. Our code will be publicly available.

## 1. Introduction

Instances with complex shapes arise in many application domains, and their morphology carries critical information. For example, the structure of gland tissues in microscopy images is essential in accessing the pathological stages for cancer diagnosis and treatment. These instances, however, are usually closely in touch with each other and have non-convex structures with parts of varying width, posing significant challenges for existing segmentation methods.

Instance segmentation approaches can be categorized as either top-down or bottom-up ones. *Top-down* approaches (*e.g.*, Mask R-CNN [7]) first localize candidate instances with bounding boxes and then predict a binary mask within each box. Although those methods achieve leading performance on natural image benchmarks [6, 13], for biomedical data where the instances have irregular shapes with various

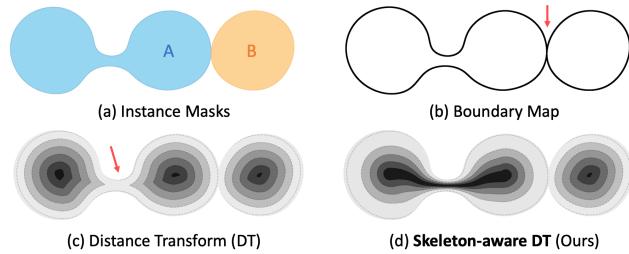


Figure 1: Skeleton-aware distance transform (SDT). Given (a) instance masks, (b) the boundary map is prone to false merge errors at object contact pixels while (c) the distance transform (DT) struggles to preserve object connectivity. (d) Our SDT can both separate touching instances and enforce object connectivity.

aspect ratios, object localization, and subsequent segmentation in the top-down paradigm tends to fail [11].

Thus, in the biomedical domain, most methods [3, 24, 4, 33, 22] follow the *bottom-up* methodology that first learns intermediate representations and then convert them into masks with standard segmentation algorithms like connected-component labeling and watershed transform. These representations are not only efficient to predict in one model forward pass but also able to capture object **geometry** (*i.e.*, precise instance boundary), which are hard for top-down methods using low-resolution features for mask generation. However, existing representations have several restrictions. For example, boundary map (or affinity map) is usually learned as a pixel-wise binary classification task, which makes the model conduct relatively local predictions and consequently become vulnerable to small errors that break the connectivity between adjacent instances (Fig. 1b). To improve the boundary map, Deep Watershed Transform (DWT) [1] predicts the Euclidean distance transform (DT) of each pixel to the instance boundary. This representation is more aware of the structure for convex objects, as the energy value for centers is significantly different from pixels close to the boundary. However, for objects with non-convex morphology (Fig. 1c), the boundary-based distance transform produces multiple local optima in the en-

	Boundary	DT	Skeleton	<b>SDT</b>
Geometry		✓		✓
Connectivity			✓	✓

Table 1: Strengths of mask representations in instance segmentation. Our SDT combines skeleton with distance transform (DT) to preserve both the geometry and connectivity of instances.

ergy landscape, which tends to break the intra-instance connectivity when applying thresholding and standard segmentation algorithms, resulting in over-segmentation.

To preserve the **connectivity** of instances while keeping the precise instance boundary, in this paper, we propose a novel representation named *skeleton-aware* distance transform (SDT). Our SDT incorporate object skeleton, a concise and connectivity-preserving representation of object structure, into the traditional boundary-based distance transform (DT) (Fig. 1d). Although object skeleton extraction has received recent attention in the vision community [26, 10, 14, 29], integrating skeleton into instance segmentation models is rarely explored. For single-object foreground segmentation, Shen *et al.* [26] jointly predicts object skeleton and the *scale* of each skeleton pixel (the diameter of the maximal disk centered at it). The mask is computed as the union of the disks centered at the predicted skeleton. However, the highly quantized scales tend to generate masks without accurate boundary localization. In comparison, our SDT produces a smooth energy landscape for each instance, with a constant highest value on skeleton pixels and the lowest value on boundary pixels. Such a design combines the merits of DT and the skeleton to capture both the geometry and topological connectivity of instances (Table 1). Similar to previous bottom-up instance segmentation methods, we learn the SDT energy map end-to-end with a fully convolutional network (FCN) model and process it into instances with watershed transform.

In quantitative evaluations, we show that our proposed SDT achieves leading performance on histopathology image segmentation for instances with various sizes and complex structures. Specifically, under the Hausdorff distance for evaluating shape similarity, our approach improves the previous state-of-the-art method by relatively 10.6%. To further demonstrate SDT’s capacity to model complex shapes, we apply it to the delineation of curvilinear structures, *e.g.*, road extraction from satellite imagery. Particularly, we reformulate the original pixel-wise binary classification task as the dual problem of instance segmentation for the background regions. We show that SDT not only outperforms previous state-of-the-art approaches but also compares favorably against DT after the reformulation.

In summary, this work makes three main contributions. First, we introduce a novel representation, *skeleton-*

*aware* DT, to capture both the geometry and connectivity of objects in instance segmentation. Second, our method achieves state-of-the-art performance on histopathology image segmentation under several metrics. Third, we apply SDT to the curvilinear structure delineation task that was not considered as instance segmentation in the literature and show that SDT also outperforms leading approaches.

## 2. Related Work

**Instance Segmentation.** Bottom-up instance segmentation approaches have become de facto for many biomedical applications due to the advantage in segmenting objects with arbitrary geometry. U-Net [24] and DCAN [4] use fully convolutional models to predict the boundary map of instances. Since the boundary map is not robust to small errors that can significantly change instance structure, shape-preserving loss [33] adds a curve fitting step in the loss function to enforce boundary connectivity. In order to further distinguish closely touching instances belong to the same semantic group, deep watershed transform (DWT) [1] predicts the distance transform (DT) representation that represents each pixel as its distance to the closest boundary. However, for complex structure with parts of varying width, the boundary-based DT tends to produce relatively low values for thin connections and consequently causes over-segmentation. In comparison with DWT, our SDT incorporates object skeleton (also known as medial axis) [2, 35, 12] that concisely captures the topological connectivity into traditional DT to enforce both the geometry and connectivity. Besides, previous works [33, 22] conduct patch-based sliding-window prediction during inference. Our SDT framework instead processes whole images to infer the complete structure of instances.

**Object Skeletonization.** Object skeleton [25] is a one-pixel wide representation of object masks that can be calculated by topological thinning [35, 12, 21] or medial axis transform [2]. The vision community has been working on direct object skeletonization from images [26, 10, 14, 29]. Among the works, only Shen *et al.* [26] shows the application of the skeleton on segmenting single-object images. We instead focus on the more challenging instance segmentation task with multiple objects closely touching each other. Object skeletons are also used to correct errors in pre-computed segmentation masks [16]. We instead use the skeleton in the direct segmentation from images.

**Curvilinear Structures Delineation.** Extracting curvilinear structures, including cracks on pavements and roads in satellite imagery, is a traditional vision task. Recent learning-based methods [19, 24] formulate it as pixel-wise classification and optimize the cross-entropy loss with deep models. However, the target structure is so thin that few pixel-level errors can break the prediction into parts with

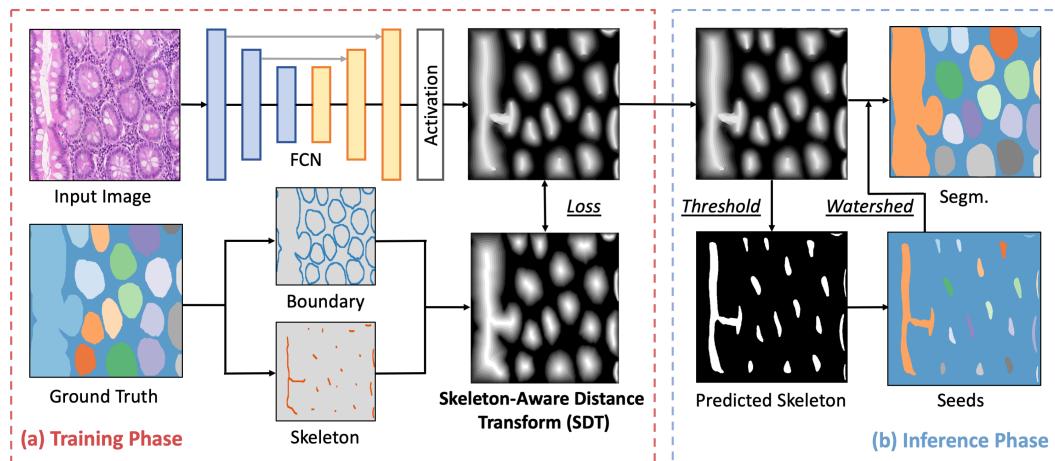


Figure 2: Overview of the SDT framework. **(a) Training Phase:** for a given input image with ground truth masks, the target SDT map is calculated conditioned on the distance to both the instance boundary and skeleton. A fully convolutional network (e.g., DeeplabV3 [5]) maps the image into the energy space to minimize the loss. **(b) Inference Phase:** we apply thresholding of the predicted SDT to generate skeleton segments, which is processed into seeds with the connected component labeling. Finally, the standard watershed transform algorithm takes the seeds and the reversed SDT energy to yield the segmentation masks.

negligible loss change. Thus, topology-aware losses [20, 9] are proposed to penalize connectivity errors in the prediction. Alternatively, we regard the task as the dual problem of instance segmentation. Specifically, we conduct segmentation for the background regions and take the instance boundaries as the final prediction. To our best knowledge, there have been no attempts from previous instance segmentation works. Capable of complex shape modeling, our SDT also achieves leading performance on this task.

### 3. Skeleton-Aware Distance Transform

In this section, we introduce the problem setup (Sec. 3.1), define the SDT energy function (Sec. 3.2) and describe important implementation choices (Sec. 3.3). An overview of the SDT framework is shown in Figure 2.

### 3.1. Problem Setup

Given the input image, we aim to design a new representation  $E$  for a deep model to learn so that the prediction  $\hat{E}$  can be decoded into instance masks with simple post-processing. Specifically, a good representation for capturing complex-structure instances should have two desired properties: *precise geometric boundary* and *robust topological connectivity*. However, existing representations for instance segmentation can hardly satisfy both properties.

We now let  $\Omega$  denote the region of an instance mask, and  $\Gamma_b$  be the boundary of the instance (pixels with other object indices in a small local neighborhood). The *boundary* (or affinity) map is a binary representation where  $E|_{\Gamma_b} = 0$  and  $E|_{\Omega \setminus \Gamma_b} = 1^1$ . Since boundary pixels only occupy a

small fraction of  $\Omega$ , few mispredictions on  $\Gamma_b$  will incur a negligible increase in the pixel-wise binary cross-entropy loss but can substantially change the connectivity between instances. To better separate adjacent instances, *distance transform* (DT) keeps  $E|_{\Gamma_b} = 0$  but represent other pixels in  $\Omega$  based on their distances to  $\Gamma_b$ . Without loss of generality, we assume  $E$  is normalized by the largest distance to the boundary. Since DT can better distinguish object center from pixels close to the boundary, previous work has indicated the improved performance for segmenting convex objects like cars even when they are closely in touch with each other [1]. However,  $E \approx 0$  for relatively thin parts in some objects (Fig. 1c), making the representation vulnerable to over-splits. The requirement of a higher threshold to separate adjacent instances and a lower threshold to keep connectivity causes inconsistency in segmentation.

Another related representation is *skeleton with associated scales* (SS), which generates masks by computing the union of disks centered at skeleton pixels (denoted as  $\Gamma_s$ ) [26].  $E|_{\Gamma_s} = 1$  ensures that skeleton pixel values are identical regardless of the distance to the boundary, which can better keep connectivity between object regions. However, joint learning of skeleton pixels and their scales is challenging, as minor errors in the predicted scales can yield masks without accurate boundary localization.

Taking the merits of DT in modeling the geometric arrangement and skeleton in preserving instance connectivity, we propose to design a new representation  $E$  that satisfies:

$$0 = E|_{\Gamma_b} < E|_{\Omega \setminus (\Gamma_b \cup \Gamma_s)} < E|_{\Gamma_s} = 1 \quad (1)$$

Here  $E|_{\Omega \setminus \Gamma_s} < E|_{\Gamma_s} = 1$  indicates that there is only one consistent with the formulation of other mask representations.

<sup>1</sup>It is identical to use the inverted values. We set  $E|_{\Gamma_b} = 0$  to keep

324 global maximum value for each instance, and the value is  
 325 assigned to a pixel if and only if the pixel is on the object  
 326 skeleton. This property avoids ambiguity in defining  
 327 the object interior and preserve connectivity. Besides,  
 328  $E|_{\Omega \setminus \Gamma_b} > E|_{\Gamma_b} = 0$  ensures that boundary is distinguishable  
 329 as the standard DT, which helps produce precise geometric  
 330 boundary. In this next part, we introduce an efficient  
 331 realization of the new representation  $E$  and describe  
 332 the learning strategy with a fully convolutional model.  
 333

### 334 3.2. SDT Energy Function

335 One way to find the desired energy function is to solve  
 336 a partial differential equation (PDE) that meets the conditions  
 337 as described in Eqn. 1. However, solving such an equation  
 338 can be computationally expensive and numerically  
 339 unstable, which is not ideal in practice. Instead, we  
 340 construct the desired energy function with the basic distance  
 341 transform morphological operation. Let  $x$  be a pixel in  
 342 the input image, and  $d$  be the metric, e.g., Euclidean distance.  
 343 The energy function for distance transform (DT) is  
 344 defined as  $E_{DT}(x) = d(x, \Gamma_b)$ , which starts from 0 at  
 345 object boundary and increases monotonically when  $x$  is away  
 346 from the boundary. Similarly, we can define an energy function  
 347  $d(x, \Gamma_s)$  representing the distance from the skeleton. It  
 348 vanishes to 0 when the pixel approaches the object skeleton.  
 349

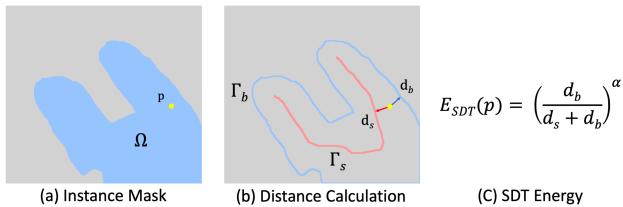
350 The key operation of our realization is to normalize  $E_{DT}$   
 351 by the sum of  $E_{DT}$  and  $d(x, \Gamma_s)$ . Then the normalized  $E_{DT}$   
 352 becomes a bounded continuous function, which still vanishes  
 353 to 0 at object boundaries as the original  $E_{DT}$ , but  
 354 becomes 1 at object skeletons as  $d(x, \Gamma_s)$  goes to zero.  
 355 This operation can be regarded as the interpolation between  
 356 the boundary-based distance transform and skeleton-based  
 357 transform, which satisfies both desired properties discussed  
 358 before (Fig. 3). Formally, we define the energy function of  
 359 the *skeleton-aware* distance transform as

$$E_{SDT}(x) = \left( \frac{d(x, \Gamma_b)}{d(x, \Gamma_s) + d(x, \Gamma_b)} \right)^\alpha, \quad \alpha > 0 \quad (2)$$

360 where  $\alpha$  controls the curvature of the energy surface<sup>2</sup>.  
 361 When  $0 < \alpha < 1$ , the function is concave and decreases  
 362 faster when being close to the boundary, and vice versa  
 363 when  $\alpha > 1$ . In the ablation studies, we demonstrate various  
 364 patterns of the model predictions given different  $\alpha$ .

365 **Learning Strategy.** Given the ground-truth SDT energy  
 366 map, there are two ways to learn it using a fully convolutional  
 367 model. The first way is to *regress* the energy map  
 368 using  $L_1$  or  $L_2$  loss. In the regression mode, the output  
 369 is a single-channel image. The second way is to quantize  
 370 the  $[0, 1]$  energy space into  $K$  bins and rephrase the  
 371 regression task into a *classification* task [1, 28], which makes the  
 372

373 <sup>2</sup>We add  $\epsilon = 10^{-6}$  to the denominator to avoid dividing by 0 for the  
 374 edge case where a pixel belongs to both instance boundary and skeleton  
 375 (*i.e.*, a one-pixel wide part of the mask).



376 Figure 3: Illustration of the SDT energy function. (a) Given an in-  
 377 stance mask  $\Omega$ , (b) we calculate the distances of a pixel to both the  
 378 skeleton and boundary. (c) Our energy function ensures a uniform  
 379 maximum value of 1 on the skeleton and minimum value of 0 on  
 380 the boundary, with a smooth interpolation in between.

381 model robust to small perturbations in the energy landscape.  
 382 For the classification mode, the model output has  $(K + 1)$   
 383 channels with one channel representing the background re-  
 384 gion. We fix the bin size to 0.1 without tweaking, making  
 385  $K = 10$ . A softmax activation is applied before calculating  
 386 the cross-entropy loss. We test both learning strategies in  
 387 the experiments to illustrate the optimal setting for SDT .  
 388

### 389 3.3. Implementation

390 **Network Architecture.** Directly learning the energy func-  
 391 tion with a fully convolutional network (FCN) can be chal-  
 392 lenging. Previous approaches either first regress an easier  
 393 direction field and then use additional layers to predict the  
 394 desired target [1], or take the multi-task learning approach  
 395 to predict additional targets [32, 33, 26].

396 Fortunately, with recent progress in FCN architectures,  
 397 it becomes feasible to learn the target energy map in an  
 398 end-to-end fashion. Specifically, in all the experiments, we  
 399 use a DeepLabV3 model [5] with a ResNet [8] backbone  
 400 to directly learn the SDT energy without additional targets  
 401 (Fig. 2, *Training Phase*). We also add a CoordConv [15]  
 402 layer before the 3rd stage in the backbone network to intro-  
 403 duce spatial information into the segmentation model.  
 404

405 **Target SDT Generation.** There is an inconsistency prob-  
 406 lem in object skeleton generation: part of the complete in-  
 407 stance skeleton can be different from the skeleton of the in-  
 408 stance part (Fig. 4). For an input image, some objects may  
 409 touch the image border due to either a restricted field of  
 410 view (FoV) of the imaging devices or spatial data aug-  
 411 mentation like the random crop. If pre-computing the skeleton,  
 412 we will get *local skeleton* (Fig. 4c) for objects with miss-  
 413 ing masks due to imaging restrictions, and *partial skeleton*  
 414 (Fig. 4b) due to spatial data augmentation, which causes  
 415 ambiguity. Therefore we calculate the local skeleton for  
 416 SDT on-the-fly after all spatial transformations instead of  
 417 pre-processing the data to prevent the model from hallu-  
 418 cinating the structure of missing objects parts out of the  
 419 currently visible region. At inference time, we always run  
 420 predictions on the whole images to avoid inconsistent pre-  
 421

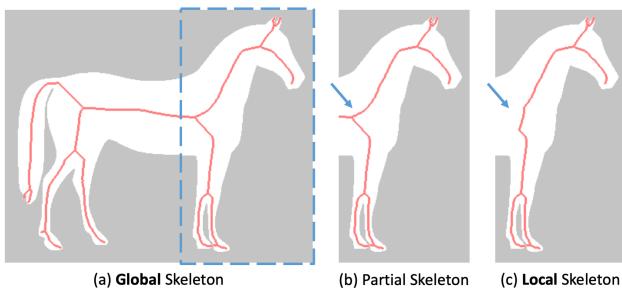


Figure 4: Skeleton generation rule. (a) Given an instance mask and the *global* skeleton, (b) the *partial* skeleton cropped from the global skeleton can be different from (c) the *local* skeleton generated from the cropped mask. For SDT, we calculate the local skeleton to prevent the model from extrapolating the unseen parts. The horse silhouette image was downloaded from *openclipart*.

dictions. We use the skeletonization algorithm in Lee *et al.* [12], which is less sensitive to small perturbations and produce skeletons with fewer branches.

**Instance Extraction from SDT.** In the SDT energy map, all boundary pixels share the same energy value and can be processed into segments by direct thresholding and connected component labeling, similar to DWT [1]. However, since the prediction is never perfect, the energy values along closely touching boundaries are usually not sharp and cause split-errors when applying a higher threshold or merge-errors when applying a lower threshold.

Therefore we utilize a skeleton-aware instance extraction (Fig. 2, *Inference Phase*) for SDT. Specifically, we set a threshold  $\theta = 0.7$  so that all pixels with the predicted energy bigger than  $\theta$  are labeled as skeleton pixels. We first perform connected component labeling of the skeleton pixels to generate seeds and run the watershed algorithm on the reversed energy map using the seeds as basins (local optima) to generate the final segmentation. We also follow previous works [4, 33] and refine the segmentation by hole-filling and removing small spurious objects.

## 4. Experiments

### 4.1. Histopathology Instance Segmentation

Accurate instance segmentation of gland tissues in histopathology images is essential for clinical analysis, especially cancer diagnosis. The diversity of object appearance, size, and shape makes the task challenging.

**Dataset and Evaluation Metric.** We use the *gland segmentation challenge dataset* [27] that contains colored light microscopy images of tissues with a wide range of histological levels from benign to malignant. There are 85 and 80 images in the training and test set, respectively, with ground truth annotations provided by pathologists. According to the challenge protocol, the test set is further divided into two

splits with 60 images of normal and 20 images of abnormal tissues for evaluation. Three evaluation criteria used in the challenge include instance-level F1 score, Dice index, and Hausdorff distance, which measure the performance of object detection, segmentation, and shape similarity, respectively. For the instance-level F1 score, an IoU threshold of 0.5 is used to decide the correctness of a prediction.

**Methods in Comparison.** We compare SDT with previous state-of-the-art methods for since the release of the dataset in 2015, including DCAN [4], multi-channel network (MCN) [32], shape-preserving loss (SPL) [33] and FullNet [22]. We also compare with suggestive annotation (SA) [34], and SA with model quantization (QSA) [31], which use multiple FCN models to select informative training samples from the dataset. With the same training settings as our SDT, we also report the performance of skeleton with scales (SS) and traditional distance transform (DT).

**Training and Inference.** Since the training data is relatively limited due to the challenges in collecting medical images, we apply pixel-level and spatial-level augmentations, including random brightness, contrast, rotation, crop, and elastic transformation, to alleviate overfitting. We set  $\alpha = 0.8$  for our SDT in Eqn. 2. We use the classification learning strategy and optimize a model with 11 output channels (10 channels for energy quantized into ten bins and one channel for background). We train the model for 20k iterations with an initial learning rate of  $5 \times 10^{-4}$  and a momentum of 0.9. The same settings are applied to DT. At inference time, we apply argmax to get the corresponding bin index of each pixel and transform the energy value to the original data range. Finally, we apply the watershed-based instance extraction rule described in Sec. 3.3.

Specifically for SS, we set the number of output channels to two, with one channel predicting skeleton probability and the other predicting scales. Since the scales are non-negative, we add a ReLU activation for the second channel and calculate the regression loss. Masks are generated by morphological dilation. We do not quantize the scales as DT and SDT since even ground-truth scales can yield masks unaligned with the instance boundary with quantization.

**Results.** Our SDT framework achieves state-of-the-art performance on 5 out of 6 evaluation metrics on the gland segmentation dataset (Table 2). With the better distinguishability of object interior and boundary, our approach can unambiguously separate closely touching instances (Fig. 5, first two rows), performs better than previous methods that use object boundary representations [4, 33]. Besides, under the Hausdorff distance for evaluating shape-similarity between ground-truth and predicted instance masks, our SDT reports an average score of 44.82 across two test splits, which improves the previous state-of-the-art approach (*i.e.*, FullNet with an average score of 50.15) by 10.6%. We also no-

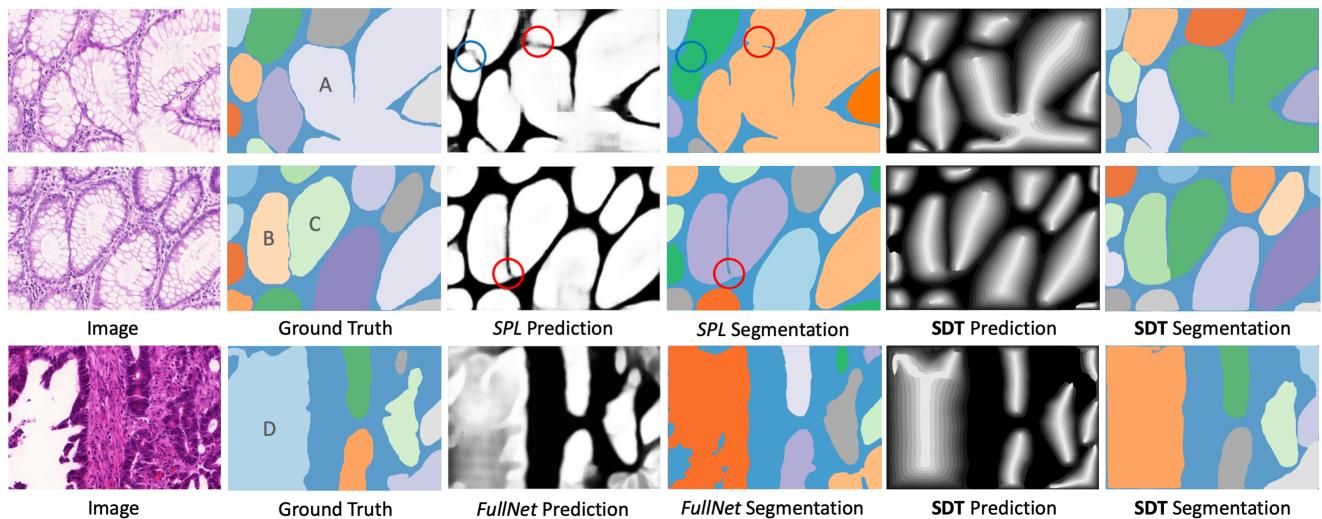


Figure 5: Visual comparison with state-of-the-art methods on histopathology image segmentation. (First 2 rows) Compared with shape-preserving loss (SPL) [33], our SDT unambiguously separates closely touching objects while preserving the structure of complicated masks. (The 3rd row) Compared with FullNet [22], our model infers the SDT energy of instance masks from a global structure perspective instead of boundary that relies on relatively local predictions, which produces high-quality instance masks for challenging cases.

561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647

tice the different sensitivities of the three evaluation metrics. Taking the instance **D** (Fig. 5, 3rd row) as an example: both SDT and FullNet [22] have 1.0 F1-score (IoU threshold 0.5) for the correct detection; SDT has a slightly higher Dice Index (0.956 vs. 0.931) for better pixel-level classification; and our SDT has significantly lower Hausdorff distance (24.41 vs. 48.81) as SDT yields a mask with much more accurate morphology.

Our SDT also compares favorably against DT under the same training settings. In a direct visual comparison between DT and SDT, we show that SDT can better enforce object connectivity because of the constant high energy value on the object skeleton (Fig. 6). In addition, we also notice that generating unbroken skeletons in the SS model with pixel-wise binary classification is quite challenging, which leads to significantly worse results (Table 2).

## 4.2. Ablation Studies

We further perform ablation studies on several important design choices of SDT. All other model hyper-parameters and the evaluation metrics are identical to Sec. 4.1.

**Loss Function.** We compare the regression mode using L1 and L2 loss with the classification mode using cross-entropy loss. There is a separate channel for background under the classification mode where the energy values are quantized into bins. However, for regression mode, if the background value is 0, then we need to use a threshold  $\tau > 0$  to decide the foreground region, which results in shrunk masks. To separate the background region from the foreground objects, we assign an energy value of  $-b$  to the background pixels ( $b \geq 0$ ). To facilitate the regression, given the pre-

Method	F1 Score $\uparrow$		Dice Index $\uparrow$		Hausdorff $\downarrow$	
	Part A	Part B	Part A	Part B	Part A	Part B
DCAN [4]	0.912	0.716	0.897	0.781	45.42	160.35
MCN [32]	0.893	0.843	0.908	0.833	44.13	116.82
SPL [33]	0.924	0.844	0.902	0.840	49.88	106.08
SA [34]	0.921	0.855	0.904	0.858	44.74	96.98
FullNet [22]	0.924	0.853	0.914	0.856	37.28	88.75
QSA [31]	0.930	0.862	0.914	<b>0.859</b>	41.78	97.39
SS	0.872	0.765	0.853	0.797	54.86	116.33
DT	0.918	0.846	0.896	0.848	41.84	90.86
<b>SDT (Ours)</b>	<b>0.931</b>	<b>0.866</b>	<b>0.919</b>	0.851	<b>32.29</b>	<b>82.40</b>

Table 2: Comparison with state-of-the-art methods on the gland segmentation challenge dataset. Our SDT achieves better or on par F1 score and Dice Index, and significantly better Hausdorff distance for evaluating *shape similarity*. DT and SS represent distance transform and skeleton with scales, respectively.

dicted value  $\hat{y}_i$  for pixel  $i$ , we apply a sigmoid function ( $\sigma$ ) and affine transformation so that:

$$\hat{y}'_i = (1 + b) \cdot \sigma(\hat{y}_i) - b \quad (3)$$

has a range of  $(-b, 1)$ . We set  $b = 0.1$  for the experiments.

We show that under the same settings, the model trained with cross-entropy loss for quantized energy reports the best results (Table 3). We also notice that the model trained with  $L_1$  loss produces a much sharper energy surface than the model trained with  $L_2$  loss, which is expected.

**Curvature.** We also compare different  $\alpha$  in Eqn. 2 that controls the curvature of the energy landscape. Table 3 shows that  $\alpha = 0.8$  achieves the best overall performance, which

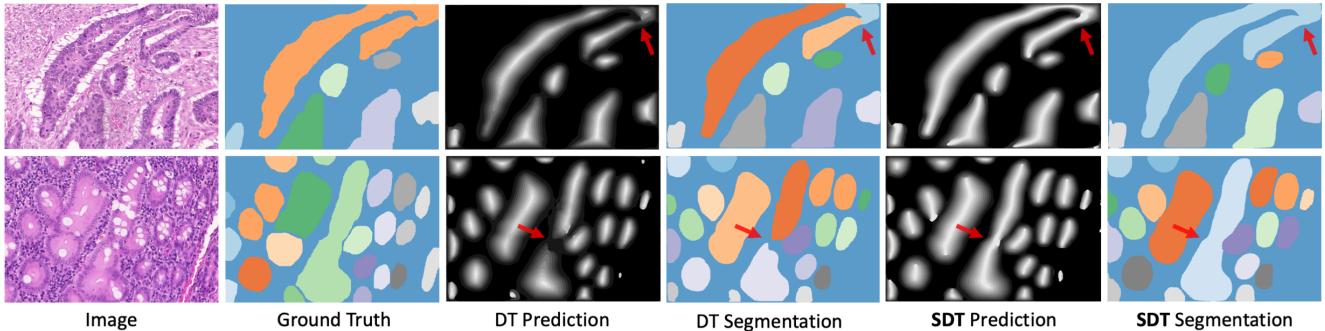


Figure 6: Direct comparison between DT and SDT. We show that using the same model and under exactly the same training settings, our proposed SDT can better preserve the instance connectivity than DT thanks to the *skeleton-aware* mask representation.

Setting	F1 Score ↑		Dice Index ↑		Hausdorff ↓	
	Part A	Part B	Part A	Part B	Part A	Part B
<i>Loss</i>						
L1	0.916	0.842	0.903	0.850	39.76	94.83
L2	0.896	0.833	0.885	0.837	49.11	110.24
CE	0.931	0.866	0.919	0.851	32.29	82.40
<i>Curvature</i>						
$\alpha = 0.6$	0.912	0.845	0.914	0.855	36.25	91.24
$\alpha = 0.8$	0.931	0.866	0.919	0.851	32.29	82.40
$\alpha = 1.0$	0.926	0.858	0.907	0.849	35.73	86.73
<i>Skeleton</i>						
Partial	0.899	0.831	0.896	0.837	47.50	105.19
Local	0.931	0.866	0.919	0.851	32.29	82.40

Table 3: Ablations studies on the gland dataset. The results suggest that the model trained with cross-entropy loss with  $\alpha = 0.8$  and local skeleton generation achieves the best performance.

is slightly better than  $\alpha = 1.0$ . Decreasing  $\alpha$  to 0.6 introduces more merges and make the results worse.

**Global/Local Skeleton Generation.** In Sec. 3.3 we describe the inconsistency problem of the global and local skeletons. In this study, we set  $\alpha = 0.8$  and let the model take pre-computed SDT energy for the training set during training. The results show that pre-computed SDT energy leads to significantly worse performance (Table 3). We argue this is because pre-computed energy not only introduces inconsistency for instances touching the image border but also restricts the diversity of SDT energy patterns.

### 4.3. Curvilinear Structure Delineation

Besides instance segmentation, our SDT representation can also be seamlessly applied to curvilinear structure delineation. Previous methods commonly regard it as a foreground segmentation task and let the network output a binary mask. Instead, we proposed to reformulate the delineation of curvilinear structures as the dual problem for the

instance segmentation of background regions (Fig. 7).

**Datasets.** We evaluate our SDT on two datasets. The first dataset is the crack delineation dataset [36] (**Crack**) that contains images of pavement cracks. The dataset is split into 104 training and 20 test images. Effective methods on the dataset can be adapted to other applications like the quality inspection of products. The second dataset is the Massachusetts Roads Dataset [18] (**Road**) that contains aerial road images for both urban and rural regions. The dataset is split into 1108 training and 49 test images, covering a wide range of road types and background appearance.

**Evaluation Metric.** We use two sets of evaluation criteria following recent works [20, 9]. One set of metrics are *Correctness* (Corr.), *Completeness* (Comp.), and *Quality* (Qual.) [30], which can be regarded as the relaxed precision, recall, and IoU, respectively. Those metrics are more robust to predictions with small shifting and width changes, which are still reasonable in representing the structures but underestimated by traditional pixel-wise metrics. According to Mosinska *et al.* [20], we use a threshold of 2 pixels.

Another set of metrics are Adapted Rand Index (ARI) and Variation of Information (VOI), which are recently used to evaluate the topological correctness of curvilinear structures [9]. ARI is defined as the maximal F-score of the foreground-restricted Rand index [23], a measure of similarity between two clusters where the background pixels of the original labels are excluded. VOI is another measure of the distance between two clusterings [17], which is a true metric obeying triangle inequality.

**Methods in Comparison.** Besides earlier methods including CrackTree [36] for the crack dataset and MNIH [19] for the road dataset, we compare our SDT with U-Net [24] that predicts the binary mask of curvilinear structures, Mosinska *et al.* [20] that uses perceptual loss in the VGG feature space and iterative refinement, as well as TopoNet [9] that add a differentiable topological loss during training. In addition, we also compare with the DT representation after

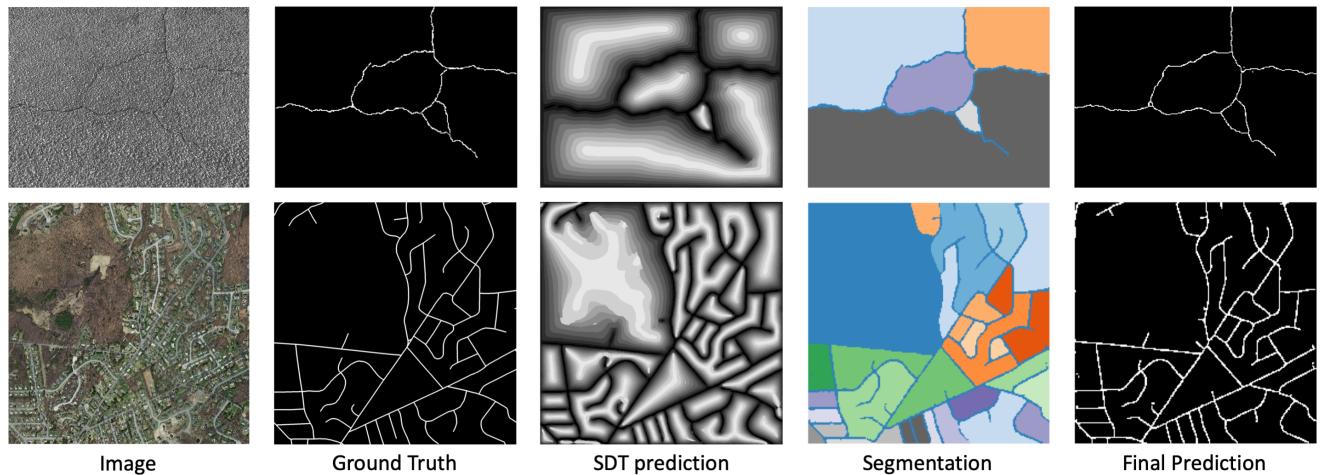


Figure 7: Curvilinear structure delineation as instance segmentation. Given an image, we predict the SDT energy and decode it into background region instances. The predicted instance boundaries become the desired curvilinear structures..

Method	Corr. $\uparrow$	Comp. $\uparrow$	Qual. $\uparrow$	ARI $\uparrow$	VOI $\downarrow$
CrackTree [36]	0.79	0.92	0.74	-	-
U-Net [24]	0.41	0.89	0.39	0.87	1.63
Mosin. [20]	0.88	<b>0.95</b>	0.85	0.89	1.11
TopoNet [9]	-	-	-	0.93	1.00
DT	0.91	0.92	0.84	<b>0.95</b>	1.01
<b>SDT (Ours)</b>	<b>0.96</b>	0.93	<b>0.89</b>	<b>0.95</b>	<b>0.81</b>

(a) Crack Dataset [36]

Method	Corr. $\uparrow$	Comp. $\uparrow$	Qual. $\uparrow$	ARI $\uparrow$	VOI $\downarrow$
MNIH [19]	0.53	0.75	0.45	-	-
U-Net [24]	0.62	0.75	0.51	0.82	2.25
Mosin. [20]	0.77	0.81	0.65	0.85	1.46
TopoNet [9]	-	-	-	<b>0.87</b>	1.23
DT	0.76	<b>0.82</b>	0.65	0.86	1.19
<b>SDT (Ours)</b>	<b>0.86</b>	0.81	<b>0.72</b>	0.85	<b>1.13</b>

(b) Road Dataset [18]

Table 4: Quantitative comparison on the curvilinear structure delineation task. Our methods outperform or are on par with previous methods on all the evaluation metrics on both the Crack and Road datasets.

reformulating the task into instance segmentation.

**Implementation Details.** The Crack dataset contains images of size of  $600 \times 800$ , and the Road dataset contains larger images of size  $1500 \times 1500$  pixels. We conduct training and inference on images resized to  $641 \times 641$  pixels. Following the settings on the gland dataset, we train the model with an initial learning rate of  $5 \times 10^{-4}$  and a momentum of 0.9. We train the model for the relatively small Crack dataset for 15k iterations and 40k iterations for the Road dataset. Please note that although the cracks and roads may not induce closed background regions, we follow Sec. 3.3 and Fig. 4 to regard image borders as instance boundaries to prevent the model from extrapolating unseen regions.

**Results.** We conduct quantitative comparisons on the two datasets in Table 4. Our SDT achieves state-of-the-art performance on 8 out of all ten evaluation metrics across two datasets. Such results indicate that the motivation of reformulating the delineation of curvilinear structures as the dual problem for the instance segmentation of background regions is a sensible choice for this task. We argue that the benefits of reformulation come from the more informative learning targets. In the previous learning settings, fore-

ground pixels occupy only a small fraction of the images, and vast background pixels are treated equally. Furthermore, since SDT is capable of modeling complex structures without ambiguity in ensuring connectivity, it also outperforms the DT baseline after the reformulation.

**Analysis of Limitation.** Although SDT works well in preserving connectivity, one limitation is its inability to handle objects divided into unconnected components (e.g., due to occlusion). This limitation does not affect the results on the histopathology image segmentation and curvilinear structure delineation, but for future application in such cases, a linking step is required to group disconnected regions.

## 5. Conclusion

In this paper, we introduce the *skeleton-aware* distance transform (SDT) to capture both the geometry and topological connectivity of instance masks with complex shapes. We hope this work can inspire more research on not only better representations of object masks but also novel models that can better predict those representations with shape encoding. We will also explore the application of SDT in the more challenging 3D instance segmentation setting.

864

## References

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

- [1] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5221–5229, 2017. 1, 2, 3, 4, 5
- [2] Harry Blum et al. *A transformation for extracting new descriptors of shape*, volume 4. MIT press Cambridge, 1967. 2
- [3] Kevin Briggman, Winfried Denk, Sebastian Seung, Moritz N Helmstaedter, and Srinivas C Turaga. Maximin affinity learning of image segmentation. In *Advances in Neural Information Processing Systems*, pages 1865–1873, 2009. 1
- [4] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016. 1, 2, 5, 6
- [5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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, 2017. 3, 4
- [6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1
- [7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [9] Xiaoling Hu, Li Fuxin, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. In *Advances in Neural Information Processing Systems*, 2019. 3, 7, 8
- [10] Wei Ke, Jie Chen, Jianbin Jiao, Guoying Zhao, and Qixiang Ye. Srn: side-output residual network for object symmetry detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1068–1076, 2017. 2
- [11] Victor Kulikov and Victor Lempitsky. Instance segmentation of biological images using harmonic embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3843–3851, 2020. 1
- [12] Ta-Chih Lee, Rangasami L Kashyap, and Chong-Nam Chu. Building skeleton models via 3-d medial surface axis thinning algorithms. *CVGIP: Graphical Models and Image Processing*, 56(6):462–478, 1994. 2, 5
- [13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

- [14] Chang Liu, Wei Ke, Fei Qin, and Qixiang Ye. Linear span network for object skeleton detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 133–148, 2018. 2
- [15] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems*, pages 9605–9616, 2018. 4
- [16] Brian Matejek, Daniel Haehn, Haidong Zhu, Donglai Wei, Toufiq Parag, and Hanspeter Pfister. Biologically-constrained graphs for global connectomics reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2
- [17] Marina Meilă. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pages 173–187. Springer, 2003. 7
- [18] Volodymyr Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013. 7, 8
- [19] Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th International conference on machine learning (ICML-12)*, pages 567–574, 2012. 2, 7, 8
- [20] Agata Mosinska, Pablo Marquez-Neila, Mateusz Koziński, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2018. 3, 7, 8
- [21] Gábor Németh, Péter Kardos, and Kálmán Palágyi. 2d parallel thinning and shrinking based on sufficient conditions for topology preservation. *Acta Cybernetica*, 20(1):125–144, Jan. 2011. 2
- [22] Hui Qu, Zhennan Yan, Gregory M Riedlinger, Subhajyoti De, and Dimitris N Metaxas. Improving nuclei/gland instance segmentation in histopathology images by full resolution neural network and spatial constrained loss. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 378–386. Springer, 2019. 1, 2, 5, 6
- [23] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971. 7
- [24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 2, 7, 8
- [25] Punam K Saha, Gunilla Borgefors, and Gabriella Sanniti di Baja. A survey on skeletonization algorithms and their applications. *Pattern recognition letters*, 76:3–12, 2016. 2
- [26] Wei Shen, Kai Zhao, Yuan Jiang, Yan Wang, Xiang Bai, and Alan Yuille. Deepskelton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images. *IEEE Transactions on Image Processing*, 26(11):5298–5311, 2017. 2, 3, 4
- [27] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, 918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

- 972 Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. 1026  
973 Gland segmentation in colon histology images: The glas 1027  
974 challenge contest. *Medical image analysis*, 35:489–502, 1028  
975 2017. 5 1029  
976 [28] Yan Wang, Xu Wei, Fengze Liu, Jieneng Chen, Yuyin Zhou, 1030  
977 Wei Shen, Elliot K Fishman, and Alan L Yuille. Deep 1031  
978 distance transform for tubular structure segmentation in ct 1032  
979 scans. In *Proceedings of the IEEE/CVF Conference on Com- 1033  
980 puter Vision and Pattern Recognition*, pages 3833–3842, 1034  
981 2020. 4 1035  
982 [29] Yukang Wang, Yongchao Xu, Stavros Tsogkas, Xiang Bai, 1036  
983 Sven Dickinson, and Kaleem Siddiqi. Deepflux for skele- 1037  
984 tons in the wild. In *Proceedings of the IEEE Conference on 1038  
985 Computer Vision and Pattern Recognition*, pages 5287– 1039  
986 5296, 2019. 2 1040  
987 [30] Christian Wiedemann, Christian Heipke, Helmut Mayer, and 1041  
988 Olivier Jamet. Empirical evaluation of automatically ex- 1042  
989 tracted road axes. *Empirical evaluation techniques in com- 1043  
990 puter vision*, pages 172–187, 1998. 7 1044  
991 [31] Xiaowei Xu, Qing Lu, Lin Yang, Sharon Hu, Danny Chen, 1045  
992 Yu Hu, and Yiyu Shi. Quantization of fully convolutional 1046  
993 networks for accurate biomedical image segmentation. In 1047  
994 *Proceedings of the IEEE conference on computer vision and 1048  
995 pattern recognition*, pages 8300–8308, 2018. 5, 6 1049  
996 [32] Yan Xu, Yang Li, Yipei Wang, Mingyuan Liu, Yubo 1050  
997 Fan, Maode Lai, I Eric, and Chao Chang. Gland instance 1051  
998 segmentation using deep multichannel neural networks. *IEEE Transactions on Biomedical Engineering*, 1052  
999 64(12):2901–2912, 2017. 4, 5, 6 1053  
1000 [33] Zengqiang Yan, Xin Yang, and Kwang-Ting Tim Cheng. A 1054  
1001 deep model with shape-preserving loss for gland instance 1055  
1002 segmentation. In *International Conference on Medical Im- 1056  
1003 age Computing and Computer-Assisted Intervention*, pages 1057  
1004 138–146. Springer, 2018. 1, 2, 4, 5, 6 1058  
1005 [34] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and 1059  
1006 Danny Z Chen. Suggestive annotation: A deep active learn- 1060  
1007 ing framework for biomedical image segmentation. In *Interna- 1061  
1008 tional conference on medical image computing and 1062  
1009 computer-assisted intervention*, pages 399–407. Springer, 2017. 5, 6 1063  
1010 [35] TY Zhang and Ching Y. Suen. A fast parallel algorithm 1064  
1011 for thinning digital patterns. *Communications of the ACM*, 1065  
1012 27(3):236–239, 1984. 2 1066  
1013 [36] Qin Zou, Yu Cao, Qingquan Li, Qingzhou Mao, and Song 1067  
1014 Wang. Cracktree: Automatic crack detection from pavement 1068  
1015 images. *Pattern Recognition Letters*, 2012. 7, 8 1069  
1016 1070  
1017 1071  
1018 1072  
1019 1073  
1020 1074  
1021 1075  
1022 1076  
1023 1077  
1024 1078  
1025 1079