

Fast Interactive Object Annotation with Curve-GCN

Huan Ling^{1,2*}

Jun Gao^{1,2*}

Amlan Kar^{1,2}

Wenzheng Chen^{1,2}

Sanja Fidler^{1,2,3}

¹University of Toronto

²Vector Institute

³NVIDIA

{linghuan, jungao, amlan, wenzheng, fidler}@cs.toronto.edu

Abstract

Manually labeling objects by tracing their boundaries is a laborious process. In [7, 2], the authors proposed Polygon-RNN that produces polygonal annotations in a recurrent manner using a CNN-RNN architecture, allowing interactive correction via humans-in-the-loop. We propose a new framework that alleviates the sequential nature of Polygon-RNN, by predicting all vertices simultaneously using a Graph Convolutional Network (GCN). Our model is trained end-to-end. It supports object annotation by either polygons or splines, facilitating labeling efficiency for both line-based and curved objects. We show that Curve-GCN outperforms all existing approaches in automatic mode, including the powerful PSP-DeepLab [8, 23] and is significantly more efficient in interactive mode than Polygon-RNN++. Our model runs at 29.3ms in automatic, and 2.6ms in interactive mode, making it 10x and 100x faster than Polygon-RNN++.

1. Introduction

Object instance segmentation is the problem of outlining all objects of a given class in an image, a task that has been receiving increased attention in the past few years [15, 36, 20, 3, 21]. Current approaches are all data hungry, and benefit from large annotated datasets for training. However, manually tracing object boundaries is a laborious process, taking up to 40sec per object [2, 9]. To alleviate this problem, a number of interactive image segmentation techniques have been proposed [28, 23, 7, 2], speeding up annotation by a significant factor. We follow this line of work.

In DEXTR [23], the authors build upon the Deeplab architecture [8] by incorporating a simple encoding of human clicks in the form of heat maps. This is a pixel-wise approach, *i.e.* it predicts a foreground-background label for each pixel. DEXTR showed that by incorporating user clicks as a soft constraint, the model learns to interactively improve its prediction. Yet, since the approach is pixel-wise, the worst case scenario still requires many clicks.

Polygon-RNN [7, 2] frames human-in-the-loop annotation as a recurrent process, during which the model sequen-

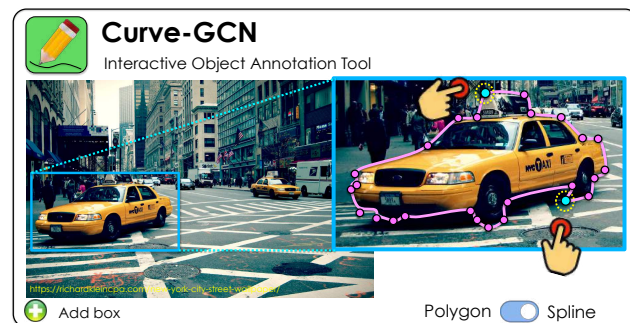


Figure 1: We propose Curve-GCN for interactive object annotation. In contrast to Polygon-RNN [7, 2], our model parametrizes objects with either polygons or splines and is trained end-to-end at a high output resolution.

tially predicts vertices of a polygon. The annotator can intervene whenever an error occurs, by correcting the wrong vertex. The model continues its prediction by conditioning on the correction. Polygon-RNN was shown to produce annotations at human level of agreement with only a few clicks per object instance. The worst case scenario here is bounded by the number of polygon vertices, which for most objects ranges up to 30-40 points. However, the recurrent nature of the model limits scalability to more complex shapes, resulting in harder training and longer inference. Furthermore, the annotator is expected to correct mistakes in a sequential order, which is often challenging in practice.

In this paper, we frame object annotation as a regression problem, where the locations of all vertices are predicted simultaneously. We represent the object as a graph with a fixed topology, and perform prediction using a Graph Convolutional Network. We show how the model can be used and optimized for interactive annotation. Our framework further allows us to parametrize objects with either polygons or splines, adding additional flexibility and efficiency to the interactive annotation process. The proposed approach, which we refer to as Curve-GCN, is end-to-end differentiable, and runs in real time. We evaluate our Curve-GCN on the challenging Cityscapes dataset [10], where we outperform Polygon-RNN++ and PSP-Deeplab/DEXTR in both automatic and interactive settings. We also show that our model outperforms the baselines in cross-domain annotation, that is, a model trained on Cityscapes is used to

*authors contributed equally

annotate general scenes [38], aerial [31], and medical imagery [16, 14]. Code is available: <https://github.com/fidler-lab/curve-gcn>.

2. Related Work

Pixel-wise methods. Interactive object segmentation has typically been formulated as a pixel-wise foreground-background segmentation. Most of the early work relies on optimization by graph-cuts to solve an energy function that depends on various color and texture cues [28, 4, 9]. The user is required to draw a box around the object, and can interact with the method by placing additional scribbles on the foreground or background, until the object is carved out correctly. However, in ambiguous cases where object boundaries blend with background, these methods often require many clicks from the user [2].

Recently, DEXTR [23] incorporated user clicks by stacking them as additional heatmap channels to image features, and exploited the powerful Deeplab architecture to perform user-guided segmentation. The annotator is expected to click on the four extreme points of the object, and if necessary, iteratively add clicks on the boundary to refine prediction. Our work differs from the above methods in that it directly predicts a polygon or spline around the object, and avoids pixel-labeling altogether. We show this to be a more efficient way to perform object instance segmentation, both in the automatic and in the interactive settings.

Contour-based methods. Another line of work to object segmentation aims to trace closed contours. Oldest techniques are based on level sets [6], which find object boundaries via front propagation by solving a corresponding partial differential equation. Several smoothing terms help the contour evolution to be well behaved, producing accurate and regularized boundaries. In [1], levelset evolution with carefully designed boundary prediction was used to find accurate object boundaries from coarse annotations. This speeds up annotation since the annotators are only required to perform very coarse labeling. [34] combines CNN feature learning with level set optimization in an end-to-end fashion, and exploits extreme points as a form of user interaction. While most level set-based methods were not interactive, [11] proposed to incorporate user clicks into the energy function. Recently, [24] proposed a structure prediction framework to learn CNN features jointly with the active contour parameters by optimizing an approximate IoU loss. Rather than relying on the regularized contour evolution which may lead to overly smooth predictions, our approach learns to perform inference using a GCN. We further tackle the human-in-the-loop scenario, not addressed in [24].

Intelligent Scissors [25] is a technique that allows the user to place “seeds” along the boundary and finds the minimal cost contour starting from the last seed up to the mouse cursor, by tracing along the object’s boundary. In the case of

error, the user is required to place more seeds.

Polygon-RNN [7, 2] adopted a similar idea of sequentially tracing a boundary, by exploiting a CNN-RNN architecture. Specifically, the RNN predicts a polygon by outputting one vertex at a time. However, the recurrent structure limits the scalability with respect to the number of vertices, and also results in slower inference times. Our work is a conceptual departure from Polygon-RNN in that we frame object annotation as a regression problem, where the locations of all vertices of the polygon are predicted simultaneously. The key advantages of our approach is that our model is significantly faster, and can be trained end-to-end using a differentiable loss function. Furthermore, our model is designed to be invariant to order, thus allowing the annotator to correct *any* vertex, and further control the influence of the correction.

Our approach shares similarities with Pixel2Mesh [33], which predicts 3D meshes of objects from single images. We exploit their iterative inference regime, but propose a different parametrization based on splines and a loss function better suited for our (2D annotation) task. Moreover, we tackle the human-in-the-loop scenario, not addressed in [33].

Splines have been used to parametrize shapes in older work on active shape models [32]. However, these models are not end-to-end, while also requiring a dataset of aligned shapes to compute the PCA basis. Furthermore, interactivity comes from the fact that the prediction is a spline that the user can modify. In our approach, every modification leads to re-prediction, resulting in much faster interactive annotation.

3. Object Annotation via Curve-GCN

In order to approximate a curved contour outlining an object, one can either draw a polygon or a spline. Splines are a more efficient form of representation as they allow precise approximation of the shape with fewer control points. Our framework is designed to enable both a polygon and a spline representation of an object contour.

We follow the typical labeling scenario where we assume that the annotator has selected the object of interest by placing a bounding box around it [7, 2]. We crop the image around this box and frame object annotation in the crop as a regression problem; to predict the locations of all control points (vertices) simultaneously, from an initialization with a fixed topology. We describe our model from representation to inference in Subsec. 3.1, and discuss training in Subsec. 3.2. In Subsec. 3.3, we explain how our model can be used for human-in-the loop annotation, by formulating both inference as well as training in the interactive regime.

3.1. Polygon/Spline-GCN

We assume our target object shapes can be well represented using N control points, which are connected to form a cycle. The induced shape is rendered by either connecting them with straight lines (thus forming a polygon), or

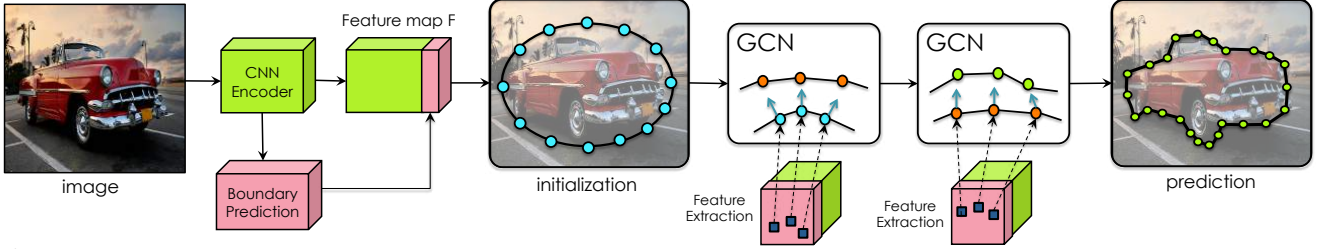


Figure 2: **Curve-GCN**: We initialize N control points (that form a closed curve) along a circle centered in the image crop with a diameter of 70% of image height. We form a graph and propagate messages via a Graph Convolutional Network (GCN) to predict a location shift for each node. This is done iteratively (3 times in our work). At each iteration we extract a feature vector for each node from the CNN’s features F , using a bilinear interpolation kernel.

higher order curves (forming a spline). We treat the location of each control point as a continuous random variable, and learn to predict these via a Graph Neural Network that takes image evidence as input. In [2], the authors exploited Gated Graph Neural Networks (GGNN) [19] as a polygon refinement step, in order to upscale the vertices output by the RNN to a higher resolution. In similar vein, Pixel2Mesh [33] exploited a Graph Convolutional Network (GCN) to predict vertex locations of a 3D mesh. The key difference between a GGNN and a GCN is in the graph information propagation; a GGNN shares propagation matrices through time akin to a gated recurrent unit (GRU), whereas a GCN has propagation steps implemented as unshared “layers”, similar to a typical CNN architecture. We adopt the GCN in our model due to its higher capacity. Hence, we name our model, *Curve-GCN*, which includes *Polygon* or *Spline-GCN*.

Notation: We initialize the nodes of the GCN to be at a static initial central position in the given image crop (Fig. 2). Our GCN predicts a location offset for each node, aiming to move the node correctly onto the object’s boundary. Let $\mathbf{cp}_i = [x_i, y_i]^T$ denote the location of the i -th control point and $V = \{\mathbf{cp}_0, \mathbf{cp}_1, \dots, \mathbf{cp}_{N-1}\}$ be the set of all control points. We define the graph to be $G = (V, E)$, with V defining the nodes and E the edges in the graph. We form E by connecting each vertex in V with its four neighboring vertices. This graph structure defines how the information propagates in the GCN. Connecting 4-way allows faster exchange of information between the nodes in the graph.

Extracting Features: Given a bounding box, we crop the corresponding area of the image and encode it using a CNN, the specific choice of which we defer to experiments. We denote the feature map obtained from the last convolutional layer of the CNN encoder applied on the image crop as F_c . In order to help the model see image boundaries, we supervise two additional branches, *i.e.* an edge branch and a vertex branch, on top of the CNN encoder’s feature map F_c , both of which consist of one 3×3 convolutional layer and one fully-connected layer. These branches are trained to predict the probability of existence of an object edge/vertex on a 28×28 grid. We train these two branches with the binary cross entropy loss. The predicted edge and vertices outputs are concatenated with F_c , to create an augmented feature

map F . The input feature for a node \mathbf{cp}_i in the GCN is a concatenation of the node’s current coordinates (x_i, y_i) , where top-left of the cropped images is $(0, 0)$ and image length is 1, and features extracted from the corresponding location in F : $f_i^0 = \text{concat}\{F(x_i, y_i), x_i, y_i\}$. Here, $F(x_i, y_i)$ is computed using bilinear interpolation.

GCN Model: We utilize a multi-layer GCN. The graph propagation step for a node \mathbf{cp}_i at layer l is expressed as:

$$f_i^{l+1} = w_0^l f_i^l + \sum_{\mathbf{cp}_j \in \mathcal{N}(\mathbf{cp}_i)} w_1^l f_j^l \quad (1)$$

where $\mathcal{N}(\mathbf{cp}_i)$ denotes the nodes that are connected to \mathbf{cp}_i in the graph, and w_0^l, w_1^l are the weight matrices. Following [5, 33], we utilize a Graph-ResNet to propagate information between the nodes in the graph as a residual function. The propagation step in one full iteration at layer l then takes the following form:

$$r_i^l = \text{ReLU}(w_0^l f_i^l + \sum_{\mathbf{cp}_j \in \mathcal{N}(\mathbf{cp}_i)} w_1^l f_j^l) \quad (2)$$

$$r_i^{l+1} = \tilde{w}_0^l r_i^l + \sum_{\mathbf{cp}_j \in \mathcal{N}(\mathbf{cp}_i)} \tilde{w}_1^l r_j^l \quad (3)$$

$$f_i^{l+1} = \text{ReLU}(r_i^{l+1} + f_i^l), \quad (4)$$

where $w_0, w_1, \tilde{w}_0, \tilde{w}_1$ are weight matrices for the residual. On top of the last GCN layer, we apply a single fully connected layer to take the output feature and predict a relative location shift, $(\Delta x_i, \Delta y_i)$, for each node, placing it into location $[x'_i, y'_i] = [x_i + \Delta x_i, y_i + \Delta y_i]$. We also perform iterative inference similar to the coarse-to-fine prediction in [33]. To be specific, the new node locations $[x'_i, y'_i]$ are used to re-extract features for the nodes, and another GCN predicts a new set of offsets using these features. This mimics the process of the initial polygon/spline iteratively “walking” towards the object’s boundaries.

Spline Parametrization: The choice of spline is important, particularly for the annotator’s experience. The two most common splines, *i.e.* the cubic Bezier spline and the uniform B-Spline [27, 12], are defined by control points which do not lie on the curve, which could potentially confuse an annotator that needs to make edits. Following [32], we use

the centripetal Catmull-Rom spline (CRS) [35], which has control points along the curve. We refer the reader to [35] for a detailed visualization of different types of splines.

For a curve segment \mathbf{S}_i defined by control points \mathbf{cp}_{i-1} , \mathbf{cp}_i , \mathbf{cp}_{i+1} , \mathbf{cp}_{i+2} and a knot sequence $t_{i-1}, t_i, t_{i+1}, t_{i+2}$, the CRS is interpolated by:

$$\mathbf{S}_i = \frac{t_{i+1}-t}{t_{i+1}-t_i} L_{012} + \frac{t-t_i}{t_{i+1}-t_i} L_{123} \quad (5)$$

where

$$L_{012} = \frac{t_{i+1}-t}{t_{i+1}-t_{i-1}} L_{01} + \frac{t-t_{i-1}}{t_{i+1}-t_{i-1}} L_{12} \quad (6)$$

$$L_{123} = \frac{t_{i+2}-t}{t_{i+2}-t_i} L_{12} + \frac{t-t_i}{t_{i+2}-t_i} L_{23} \quad (7)$$

$$L_{01} = \frac{t_i-t}{t_i-t_{i-1}} \mathbf{cp}_{i-1} + \frac{t-t_{i-1}}{t_i-t_{i-1}} \mathbf{cp}_i \quad (8)$$

$$L_{12} = \frac{t_{i+1}-t}{t_{i+1}-t_i} \mathbf{cp}_i + \frac{t-t_i}{t_{i+1}-t_i} \mathbf{cp}_{i+1} \quad (9)$$

$$L_{23} = \frac{t_{i+2}-t}{t_{i+2}-t_{i+1}} \mathbf{cp}_{i+1} + \frac{t-t_{i+1}}{t_{i+2}-t_{i+1}} \mathbf{cp}_{i+2}, \quad (10)$$

and $t_{i+1} = \|\mathbf{cp}_{i+1} - \mathbf{cp}_i\|_2^\alpha + t_i$, $t_0 = 0$. Here, α ranges from 0 to 1. We choose $\alpha = 0.5$ following [32], which in theory produces splines without cusps or self-intersections [35]. To make the spline a closed and C^1 -continuous curve, we add three additional control points:

$$\mathbf{cp}_N = \mathbf{cp}_0 \quad (11)$$

$$\mathbf{cp}_{N+1} = \mathbf{cp}_0 + \frac{\|\mathbf{cp}_{N-1} - \mathbf{cp}_0\|_2}{\|\mathbf{cp}_1 - \mathbf{cp}_0\|_2} (\mathbf{cp}_1 - \mathbf{cp}_0) \quad (12)$$

$$\mathbf{cp}_{-1} = \mathbf{cp}_0 + \frac{\|\mathbf{cp}_1 - \mathbf{cp}_0\|_2}{\|\mathbf{cp}_{N-1} - \mathbf{cp}_0\|_2} (\mathbf{cp}_{N-1} - \mathbf{cp}_0). \quad (13)$$

3.2. Training

We train our model with two different loss functions. First, we train with a Point Matching Loss which we introduce in Subsec. 3.2.1, and then fine-tune it with a Differentiable Accuracy Loss described in Subsec. 3.2.2. Details and ablations are provided in Experiments.

3.2.1 Point Matching Loss

Typical point-set matching losses, such as the Chamfer Loss, assumed unordered sets of points (*i.e.* they are permutation invariant). A polygon/spline, however, has a well defined ordering, which an ideal point set matching loss would obey. Assuming equal sized and similarly ordered (clockwise or counter-clockwise) prediction and ground truth point sets, denoted as $\mathbf{p} = \{p_0, p_1, \dots, p_{K-1}\}$, and $\mathbf{p}' = \{p'_0, p'_1, \dots, p'_{K-1}\}$ respectively (K is the number of points), we define our matching loss as:

$$L_{\text{match}}(\mathbf{p}, \mathbf{p}') = \min_{j \in [0, \dots, K-1]} \sum_{i=0}^{K-1} \|p_i - p'_{(j+i)\%K}\|_1 \quad (14)$$

Notice that this loss explicitly ensures an order in the vertices in the loss computation. Training with an unordered point set loss function, while maintaining the topology of the polygon could result in self-intersections, while the ordered loss function discourages it.

Sampling equal sized point sets. Since annotations may vary in the number of vertices, while our model always assumes N , we sample additional points along boundaries of both ground-truth polygons and our predictions. For Polygon-GCN, we uniformly sample K points along edges of the predicted polygons, while for Spline-GCN, we sample K points along the spline by uniformly ranging t from t_i to t_{i+1} . We also uniformly sample the same number of points along the edges of the ground-truth polygon. We use $K = 1280$ in our experiments. Sampling more points would have a higher computational cost, while sampling fewer points would make curve approximation less accurate. Note that the sampling only involves interpolating the control points, ensuring differentiability.

3.2.2 Differentiable Accuracy Loss

Note that training with the point matching loss results in overly smooth predictions. To perfectly align the predicted polygon and the ground-truth silhouette, we employ a differentiable rendering loss, which encourages masks rendered from the predicted control points to agree with ground-truth masks by directly optimizing for accuracy. This has been used previously to optimize 3D mesh vertices to render correctly onto a 2D image [17, 22].

The rendering process can be described as a function R ; $M(\theta) = R(\mathbf{p}(\theta))$, where \mathbf{p} is the sampled point sequence on the curve, and M is the corresponding mask rendered from \mathbf{p} . The predicted and the ground-truth masks can be compared by computing their difference with the L_1 loss:

$$L_{\text{render}}(\theta) = \|M(\theta) - M_{\text{gt}}\|_1 \quad (15)$$

Note that L_{render} is exactly the pixel-wise accuracy of the predicted mask $M(\theta)$ with respect to the ground truth M_{gt} . We now describe the method for obtaining M in the forward pass and back-propagating the gradients through the rendering process R , from $\frac{\partial L}{\partial M}$ to $\frac{\partial L}{\partial \mathbf{p}}$ in the backward pass.

Forward Pass: We render \mathbf{p} into a mask using OpenGL. As shown in Fig. 3, we decompose the shape into triangle fans \mathbf{f}_j and assign positive or negative values to their area based on their orientation. We render each face with the assigned value, and sum over the rendering of all the triangles to get the final mask. We note that this works for both convex and concave polygons [29].

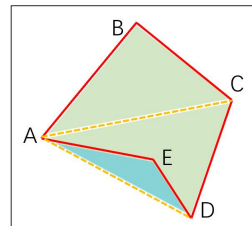


Figure 3: We decompose the polygon ABCDE into 3 triangle fans ABC, ACD and ADE, and render them separately. We assign positive value for clock wise triangles (ABC, ACD) and negative value for the others (ADE). Finally we sum over all the renderings. The sum retains only the interior of the polygon.

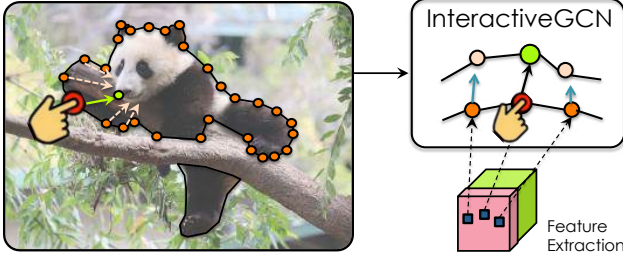


Figure 4: **Human-in-the-Loop:** An annotator can choose any wrong control point and move it onto the boundary. Only its immediate neighbors ($k = 2$ in our experiments) will be re-predicted based on this interaction.

Backward Pass: The rendering process is non-differentiable in OpenGL due to rasterization, which truncates all float values to integers. However, following [22], we compute its gradient with first order Taylor expansion. We reuse the triangle fans from the decomposition in the forward pass (see Fig. 3) and analyze each triangle fan separately. Taking a small shift of the fan \mathbf{f}_j , we calculate the gradient w.r.t. the j -th triangle as:

$$\frac{\partial M_j}{\partial \mathbf{f}_j} = \frac{R(\mathbf{f}_j + \Delta t) - R(\mathbf{f}_j)}{\Delta t}, \quad (16)$$

where M_j is the mask corresponding to the fan \mathbf{f}_j . Here, Δt can be either in the x or y direction. For simplicity, we let Δt to be a 1 pixel shift, which alleviates the need to render twice, and lets us calculate gradients by subtracting neighboring pixels. Next, we pass the gradient $\frac{\partial M_j}{\partial \mathbf{f}_j}$ to its three vertices $\mathbf{f}_{j,0}$, $\mathbf{f}_{j,1}$ and $\mathbf{f}_{j,2}$:

$$\frac{\partial M_j}{\partial \mathbf{f}_{j,k}} = \sum_i w_k^i \frac{\partial M_j^i}{\partial \mathbf{f}_j} \quad k \in [0, 1, 2] \quad (17)$$

where we sum over all pixels i . For the i -th pixel M_j^i in the rendered image M_j , we compute its weight w_0^i , w_1^i and w_2^i with respect to the vertices of the face \mathbf{f}_j as its barycentric coordinates. For more details, please refer to [22].

3.3. Annotator in The Loop

The drawback of Polygon-RNN is that once the annotator corrects one point, *all* of the subsequent points will be affected due to the model’s recurrent structure. This is often undesirable, as the changes can be drastic. In our work, we want the flexibility to change *any* point, and further constrain that only the neighboring points can change. As in Polygon-RNN, the correction is assumed to be in the form of drag-and-drop of a point.

To make our model interactive, we train another GCN that consumes the annotator’s correction and predicts the relative shifts of the other control points. We refer to it as the *InteractiveGCN*. We keep the network’s architecture the same as the original GCN, except that we now append two additional dimensions to the corrected node’s (say node i) input feature, representing the annotator’s correction:

$$\mathbf{f}_i^0 = \text{concat}\{F(x_i, y_i), x_i, y_i, \Delta x_i, \Delta y_i\}, \quad (18)$$

Algorithm 1 Learning to Incorporate Human-in-the-Loop

```

1: while not converged do
2:   (rawImage, gtCurve) = Sample(Dataset)
3:   (predCurve, F) = Predict(rawImage)
4:   data = []
5:   for i in range(c) do
6:     corrPoint = Annotator(predictedCurve)
7:     data += (predCurve, corrPoint, gtCurve, F)
8:     predCurve = InteractiveGCN(predCurve, corrPoint)
9:   ▷ Do not stop gradients
10:  TrainInteractiveGCN(data)

```

where $(\Delta x_i, \Delta y_i)$ is the shift given by the annotator. For all other nodes, we set $(\Delta x_i, \Delta y_i)$ to zero. We do not perform iterative inference here. Our InteractiveGCN allows a radius of influence by simply masking predictions of nodes outside the radius to 0. In particular, we let k neighbors on either side of node i to be predicted, *i.e.*, $\mathbf{cp}_{(i-k)\%N}, \dots, \mathbf{cp}_{(i-1)\%N}, \mathbf{cp}_{(i+1)\%N}, \dots, \mathbf{cp}_{(i+k)\%N}$. We set $k = 2$ in our experiments, while noting that in principle, the annotator could vary k at test time.

We train InteractiveGCN by mimicking an annotator that iteratively moves wrong control points onto their correct locations. We assume that the annotator always chooses to correct the worst predicted point. This is computed by first aligning the predicted polygon with GT, by finding the minimum of our point matching loss (Sec. 3.2.1). We then find the point with the largest manhattan distance to the corresponding GT point. The network is trained to move the neighboring points to their corresponding ground-truth positions. We then iterate between the annotator choosing the worst prediction, and training to correct its neighbors. In every iteration, the GCN first predicts the correction for the neighbors based on the last annotator’s correction, and then the annotator corrects the next worst point. We let the gradient back-propagate through the iterative procedure, helping the InteractiveGCN to learn to incorporate possibly many user interactions. The training procedure is summarized in Alg. 1, where c denotes the number of iterations.

4. Experimental Results

In this section, we extensively evaluate our Curve-GCN for both in-domain and cross-domain instance annotation. We use the Cityscapes dataset [10] as the main benchmark to train and test our model. We analyze both automatic and interactive regimes, and compare to state-of-the-art baselines for both. For cross-domain experiments, we evaluate the generalization capability of our Cityscapes-trained model on the KITTI dataset [13] and four out-of-domain datasets, ADE20K [38], Aerial Rooftop [31], Cardiac MR [30], and ssTEM [14], following Polygon-RNN++ [2].

To indicate whether our model uses polygons or splines, we name them Polygon-GCN and Spline-GCN, respectively.

Model	Bicycle	Bus	Person	Train	Truck	Motorcycle	Car	Rider	Mean
Polygon-RNN++	57.38	75.99	68.45	59.65	76.31	58.26	75.68	65.65	67.17
Polygon-RNN++ (with BS)	63.06	81.38	72.41	64.28	78.90	62.01	79.08	69.95	71.38
PSP-DeepLab	67.18	83.81	72.62	68.76	80.48	65.94	80.45	70.00	73.66
Polygon-GCN (MLoss)	63.68	81.42	72.25	61.45	79.88	60.86	79.84	70.17	71.19
+ DiffAcc	66.55	85.01	72.94	60.99	79.78	63.87	81.09	71.00	72.66
Spline-GCN (MLoss)	64.75	81.71	72.53	65.87	79.14	62.00	80.16	70.57	72.09
+ DiffAcc	67.36	85.43	73.72	64.40	80.22	64.86	81.88	71.73	73.70

Table 1: **Automatic Mode on Cityscapes.** We compare our Polygon and Spline-GCN to Polygon-RNN++ and PSP-DeepLab. Here, *BS* indicates that the model uses beam search, which we do not employ.

Model	mIOU	F at 1px	F at 2px
Polymn++ (BS)	71.38	46.57	62.26
PSP-DeepLab	73.66	47.10	62.82
Spline-GCN	73.70	47.72	63.64
DEXTR	79.40	55.38	69.84
Spline-GCN-EXTR	79.88	57.56	71.89

Table 2: **Different Metrics.** We report IoU & F boundary score. We favorably cross-validate PSP-DeepLab and DEXTR for each metric on val. *Spline-GCN-EXTR* uses extreme points as additional input as in DEXTR.

Model	Spline	Polygon
GCN	68.55	67.79
+ Iterative Inference	70.00	70.78
+ Boundary Pred.	72.09	71.19
+ DiffAcc	73.70	72.66

Table 3: **Ablation study on Cityscapes.** We use 3 steps when performing iterative inference. *Boundary Pred* adds the boundary prediction branch to our CNN.

Model	Time(ms)
Polygon-RNN++	298.0
Polygon-RNN++ (Corr.)	270.0
PSP-Deeplab	71.3
Polygon-GCN	28.7
Spline-GCN	29.3
Polygon-GCN (Corr.)	2.0
Spline-GCN (Corr.)	2.6

Table 4: **Avg. Inference Time** per object. We are $10\times$ faster than Polygon-RNN++ in forward pass, and $100\times$ for every human correction.

Image Encoder: Following Polygon-RNN++ [2], we use the ResNet-50 backbone architecture as our image encoder.

Training Details: We first train our model via the matching loss, followed by fine-tuning with the differentiable accuracy loss. The former is significantly faster, but has less flexibility, *i.e.* points are forced to exactly match the GT points along the boundary. Our differentiable accuracy loss provides a remedy as it directly optimizes for accuracy. However, since it requires a considerably higher training time we employ it only in the fine-tuning stage. For speed issues we use the matching loss to train the InteractiveGCN. We use a learning rate of $3e-5$ which we decay every 7 epochs.

We note that the Cityscapes dataset contains a significant number of occluded objects, which causes many objects to be split into disconnected components. Since the matching loss operates on single polygons, we train our model on single component instances first. We fine-tune with the differentiable accuracy loss on *all* instances.

Baselines: Since Curve-GCN operates in two different regimes, we compare it with the relevant baselines in each. For the automatic mode, we compare our approach to Polygon-RNN++ [2], and PSP-DeepLab [8, 37]. We use the provided DeepLab-v2 model by [23], which is pre-trained on ImageNet, and fine-tuned on PASCAL for semantic segmentation. We stack Pyramid scene parsing [37] to enhance performance. For the interactive mode, we benchmark against Polygon-RNN++ and DEXTR [23]. We fine-tune both PSP-DeepLab and DEXTR on the Cityscapes dataset. We cross-validate their thresholds that decide between foreground/background on the validation set.

Evaluation Metrics: We follow Polygon-RNN [7] to evaluate performance by computing Intersection-over-Union (IoU) of the predicted and ground-truth masks. However, as noted in [18], IoU focuses on the full region and is less

sensitive to the inaccuracies along the object boundaries. We argue that for the purpose of object annotation boundaries are very important – even slight deviations may not escape the eye of an annotator. We thus also compute the Boundary F score [26] which calculates precision/recall between the predicted and ground-truth boundary, by allowing some slack wrt misalignment. Since Cityscapes is finely annotated, we report results at stringent thresholds of 1 and 2 pixels.

4.1. In-Domain Annotation

We first evaluate our model when both training and inference are performed on Cityscapes [10]. This dataset contains 2975/500/1525 images for training, validation and test, respectively. For a fair comparison, we follow the same split and data preprocessing procedure as in Polygon-RNN++ [7].

Automatic Mode: Table 1 reports results of our Curve-GCN and compares with baselines, in terms of IoU. Note that PSP-DeepLab uses a more powerful image encoder, which is pretrained on PASCAL for segmentation. Our Spline-GCN outperforms Polygon-RNN++ and is on par with PSP-DeepLab. It also wins over Polygon-GCN, likely because most Cityscapes objects are curved. The results also show the significance of our differentiable accuracy loss (diffAcc) which leads to large improvements over the model trained with the matching loss alone (denoted with MLoss in Table). Our model mostly loses against PSP-DeepLab on the *train* category, which we believe is due to the fact that trains in Cityscapes are often occluded and broken into multiple components. Since our approach predicts only a single connected component, it struggles in such cases.

Table 2 compares models with respect to F boundary metrics. We can observe that while Spline-GCN is on par with PSP-DeepLab under the IoU metric, it is significantly better in the more precise F score. This means that our model more accurately aligns with the object boundaries. We show



Figure 5: **Automatic Mode on Cityscapes.** The input to our model are bounding boxes for objects.

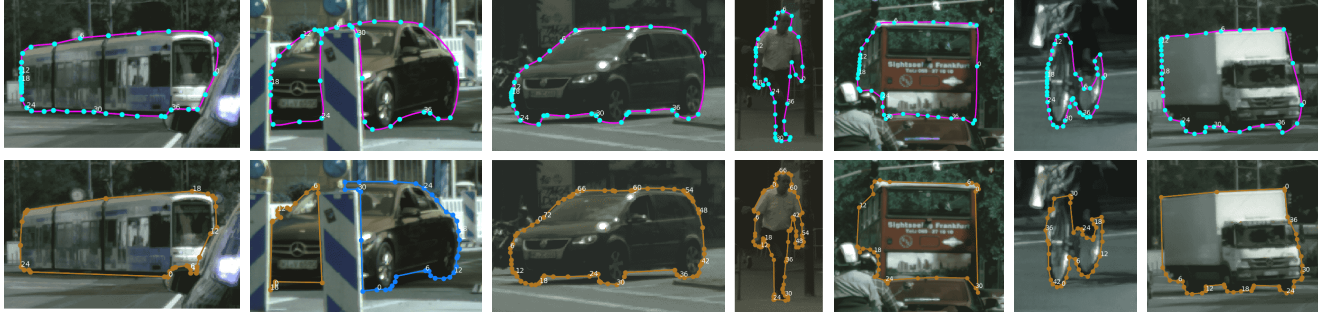


Figure 6: **Automatic mode on Cityscapes.** We show results for individual instances. **(top)** Spline-GCN, **(bottom)** ground-truth. We can observe that our model fits object boundaries accurately, and surprisingly finds a way to “cheat” in order to annotate multi-component instances.

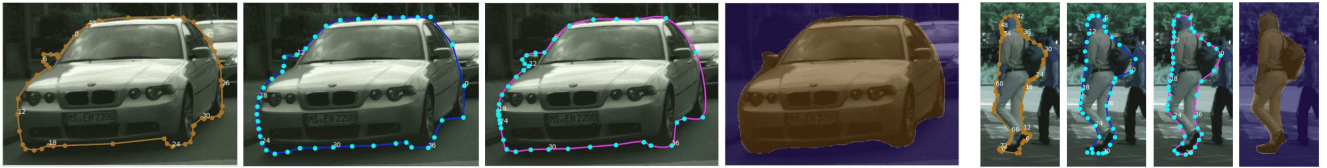


Figure 7: **Comparison in Automatic Mode.** From left to right: ground-truth, Polygon-GCN, Spline-GCN, PSP-DeepLab.

qualitative results in Fig 5, 6, and 7.

Ablation Study: We study each component of our model and provide results for both Polygon and Spline-GCN in Table 3. Performing iterative inference leads to a significant boost, and adding the boundary branch to our CNN further improves performance.

Additional Human Input: In DEXTR [23], the authors proposed to use 4 extreme points on the object boundary as an effective information provided by the annotator. Compared to just a box, extreme points require 2 additional clicks. We compare to DEXTR in this regime, and follow their strategy in how this information is provided to the model. To be specific, points (in the form of a heat map) are stacked with the image, and passed to a CNN. To compare with DEXTR, we use DeepLab-v2 in this experiment, as they do. We refer to our models with such input by appending *EXTR*.

We notice that the image crops used in Polygon-RNN, are obtained by extracting an image inside a *square box* (and not the actual box provided by the annotator). However, due to significant occlusion in Cityscapes, doing so leads to ambiguities, since multiple objects can easily fall in the same box. By providing 4 extreme points, the annotator more accurately points to the target object. To verify how much accuracy is really due to the additional two clicks, we also test an instantiation of our model to which the four corners of the bounding box are provided as input. This is still a 2-click (box) interaction from the user, however, it reduces the ambiguity of which object to annotate. We refer

to this model by appending *BOX*.

Since DEXTR labels pixels and thus more easily deals with multiple component instances, we propose another instantiation of our model which still exploits 4 clicks on average, yet collects these differently. Specifically, we request the annotator to provide a box around *each component*, rather than just a single box around the full object. On average, this leads to 2.4 clicks per object. This model is referred to with *MBOX*. To match the 4-click budget, our annotator clicks on the worst predicted boundary point for each component, which leads to 3.6 clicks per object, on average.

Table 5 shows that in the extreme point regime, our model is already better than DEXTR, whereas our alternative strategy is even better, yielding an 0.8% improvement overall with fewer clicks on average. Our method also significantly outperforms DEXTR in the boundary metrics (Table 2).

Interactive Mode: We simulate an annotator correcting vertices, following the protocol in [2]. In particular, the annotator iteratively makes corrections until the IoU is greater than a threshold T , or the model stops improving its prediction. We consider the predicted curve achieving agreement above T as a satisfactory annotation.

Plots 8 and 9 show IoU vs number of clicks at different thresholds T . We compare to Polygon-RNN++. Our results show significant improvements over the baseline, highlighting our model as a more efficient annotation tool. We further analyze performance when using 40 vs 20 control points. The version with fewer control points is slightly worse in

Model	Bicycle	Bus	Person	Train	Truck	Mcycle	Car	Rider	Mean	# clicks
Spline-GCN-BOX	69.53	84.40	76.33	69.05	85.08	68.75	83.80	73.38	76.29	2
PSP-DEXTR	74.42	87.30	79.30	73.51	85.42	73.69	85.57	76.24	79.40	4
Spline-GCN-EXTR	75.09	87.40	79.88	72.78	86.76	73.93	86.13	77.12	79.88	4
Spline-GCN-MBOX	70.45	88.02	75.87	76.35	82.73	70.76	83.32	73.49	77.62	2.4
+ One click	73.28	89.18	78.45	79.89	85.02	74.33	85.15	76.22	80.19	3.6

Table 5: **Additional Human Input.** We follow DEXTR [23] and provide a budget of 4 clicks to the models. Please see text for details.

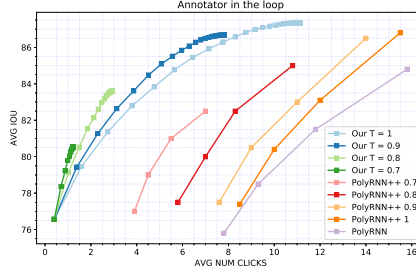
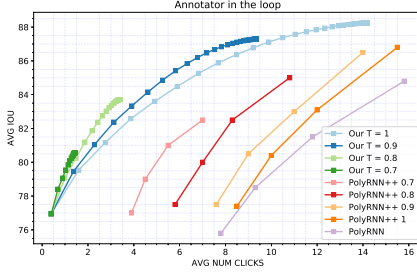


Figure 8: **Interactive Mode on Cityscapes:** (left) 40 control points, (right) 20 control points.

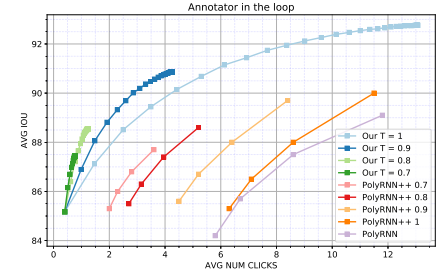


Figure 9: **Inter. Mode on KITTI:** 40 cps

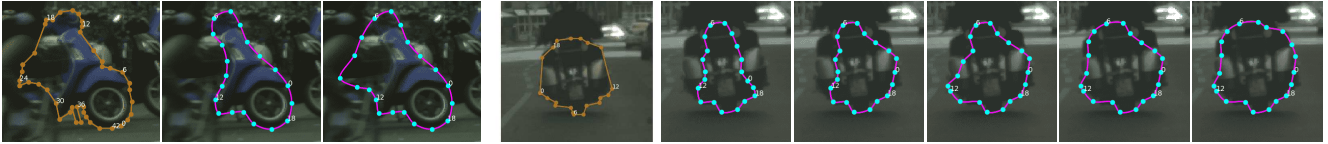


Figure 10: **Annotator in the Loop:** GT, 2nd column is the initial prediction from Spline-GCN, and the following columns show results after (simulated) annotator's corrections. Our corrections are local, and thus give more control to the annotator. However, they sometimes require more clicks (right).

Model	KITTI	ADE	Rooftop	Card.MR	ssTEM
Square Box (Perfect)	-	69.35	62.11	79.11	66.53
Ellipse (Perfect)	-	69.53	66.82	92.44	71.32
Polygon-RNN++ (BS)	83.14	71.82	65.67	80.63	53.12
PSP-DeepLab	83.35	72.70	57.91	74.11	47.65
Spline-GCN	84.09	72.94	68.33	78.54	58.46
+ finetune	84.81	77.35	78.21	91.33	-
Polygon-GCN	83.66	72.31	66.78	81.55	60.91
+ finetune	84.71	77.41	75.56	90.91	-

Table 6: **Automatic Mode on Cross-Domain.** We outperform PSP-DeepLab out-of-the-box. Fine-tuning on 10% is effective.



Table 7: **Automatic Mode for Cross-Domain.** (top) Out-of-the-box output of Cityscapes-trained models, (bottom) fine-tuned with 10% of data from new domain.

automatic mode (see Appendix), however, it is almost on par in the interactive mode. This may suggest that coarse-to-fine interactive correction may be the optimal approach.

Inference Times: Timings are reported in Table 4. Our model is an order of magnitude faster than Polygon-RNN++, running at 29.3 ms, while Polygon-RNN++ requires 298.0ms. In the interactive mode, our model re-uses the computed image features computed in the forward pass, and thus only requires 2.6ms to incorporate each correction. On the other hand, Polygon-RNN requires to run an RNN after every correction, thus still requiring 270ms.

4.2. Cross-Domain Evaluation

We now evaluate the ability of our model to generalize to new datasets. Generalization is crucial, in order to effectively annotate a variety of different imagery types. We further show that by fine-tuning on only a small set of the new dataset (10%) leads to fast adaptation to new domains.

We follow [2] and use our Cityscapes-trained model and test it on KITTI [13] (in-domain driving dataset),

ADE20k [38] (general scenes), Rooftop [31] (aerial imagery), and two medical datasets [30, 16, 14].

Quantitative Results. Table 6 provides results. We adopt simple baselines from [2]. We further fine-tune (with diffAcc) the models with 10% randomly sampled training data from the new domain. Note that ssTEM does not have a training split, and thus we omit this experiment for this dataset. Results show that our model generalizes better than PSP-DeepLab, and that fine-tuning on very little annotated data effectively adapts our model to new domains. Fig. 7 shows a few qualitative results before and after fine-tuning.

5. Conclusion

We presented Curve-GCN for efficient interactive annotation. Our model improves over the state-of-the-art and is significantly faster. We further allow interactive corrections that only have local effect, giving more control to the annotators. This leads to the better overall annotation strategy. We will release an annotation tool running our model, in order to facilitate faster collection of computer vision datasets.

References

- [1] D. Acuna, A. Kar, and S. Fidler. Devil is in the edges: Learning semantic boundaries from noisy annotations. In *CVPR*, 2019.
- [2] D. Acuna, H. Ling, A. Kar, and S. Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *CVPR*, 2018.
- [3] M. Bai and R. Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017.
- [4] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *ICCV*, 2001.
- [5] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. In *CVPR*, 2017.
- [6] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *IJCV*, 22(1):61–79, 1997.
- [7] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler. Annotating object instances with a polygon-rnn. In *CVPR*, 2017.
- [8] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40(4):834–848, 2018.
- [9] L.-C. Chen, S. Fidler, A. Yuille, and R. Urtasun. Beat the muturkers: Automatic image labeling from weak 3d supervision. In *CVPR*, 2014.
- [10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.
- [11] D. Cremers, O. Fluck, M. Rousson, and S. Aharon. A probabilistic level set formulation for interactive organ segmentation. In *SPIE*, 2007.
- [12] J. Gao, C. Tang, V. Ganapathi-Subramanian, J. Huang, H. Su, and L. J. Guibas. Deepspline: Data-driven reconstruction of parametric curves and surfaces. *arXiv preprint arXiv:1901.03781*, 2019.
- [13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012.
- [14] S. Gerhard, J. Funke, J. Martel, A. Cardona, and R. Fetter. Segmented anisotropic ssTEM dataset of neural tissue. 11 2013.
- [15] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. *ICCV*, 2017.
- [16] A. H. Kadish, D. Bello, J. P. Finn, R. O. Bonow, A. Schaechter, H. Subacius, C. Albert, J. P. Daubert, C. G. Fonseca, and J. J. Goldberger. Rationale and Design for the Defibrillators to Reduce Risk by Magnetic Resonance Imaging Evaluation (DETERMINE) Trial. *J Cardiovasc Electrophysiol*, 20(9):982–7, 2009.
- [17] H. Kato, Y. Ushiku, and T. Harada. Neural 3d mesh renderer. In *ECCV*, 2018.
- [18] P. Krähenbühl and V. Koltun. Geodesic object proposals. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *ECCV*, pages 725–739, 2014.
- [19] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel. Gated graph sequence neural networks. In *ICLR*, 2016.
- [20] S. Liu, J. Jia, S. Fidler, and R. Urtasun. Sequential grouping networks for instance segmentation. In *ICCV*, 2017.
- [21] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018.
- [22] M. M. Loper and M. J. Black. Opendr: An approximate differentiable renderer. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *ECCV*, pages 154–169, 2014.
- [23] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. Deep extreme cut: From extreme points to object segmentation. In *CVPR*, 2018.
- [24] D. Marcos, D. Tuia, B. Kellenberger, L. Zhang, M. Bai, R. Liao, and R. Urtasun. Learning deep structured active contours end-to-end. In *CVPR*, pages 8877–8885, 2018.
- [25] E. N. Mortensen and W. A. Barrett. Intelligent scissors for image composition. In *SIGGRAPH*, pages 191–198, 1995.
- [26] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [27] H. Prautzsch, W. Boehm, and M. Paluszny. *Bézier and B-spline techniques*. Springer Science & Business Media, 2013.
- [28] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.
- [29] D. Shreiner and T. K. O. A. W. Group. *OpenGL Programming Guide: The Official Guide to Learning OpenGL, Versions 3.0 and 3.1*. Addison-Wesley Professional, 7th edition, 2009.
- [30] A. Suinesiaputra, B. R. Cowan, A. O. Al-Agamy, M. A. Elattar, N. Ayache, A. S. Fahmy, A. M. Khalifa, P. Medrano-Gracia, M.-P. Jolly, A. H. Kadish, et al. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac mr images. *Medical image analysis*, 18(1):50–62, 2014.
- [31] X. Sun, C. M. Christoudias, and P. Fua. Free-shape polygonal object localization. In *European Conference on Computer Vision*, pages 317–332. Springer, 2014.
- [32] J. H. Tan and U. R. Acharya. Active spline model: a shape based model interactive segmentation. *Digital Signal Processing*, 35:64–74, 2014.
- [33] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y. Jiang. Pixel2mesh: Generating 3d mesh models from single RGB images. In *ECCV*, volume 11215, pages 55–71. Springer, 2018.
- [34] Z. Wang, D. Acuna, H. Ling, A. Kar, and S. Fidler. Object instance annotation with deep extreme level set evolution. In *CVPR*, 2019.
- [35] C. Yuksel, S. Schaefer, and J. Keyser. Parameterization and applications of catmull-rom curves. *Computer-Aided Design*, 43(7):747–755, 2011.
- [36] Z. Zhang, S. Fidler, and R. Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *CVPR*, 2016.
- [37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.

- [38] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.