

One Sketch for All: One-Shot Personalized Sketch Segmentation

Anran Qi, Yulia Gryaditskaya, Tao Xiang, Yi-Zhe Song
SketchX Lab, CVSSP, University of Surrey

Abstract—We present the first one-shot personalized sketch segmentation method. We aim to segment all sketches belonging to the same category provisioned with a single sketch with a given part annotation while (i) preserving the parts semantics embedded in the exemplar, and (ii) being robust to input style and abstraction. We refer to this scenario as *personalized*. With that, we importantly enable a much-desired personalization capability for downstream fine-grained sketch analysis tasks. To train a robust segmentation module, we deform the exemplar sketch to each of the available sketches of the same category. Our method generalizes to sketches not observed during training. Our central contribution is a sketch-specific hierarchical deformation network. Given a multi-level sketch-strokes encoding obtained via a graph convolutional network, our method estimates rigid-body transformation from the reference to the exemplar, on the upper level. Finer deformation from the exemplar to the globally warped reference sketch is further obtained through stroke-wise deformations, on the lower-level. Both levels of deformation are guided by mean squared distances between the keypoints learned without supervision, ensuring that the stroke semantics are preserved. We evaluate our method against the state-of-the-art segmentation and perceptual grouping baselines re-purposed for the one-shot setting and against two few-shot 3D shape segmentation methods. We show that our method outperforms all the alternatives by more than 10% on average. Ablation studies further demonstrate that our method is robust to *personalization*: changes in input part semantics and style differences.

Index Terms—sketch, segmentation, few-shot, deformation.

I. INTRODUCTION

With the appearance of large-scale sketch datasets and recent advances in deep learning, sketch-related research thrives [1], [2], [3], [4], [5], [6], [7]. Sketch segmentation in particular is an important capability that underpins the recent focus of fine-grained sketch analysis, *e.g.*, part-based sketch-based modeling and retrieval [8], [9], or fine-grained sketch editing [10], [11].

Existing methods aiming at semantic sketch segmentation rely on the availability of large-scale carefully annotated sketch datasets. Obtaining such annotations for new categories is however an extremely labor intensive task. *Moreover, current datasets provide fixed part sketches labeling, disregarding the subjective nature of segmentations [12], [13] – this impedes downstream applications, not allowing the number of meaningful labels to be task-specific.*

In this paper, for the first time, we address the problem of *personalized* sketch segmentation under a one-shot setting, as

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

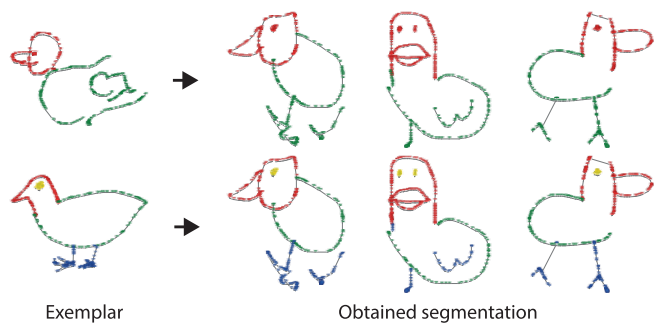


Fig. 1: Given an exemplar annotated sketch we are able to transfer the labels to any amount of sketches of the same category, and can achieve task-specific segmentation.

an attempt to facilitate the fine-grained analysis of sketches. Namely, given *one* input sketch that is part annotated, we seek to label any number of gallery sketches in a consistent manner by conforming to the same user-defined semantic interpretation, and robustly to differences in drawing style and abstraction between the input and gallery. We refer to this scenario as *personalized sketch segmentation*. Fig. 1 offers two working examples where two sketches of different styles and part decomposition were used to segment unseen sketches of the same category consistent with each input semantic segmentation.

The problem of one-shot personalized sketch segmentation is however non-trivial. The key challenge lies with deriving means of effectively transferring the semantic labels in the exemplar to all target sketches so that part semantics are preserved. Diversity in sketch depictions among humans (*e.g.*, the ducks in Fig. 1) makes this task even more challenging.

In our work, we exploit *deformation*, and train the segmentation module by morphing the exemplar sketch to available sketches of a given category. To increase robustness, we propose a hierarchical deformation model specific to sketch data. During training, we first globally align all sketches with the exemplar and then obtain a fine-level warping of the exemplar towards the globally aligned sketches. Such an approach simplifies the task of the segmentation network: The segmentation network does not need to deal with large geometric differences. At inference time, we first perform a global alignment with the exemplar, then predict the segmentation and finally apply an inverse global transformation operator.

Our first key contribution is in exploiting the connectivity information encoded in sketch strokes, in both parts of our segmentation model: deformation and segmentation. To embed

this information we obtain a multi-level sketch encoding using a graph convolution network (GCN). The graph consists of dynamic and static edges. The role of dynamic edges is to encode global information embedded into the holistic sketch. The static edges connect the consequent points along strokes, encoding stroke-level information. Such representation allows to retain both sketch- and stroke-level embeddings in a synergistic fashion.

The sketch deformation module, being our central contribution, builds around this sketch representation. On the holistic sketch level, we first predict a global rigid-body transformation with respect to an exemplar sketch, relying on an analytic solver. We observe that the conventional Chamfer distance [14], [15] between an input sketch and an exemplar is not sufficient to predict the global transformation which preserves semantic part correspondences, due to the sparse nature of sketches. We instead jointly train for unsupervised keypoints prediction [16] which are used to derive global transformation. Such rigid-body deformation however only *roughly* aligns the gallery sketch to the exemplar. To achieve finer alignment, we impose stroke-level deformations rather than individually on stroke points. To avoid unrealistic stroke distortions we limit the transformations to stroke-level rotation, translation and scaling. To achieve a globally consistent deformation on this level we leverage both Chamfer distance and mean square error on keypoints. The distances are computed between the fine-level deformed exemplar and a holistic sketch-level deformed gallery sketch.

Given the deformed exemplar we train a segmentation module in a standard fashion via a cross-entropy loss, which given the full sketch encoding predicts a label for each sketch point. We further observe that sketch strokes are frequently fully contained within one semantic part. We thus further condition the segmentation label predictor on the stroke encoding, eliminating the need for a labels refinement step, such as graph-cuts [17] or conditional random fields [18], [19].

In summary, (i) we propose for the first time the problem of one-shot sketch segmentation; (ii) we show that by transferring the semantic labeling from an input sketch to target ones, a much-desired personalized segmentation capability can be achieved; (iii) we propose a hierarchical sketch deformation framework that faithfully deforms the input sketch to each reference, as means to assist the transfer; (iv) we conduct multiple ablation studies demonstrating the robustness of our proposed method in terms of variations in part semantics and sketching style; (v) we show an advantage of our method over alternative solutions for a few-shot sketch segmentation.

II. RELATED WORK

a) Sketch segmentation: Prior works on sketch segmentation can be divided into two categories based on the segmentation goals: strokes perceptual grouping [20], [21], [22], [23] and semantic segmentation [24], [25], [18], [26], [17], [27], [28], [29], [30], [31], [32], [33], [34]. Our work belongs to the second category. Yet, we, for the first time, consider the problem of a *one-shot* semantic sketch segmentation.

Appearance of large-scale annotated sketch datasets [25], [22], [35], [17] fostered research on supervised semantic

sketch segmentation with deep learning. These methods can be classified into several groups by the used sketch representation: image-based [17], [19], [36], point-based [32], [37], ordered point sequences-based [27], [30], joint pixel-/point-based [31] methods, and graph-based [33]. *Image-based* methods represent a sketch as a raster image, and build on a success of convolutional neural networks (CNNs) in learning descriptive features. Zhu et al. [19] combine a CNN-based segmentation with a CRF-based refining. *Point-based* methods [32], [37] represent a sketch as an ordered point cloud. Such works build on the point-cloud architectures, first designed for 3D shape analysis and representation [38], [39], [40], [41]. Point-based representation allows to reduce models complexity, compared to imaged-based representations, due to the sparsity of lines in sketches. For the task of sketch recognition, Wang et al. [28] proposed a sketch-dedicated point cloud architecture. Both Wang et al. [28] and Hahnlein et al. [32] take as points features not only points spatial coordinates, but also strokes order. A number of works [27], [30], [42] exploit *recurrent neural networks (RNNs)* to translate sequence of strokes into their semantic parts. Such architectures were first designed for the the task of sketch recognition [43], [44], [45]. We though did not observe a correlation between a semantic stroke label and a stroke number in a general case. Wang et al. [31] *fuses* a prediction of the dedicated image-based architecture with the one obtained from a point-based architecture [38]. Recently, Zhu et al. [34] applied a similar idea to a sketch recognition task. Despite *graph-based* sketch representation was commonly used in earlier works on sketch analysis, just recently graph convolution networks (GCNs) were adopted for sketch processing. Yang et al. [33] proposed a two branches GCN for a supervised sketch segmentation task, implementing graph convolutions as was proposed in [41] and [46]. Su et al. [11] exploited GCN with a sparse number of nodes for the partial sketch completion task. Yang et al. [42] exploit RNN with conjunction with GCN for a sketch recognition task. In our work, we leverage a GCN for the task of a few-shot sketch segmentation, leveraging a multi-level sketch encoding.

b) 3D shape segmentation: Concurrently, two approaches for a few-shot 3D shape segmentation were recently proposed [15], [14], relying on the idea of being able to morph a template shape to an arbitrary target shape. Yuan et al. [15] directly transfers a label from the morphed template to the target shape by proximity. Wang et al. [14] instead learns a continuous probability distribution function that learns to assign to each point in space a probability of having a certain semantic label, conditioned on a global shape feature vector. Yuan et al. [15] exploit mesh connectivity, while Wang et al. [14] rely on a point-based shape representation.

Chen et al. [47] proposed an autoencoder for unsupervised consistent segmentation of shapes from the same class, and demonstrated how such architecture can be tuned for one-shot learning. The autoencoder consist of several branches, which are trained to encode complementary shape parts. Instead of decomposing the shape into parts, Chen et al. [16] study the problem of unsupervised prediction of semantically consist keypoints across all shapes of the same class. Dense predicted keypoints correspondences are used to transfer semantic labels

from one shape to another.

Our work extends these ideas to a few shot sketch segmentation, taking into account points connectivity and sketch sparsity, designed to be robust to arbitrary global sketch rotation and reflection.

III. METHOD

Our goal is, given an annotated exemplar sketch, to transfer its semantic part labels to an arbitrary target sketch of the same class as the exemplar sketch. We model this segmentation task as a two step process, comprising exemplar morphing towards reference sketches (Sec. III-A) and a robust labels prediction (Sec. III-E). We leverage a graph convolutional network for multi-level sketch encoding (Sec.III-D).

A. Deformation model

We expect the deformation from the exemplar to the reference to comply with the following rules: (i) The semantic meaning of the stroke segment label should stay invariant under the deformation (*e.g.*, a segment of the plane wing should not be deformed into a body, if they have different labels in the exemplar); (ii) The relative part relationship should stay invariant under the deformation (*e.g.*, an eye should not move outside a head).

We represent a sketch as an N-point set $V = \{v_i = (v_i^x, v_i^y)\}_{i=1,2,\dots,N}$, where v_i^x and v_i^y are the 2D absolute coordinates of the point v_i .

We model the deformation process between the exemplar sketch E and the unlabeled reference sketch X from the training set hierarchically. First, the global deformation that accounts for a global sketch rotation and reflection aligns each sketch in the training batch with the exemplar. Then, the exemplar is morphed to each of the globally morphed unlabeled sketches. As mention in the introduction and shown in the ablation studies section, this hierarchically two-directional deformation allows to increase the accuracy of segmentation results by simplifying the task of the segmentation module, lifting the requirement of learning rotation invariant segmentation. The full deformation network is shown in Fig. 2, and explained in detail below.

First, a sketch-level transformation, consisting of rotation/reflection $\hat{R} \in \mathbb{R}^{2 \times 2}$ and translation $\hat{\mathbf{t}} \in \mathbb{R}^{2 \times 1}$, aligns an unlabeled sketch towards an exemplar sketch:

$$\hat{x}_i = \hat{R}x_i + \hat{\mathbf{t}}, \forall x_i \in X, \quad (1)$$

where the reference sketch $X \subset \mathbb{R}^{2 \times N}$ consists of N 2D stroke point coordinates. Then, $\hat{X} = \bigcup_i \hat{x}_i \subset \mathbb{R}^{2 \times N}$ is the reference sketch globally aligned with the exemplar sketch.

Second, a stroke-level transformation morphs the strokes of the exemplar sketch towards the globally aligned reference sketch \hat{X} . The stroke level deformation comprises per stroke rotation $R_{\hat{X}} \subset \mathbb{R}^{2 \times 2}$, scaling $\sigma_{\hat{X}} = [\sigma_{\hat{X}}^x, \sigma_{\hat{X}}^y] \subset \mathbb{R}^2$ and translation $\mathbf{t}_{\hat{X}} \subset \mathbb{R}^{2 \times 1}$. The deformation model thus can be written as follows:

$$\hat{e}_i = R_{j,\hat{X}} \begin{bmatrix} \sigma_{j,\hat{X}}^x & 0 \\ 0 & \sigma_{j,\hat{X}}^y \end{bmatrix} e_i + \mathbf{t}_{j,\hat{X}}, \forall e_i \in \mathbf{s}_j \in E, \quad (2)$$

where \mathbf{s}_j is the j -th stroke of the the exemplar sketch $E \subset \mathbb{R}^{2 \times N}$, and e_i is a 2D stroke point coordinate. Then, $\hat{E}_{\hat{X}} = \bigcup_i \hat{e}_i \subset \mathbb{R}^{2 \times N}$ is the exemplar sketch aligned with the globally aligned reference sketch \hat{X} .

B. Sketch-level transformation estimation

We formulate the global sketch-level alignment task as a shape-matching problem [48], that allows to analytically compute a rigid transformation between the two sets of corresponding points:

$$(\hat{R}, \hat{\mathbf{t}}) = \underset{R,t}{\operatorname{argmin}} \frac{1}{M} \sum_{j=1}^M \| (Rx_i + t) - e_j \|^2, \quad (3)$$

where M is the total number of point pairs, x_i and e_j are the points in the reference and exemplar sketches.

To obtain point correspondences between the two sketches, we adopt the recent method by Chen et al. [16], which estimates semantically consistent sets of keypoints in unsupervised manner. We first obtain feature representation for each sketch point, using the encode we describe in Sec. III-D. The shared multi-layer perceptron (MLP) Ω_{θ_1} is then trained to predict for each point v_i a probability map $P = \{p_1, p_2, \dots, p_M\}$. Each element p_j represents the probability of the point v_i to be a j -th keypoint k_j . The keypoints are thus computed as $k_j = \sum_{i=1}^L v_i p_j^i$. Note that the extracted keypoints are not selected from the sketch stroke points but are enforced to lie close to the input strokes by defining the Chamfer loss between the the input points $V \subset \mathbb{R}^{2 \times N}$ and predicted keypoints $K_V \subset \mathbb{R}^{2 \times M}$, which we denote as $\mathcal{L}_{CD}(V, K_V)$. In Fig. 9 we demonstrate example extracted keypoints. Since we search for the transformation optimal in a least square sense the method is tolerant to small imprecision in the keypoints locations.

C. Stroke-level transformation estimation

Given an aligned reference sketch \hat{X} and an exemplar sketch E , we next predict stroke-level transformations $(R_j, \mathbf{t}_j, \sigma_j)$ for each sketch stroke $\mathbf{s}_j \in E$. To predict stroke level transformation we use a network Φ_{θ_2} consisting of successive multi-layer perceptrons (MLPs), followed by a ReLu activation function. The stroke transformation is calculated as

$$(R_j, \mathbf{t}_j, \sigma_j) = \Phi_{\theta_2}([\mathcal{F}_{\hat{X}}^{sketch}, \mathcal{F}_E^{sketch}, \mathcal{F}_{\mathbf{s}_j \in E}^{stroke}]) \quad (4)$$

where $[*,*]$ denotes the vector concatenation operation, and $\mathcal{F}_{\hat{X}}^{sketch}, \mathcal{F}_E^{sketch}$ are the global embedding vectors of a globally aligned reference sketch \hat{X} and an exemplar E ; $\mathcal{F}_{\mathbf{s}_j}^{stroke}$ is an embedding vector of the j -th stroke of the exemplar E .

To train this deformation we compute the Chamfer loss between the deformed template \hat{E} , obtained by substituting the prediction result in Eq. 4 to Eq. 2, and the globally aligned reference sketch \hat{X} : $\mathcal{L}_{CD}(\hat{E}_{\hat{X}}, \hat{X})$. We also compute the mean square error (MSE) distance between their keypoints: $\mathcal{L}_{MSE}(K_{\hat{E}_{\hat{X}}}, K_{\hat{X}})$.

We assume that sketch-level deformation compensates for the large distances between structure points of the exemplar

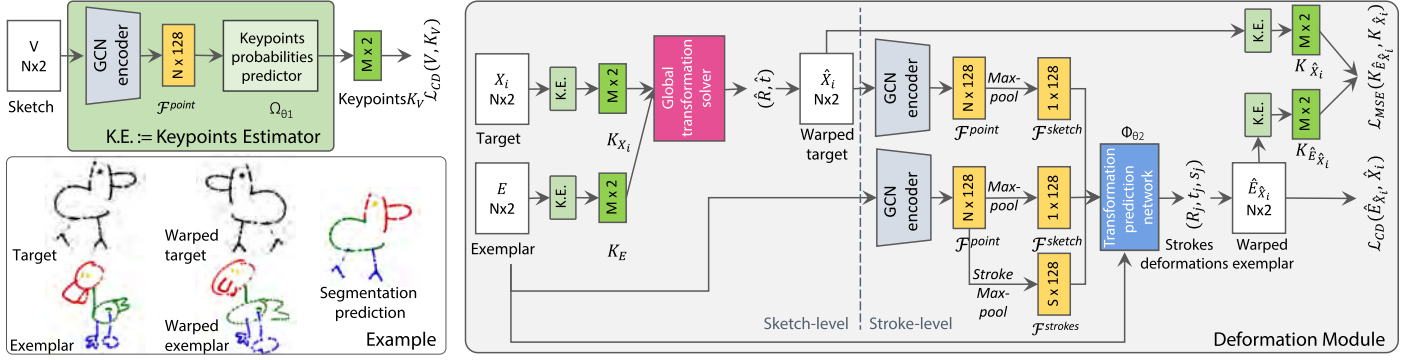


Fig. 2: Deformation module architecture (Section III-A) and a working example of the the deformation process with a segmentation prediction result by the segmentation module τ_{θ_3} , described in Section III-E.

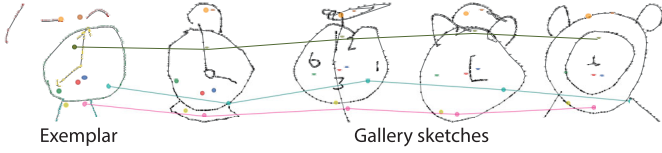


Fig. 3: Example keypoints predictions. For the visualization purposes we show only 9 keypoints. The exemplar lines trace the position of a keypoint across sketches.

and reference sketches, and only small deformations are required on a stroke level. During training we constrain each rotation matrix $R_j \subset \mathbb{R}^{2 \times 2}$ to be close to an orthogonal matrix: $\mathcal{L}_{orth,j} = \|I - R_j R_j^T\|^2$. Then, since the rotation matrix can be expressed in terms of one angle α_j , we use a soft constraint on the stroke rotation to not exceed 30 degrees. We achieve this by constraining each of the matrix elements to lie in the respective range. For instance, we constraint the first row and column element r_{11}^j of the rotation matrix R_j , which encodes $\cos(\alpha_j)$ to lie in the interval $[\frac{\sqrt{3}}{2}, 1]$:

$$\mathcal{L}_{rot,j}(r_{11}) = \max(0, \frac{\sqrt{3}}{2} - r_{11}^j) + \max(0, r_{11}^j - 1.0). \quad (5)$$

Similarly, we use a soft constraint on the stroke scaling $\sigma = [\sigma^x, \sigma^y] \subset \mathbb{R}^2$ to belong to the interval $[0.5, 2.0]$ with $\mathcal{L}_{\sigma,j}$. Finally, we constraint strokes translation vectors \mathbf{t}_j to have a small norm: $\mathcal{L}_{t,j} = \|\mathbf{t}_j\|$.

Constraining stroke level deformations, and exploiting key-points, allows us to avoid erroneous deformations, and ensures that the deformations comply with the two rules listed in Sec. III-A.

The full loss used to train stroke-level deformations is defined as

$$\begin{aligned} \mathcal{L}_{strokes} = & \beta \mathcal{L}_{CD}(\hat{E}_{\hat{X}}, \hat{X}) + \gamma \mathcal{L}_{MSE}(K_{\hat{E}_{\hat{X}}}, K_{\hat{X}}) \\ & + \gamma \frac{1}{\#E} \sum_j^{\#E} (\mathcal{L}_{orth,j} + \mathcal{L}_{rot,j} + \mathcal{L}_{\sigma,j} + \mathcal{L}_{t,j}), \end{aligned} \quad (6)$$

where $\#E$ is the number of strokes in the exemplar sketch.

D. Sketch Encoding

We exploit graph convolutional network (GCN) to obtain sketch-level, stroke-level, and point-level features. The network architecture we use is similar to the architecture of the

global branch used in [33], building on the study of GCNs by Li et al. [46]. We found a single branch to perform better than the full architecture proposed in [33].

The network consists of 4 layers with residual connections [46]. At each layer we construct a sketch graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = V$ is a set of all sketch points, while \mathcal{E} consists of two types of edges: static and dynamic.

Dynamic edges result in large receptive field and improved performance, when combined with residual connections [46]. We construct dynamic edges using a Dilated k-NN strategy proposed in [46]. It first selects $k \times d$ nearest neighbors at each layer, and then constructs edges by selecting every d -th neighbor. We also implement the stochastic dilation, which, with a probability $\epsilon = 0.2$, selects d neighbors uniformly from $k \times d$ nearest neighbors, instead. Following [33], we select $k = 4$, and set the dilation rate d to 1, 4, 8, 16 for successive layers.

Since points connectivity in strokes encodes an important information about a sketch, similar to [33], we combine the dynamically constructed edges with static edges that are obtained by gallconnecting consequent strokes points.

We use the convolutional operation, proposed in [41], to extract point-level features \mathcal{F}_i^{point} . Then, the stroke level features are defines as

$$\mathcal{F}_{s_j \in V}^{stroke} = \max_{i: v_i \in s_j} \mathcal{F}_i^{point}, \quad (7)$$

where s_j denotes j -th strokes of a sketch with a point set V , and v_i are all the points belonging to the stroke s_j . Similarly, the sketch embedding vector is defined as

$$\mathcal{F}_V^{sketch} = \max_{i: v_i \in V} \mathcal{F}_i^{point}, \quad (8)$$

where a max-pooling is performed over all sketch points.

E. Segmentation

To obtain a label for each segment we train a label probability function τ , which takes as an input a sketch point coordinate $v_i = (v_i^x, v_i^y)$, a sketch embedding vector \mathcal{F}_V^{sketch} , and a stroke embedding vector $\mathcal{F}_{s_j \in V: v_i \in s_j}^{stroke}$. We condition the predictor on the stroke, since the points which belong to the same stroke are likely to have the same label. The label

probability function consists of the successive MLP layers with ReLU activation functions:

$$\tau(v_i) := \tau_{\theta_3}(v_i, \mathcal{F}_{s_j \in V: v_i \in s_j}^{stroke}, \mathcal{F}_V^{sketch}), \quad (9)$$

such that $\tau : \mathbb{R}^{1 \times (2+2K)} \rightarrow [0, 1]^L$, where L is a number of labels, and $K = 128$ is a length of a sketch/stroke embedding vectors.

At training time, for each $\hat{E}_{\hat{X}_t}$, where t goes over all the reference sketches in a batch, we compute the cross entropy classification loss at each point $\hat{e}_{i, \hat{X}_t} \in \hat{E}_{\hat{X}_t}$ with a label l_{it} :

$$\mathcal{L}_{CE} = \sum_{i=1}^N \mathcal{L}_{\text{cross entropy}}(\tau_{\theta_3}(\hat{e}_{i, \hat{X}_t}), l_{it}). \quad (10)$$

F. Training and Losses

We train our segmentation network in the end-to-end manner, where the keypoints prediction network Ω_{θ_1} , the stroke morphing module Φ_{θ_2} , and the label probability distribution function τ_{θ_3} are trained jointly. The full loss is defined as

$$\begin{aligned} \mathcal{L} = & \alpha \frac{1}{|\mathcal{B}|} \sum_{t=1}^{|\mathcal{B}|} \mathcal{L}_{CD}(X_t, K_{X_t}) + \\ & \frac{1}{|\mathcal{B}|} \sum_{t=1}^{|\mathcal{B}|} \left(\mathcal{L}_{\text{strokes}}(\hat{E}_{\hat{X}_t}, \hat{X}_t) + \delta \mathcal{L}_{CE}(\hat{E}_{\hat{X}_t}) \right), \end{aligned} \quad (11)$$

where $|\mathcal{B}|$ is the number of sketches in each batch.

IV. EXPERIMENTS

A. Datasets

We evaluate our method on four sketch datasets: SPG [22], Huang14 [25], TU-Berlin [35], [17] and creative birds [36]. The SPG dataset consists of 25 categories with 800 sketches each, annotated with stroke-level semantic labels. The sketches come from the QuickDraw dataset [43] – a vector sketches dataset, collected via an online game where the players are asked to draw objects within 20 seconds. Annotated TU-Berlin dataset [17] contains 5 categories with 80 sketches each from the TU-Berlin dataset [35]. The participants were asked to produce a sketch of a given category within 30 minutes window. The labeling is obtained through crowd-sourcing. Huang14 dataset [25] consists of 10 categories with 30 sketches each, drawn by 3 participants from a reference photo. Creative birds and creatures [36] is a challenging dataset, featuring non-canonical representation of birds and arbitrary creatures. The participants are provided with a first stroke, and are asked to place an eye where they like, and to visualize how the stroke and the eye can be incorporated into a creative sketch, *e.g.*, of a bird. We evaluate on the creative birds only. The segmentation labels vary significantly among people on the creative creatures, and pose a poor ground-truth for our studies.

B. Alternative solutions

As discussed in Sec. II, we are the first to consider the problem of one/a few shot sketch segmentation. We therefore compare with two state-of-the-art supervised semantic sketch segmentation method [33] and perceptual grouping method [22], trained with one or a few example sketches, and few-shot segmentation methods for 3D shapes [14], [16]. **SPGG** [22] exploits a sequence-to-sequence variational autoencoder to obtain sketch encoding, and aims at globally consistent segments grouping. This method does not predict a label for the group, for the evaluation we assign each group the label, based on the overlap with the ground-truth grouping. **SGCN** [33] encodes a sketch with a GCN, consisting of two branches with static and dynamic convolutions. It is trained with the cross-entropy loss. This work gives the state-of-art segmentation results under supervised segmentation setting. In our work we use a similar sketch encoding architecture, it thus is a strong baseline for our method. **FLSS** [14] serves as a main baseline for our method, and addresses a few-shot 3D shape segmentation. Unlike us, it assumes that all shapes have similar global shape alignment, and models morphing from the template annotated 3D shape to the reference shape by predicting per point offset vectors. **ISPP** [16] is an encoder-decoder based architecture for semantically meaningful keypoints selection on a 3D point cloud. In our work we exploit this architecture to supervise sketch morphing. To perform a label transfer directly, we, first, for each point in the target sketch, find the closest keypoint in Euclidean space. Then, we select the closest keypoint from the exemplar sketch to the selected keypoint in the feature space and transfer its label. The original ISPP [16] method relies on the PointNet++ encoder. In all our comparisons, we instead use our GCN encoder as it results in better performance (we refer the reader to the supplemental material for the detailed evaluation).

C. Implementation Details

To obtain an N point-set sketch representation, we first simplify the sketches with Ramer-Douglas-Peucker algorithm to nearly 256 points. Then, if there are less than 256 points, we use a simple sampling strategy of dividing random segments in the middle till we have roughly 256 points. In case if there are still more points we randomly skip some points. We use $M-256$ keypoints in our work. For SPGG we used the original points sampling. We set $\alpha = 1.0$, $\delta = 0.02$ in Eq. 11, $\beta = 0.2$ and $\gamma = 50$ in Eq. 6.

For all methods we perform data augmentation by rotating sketches by a randomly chosen angle from the interval $[-\frac{\pi}{12}, \frac{\pi}{12}]$. We use Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) with a learning rate $5e-5$ and a batch size 24. At inference, to obtain the labeling via Eq. 9, we first estimate our hierarchical deformation, then the label of a point v_i is obtained as follows $\tau(v_i) = \tau_{\theta_3}(v_i, \mathcal{F}_{s_j \in \hat{X}: v_i \in s_j}^{stroke}, \mathcal{F}_{\hat{E}}^{sketch})$.

D. Evaluation

We evaluate with traditional segmentation metrics: (a) *pixel/point accuracy (P-metric)* – the fraction of points that are

category	P-metric										C-metric									
	SPGG [22]		SGCN [33]		FLSS [14]		ISPP [16]		Ours		SPGG [22]		SGCN [33]		FLSS [14]		ISPP [16]		Ours	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
airplane	24.1	9.1	66.6	14.0	56.3	11.2	50.8	14.2	86.0	4.9	20.0	7.4	57.3	17.8	34.6	15.5	23.0	12.7	80.6	8.5
alarm clock	27.6	5.9	79.7	9.9	59.7	10.4	59.4	11.5	86.4	9.1	10.2	1.1	68.4	15.7	36.5	17.4	32.9	17.8	76.0	15.7
ambulance	29.6	9.2	78.1	3.4	61.5	12.0	60.1	10.3	87.1	3.7	14.6	6.5	66.9	7.2	46.2	10.4	33.7	8.7	81.4	6.2
angel	3.5	1.0	54.2	12.0	47.6	9.4	57.8	6.0	70.7	11.5	1.6	1.0	45.7	13.2	22.1	10.3	31.4	7.2	64.9	7.8
ant	10.7	1.5	44.2	17.5	41.7	14.5	47.3	12.6	60.8	18.0	7.8	4.9	35.2	9.5	22.5	12.6	27.2	15.0	50.6	18.6
apple	57.5	10.2	83.4	10.7	82.0	8.6	78.2	7.6	94.3	5.4	35.9	8.5	78.3	16.9	59.5	13.1	56.7	14.9	87.1	11.9
backpack	33.6	6.1	59.2	3.9	35.9	3.8	33.7	6.0	64.6	9.2	21.0	10.1	46.6	2.1	6.4	3.8	8.0	4.5	50.2	11.0
basket	30.9	3.5	68.7	15.7	65.9	14.2	55.2	15.1	79.1	10.3	25.4	7.6	61.1	12.7	41.6	20.4	28.8	18.2	70.1	11.3
bulldozer	40.7	6.6	53.4	15.7	56.0	9.2	67.9	5.1	69.1	11.0	22.1	11.4	43.1	18.4	38.8	11.2	49.1	8.9	58.5	13.4
butterfly	43.9	3.7	78.2	9.3	70.2	7.5	65.0	8.1	91.7	3.7	38.8	4.2	67.4	13.2	54.1	9.1	38.9	14.1	86.2	5.8
cactus	40.1	8.0	84.6	4.6	41.9	9.1	47.7	11.7	89.2	6.8	21.9	11.5	80.4	8.2	18.9	12.2	14.1	5.6	83.3	9.4
calculator	28.8	3.3	89.2	4.6	67.3	4.1	52.7	12.1	92.6	2.8	22.3	4.3	87.7	4.8	44.2	6.2	24.5	7.7	90.1	5.0
campfire	32.3	8.1	91.2	3.2	80.7	4.2	73.5	5.0	93.9	1.6	17.6	6.9	88.4	4.6	71.4	11.4	57.1	7.2	89.0	4.2
candle	21.6	3.5	89.8	5.7	86.7	4.5	85.2	1.7	96.3	1.7	25.9	5.7	81.0	10.0	71.9	10.3	67.8	7.4	93.9	2.2
coffee cup	48.0	10.9	73.6	10.6	73.7	5.9	66.2	7.1	82.6	6.9	30.7	10.4	76.2	13.6	54.6	11.8	38.3	15.5	81.3	5.7
crab	28.9	5.5	56.2	13.8	49.5	10.3	48.6	13.2	75.4	12.9	21.6	6.6	51.9	11.8	27.0	8.4	21.6	10.5	69.9	14.8
drill	44.9	12.0	71.3	8.2	80.6	1.9	84.1	1.5	88.7	8.0	26.5	8.2	56.4	5.3	55.1	1.9	68.1	8.6	79.8	10.3
duck	30.3	7.3	61.2	10.5	53.6	4.5	71.2	6.0	89.6	4.1	22.6	5.5	53.9	9.3	26.5	8.4	48.4	10.5	84.0	7.5
face	12.5	1.4	69.9	14.4	38.3	6.8	41.8	10.8	83.3	6.9	17.4	5.4	55.2	16.0	12.4	8.2	16.6	11.5	72.2	11.7
flower	17.6	3.8	75.6	14.2	62.6	3.5	58.1	3.4	88.3	2.0	27.0	3.7	72.7	11.2	36.4	6.5	27.3	3.7	87.1	3.2
house	22.3	8.5	82.2	9.3	57.8	10.7	58.4	9.3	89.4	2.4	16.9	5.4	81.7	7.9	34.7	13.4	32.8	13.3	85.3	2.4
ice cream	30.4	8.9	82.5	5.7	75.2	4.3	72.9	1.0	86.5	8.3	21.0	7.0	79.2	5.9	62.2	5.6	60.0	2.8	80.6	8.3
pig	20.4	1.8	66.8	20.8	37.1	12.2	45.8	9.7	76.6	14.6	20.1	1.9	55.6	20.9	14.8	10.2	20.6	8.6	64.5	18.1
pineapple	27.4	8.6	76.9	13.5	66.6	5.1	56.5	8.4	80.8	8.3	25.3	2.3	74.4	12.3	43.5	11.4	35.0	7.3	75.8	6.8
suitcase	28.7	9.2	89.2	1.6	82.4	5.7	81.7	2.6	93.8	1.6	17.3	7.6	90.7	1.5	72.9	7.2	61.0	5.3	93.2	1.7
Average	29.5	6.3	73.0	10.1	61.2	7.7	60.8	8.0	83.9	7.0	21.3	6.2	66.2	10.8	40.4	10.3	36.9	9.9	77.4	8.9
Average ref.	28.9	6.5	73.3	10.2	70.1	9.2	70.8	11.4	84.1	7.0	25.1	6.4	66.9	10.8	60.1	10.6	61.3	12.1	78.2	8.4
Airplane [35]	20.4	5.4	54.2	14.9	55.3	12.2	57.1	12.8	64.1	13.1	13.1	5.7	40.2	10.4	36.2	6.9	32.5	11.4	53.2	10.1
Airplane [35] ref.	20.8	5.8	54.8	13.6	62.0	13.6	64.2	17.4	65.3	13.8	13.8	5.9	42.2	9.8	52.6	15.8	53.3	19.8	53.6	8.7
Airplane [25]	17.3	6.4	43.7	3.1	45.7	7.0	40.5	9.9	51.9	6.8	14.5	4.2	26.1	2.9	22.2	8.2	13.1	9.2	29.3	7.9
Airplane [25] ref.	16.8	5.2	44.7	5.0	50.7	8.1	44.4	10.9	53.1	6.8	14.9	4.5	30.3	7.6	28.8	7.2	19.2	10.8	33.4	11.3
Creative birds	13.8	4.2	12.5	4.6	25.4	3.9	26.8	2.5	30.3	2.9	12.8	3.5	12.5	2.6	15.1	2.2	12.4	2.8	19.0	2.1
Creative birds ref.	14.5	4.4	12.5	4.6	29.5	7.1	29.6	3.9	30.4	1.7	13.6	3.9	12.5	2.5	20.8	5.3	20.6	5.8	20.1	1

TABLE I: Numerical evaluation on the SPG dataset [22]: first 25 categories; on the ‘airplane’ category from TUBerlin [35] and Huang14 [25] datasets; on creative birds [36]. μ denotes the average accuracy over 5 runs with 5 randomly chosen templates, and σ is the standard deviation of the 5 runs results. The five categories in bold are the categories we use for the detailed analysis and ablation studies. ‘ref.’ refers to the results refined by recomputing the label per point based on the dominant stroke label.

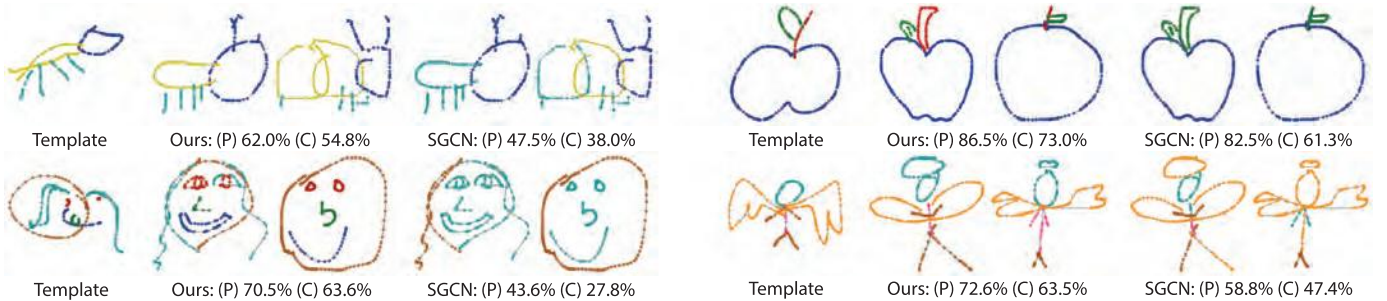


Fig. 4: Comparison between our method and one-shot setting for SGCN [33]. Our method is generally more accurate. The most advantage can be observed on more abstract exemplars (e.g., face) or complex categories (e.g., angel). The numbers indicate (P) point and (C) component average accuracy over the category for the shown template. The points show the sampled points, and the grey thin lines demonstrate their connectivity into strokes.

assigned with a correct label and (b) *component accuracy* (C-metric) – the number of correctly labeled components divided by the total number of components.

It is challenging to develop a faithful evaluation of the few-shot segmentation, since each ground-truth has only one set of labels, which can have different granularity within the same category and dataset. There are two challenging cases: (1) The ground-truth labeling of the target is more fine-grained than the labeling in the exemplar: e.g., a user does not draw or mark a window for an airplane in the exemplar, while the target sketch ground-truth contains a window labeling; (2) The exemplar

labeling is more fine-grained than the ground-truth labeling of the target: e.g., the user labels the plane tail in the exemplar, while the target sketch ground-truth considers it to be a body of the plane. The first case we address automatically, by ignoring during the evaluation the points of the target sketch with the label not-existing in the exemplar. The second case is more challenging since it is impossible to automatically detect when the ground-truth labeling has a different labeling granularity. In the supplemental we provide a more restrictive evaluation on subsets of sketches that have the same set of labels as an exemplar. We though do not observe much differences between

the two evaluation methods.

E. One-shot segmentation: Performance on average

We first evaluate a one-shot segmentation scenario. For each category we randomly select 5 sketches as exemplars, and report an average accuracy over the 5 runs. We train on the training subset of SPG dataset [22] and evaluate on the test subset of SPG dataset and the overlapping categories from Huang14 [25] and TU-Berlin [35], [17] datasets, demonstrating that the segmentation can generalize to sketches from different distributions and to not observed during training sketches. Since Huang14 and TU-Berlin have slightly different set of labels, namely there is no ‘window’ label in both datasets, we change in all used exemplars the ‘window’ label to the ‘body’ label. Similarly, ‘airplane_horistab’ and ‘airplane_vertstab’, labels are merged to a single ‘tail’ label. We also train and test on the creative birds [36] dataset.

Tab. I shows the numerical evaluation of all considered methods. Figures 4, 5 show visual comparisons. Our method results in highest point and component accuracy on all the evaluated datasets and categories. On average on the SPG dataset our method results in 10.8% and 11.2% higher point and component accuracy than the second best method SGCN [33]. On the ‘duck’ category it results in up to 28.4% accuracy increase. FLSS [14] and ISPP [16] have similar performance, and FLSS performs slightly better. On average on the SPG dataset our method results in 22.6% and 37.1% higher point and component accuracy values than FLSS. On the ‘creative birds’ dataset SGCN [33] performs the worst among all the methods, not being able to account for diversity in the dataset. ISPP [16] method results in the second best, after our method, performance on this dataset.

Since the labeling in the considered datasets is defined per stroke (the strokes are broken into multiple at data-annotation stage if is needed), we additionally can perform an easy label refinement step, by assigning to a point a label dominant to the stroke the point belongs to (Tab. I ‘ref.’). Note though that in sketches found in the wild the stroke can have several labels, and such refinement step can reduce the segmentation accuracy. It can be seen that under this setting our method also performs the best. Only on the creative birds dataset our method gives lower C-metric than FLSS and ISPP, but still results in higher P-metric. On the SPG dataset FLSS and ISPP are still losing to SGCN and our methods, where our method gives 10.8% and 11.3% higher point and component accuracy than SGCN. After refinement, on the SPG dataset, the accuracy of our method increases just by 0.2/0.8 points on P/C-metrics, compared to by 10.0/24.4 points for ISSP and 8.9/19.7 points for FLSS. Importantly, these results show that our method is capable of directly taking points connectivity into strokes, but does not pose a strict requirement of one label per stroke.

F. One-shot segmentation: Robustness to number of parts, complexity and diversity

For the remaining experiments, we select 5 categories from the PSG dataset [22] of varying complexity, by selecting categories with different maximum number of parts: *Apple*

has at most three semantic parts, *Duck* – four, *Ambulance* – five, *Face* – seven, and *Pig* – eight.

We first evaluate the robustness of our method under different number of part labels, and compare to the second best method on the SPG dataset – SGCN [33]. For each of the five categories we randomly select 3 templates with

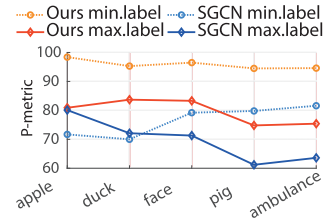


Fig. 6: Performance based on number of parts.

minimum and maximum number of labels. Fig. 6 plots an average accuracy over three runs, and shows that our method consistently outperforms SGCN. Our method achieves an average over the five classes accuracy of 95.9% on a simpler task with little parts (yellow dashed line). Finally, our method is more robust on complex sketches: as the sketches complexity and diversity increase from ‘apple’ to ‘pig’ categories under the challenging task with many labels, SGCN performance degrades by 18.9 points (solid blue line), versus just 6.1 points with our method (solid red line).

G. One-shot vs. few-shot

Tab. II shows that the performance improves if there are several templates available, and our results consistently outperform SGCN. For these experiment we select 5 templates for each category with consistent semantic parts, selecting the most frequent number of parts in each class ground-truth labeling. For the results on 1 template, we average the labeling accuracy over labeling results with each template individually. For the 3 templates, we average over all unique subsets of 5 templates. When multiple templates are provided, to choose the best we perform the hierarchical deformation estimation and select the one that results in the smallest Chamfer distance.

	Category	1 template		3 templates		5 templates	
		(P)	(C)	(P)	(C)	(P)	(C)
Ours	Average 5	85.7	76.0	91.3	85.4	92.4	87.8
SGCN [33]	Average 5	77.9	63.4	86.5	79.4	86.5	79.2

TABLE II: One shot vs. few shot. See Sec.IV-G for the details.

H. Generalization to unseen sketches of the same category

All results in our work are evaluated on sketches that are not observed during training. Yet, the target unlabeled sketches in practice can be used for training. Our results in Table III demonstrate that the performance is the same whether the target sketches are used for training directly or not, showing good generalization properties of our method.

I. Discussion

Our first advantage over SGCN lies in the ability to generate structural variations on the exemplar, making the training robust towards different sketch abstractions and styles. Second, our global deformation step from the reference to the exemplar

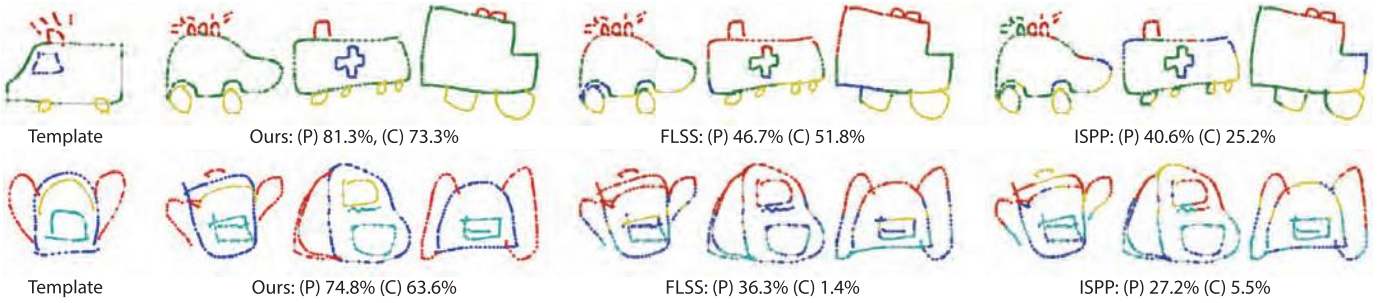


Fig. 5: Comparison of our method with FLSS [14] and ISPP [16] on the example templates from ‘ambulance’ and ‘backpack’ categories of the SPG dataset [22]. The numbers indicate (P) point and (C) component average accuracy for the shown template. The points show the sampled points, and the grey thin lines demonstrate their connectivity into strokes.

P-metric		C-metric					
Training		Test		Training		Test	
μ	σ	μ	σ	μ	σ	μ	σ
85.9	6.7	86.2	6.9	77.7	10.3	77.8	11.1

TABLE III: Comparison of the segmentation results on the unlabeled sketches from the training set (‘Training’) and on unseen sketches (‘Test’). The results are averaged over the five selected categories, over the runs with the same templates as in Tab. I.



Fig. 7: The deformation process: our method vs. FLSS [14].

lifts the requirement on the segmentation module to be rotation/reflection invariant. Our advantage over FLSS and ISPP methods lies in the ability to account for points connectivity into strokes, resulting in more meaningful labeling. Compared to ISPP, our robust segmentation network makes our algorithm more tolerant towards mistakes in the keypoints prediction. Finally, our sketch dedicated deformation model allows to better preserve sketch structure than FLSS, as shown in Fig. 7.

Our one-shot segmentation performance does not yet reach the performance of the fully supervised methods. For instance, the state-of-the-art SGCN under the fully supervised non-personalized (all sketches have different segmentation granularity) setting achieves accuracy of 96.8%/94.3% on P/C metrics. Nevertheless, as it can be seen in Sec. IV-G the performance of the personalized sketch segmentation quickly increases when there are more templates available. Moreover, our solution consistently achieves higher accuracy in one-shot and few-shot personalized segmentation scenarios than the state-of-the-art segmentation method, demonstrating the efficiency of our deformation module.

J. Ablation studies

a) Global transformation (G.T.) and keypoints loss:

Tab. IV (‘Ours’ vs. ‘No G.T.’) shows the importance of

Ours full		No G.T.		Chamfer G.T.		Reverse G.T.		No \mathcal{F}^{stroke}	
(P)	(C)	(P)	(C)	(P)	(C)	(P)	(C)	(P)	(C)
86.2	77.8	68.5	60.1	49.6	29.1	57.9	44.1	60.8	36.0

TABLE IV: Ablation studies: The results are averaged over the five selected categories, over the runs with the same templates as in Tab. I. ‘No G.T.’: We skip the step of global transformation (G.T.), and perform only stroke-level deformations. ‘Chamfer G.T.’: The global transformation is estimated with the Chamfer distance, instead of relying on the distances between keypoints. ‘Reverse G.T.’: We estimate the global transformation from an exemplar towards a reference. ‘No \mathcal{F}^{stroke} ’: We remove \mathcal{F}^{stroke} in Eq. 9.

P-metric				C-metric			
PointNet++		Ours		PointNet++		Ours	
μ	σ	μ	σ	μ	σ	μ	σ
39.9	9.7	86.2	6.9	17.2	6.6	77.8	11.1

TABLE V: PointNet++ encoder [39] vs. our GCN encoder. The results are averaged over the five selected categories, over the runs with the same templates as in Tab. I.

hierarchical estimation: If we remove the step of global transformation the average over the 5 representative categories drops by 17.7 points in terms of the point- and component-based accuracy. Similarly, we demonstrate the importance of globally warping reference sketches towards an exemplar, rather than an exemplar towards reference sketches (‘Ours’ vs. ‘Reverse G.T.’).

Further, Tab. IV (‘Ours’ vs. ‘Chamfer G.T.’) shows the importance of relying on keypoints instead of the Chamfer distance between the points of two sparse sketches. If the Chamfer distance is used, the alignment does not necessary respect the semantics of strokes.

b) *Stroke-level information for segmentation:* We show that accounting for stroke-level information in the segmentation module (Sec.III-E, Eq. 4) has high impact on the accuracy of the prediction (Tab. IV (‘Ours’ vs. ‘No \mathcal{F}^{stroke} ’)).

c) *Graph-based vs. point cloud-based encoder:* We demonstrate an advantage of a Graph Convolutional Network (GCN) over point cloud encoders for the few shot sketch segmentation problem. We exploit here the PointNet++ [39] encoder, which we use instead of GCN, keeping the architecture otherwise the same. Table V and Fig. 8 show that point cloud encoder is not capable of capturing well stroke-level information, resulting in poor segmentation performance.

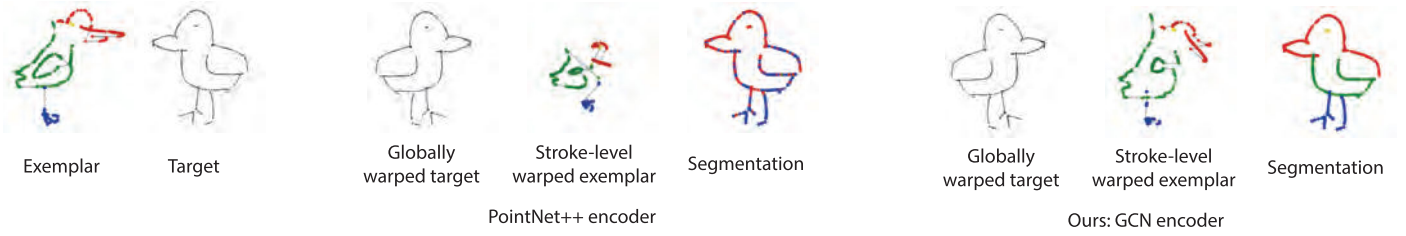


Fig. 8: Visual comparison of the deformation and segmentation results, depending on the used sketch encoder: PointNet++ [39] or our GCN encoder.

No $\mathcal{L}_{rot}, \mathcal{L}_{\sigma}, \mathcal{L}_t$		No \mathcal{L}_{rot}		No \mathcal{L}_{σ}		No \mathcal{L}_t		Ours full	
(P)	(C)	(P)	(C)	(P)	(C)	(P)	(C)	(P)	(C)
61.4	42.96	81.04	69.62	77.36	67.82	63.52	44.14	86.18	77.84

TABLE VI: The role of constraints on stroke-level deformation. The results are averaged over the five selected categories, over the runs with the same templates as in Tab. I.

d) *Soft constraints on stroke-level deformation:* To evaluate the role of our soft constraints on stroke-level deformation, we first disable all the constraints $\mathcal{L}_{rot}, \mathcal{L}_{\sigma,j}, \mathcal{L}_{t,j}$, apart from \mathcal{L}_{orth} in Eq. 6. We then remove each of the three terms individually. Table VI demonstrates the importance of these constraints. \mathcal{L}_t is the most important since it limits how far strokes can move from their original positions, ensuring global sketch structure maintenance.

Despite such constraints help to preserve the relative part relationship, they limit the space of achievable deformations. For instance, Fig. 7 shows that our model just roughly aligns the exemplar to the reference. To improve on our results, future work should investigate alternative deformation models that can preserve the relative part relationship while achieving better alignment with the reference sketch.

V. CONCLUSION

We present the first one-shot personalized sketch segmentation method and study a set of alternatives solutions adopting the state-of-the-art segmentation and perceptual grouper networks, and two 3D shape few-shot segmentation networks. We address this problem by estimating the deformation from an exemplar sketch towards a reference sketch and training a robust part label predictor network on the warped exemplars. Our key contributions lie in proposing a hierarchical deformation model that works at both sketch- and stroke-level. Our hierarchical two-ways deformation model allows to explicitly account for ambiguity in global sketch orientation, resulting in more robust segmentation results. We also demonstrate the importance of taking stroke connectivity into consideration and compare point cloud and graph-based encoders. We show that our method by far outperforms all existing alternatives, showing robust performance on the highly abstract exemplars and complex categories.

REFERENCES

- [1] Q. Liu, L. Xie, H. Wang, and A. L. Yuille, "Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.
- [2] A. Ghosh, R. Zhang, P. K. Dokania, O. Wang, A. A. Efros, P. H. S. Torr, and E. Shechtman, "Interactive sketch & fill: Multiclass sketch-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.
- [3] C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, and C. Zou, "Sketchycoco: Image generation from freehand scene sketches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
- [4] Q. Zheng, Z. Li, and A. Bargteil, "Learning to shadow hand-drawn sketches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
- [5] H. Lin, Y. Fu, X. Xue, and Y.-G. Jiang, "Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
- [6] R. Liu, Q. Yu, and S. Yu, "Unsupervised sketch to photo synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020.
- [7] F. Liu, C. Zou, X. Deng, R. Zuo, Y.-K. Lai, C. Ma, Y.-J. Liu, and H. Wang, "Scenesketcher: Fine-grained image retrieval with scene sketches," in *Proc. Eur. Conf. Comput. Vis.*, 2020.
- [8] R. K. Jones, T. Barton, X. Xu, K. Wang, E. Jiang, P. Guerrero, N. J. Mitra, and D. Ritchie, "Shapeassembly: Learning to generate programs for 3d shape structure synthesis," *ACM Trans. Graph.*, vol. 39, no. 6, 2020.
- [9] K. Li, K. Pang, Y.-Z. Song, T. Hospedales, H. Zhang, and Y. Hu, "Fine-grained sketch-based image retrieval: The role of part-aware attributes," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016.
- [10] C. Zou, Q. Yu, R. Du, H. Mo, Y.-Z. Song, T. Xiang, C. Gao, B. Chen, and H. Zhang, "Sketchyscene: Richly-annotated scene sketches," in *Proc. Eur. Conf. Comput. Vis.*, September 2018.
- [11] G. Su, Y. Qi, K. Pang, J. Yang, Y.-Z. Song, and C. SketchX, "Sketch-healer: A graph-to-sequence network for recreating partial human sketches," in *Proc. Brit. Mach. Vis. Conf.*, 2020.
- [12] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001.
- [13] F. Perteneder, M. Bresler, E.-M. Grossauer, J. Leong, and M. Haller, "cluster: Smart clustering of free-hand sketches on large interactive surfaces," in *Proc. ACM Symp. User Interface Softw. Technol.*, 2015.
- [14] L. Wang, X. Li, and Y. Fang, "Few-shot learning of part-specific probability space for 3d shape segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
- [15] S. Yuan and Y. Fang, "Ross: Robust learning of one-shot 3d shape segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020.
- [16] N. Chen, L. Liu, Z. Cui, R. Chen, D. Ceylan, C. Tu, and W. Wang, "Unsupervised learning of intrinsic structural representation points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
- [17] L. Li, H. Fu, and C.-L. Tai, "Fast sketch segmentation and labeling with deep learning," *IEEE Comput. Graph. Appl.*, vol. 39, no. 2, 2018.
- [18] R. G. Schneider and T. Tuytelaars, "Example-based sketch segmentation and labeling using crfs," *ACM Trans. Graph.*, vol. 35, no. 5, 2016.
- [19] X. Zhu, Y. Xiao, and Y. Zheng, "2d freehand sketch labeling using cnn and crf," *Multimed. Tools. Appl.*, vol. 79, no. 1, 2020.
- [20] Y. Qi, J. Guo, Y. Li, H. Zhang, T. Xiang, and Y.-Z. Song, "Sketching by perceptual grouping," in *Proc. IEEE Int. Conf. on Image Process.*, 2013.
- [21] Y. Qi, Y.-Z. Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, and J. Guo, "Making better use of edges via perceptual grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.
- [22] K. Li, K. Pang, J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, and H. Zhang, "Universal sketch perceptual grouping," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

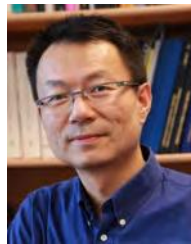
- [23] C. Liu, E. Rosales, and A. Sheffer, "Strokeaggregator: Consolidating raw sketches into artist-intended curve drawings," *ACM Trans. Graph.*, vol. 37, no. 4, 2018.
- [24] Z. Sun, C. Wang, L. Zhang, and L. Zhang, "Free hand-drawn sketch segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2012.
- [25] Z. Huang, H. Fu, and R. W. Lau, "Data-driven segmentation and labeling of freehand sketches," *ACM Trans. Graph.*, vol. 33, no. 6, 2014.
- [26] X. Zhu, Y. Xiao, and Y. Zheng, "Part-level sketch segmentation and labeling using dual-cnn," in *Proc. IEEE Int. Conf. on Image Process.*, 2018.
- [27] X. Wu, Y. Qi, J. Liu, and J. Yang, "Sketchsegnet: A rnn model for labeling sketch strokes," in *IEEE Int. Workshop Mach. Learn. Signal Process.*, 2018.
- [28] X. Wang, X. Chen, and Z. Zha, "Sketchpointnet: A compact network for robust sketch recognition," in *Proc. IEEE Int. Conf. on Image Process.*, 2018.
- [29] K. Li, K. Pang, Y.-Z. Song, T. Xiang, T. M. Hospedales, and H. Zhang, "Toward deep universal sketch perceptual grouper," *IEEE Trans. Image Process.*, vol. 28, no. 7, 2019.
- [30] Y. Qi and Z.-H. Tan, "Sketchsegnet+: An end-to-end learning of rnn for multi-class sketch semantic segmentation," *IEEE Access*, vol. 7, 2019.
- [31] F. Wang, S. Lin, H. Wu, H. Li, R. Wang, X. Luo, and X. He, "Spfusionnet: Sketch segmentation using multi-modal data fusion," in *Proc. IEEE Int. Conf. on Multimedia and Exposition*, 2019.
- [32] F. Hähnlein, Y. Gryaditskaya, and A. Bousseau, "Bitmap or vector? a study on sketch representations for deep stroke segmentation," in *Journées Françaises d'Informatique Graphique et de Réalité virtuelle*, 2019.
- [33] L. Yang, J. Zhuang, H. Fu, X. Wei, K. Zhou, and Y. Zheng, "Sketchgnn: Semantic sketch segmentation with graph neural networks," *ACM Trans. Graph.*, vol. 40, no. 3, pp. 1–13, 2021.
- [34] X. Zhu, Y. Xiao, Y. Zheng, G. Tan, and S. Zhou, "Sketchppnet: A joint pixel and point convolutional neural network for low resolution sketch image recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020.
- [35] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, 2012.
- [36] S. Ge, V. Goswami, C. L. Zitnick, and D. Parikh, "Creative sketch generation," in *Int. Conf. Learn. Represent.*, 2021.
- [37] F. Wang, S. Lin, H. Li, H. Wu, T. Cai, X. Luo, and R. Wang, "Multi-column point-cnn for sketch segmentation," *Neurocomputing*, vol. 392, 2020.
- [38] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.
- [39] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5105–5114.
- [40] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "Pointnnc: Convolution on χ -transformed points," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 828–838.
- [41] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, 2019.
- [42] L. Yang, A. Sain, L. Li, Y. Qi, H. Zhang, and Y.-Z. Song, "S 3 net: Graph representational network for sketch recognition," in *Proc. IEEE Int. Conf. on Multimedia and Exposition*, 2020, pp. 1–6.
- [43] D. Ha and D. Eck, "A neural representation of sketch drawings," in *Int. Conf. Learn. Represent.*, 2018.
- [44] L. Li, C. Zou, Y. Zheng, Q. Su, H. Fu, and C.-L. Tai, "Sketch-R2CNN: An rnn-rasterization-cnn architecture for vector sketch recognition," *IEEE Trans. Vis. Comput. Graph.*, 2020.
- [45] Q. Jia, X. Fan, M. Yu, Y. Liu, D. Wang, and L. J. Latecki, "Coupling deep textural and shape features for sketch recognition," in *Proc. of ACM International Conference on Multimedia*, 2020.
- [46] G. Li, M. Muller, A. Thabet, and B. Ghanem, "Deepgcns: Can gcns go as deep as cnns?" in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.
- [47] Z. Chen, K. Yin, M. Fisher, S. Chaudhuri, and H. Zhang, "Bae-net: Branched autoencoder for shape co-segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.
- [48] O. Sorkine, "Least-squares rigid motion using svd," *Technical notes*, vol. 120, no. 3, 2009.



Anran Qi is a PhD student of SketchX Lab, Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. Her sketch focus on sketch oriented or aided 3D shaped research topic, including sketch-based 3D shape retrieval and 3D shape reconstruction.



Yulia Gryaditskaya is a Senior Research Fellow in Computer Vision and Machine Learning at the Centre for Vision Speech and Signal Processing (CVSSP). Previously, she was a postdoctoral researcher at Inria, Sophia Antipolis, France. She obtained her PhD in 2017 on Computer Vision and Graphics from the Max-Planck institute for informatics, Saarbruecken, Germany. She got her diploma from Lomonosov Moscow State University, Moscow, Russia. Her research interests cover sketch-based modeling, sketch beautification, geometric deep learning, sketch classification, sketch generation, high dynamic range image and video capture, tone-mapping and calibration, depth estimation from the structured light-fields, materials representation and editing.



Tao Xiang received the PhD degree in electrical and computer engineering from the National University of Singapore in 2001. He is currently a Professor in the Department of Electrical and Electronic Engineering, University of Surrey and a principal researcher at Samsung AI center. His research interests include computer vision, machine learning, and data mining. He has published over 150 papers in international journals and conferences.



Yi-Zhe Song Yi-Zhe Song is a Professor of Computer Vision and Machine Learning, and Director of SketchX Lab at the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. Previously, he was a Senior Lecturer at the Queen Mary University of London, and a Research and Teaching Fellow at the University of Bath. He obtained his PhD in 2008 on Computer Vision and Machine Learning from the University of Bath, and received a Best Dissertation Award from his MSc degree at the University of Cambridge in 2004, after getting a First Class Honours degree from the University of Bath in 2003. He is a Program Chair for British Machine Vision Conference (BMVC) 2021, and regularly serves as Area Chair (AC) for flagship computer vision and machine learning conferences, most recently at ICCV'21, and BMVC'20. He is a Senior Member of IEEE, a Fellow of the Higher Education Academy, as well as full member of the EPSRC review college, the UK's main agency for funding research in engineering and the physical sciences.

VI. SUPPLEMENTAL WEB-PAGES

We provide the supplemental web-pages that show the 5 templates, used for each category to compute the results in Table I in the main document, and the representative segmentation results for each method.

VII. ONE-SHOT SEGMENTATION

A. Alternative evaluation

As we describe in Section IV-D in the main document:

It is challenging to develop a faithful evaluation of the few-shot segmentation, since each ground-truth has only one set of labels, which can have different granularity within the same category and dataset. There are two challenging cases: (1) The ground-truth labeling of the target is more fine-grained than the labeling in the exemplar: *e.g.*, a user does not draw or mark a window for an airplane in the exemplar, while the target sketch ground-truth contains a window labeling; (2) The exemplar labeling is more fine-grained than the ground-truth labeling of the target: *e.g.*, the user labels the plane tail in the exemplar, while the target sketch ground-truth considers it to be a body of the plane. The first case we address automatically, by ignoring during the evaluation the points of the target sketch with the label not-existing in the exemplar. The second case is more challenging, since it is not possible to automatically distinguish the difference in amount of labels due to labeling granularity, with due to a different amount of semantic parts.

We thus provide in Table VII a more restrictive evaluation on subsets of sketches that have the same set of labels as an exemplar. Compared to the evaluation in the main paper, numerical results in Table VII do not account for the cases when the target sketch has less parts and only a part of labels has to be transferred. This is the reason why in the main document we use a less restrictive evaluation strategy. It can be seen that similarly to the results in the main document our approach outperforms the alternative solutions.

The remaining experiments in this document use the evaluation strategy used in the main document.

B. Detailed numerical evaluation after label refinement

As we mention in the main document:

Since the labeling in the considered datasets is defined per stroke (the strokes are broken into multiple at data-annotation stage if is needed), we additionally can perform an easy label refinement step, by assigning to a point a label dominant to the stroke the point belongs to (Tab. I ‘ref.’). Note though that in sketches found in the wild the stroke can have several labels, and such refinement step can reduce the segmentation accuracy. It can be seen that under this setting our method also performs the best. Only on creative birds dataset our method gives lower C-metric than FLSS and ISPP, but still results

in higher P-metric. On the SPG dataset FLSS and ISPP are still losing to SGCN and our methods, where our method gives 10.8% and 11.3% higher point and component accuracy than SGCN. After refinement, on the SPG dataset, an accuracy of our method increases just by 0.2/0.8 points on P/C-metrics, compared to by 10.0/24.4 points for ISPP and 8.9/19.7 points for FLSS. Importantly, these results show that our method is capable of directly taking points connectivity into strokes, but does not pose a strict requirement of one label per stroke.

In Table VIII we provide the detailed numerical results per category. While on average our method outperforms competing approaches after refinement, our method is outperformed by ISPP method on the ‘bulldozer’ category and tightly follows FLSS on the ‘suitcase’ category. The worse performance of our method than the ISPP method on the ‘bulldozer’ category can be explained by the fact that we solve jointly for the keypoints and stroke-level transformations. In this case, the prediction of keypoints sometimes can degrade, resulting in the method not being able to correctly estimate the global reflection between the two sketches, *e.g.* ‘bulldozer’ facing right or left. In Section VIII-D we evaluate a separate training strategy, where the keypoints prediction network is trained separately. While separate training does increase the performance on the ‘bulldozer’ category by 15.3 points, in overall, the joint training strategy results in more stable performance across categories, showing better results on more categories. Please see Section VIII-D for the further comparison of these two strategies.

C. One-shot vs. few-shot

In the main paper we show in Table II that the performance improves if there are several templates available, and our results consistently outperform SGCN.

Here in Table IX we show the numerical evaluation per category.

VIII. ABLATION STUDIES

A. ISPP: GCN vs PointNet++ encoder

Table X shows that when the PointNet++ encoder is used as was proposed in the original paper, the ISPP method performance on one shot sketch segmentation consistently drops: The point accuracy reduces on average over the five categories by 3.6 points, and the component accuracy – by 5 points.

B. Segmentation module

As we mention in Section IV-C in the main document:

At inference, to obtain the labeling via Eq. 9, we first estimate our hierarchical deformation, then the label of a point v_i is obtained as follows $\tau(v_i) = \tau_{\theta_3}(v_i, \mathcal{F}_{s_j \in \tilde{X}: v_i \in s_j}^{stroke}, \mathcal{F}_{\tilde{E}}^{sketch})$.

Here we compare this strategy with the strategy of passing in an encoding of a globally warped target sketch $\mathcal{F}_{\tilde{X}}^{sketch}$,

Category	P-metric										C-metric									
	SPGG [22]		SGCN [33]		FLSS [14]		ISPP [16]		Ours		SPGG [22]		SGCN [33]		FLSS [14]		ISPP [16]		Ours	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
airplane	22.1	8.6	66.0	14.6	56.7	12.3	51.1	13.5	83.4	8.7	21.8	7.8	56.7	17.5	38.3	15.0	26.0	15.3	76.2	12.7
alarm clock	23.0	5.3	81.8	11.4	60.7	12.9	59.5	13.7	86.5	12.3	8.7	2.0	72.5	15.9	42.8	11.8	36.1	17.8	77.1	19.7
ambulance	28.5	10.3	76.6	7.2	62.0	11.9	60.7	11.7	85.8	6.3	15.1	7.2	64.8	11.0	49.6	8.4	32.5	9.1	79.0	10.2
angel	1.8	1.8	57.2	9.8	45.0	19.0	61.3	5.4	71.7	14.7	1.9	1.9	46.3	13.2	25.3	17.0	33.0	10.7	65.1	13.2
ant	7.5	4.7	44.8	19.5	42.4	17.1	47.6	11.5	61.8	22.1	9.9	6.1	34.5	13.4	23.4	11.7	23.0	14.2	46.4	21.0
apple	52.1	11.5	83.4	8.2	80.7	6.5	78.9	7.1	93.4	4.3	32.9	8.1	78.9	13.5	60.3	11.8	60.9	17.4	85.6	8.7
backpack	30.0	8.7	55.4	6.5	35.2	5.0	34.4	5.2	63.7	13.0	14.0	7.2	45.2	8.5	5.5	4.1	14.5	7.4	48.0	10.9
basket	21.9	8.4	67.7	16.5	67.3	16.5	55.9	18.1	81.0	12.8	22.3	11.4	61.3	13.2	44.0	21.3	32.5	21.3	74.9	16.3
bulldozer	34.7	15.5	51.8	13.1	53.7	5.5	66.7	2.9	69.6	12.3	20.8	11.9	40.8	13.3	33.7	2.7	46.6	4.5	58.9	13.0
butterfly	42.0	6.1	79.6	11.3	68.0	8.9	63.7	8.9	93.2	3.7	35.4	6.2	71.4	17.3	47.1	26.0	33.7	19.9	90.4	5.2
cactus	30.3	11.5	86.6	13.4	49.0	10.9	52.3	9.3	90.3	10.5	25.1	11.7	86.4	12.4	27.4	20.2	16.6	11.5	89.9	11.2
calculator	25.3	5.2	90.5	5.9	62.3	10.0	49.2	14.3	94.0	3.8	18.5	10.2	89.3	7.0	44.4	6.7	21.1	12.4	91.4	6.3
campfire	30.8	8.1	91.7	3.2	81.1	4.0	74.0	5.1	94.5	1.7	17.6	6.4	89.9	4.6	72.7	10.9	58.3	6.9	90.6	4.2
candle	21.0	4.4	90.6	5.5	87.0	2.6	86.4	2.0	97.0	1.8	27.0	4.0	80.6	9.2	70.4	7.6	69.4	7.4	96.2	2.2
coffee cup	44.6	11.9	79.4	9.2	71.5	7.5	74.2	3.1	87.8	4.3	24.9	12.1	81.6	8.0	53.4	6.1	55.0	4.5	85.5	4.3
crab	25.7	3.7	60.1	22.6	49.1	12.8	48.7	13.2	75.4	14.5	21.0	3.8	57.9	23.7	28.3	9.7	25.4	16.7	70.2	17.8
drill	44.5	11.8	71.2	7.9	80.9	1.9	84.9	1.3	88.6	8.2	26.7	6.9	55.2	5.2	53.4	2.1	68.6	9.9	79.7	8.4
duck	27.3	5.5	66.9	12.7	56.0	10.0	72.7	5.3	91.1	4.3	20.6	8.8	58.0	15.4	30.2	16.4	51.4	10.2	86.6	7.5
face	11.0	3.8	60.1	18.7	37.8	9.1	39.6	11.0	70.4	13.9	16.5	7.8	45.4	15.4	11.6	4.8	16.2	10.0	61.7	12.2
flower	18.1	3.4	74.8	14.5	63.0	2.1	58.0	2.8	88.6	2.0	25.6	5.1	71.0	11.1	36.2	7.9	28.0	1.7	90.0	2.0
house	23.5	9.6	79.9	9.6	59.5	9.3	58.9	8.5	90.8	3.8	19.8	7.5	78.8	9.2	36.7	7.2	32.3	9.8	86.9	4.4
ice cream	30.3	9.8	83.9	6.0	75.8	4.8	73.0	1.5	88.8	7.4	22.3	7.0	81.1	7.5	63.7	6.7	62.5	3.5	84.3	8.5
pig	21.7	1.1	68.9	26.4	34.3	17.2	51.2	5.6	77.3	26.2	22.4	3.6	59.7	25.4	15.0	12.4	26.2	4.5	69.1	26.2
pineapple	29.4	6.2	76.0	11.9	65.1	3.9	58.5	5.3	79.2	7.2	29.4	3.1	72.8	10.8	40.4	7.5	36.1	6.4	73.1	7.0
suitcase	30.4	14.2	89.3	2.3	82.7	4.8	81.2	4.3	93.7	3.0	16.3	6.1	91.0	2.5	72.4	7.4	60.0	7.2	92.6	3.4
Average	27.1	7.6	73.4	11.5	61.1	9.1	61.7	7.6	83.9	8.9	20.7	7.0	66.8	12.2	41.0	10.5	38.6	10.4	78.0	10.3
Airplane [35]	20.0	6.3	53.0	13.9	55.9	19.1	57.8	14.0	67.2	19.7	15.2	8.5	40.1	8.2	36.4	8.3	28.6	5.0	58.9	23.9
Airplane[25]	16.4	10.2	32.7	19.1	29.7	23.1	35.4	23.8	37.9	20.4	9.6	6.0	15.6	11.3	13.7	11.9	6.7	8.1	26.2	15.0
Creative birds	13.9	4.1	13.8	5.9	28.2	9.3	28.3	3.3	29.9	2.6	15.3	2.2	14.4	1.9	16.5	7.7	12.6	3.8	19.1	4.7

TABLE VII: Numerical evaluation on the SPG dataset [22]: first 25 categories; on the ‘airplane’ category from TUBerlin [35] and Huang14 [25] datasets; on creative birds [36]. μ denotes the average accuracy over 5 runs with 5 randomly chosen templates, and σ is the standard deviation of the 5 runs results. The evaluation in this table is done only on those sketches that have the same semantic parts as an exemplar sketch.

instead of an encoding of a stroke-level warped exemplar $\mathcal{F}_{\hat{E}}^{sketch}$. Table XI shows that this strategy slightly loses the one we use in the main paper.

C. Chamfer distance in the stroke-level deformation

Finally, we evaluate the role of the Chamfer distance in Equation 6. Table XII shows the segmentation accuracy if the stroke level-deformation is guided only by the mean square distance between the keypoints of the deformed template \hat{E} and the keypoints of the globally deformed sketch \hat{X} : $\mathcal{L}_{MSE}(K_{\hat{E}}, K_{\hat{X}})$. It can be seen that using both losses $\mathcal{L}_{MSE}(K_{\hat{E}}, K_{\hat{X}})$ and $\mathcal{L}_{CD}(\hat{X}, \hat{E}_{\hat{X}})$ gives a slight advantage over using the keypoints loss only.

D. Two steps training: Isolated training for keypoints

In this section we evaluate the overall performance of our method, if we train in two steps. First, we train a keypoints estimation module with our GCN sketch encoder. Then, we train the deformation and segmentation modules. In this case the GCN encoders are trained separately at each step. Table XIII provides the comparison between SGCN [33], FLSS [14], ISPP [16], ours joint training strategy used in the main document (Ours Joint), and a two steps training (Ours Separate). It can be seen that on average separate training results in a slightly better average segmentation accuracy with P-metric of 84% vs. 83.9%, and C-metric of 77.6% vs. 77.4%. Nevertheless, (Ours Joint) strategy gives higher points accuracy than (Ours Separate) on 14 out of 25 categories on the SPG dataset. Moreover, (Ours Joint) consistently outperforms all

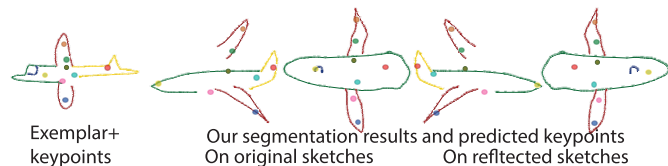


Fig. 9: Keypoints and segmentation results. We visualize 8 keypoints, while use 256 for deformations computations.

other methods, while (Ours Separate) gives lower accuracy than SGCN on the ‘backpack’ and ‘house’ categories. We observe that the stroke-level deformation benefits from joint training, although, for some categories, it comes at cost of decreased performance of the keypoints prediction step (e.g. the ‘bulldozer’ category). Joint strategy results in a more robust performance across the categories with the standard deviation of point accuracy equal to 9.6% versus 10.1% for the separate training strategy (Table XIII).

E. Keypoints sensitivity to rotations and robustness of their prediction

As demonstrated in the supplemental web-pages and in Fig. 9, keypoints prediction is robust to rotations, not affecting the segmentation performance. The mean μ and standard deviation σ of mean L_2 -distances between the keypoints from the original sketch and its reflected version (after reflecting back), on the ablation categories is $\mu = 0.058$, $\sigma = 0.008$. All sketches are normalized to fit the $[-0.5, 0.5]$ bounding box.

REFERENCES

- [1] Q. Liu, L. Xie, H. Wang, and A. L. Yuille, “Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.

Category	P-metric										C-metric									
	SPGG [22]		SGCN [33]		FLSS [14]		ISPP [16]		Ours		SPGG [22]		SGCN [33]		FLSS [14]		ISPP [16]		Ours	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
airplane	23.8	10.5	66.8	13.7	71.5	12.8	57.4	20.7	86.5	4.6	23.4	9.9	59.1	15.4	62.4	15.7	50.3	20.7	81.0	7.7
alarm clock	28.1	8.4	79.9	9.9	78.9	11.4	76.5	10.7	86.6	8.9	17.0	7.4	68.9	15.6	65.7	16.8	63.8	14.9	76.4	14.9
ambulance	28.7	9.3	78.3	3.1	70.5	13.5	68.3	9.5	87.3	3.3	16.1	7.9	67.6	7.3	62.8	11.5	48.5	8.3	81.2	5.4
angel	2.3	0.8	54.3	12.2	54.3	13.9	69.9	6.6	71.0	11.4	2.1	1.2	46.5	13.8	44.3	16.6	59.2	7.7	67.1	7.2
ant	10.1	1.3	44.1	17.4	41.5	14.3	48.6	14.5	60.5	17.9	9.9	6.0	35.4	9.3	30.8	10.7	40.4	12.9	51.0	18.0
apple	59.1	8.8	82.6	11.4	90.6	5.0	91.3	4.9	94.3	5.4	40.8	4.7	78.9	16.2	73.9	12.4	80.1	13.7	87.4	11.5
backpack	34.5	5.8	59.4	4.1	48.2	6.0	44.1	10.4	65.3	9.4	27.0	4.0	47.7	2.5	35.3	5.9	34.0	7.0	52.5	10.6
basket	28.6	2.6	68.9	15.8	77.5	14.9	66.9	21.3	79.2	11.1	27.3	8.2	61.9	13.0	68.4	17.9	58.3	23.4	70.4	11.3
bulldozer	42.2	7.6	53.6	15.6	59.8	10.7	78.6	7.2	69.2	10.9	27.2	10.7	44.7	17.7	53.4	11.5	72.7	6.6	59.0	12.8
butterfly	44.6	3.8	78.0	9.9	82.0	13.8	74.0	8.4	91.8	3.5	43.3	5.3	68.9	14.0	73.9	19.8	62.6	8.7	87.3	5.1
cactus	37.9	8.2	84.6	4.6	36.8	15.5	49.1	21.4	89.0	7.4	20.9	10.6	80.5	8.3	30.7	15.6	36.9	21.4	84.1	10.2
calculator	24.8	2.2	89.2	4.6	83.1	4.3	69.7	20.8	92.8	2.9	27.7	3.5	88.0	4.5	69.0	6.9	57.4	17.5	90.5	5.5
campfire	33.5	9.6	92.1	3.2	91.3	6.7	79.9	12.6	94.0	1.6	22.2	6.7	88.6	4.6	88.4	9.2	75.3	11.5	88.6	3.9
candle	19.1	5.4	89.9	5.7	95.4	1.8	94.1	2.5	96.4	1.4	28.3	4.7	81.6	10.2	87.2	5.9	82.7	8.6	94.2	2.0
coffee cup	50.8	11.0	78.6	11.2	80.8	6.7	79.3	10.9	82.4	7.8	34.6	9.2	77.9	14.9	76.2	9.3	66.0	13.7	82.2	5.0
crab	28.4	5.5	56.2	13.7	49.0	14.9	51.2	15.9	75.6	13.1	25.5	5.6	52.2	11.7	39.2	16.0	40.2	15.3	70.3	15.1
drill	46.6	14.3	71.5	8.0	82.1	1.0	85.0	4.2	89.1	7.8	31.9	10.1	57.6	5.2	55.8	3.1	77.0	10.5	80.7	9.0
duck	28.4	6.8	61.1	10.4	54.2	5.8	82.8	7.6	89.7	3.8	25.2	4.7	54.0	9.4	41.9	8.2	71.4	9.0	83.5	7.3
face	12.3	1.8	70.0	14.3	57.1	10.7	58.8	21.5	83.7	6.8	20.0	5.5	55.8	16.1	39.1	7.2	46.5	18.7	74.7	9.6
flower	14.1	1.8	75.7	14.4	73.6	2.7	78.7	4.0	89.2	1.9	31.5	1.4	73.9	10.4	58.2	4.2	72.7	9.6	89.9	2.1
house	20.0	7.5	82.3	9.3	76.3	7.7	72.1	5.8	89.5	2.5	20.7	7.5	82.1	7.9	68.0	8.4	66.8	6.1	86.2	2.2
ice cream	28.5	9.2	82.6	5.8	78.7	11.5	81.9	4.9	86.7	8.5	27.8	9.7	79.3	6.2	76.0	10.3	81.1	4.2	81.0	8.5
pig	21.5	2.5	67.0	20.8	44.5	17.9	60.2	15.2	77.4	12.6	25.3	0.6	56.1	20.8	34.5	15.1	45.7	12.7	66.0	16.9
pineapple	25.2	7.9	76.9	13.6	81.0	4.4	64.5	17.7	81.0	8.2	30.1	4.8	74.8	12.5	72.6	3.8	60.7	13.0	76.9	6.8
suitcase	28.3	9.8	89.3	1.5	94.7	2.4	88.1	5.7	93.7	1.7	22.4	9.4	90.8	1.6	93.7	2.9	81.4	7.8	93.0	2.0
Average	28.9	6.5	73.3	10.2	70.1	9.2	70.8	11.4	84.1	7.0	25.1	6.4	66.9	10.8	60.1	10.6	61.3	12.1	78.2	8.4
Airplane [35]	20.8	5.8	54.8	13.6	62.0	13.6	64.2	17.4	65.3	13.8	13.8	5.9	42.2	9.8	52.6	15.8	53.3	19.8	53.6	8.7
Airplane [25]	16.8	5.2	44.7	5.0	50.7	8.1	44.4	10.9	53.1	6.8	14.9	4.5	30.3	7.6	28.8	7.2	19.2	10.8	33.4	11.3
Creative birds	14.5	4.4	12.5	4.6	29.5	7.1	29.6	3.9	30.4	1.7	13.6	3.9	12.5	2.5	20.8	5.3	20.6	5.8	20.1	1.0

TABLE VIII: Numerical evaluation on the SPG dataset [22]: first 25 categories; on the ‘airplane’ category from TUBerlin [35] and Huang14 [25] datasets; on creative birds [36]. μ denotes the average accuracy over 5 runs with 5 randomly chosen templates, and σ is the standard deviation of the 5 runs results. The results after refining each point label with a label dominant for each stroke.

Category	1 template (P) (C)	3 templates (P) (C)	5 templates	
			(P)	(C)
Ours	ambulance	86.0 77.4	92.8 90.5	93.2 91.1
	apple	83.4 69.6	89.1 81.3	89.8 80.3
	duck	88.2 78.3	91.3 84.8	92.9 89.0
	face	86.0 77.6	91.9 85.0	93.8 89.9
	pig	85.2 76.8	91.4 85.6	92.3 88.8
SGCN [33]	ambulance	77.5 63.4	90.1 86.3	92.1 87.7
	apple	82.1 65.3	84.0 72.0	86.0 77.0
	duck	74.0 59.0	86.2 80.8	77.2 68.8
	face	78.4 64.6	86.8 78.9	89.9 82.1
	pig	77.4 64.7	85.6 79.2	87.2 80.4

TABLE IX: One shot vs. few shot. See Sec.VII-C for the details.

Category	P-metric				C-metric			
	ISPP PointNet++		ISPP GCN		ISPP PointNet++		ISPP GCN	
	μ	σ	μ	σ	μ	σ	μ	σ
ambulance	58.3	8.2	60.1	10.3	28.5	10.8	33.7	8.7
apple	73.9	9.9	78.2	7.6	50.5	18.3	56.7	14.9
duck	67.2	10.6	71.2	6.0	45.9	14.8	48.4	10.5
face	39.2	8.2	41.8	10.8	11.0	6.2	16.6	11.5
pig	40.7	9.1	45.8	9.7	15.1	7.5	20.6	8.6
Average	55.9	9.2	59.4	8.9	30.2	11.5	35.2	10.8

TABLE X: Segmentation accuracy comparison for the ISPP [16] method, when the originally proposed PointNet++ encoder is used instead of our GCN encoder.

- [2] A. Ghosh, R. Zhang, P. K. Dokania, O. Wang, A. A. Efros, P. H. S. Torr, and E. Shechtman, “Interactive sketch & fill: Multiclass sketch-to-image translation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.
- [3] C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, and C. Zou, “Sketchycoco: Image generation from freehand scene sketches,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
- [4] Q. Zheng, Z. Li, and A. Bargteil, “Learning to shadow hand-drawn

Category	P-metric				C-metric			
	$\mathcal{F}_{\hat{X}}^{sketch}$		$\mathcal{F}_{\hat{E}}^{sketch}$		$\mathcal{F}_{\hat{X}}^{sketch}$		$\mathcal{F}_{\hat{E}}^{sketch}$	
	μ	σ	μ	σ	μ	σ	μ	σ
ambulance	86.7	4.3	87.1	3.7	80.2	7.5	81.4	6.2
apple	94.1	5.6	94.3	5.4	86.6	12.5	87.1	11.9
duck	89.1	4.5	89.6	4.1	83.6	7.6	84.0	7.5
face	82.2	6.4	83.3	6.9	70.1	11.5	72.2	11.7
pig	76.4	16.4	76.6	14.6	64.4	20.0	64.5	18.1

TABLE XI: Numerical evaluation of alternative strategies in the segmentation module.

Category	P-metric				C-metric			
	No \mathcal{L}_{CD}		Ours full		No \mathcal{L}_{CD}		Ours full	
	μ	σ	μ	σ	μ	σ	μ	σ
ambulance	88.4	2.4	87.1	3.7	82.5	4.3	81.4	6.2
apple	90.8	7.0	94.3	5.4	79.9	14.2	87.1	11.9
duck	83.1	11.2	89.6	4.1	74.7	15.5	84.0	7.5
face	81.5	6.9	83.3	6.9	69.3	8.2	72.2	11.7
pig	75.9	15.8	76.6	14.6	62.4	20.0	64.5	18.1
Average	83.9	8.7	86.2	6.9	73.8	12.4	77.8	11.1

TABLE XII: The role of Chamfer distance for stroke-level deformation estimation.

- sketches,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
- [5] H. Lin, Y. Fu, X. Xue, and Y.-G. Jiang, “Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
- [6] R. Liu, Q. Yu, and S. Yu, “Unsupervised sketch to photo synthesis,” in *Proc. Eur. Conf. Comput. Vis.*, 2020.
- [7] F. Liu, C. Zou, X. Deng, R. Zuo, Y.-K. Lai, C. Ma, Y.-J. Liu, and H. Wang, “Scenesketcher: Fine-grained image retrieval with scene sketches,” in *Proc. Eur. Conf. Comput. Vis.*, 2020.
- [8] R. K. Jones, T. Barton, X. Xu, K. Wang, E. Jiang, P. Guerrero, N. J. Mitra, and D. Ritchie, “Shapeassembly: Learning to generate programs for 3d shape structure synthesis,” *ACM Trans. Graph.*, vol. 39, no. 6,

Category	P-metric										C-metric									
	SGCN [33]		FLSS [14]		ISPP [16]		Ours (Joint)		Ours (Separate)		SGCN [33]		FLSS [14]		ISPP [16]		Ours (Joint)		Ours (Separate)	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
airplane	66.6	14.0	56.3	11.2	50.8	14.2	86.0	4.9	85.3	7.7	57.3	17.8	34.6	15.5	23.0	12.7	80.6	8.5	79.6	11.3
alarm clock	79.7	9.9	59.7	10.4	59.4	11.5	86.4	9.1	85.4	14.4	68.4	15.7	36.5	17.4	32.9	17.8	76.0	15.7	77.1	22.3
ambulance	78.1	3.4	61.5	12.0	60.1	10.3	87.1	3.7	89.1	4.6	66.9	7.2	46.2	10.4	33.7	8.7	81.4	6.2	85.6	7.3
angel	54.2	12.0	47.6	9.4	57.8	6.0	70.7	11.5	69.8	13.0	45.7	13.2	22.1	10.3	31.4	7.2	64.9	7.8	62.5	6.8
ant	44.2	17.5	41.7	14.5	47.3	12.6	60.8	18.0	70.1	10.1	35.2	9.5	22.5	12.6	27.2	15.0	50.6	18.6	55.5	11.5
apple	83.4	10.7	82.0	8.6	78.2	7.6	94.3	5.4	94.2	5.3	78.3	16.9	59.5	13.1	56.7	14.9	87.1	11.9	86.8	11.7
backpack	59.2	3.9	35.9	3.8	33.7	6.0	64.6	9.2	55.8	13.8	46.6	2.1	6.4	3.8	8.0	4.5	50.2	11.0	41	13.8
basket	68.7	15.7	65.9	14.2	55.2	15.1	79.1	10.3	72.4	14.1	61.1	12.7	41.6	20.4	28.8	18.2	70.1	11.3	61.5	15.5
bulldozer	53.4	15.7	56.0	9.2	67.9	5.1	69.1	11.0	84.4	5.0	43.1	18.4	38.8	11.2	49.1	8.9	58.5	13.4	77.4	6.4
butterfly	78.2	9.3	70.2	7.5	65.0	8.1	91.7	3.7	86.9	7.3	67.4	13.2	54.1	9.1	38.9	14.1	86.2	5.8	79.3	10.5
cactus	84.6	4.6	41.9	9.1	47.7	11.7	89.2	6.8	91.0	5.2	80.4	8.2	18.9	12.2	14.1	5.6	83.3	9.4	85.7	9.5
calculator	89.2	4.6	67.3	4.1	52.7	12.1	92.6	2.8	92.3	2.9	87.7	4.8	44.2	6.2	24.5	7.7	90.1	5.0	88	6.5
campfire	91.2	3.2	80.7	4.2	73.5	5.0	93.9	1.6	95.0	1.4	88.4	4.6	71.4	11.4	57.1	7.2	89.0	4.2	92.8	2.8
candle	89.8	5.7	86.7	4.5	85.2	1.7	96.3	1.7	96.7	1.5	81.0	10.0	71.9	10.3	67.8	7.4	93.9	2.2	94.8	2.1
coffee cup	73.6	10.6	73.7	5.9	66.2	7.1	82.6	6.9	82.3	18.9	76.2	13.6	54.6	11.8	38.3	15.5	81.3	5.7	78.8	22.9
crab	56.2	13.8	49.5	10.3	48.6	13.2	75.4	12.9	72.1	13.4	51.9	11.8	27.0	8.4	21.6	10.5	69.9	14.8	66.4	14.9
drill	71.3	8.2	80.6	1.9	84.1	1.5	88.7	8.0	96.7	1.0	56.4	5.3	55.1	1.9	68.1	8.6	79.8	10.3	95.6	1.3
duck	61.2	10.5	53.6	4.5	71.2	6.0	89.6	4.1	91.4	3.0	53.9	9.3	26.5	8.4	48.4	10.5	84.0	7.5	86.6	5.8
face	69.9	14.4	38.3	6.8	41.8	10.8	83.3	6.9	83.0	5.0	55.2	16.0	12.4	8.2	16.6	11.5	72.2	11.7	72.7	6.6
flower	75.6	14.2	62.6	3.5	58.1	3.4	88.3	2.0	89.0	4.2	72.7	11.2	36.4	6.5	27.3	3.7	87.1	3.2	88.4	3.4
house	82.2	9.3	57.8	10.7	58.4	9.3	89.4	2.4	75.2	17.4	81.7	7.9	34.7	13.4	32.8	13.3	85.3	2.4	68	17.5
ice cream	82.5	5.7	75.2	4.3	72.9	1.0	86.5	8.3	84.5	10.1	79.2	5.9	62.2	5.6	60.0	2.8	80.6	8.3	79.1	10.2
pig	66.8	20.8	37.1	12.2	45.8	9.7	76.6	14.6	78.9	10.2	55.6	20.9	14.8	10.2	20.6	8.6	64.5	18.1	64.3	16.8
pineapple	76.9	13.5	66.6	5.1	56.5	8.4	80.8	8.3	85.4	5.8	74.4	12.3	43.5	11.4	35.0	7.3	75.8	6.8	78.3	6.3
suitcase	89.2	1.6	82.4	5.7	81.7	2.6	93.8	1.6	93.7	1.0	90.7	1.5	72.9	7.2	61.0	5.3	93.2	1.7	93.1	1.1
Average	73.0	10.1	61.2	7.7	60.8	8.0	83.9	7.0	84.0	7.9	66.2	10.8	40.4	10.3	36.9	9.9	77.4	8.9	77.6	9.8
Min	44.2		35.9		33.7		60.8		55.8		35.2		6.4		8.0		50.2		41	
Max	91.2		86.7		85.2		96.3		96.7		90.7		72.9		68.1		93.9		95.6	
Std.	12.9		15.5		13.6		9.6		10.1		15.6		19.0		17.2		12.2		13.5	

TABLE XIII: The comparison of training strategies for our proposed method. Ours (Joint) refers to the joint training strategy used in the main document. Ours (Separate) refers to a two a two-steps training strategy, where we first train the keypoints prediction network, as described in Section VIII-D. We also compute the minimum average accuracy across categories (Min), the maximum average accuracy across categories (Max), and the standard deviations across categories (Std.). These numbers allow to evaluate how consistent are the segmentation results of each method across different categories.

- 2020.
- [9] K. Li, K. Pang, Y.-Z. Song, T. Hospedales, H. Zhang, and Y. Hu, "Fine-grained sketch-based image retrieval: The role of part-aware attributes," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016.
- [10] C. Zou, Q. Yu, R. Du, H. Mo, Y.-Z. Song, T. Xiang, C. Gao, B. Chen, and H. Zhang, "Sketchscene: Richly-annotated scene sketches," in *Proc. Eur. Conf. Comput. Vis.*, September 2018.
- [11] G. Su, Y. Qi, K. Pang, J. Yang, Y.-Z. Song, and C. SketchX, "Sketch-healer: A graph-to-sequence network for recreating partial human sketches," in *Proc. Brit. Mach. Vis. Conf.*, 2020.
- [12] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001.
- [13] F. Perteneder, M. Bresler, E.-M. Grossauer, J. Leong, and M. Haller, "cluster: Smart clustering of free-hand sketches on large interactive surfaces," in *Proc. ACM Symp. User Interface Softw. Technol.*, 2015.
- [14] L. Wang, X. Li, and Y. Fang, "Few-shot learning of part-specific probability space for 3d shape segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
- [15] S. Yuan and Y. Fang, "Ross: Robust learning of one-shot 3d shape segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020.
- [16] N. Chen, L. Liu, Z. Cui, R. Chen, D. Ceylan, C. Tu, and W. Wang, "Unsupervised learning of intrinsic structural representation points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
- [17] L. Li, H. Fu, and C.-L. Tai, "Fast sketch segmentation and labeling with deep learning," *IEEE Comput. Graph. Appl.*, vol. 39, no. 2, 2018.
- [18] R. G. Schneider and T. Tuytelaars, "Example-based sketch segmentation and labeling using crfs," *ACM Trans. Graph.*, vol. 35, no. 5, 2016.
- [19] X. Zhu, Y. Xiao, and Y. Zheng, "2d freehand sketch labeling using cnn and crf," *Multimed. Tools. Appl.*, vol. 79, no. 1, 2020.
- [20] Y. Qi, J. Guo, Y. Li, H. Zhang, T. Xiang, and Y.-Z. Song, "Sketching by perceptual grouping," in *Proc. IEEE Int. Conf. on Image Process.*, 2013.
- [21] Y. Qi, Y.-Z. Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, and J. Guo, "Making better use of edges via perceptual grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.
- [22] K. Li, K. Pang, J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, and H. Zhang, "Universal sketch perceptual grouping," in *Proc. Eur. Conf. Comput. Vis.*, 2018.
- [23] C. Liu, E. Rosales, and A. Sheffer, "Strokeagggregator: Consolidating raw sketches into artist-intended curve drawings," *ACM Trans. Graph.*, vol. 37, no. 4, 2018.
- [24] Z. Sun, C. Wang, L. Zhang, and L. Zhang, "Free hand-drawn sketch segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2012.
- [25] Z. Huang, H. Fu, and R. W. Lau, "Data-driven segmentation and labeling of freehand sketches," *ACM Trans. Graph.*, vol. 33, no. 6, 2014.
- [26] X. Zhu, Y. Xiao, and Y. Zheng, "Part-level sketch segmentation and labeling using dual-cnn," in *Proc. IEEE Int. Conf. on Image Process.*, 2018.
- [27] X. Wu, Y. Qi, J. Liu, and J. Yang, "Sketchsegnet: A rnn model for labeling sketch strokes," in *IEEE Int. Workshop Mach. Learn. Signal Process.*, 2018.
- [28] X. Wang, X. Chen, and Z. Zha, "Sketchpointnet: A compact network for robust sketch recognition," in *Proc. IEEE Int. Conf. on Image Process.*, 2018.
- [29] K. Li, K. Pang, Y.-Z. Song, T. Xiang, T. M. Hospedales, and H. Zhang, "Toward deep universal sketch perceptual grouper," *IEEE Trans. Image Process.*, vol. 28, no. 7, 2019.
- [30] Y. Qi and Z.-H. Tan, "Sketchsegnet+: An end-to-end learning of rnn for multi-class sketch semantic segmentation," *IEEE Access*, vol. 7, 2019.
- [31] F. Wang, S. Lin, H. Wu, H. Li, R. Wang, X. Luo, and X. He, "Spfusionnet: Sketch segmentation using multi-modal data fusion," in *Proc. IEEE Int. Conf. on Multimedia and Exposition*, 2019.
- [32] F. Hähnlein, Y. Gryaditskaya, and A. Bousseau, "Bitmap or vector? a study on sketch representations for deep stroke segmentation," in *Journées Françaises d'Informatique Graphique et de Réalité virtuelle*, 2019.
- [33] L. Yang, J. Zhuang, H. Fu, X. Wei, K. Zhou, and Y. Zheng, "Sketchgnn: Semantic sketch segmentation with graph neural networks," *ACM Trans. Graph.*, vol. 40, no. 3, pp. 1–13, 2021.
- [34] X. Zhu, Y. Xiao, Y. Zheng, G. Tan, and S. Zhou, "Sketchppnet: A joint pixel and point convolutional neural network for low resolution sketch

- image recognition,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020.
- [35] M. Eitz, J. Hays, and M. Alexa, “How do humans sketch objects?” *ACM Trans. Graph.*, vol. 31, no. 4, 2012.
- [36] S. Ge, V. Goswami, C. L. Zitnick, and D. Parikh, “Creative sketch generation,” in *Int. Conf. Learn. Represent.*, 2021.
- [37] F. Wang, S. Lin, H. Li, H. Wu, T. Cai, X. Luo, and R. Wang, “Multi-column point-cnn for sketch segmentation,” *Neurocomputing*, vol. 392, 2020.
- [38] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.
- [39] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5105–5114.
- [40] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, “Pointcnn: Convolution on χ -transformed points,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 828–838.
- [41] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic graph cnn for learning on point clouds,” *ACM Trans. Graph.*, vol. 38, no. 5, 2019.
- [42] L. Yang, A. Sain, L. Li, Y. Qi, H. Zhang, and Y.-Z. Song, “S 3 net: Graph representational network for sketch recognition,” in *Proc. IEEE Int. Conf. on Multimedia and Exposition*, 2020, pp. 1–6.
- [43] D. Ha and D. Eck, “A neural representation of sketch drawings,” in *Int. Conf. Learn. Represent.*, 2018.
- [44] L. Li, C. Zou, Y. Zheng, Q. Su, H. Fu, and C.-L. Tai, “Sketch-R2CNN: An rnn-rasterization-cnn architecture for vector sketch recognition,” *IEEE Trans. Vis. Comput. Graph.*, 2020.
- [45] Q. Jia, X. Fan, M. Yu, Y. Liu, D. Wang, and L. J. Latecki, “Coupling deep textural and shape features for sketch recognition,” in *Proc. of ACM International Conference on Multimedia*, 2020.
- [46] G. Li, M. Muller, A. Thabet, and B. Ghanem, “Deepgens: Can gcns go as deep as cnns?” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.
- [47] Z. Chen, K. Yin, M. Fisher, S. Chaudhuri, and H. Zhang, “Bae-net: Branched autoencoder for shape co-segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.
- [48] O. Sorkine, “Least-squares rigid motion using svd,” *Technical notes*, vol. 120, no. 3, 2009.