Geometry-Aware End-to-End Skeleton Detection

Weijian Xu
wex041@eng.ucsd.edu
Gaurav Parmar
gparmar@ucsd.edu
Zhuowen Tu
ztu@ucsd.edu

University of California San Diego
9500 Gilman Drive,
La Jolla, CA, USA

Abstract

In this paper, we propose a new skeleton detection method that is geometry-aware and can be learned in an end-to-end fashion. Recent approaches in this area are based primarily on the holistically-nested edge detector (HED) that is learned in a fundamentally bottom-up fashion by minimizing a pixel-wise cross-entropy loss. Here, we introduce a new objective function inspired by the Hausdorff distance that carries both global and local shape information and is made differentiable through an end-to-end neural network framework. When compared with the existing approaches on several widely adopted skeleton benchmarks, our method achieves state-of-the-art results under the standard F-measure. This sheds some light towards directly incorporating shape and geometric constraints in an end-to-end fashion for image segmentation and detection problems — a viewpoint that has been mostly neglected in the past.

1 Introduction

An object skeleton is a compact visual representation that captures the centerline and the symmetric axis of an object [1]. A wide range of computer vision applications have adopted the skeleton representation in their systems, including pose estimation [28, 33, 36], object segmentation [11, 20], scene text detection [3], and character recognition [34]. It is generally observed that an object skeleton exhibits both global (*e.g*. medial axis [29, 30, 41]) and local (*e.g*. continuity, local symmetry, and junctions [7, 18]) geometric properties.

Object skeleton extraction is a long standing problem in computer vision. Early approaches [10, 16, 17, 38] are based on classic image processing techniques using morphological operators. Learning-based approaches, particularly the ones based on deep convolutional neural networks (CNNs), have become increasingly popular. A number of recent methods [14, 19, 26, 27, 35, 40] build their skeleton extraction systems on top of the holistically-nested edge detector (HED) [37], making use of a balanced cross-entropy loss within a deeply-supervised [15] fully convolutional neural networks (FCN) [22] framework. These skeleton extraction algorithms demonstrate various properties focusing on different aspects, *e.g*. scale separation and aggregation [26, 27], structural architecture design [14, 19, 40], and rich intermediate representation [35].

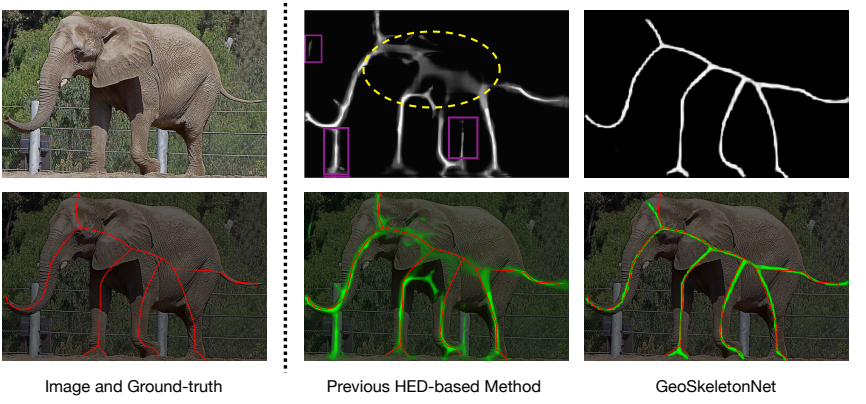

Figure 1: Qualitative comparison between a HED-based skeleton detection method [40] and our proposed method. Left: an input image and its ground-truth skeleton. Middle: predicted skeleton map (the first row) and a superimposed version with the ground-truth in the input image (the second row); the yellow dashed circle indicates an unsatisfactory area with a blurry effect on elephant's back; purple rectangles show some false positives. Right: predicted skeleton map (the first row) by our method and a superimposed version with the ground-truth in the input image (the second row).

HED was originally designed to perform end-to-end edge detection. It adopts an image-to-image prediction framework by learning a family of nested edge features beyond image gradients. The weighted cross-entropy loss in HED is customized for edge detection, creating a performance gap when applied to skeleton detection. The pixel-wise cross-entropy loss is most effective for the semantic labeling [5] and edge detection task [37], making dense pixel-wise prediction based primarily on local image contents. A skeleton, however, observes strong geometric properties with structural information capturing the long-range contextual shape information (*e.g.* symmetry). Figure 1 shows a typical example where the result by a standard HED-based skeleton detector [40] outputs a blurry skeleton for the main body of the elephant. In addition, a pixel-wise loss makes an independence assumption for each pixel, rendering violations to the global and local geometric constraints that are commonly observed for object skeletons. HED-based models learn rich hierarchical edge features, but the learning process is not made geometry-aware explicitly. Existing methods in this domain instead rely on a separate post-processing step [40] which often has a limited scope of success and cannot correct large mistakes. Figure 1 again shows an failure case where significant false positive and false negative are present, which is difficult to be fixed by a standard post-processing algorithm.

Motivated by the above observations, we develop a new convolutional neural network based skeleton detector (named as *GeoSkeletonNet*) by introducing a geometry-aware objective. Specifically, the training objective (still learned end-to-end) consists of a weighted Hausdorff distance and a geometrically weighted cross-entropy loss, providing the global and local geometric constraints. In addition, an extra patch-based point loss is added to the overall objective in order to employ the local geometric constraints. Our proposed algorithm mitigates the limitation in the existing skeleton detection methods [19, 26, 27, 40] that do not take into account explicit geometric constraints in training. Figure 1 shows the advantage of our geometry-aware framework when compared with a standard learning-based object skeleton extraction system.

The main properties of our GeoSkeletonNet are summarized as follows: (1) we propose

to incorporate a geometry-aware objective property within an end-to-end the skeleton detection framework, (2) we adopt a weighted Hausdorff distance and design a geometrically weighted cross-entropy loss, while utilizing a patch-based point loss for local constraints, (3) GeoSkeletonNet demonstrates state-of-the-art results on five skeleton detection benchmarks, outperforming recent competing methods in this domain [14, 19, 26, 27, 35, 40].

2 Related Work

The task of skeleton detection has been long studied [1], both in computer vision [17] and medical imaging [29]. We refer to a recent paper [35] for a relative comprehensive discussion about the literature in this subject. Here, we primarily discuss some recent deep learning based skeleton detection algorithms.

Existing Skeleton Detection Methods. Most recent skeleton detection algorithms [14, 19, 26, 35, 40] are built on top of the holistically-nested edge detector (HED) [37]. Shen *et al*. [26] propose to enforce the side output of deep supervision with a specific receptive field to match the skeleton at the corresponding scale. Ke *et al*. [14] attempt to apply bi-directional residual learning to side outputs, aiming at a larger receptive field. Liu *et al*. [19] generalize the residual unit in [14] by employing the linear span unit, which improves the expressiveness of side outputs. Zhao *et al*. [40] design a hierarchical fusion architecture in the deep supervision to further enrich the representation of side outputs. However, all above approaches merely modify the network structure, especially the deeply supervised part, yet still suffer from the side effects of the pixel-wise loss. Recently, Wang *et al*. [35] propose to change the skeleton prediction from the probability-based map to the flux-based vector field. The flux representation encodes the local geometric relationship between image pixels and skeletal points, but this representation is difficult to learn accurately, which leads to many discontinuities in the predicted skeleton map after post-processing. In contrast, our approach keeps the probability-based skeleton map, while injecting the local and global geometric relation between the prediction and the ground-truth into the objective function, which boosts the overall performance and maintain the visual continuity.

Geometry-aware Distances in Vision. In computer vision, geometry-aware distances has been widely adopted, especially in object matching [6], face recognition [12] and image retrieval [9]. In the deep learning era, geometry-aware distances have been employed in tasks such as 3D object reconstruction [8] and object localization [24]: Fan *et al*. attempts to build a shape reconstruction framework based on point cloud representation, which minimizes the Chamfer distance and the Earth-mover distance between the prediction and the ground-truth. Ribera *et al*. [24] proposes to relax the Hausdorff distance and optimize it on the location probability map in the object localization task. Inspired by [24], we adopt the weighted Hausdorff distance in the objective function for skeleton detection. Besides, we augment the objective with a geometrically weighted cross-entropy loss and a patch-based point loss, which provides additional global and local geometric constraints.

3 Method

3.1 Problem Formulation

Consider a training dataset $\{(\mathrm{X}_k, \Gamma_k), k = 1, 2, ..., K\}$, where X_k and Γ_k respectively refer to the k-th input image and its corresponding ground-truth skeleton. Note that Γ_k is an

explicit vectorized representation encoding the object centerline. K is the total number of training images. In the literature, Γ was represented in various forms, *e.g.* the medial axis [29] or the shock graph [30] representation. The simplest form of Γ_k can be in a parametric representation $\Gamma_k = (i(s), j(s) : s \in [0,1])$ where $(i(s), j(s))$ indicates each 2D skeleton point $(i(s), j(s))$ parameterized by s. For notational simplicity, we drop k by considering only one image X in the training set, $\{(\mathrm{X}, \Gamma)\}$.

To facilitate our training process, skeleton Γ is converted into a label map: $\mathrm{Y} = (y_{(i,j)}; (i,j) \in \Lambda)$, where $y_{(i,j)} = 1$ if pixel (i,j) is on the skeleton, and $y_{(i,j)} = 0$ otherwise. Thus, our training set is simplified to $\{(\mathrm{X}, \mathrm{Y})\}$, where X refers to the input image and Y denotes the corresponding ground-truth label map. For X, its image lattice that includes all the pixels is defined as $\Lambda = \{(i,j), i = 1..Height, j = 1..Width\}$ where $Height$ and $Width$ refer to the height and width of image X respectively.

3.2 Geometry-aware Objective Function

3.2.1 Revisit the Weighted Cross-entropy Loss

Given a training image X together with its ground-truth label map Y, our goal is to learn, in an end-to-end fashion, a neural network model to predict $\hat{\mathrm{Y}} = (\hat{y}_{(i,j)}; (i,j) \in \Lambda)$ where $\hat{y}_{(i,j)} \in \{0,1\}$ that is as faithful as possible to the ground-truth $\mathrm{Y} = (y_{(i,j)}; (i,j) \in \Lambda)$. We further define a positive set $Y_+ = \{(i,j); y_{(i,j)} = 1\}$ and a negative set $Y_- = \{(i,j); y_{(i,j)} = 0\}$, to have separate notations for the skeleton and background pixels.

To make the learning process differentiable, the hard prediction map $\hat{\mathrm{Y}}$ is relaxed by a soft probability map $\mathrm{P} = (p_{(i,j)}; (i,j) \in \Lambda)$. Typically, a neural network model can be learned through a pixel-wise cross-entropy loss between the predicted probability map P and the ground-truth Y. Specifically, in [14, 19, 26, 27, 40], a weighted cross-entropy (WC) proposed by [37] is used to tackle the problem of an imbalanced dataset:

$$\mathcal{L}_{\mathrm{WC}} = -\beta \sum_{a \in Y_+} \log p_a - (1-\beta) \sum_{a \in Y_-} \log(1-p_a), \tag{1}$$

where $\beta = |Y_-| / |\Lambda|$ and $1-\beta = |Y_+| / |\Lambda|$. However, the pixel-wise loss in Equation 1 basically evaluates all pixels independently and is absent of explicit geometric constraints, which are important prior for tasks like skeleton extraction. A result obtained by such a loss is illustrated in Figure 1, showing problems in localization, precision, and structural consistency.

3.2.2 Weighted Hausdorff Distance

To combat the problem described above, we introduce geometry-aware loss in training to take into account both global and local geometry of the learned skeletons. Following our previous notations, let Y_+ and $\hat{Y}_+$ be the set of skeleton pixels for the ground-truth and prediction respectively. We adopt a Hausdorff distance (HD) to capture the geometric relation between these two sets. For two point sets $\hat{Y}_+$ and Y_+, the Hausdorff distance is computed as:

$$D_{\mathrm{H}} = \max\Big(\max_{b \in \hat{Y}_+} \min_{a \in Y_+} d(b,a), \max_{a \in Y_+} \min_{b \in \hat{Y}_+} d(a,b)\Big). \tag{2}$$

where $d(x,y)$ is the Euclidean distance between point a and b. To increase the robustness of the Hausdorff distance measure against the outliers due to the max operation, a variant of the

Hausdorff distance, the average Hausdorff distance (AHD) is adopted:

$$D_{\mathrm{AH}} = \frac{1}{\left|\hat{Y}_+\right|} \sum_{b \in \hat{Y}_+} \min_{a \in Y_+} d(b,a) + \frac{1}{|Y_+|} \sum_{a \in Y_+} \min_{b \in \hat{Y}_+} d(a,b). \tag{3}$$

Adding geometric constraints such as Equation 2 and 3 to a problem that exhibits a strong shape prior seems to be natural step to take. However, making geometry-aware loss in an end-to-end learning framework has been under-explored, due primarily to the difficulty in making the loss differentiable through back-propagation. In this paper, we combat this issue by adopting a weighted Hausdorff distance (WHD) that was originally proposed in [24] for object detection/localization:

$$D_{\mathrm{WH}} = \frac{1}{|\tilde{Y}_+| + \varepsilon} \sum_{b \in \Lambda} p_b \min_{a \in Y_+} d(b,a) + \frac{1}{|Y_+|} \sum_{a \in Y_+} \min_{b \in \Lambda} \frac{d(a,b) + \varepsilon}{(p_b)^{\alpha} + \varepsilon / d_{max}}, \tag{4}$$

where $|\tilde{Y}_+| = \sum_{b \in \Lambda} p_b$ is an estimate of the number of positive skeletal points in prediction. ε is a small positive number (*e.g.* 10^{-6}) to avoid zero numerator and denominator and d_{max} is the length of diagonal of the skeleton map. When p_b takes one of the two extreme values $\in \{0,1\}$, the weighted Hausdorff distance reduces to the average Hausdorff distance [24]. Note the difference between our method and that in [24] where the main focus of [24] is to use a WHD to better localize the object whereas our motivation is to employ WHD as a geometry-aware loss for end-to-end learning of image-to-image prediction.

The weighted Hausdorff distance enjoys the benefit of capturing a shape constraint beyond pixel-wise prediction, encouraging both local and global shape match between the predicted and the ground-truth skeletons. This is a much needed property in the current end-to-end skeleton learning frameworks but remains largely absent in the previous literature.

3.2.3 Patch-based Point Loss

Including the WHD in the objective function reduces the blurry artifacts and makes the predictions better localized. However, directly training on WHD is unstable and disconnected skeleton segments are still present. To address this issue, we add an additional patch-based point loss (PPL) term. PPL aims to minimize the difference between the number of points in P above a specified threshold λ_T and the number of points in Y. To prevent the predicted skeleton from becoming too thick and to enforce local geometric regularities, we apply the proposed point loss in a patch-wise manner.

We divide the image into patches in a grid like manner with the patch size $M \times M$. Each patch coordinate set $\Lambda_{i,j}$ can be represented as a strict subset of Λ such that $\Lambda = \bigcup \Lambda_{i,j}$. $\tilde{p_b}$ represents the probability at the position b if it greater than λ_T, else it is 0. Thus, the patch-based point loss term $\mathcal{L}_{\mathrm{PPL}}$ is formulated as:

$$\mathcal{L}_{\mathrm{PPL}} = \sum_{\Lambda_{i,j}} \left| \sum_{b \in \Lambda_{i,j}} \tilde{p_b} - \left|\Lambda_{i,j} \cap Y_+\right| \right|. \tag{5}$$

3.2.4 Geometrically Weighted Cross-entropy Loss

Based on the weighted Hausdorff distance in Equation 4, we further adapt the weighted cross-entropy in Equation 1 to incorporate geometric awareness. For the predicted probabilities corresponding to negative ground-truth points, we scale each pixel-wise cross-entropy

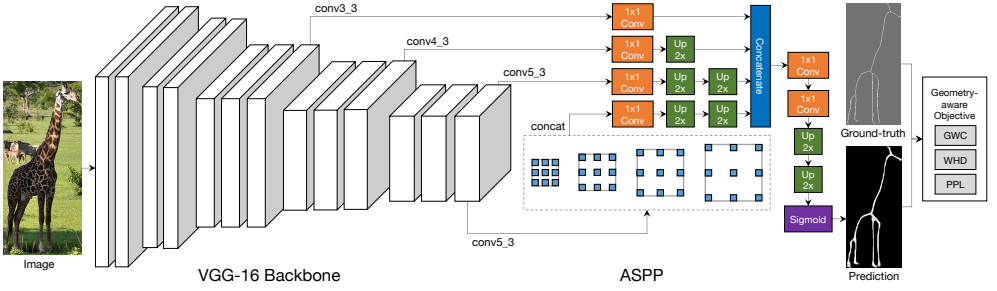

Figure 2: Schematic illustration of the network architecture of our GeoSkeletonNet framework.

term by multiplying with a geometric distance between current point and its nearest positive point in the ground-truth:

$$\mathcal{L}_{\text{GWC}} = -\beta \sum_{a \in Y_+} \log p_a - (1-\beta) \sum_{b \in Y_-} \min_{a \in Y_+} d(a,b)^{\gamma} \log(1-p_b). \tag{6}$$

where γ is a hyper-parameter to adjust the effect of distance. The second term in $\mathcal{L}_{\text{GWC}}$ resembles the first term in D_{WH}, which brings similar benefits such as removing unwanted blurs and background noise. In practice, this geometrically weighted cross-entropy loss works significantly well in enhancing the overall performance.

Combining the weighted Hausdorff distance, patch-based point loss and the geometrically weighted cross-entropy loss, the final objective function $\mathcal{L}$ is represented as:

$$\mathcal{L} = \lambda_1 D_{\text{WH}} + \lambda_2 \mathcal{L}_{\text{PPL}} + \lambda_3 \mathcal{L}_{\text{GWC}}. \tag{7}$$

3.3 Network Architecture

Figure 2 displays the neural network architecture for our model. We follow the network design from [35]: The VGG-16 [31] network is used as the feature backbone for fair comparison with other approaches. On top of the last convolutional layer (conv5_3) of VGG network, the atrous spatial pyramid pooling (ASPP) [4] is applied to enlarge the receptive field. Then, to construct a multi-scale intermediate feature map, we fuse the ASPP output and VGG side outputs (conv3_3, conv4_3, conv5_3) after 1x1 convolutions and bilinear upsampling kernels. We convert the intermediate feature map to a single channel probability map with original image size as prediction.

4 Experiments

4.1 Datasets

We evaluate our method on five major datasets for skeleton detection: SK-LARGE [27], SK-SMALL [26], SYM-PASCAL [14], SYMMAX300 [32] and WH-SYMMAX [25]. Images in SK-LARGE and SK-SMALL are selected from MS COCO dataset [13] with single object, while the ones in SYM-PASCAL may have multiple objects. SYMMAX300 and WH-SYMMAX are adapted from the BSDS dataset [23] and the Weizmann Horse dataset [2] respectively.

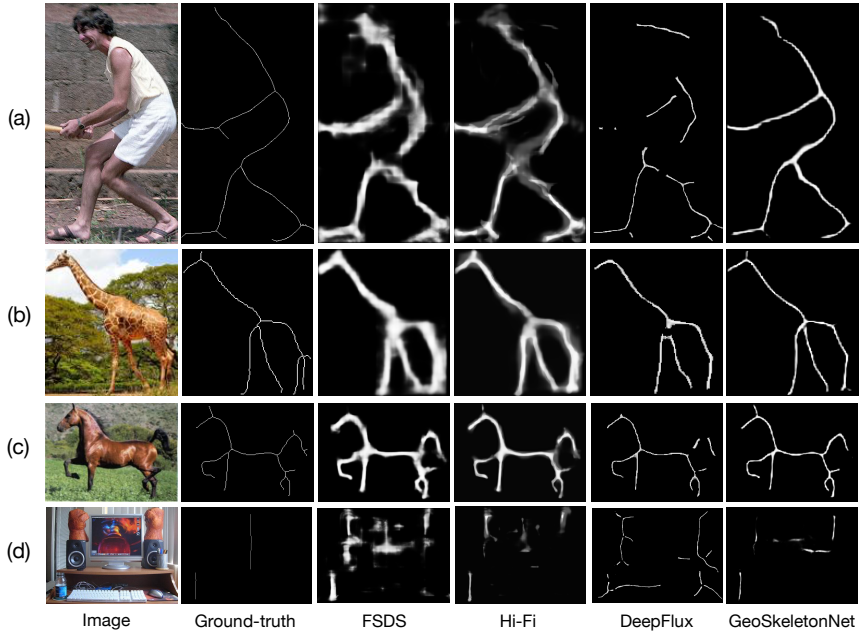

Figure 3: Qualitative comparison with the existing methods. We show four examples from four benchmark datasets including (a), (b), (c) and (d) that are from SK-LARGE, SK-SMALL, WH-SYMMAX and SYM-PASCAL respectively. The skeleton maps are the predictions by the competing method and ours, before the non-maximum suppression operation (NMS).

4.2 Evaluation Protocol

PR Curve and F-measure ODS. After obtaining the predicted probability map P from the network, we apply the standard non-maximum suppression (NMS) to P and threshold it by $\delta \in \{0.01, ..., 0.99\}$ to create the actual skeleton map $\hat{\mathrm{Y}}_\delta$. We evaluate the performance of the model by the Precision-Recall (PR) curve of $\hat{\mathrm{Y}}_\delta$ in the dataset over all thresholds. In addition, the optimal F-measure on the PR curve, named as F-measure at Optimal Dataset Scale (ODS), is used as evaluation metric as well.

4.3 Implementation Details

Training Settings. We build our network and pipeline in PyTorch and run the experiments on NVIDIA TITAN X GPUs. In the experiments, we use the Adam optimizer with the learning rate of 0.0001 and momentum coefficients $\beta_1 = 0.9, \beta_2 = 0.999$. Batch size is set as 1 due to the GPU memory limitation, but we track the *average gradients* every 10 batches and update the weights once, which indicates an equivalent batch size of 10. In the loss terms, we set $\alpha = 4$, $\varepsilon = 10^{-6}$, as recommended by [24], and $\gamma = 0.5$ due to the ablation study. In $\mathcal{L}_{\mathrm{PPL}}$, the patch size $M = 32$ and the threshold $\lambda_T = 0.95$. Besides, during training, we optimize the model on $\mathcal{L}_{\mathrm{GWC}}$ (i.e. $\lambda_1 = \lambda_2 = 0, \lambda_3 = 1$) for 30 epochs and then fine-tune the model on $D_{\mathrm{WH}} + \mathcal{L}_{\mathrm{PPL}}$ (i.e. $\lambda_3 = 0, \lambda_1 = \lambda_2 = 0.01$) for 5 epochs. The 2-stage procedure leads to a more stable training process.

Methods	SK-LARGE	SK-SMALL	WH-SYMMAX	SYM-PASCAL	SYMMAX300
MIL [32]	0.353	0.392	0.365	0.174	0.362
HED [37]	0.497	0.541	0.732	0.369	0.427
RCF [21]	0.626	0.613	0.751	0.392	-
FSDS [26]	0.633	0.623	0.769	0.418	0.467
LMSDS [27]	0.649	0.621	0.779	-	-
SRN [4]	0.678	0.632	0.780	0.443	0.446
LSN [19]	0.668	0.633	0.797	0.425	0.480
Hi-Fi [40]	0.724	0.681	0.805	0.454	-
DeepFlux [35]	0.732	0.695	0.840	0.502	0.491
GeoSkeletonNet	**0.757**	**0.727**	**0.849**	**0.520**	**0.501**

Table 1: Test F-measure ODS comparison on all skeleton datasets. The best numbers are in bold.

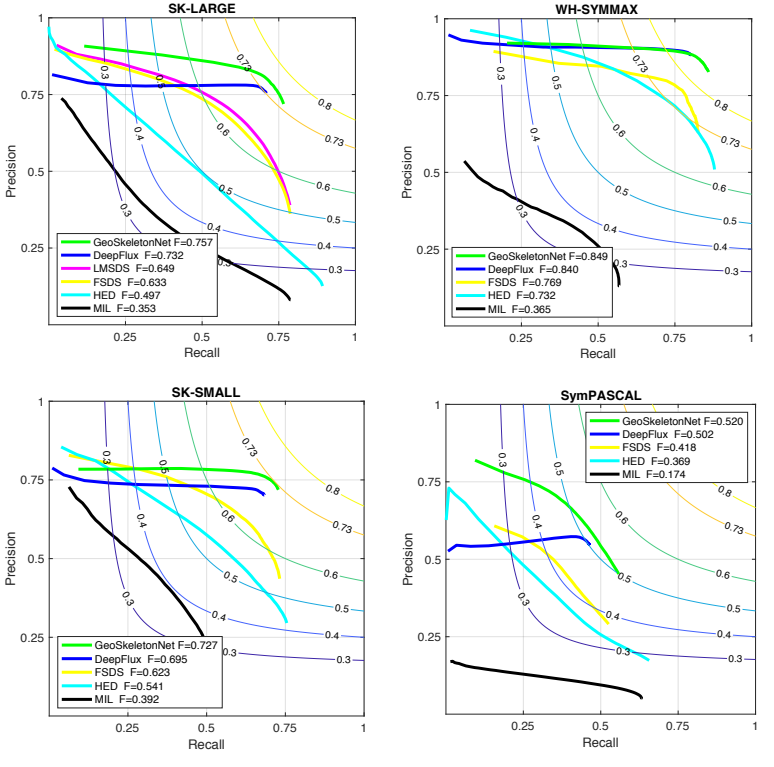

Figure 4: Precision-recall curves on four skeleton datasets.

Resolution Normalization. In most of the skeleton datasets, the images have quite scattered resolutions, which bring a long series of various scales and increase the difficulty in model training. Therefore, in the training stage before data augmentation, we resize the image and ground-truth from size $H \times W$ to $\sqrt{\frac{KH}{W}} \times \sqrt{\frac{KW}{H}}$, which keeps the original aspect ratio and normalize the number of pixels to a fixed value K. We also apply a standard thinning algorithm [39] on the resized ground-truth to avoid unnecessary thickness. In test stage, we still feed the normalized image into network to obtain the prediction, then resize the prediction map back to $H \times W$ for evaluation. We use $K = 180,000$ for the SYM-PASCAL dataset and $K = 60,000$ for the other datasets.

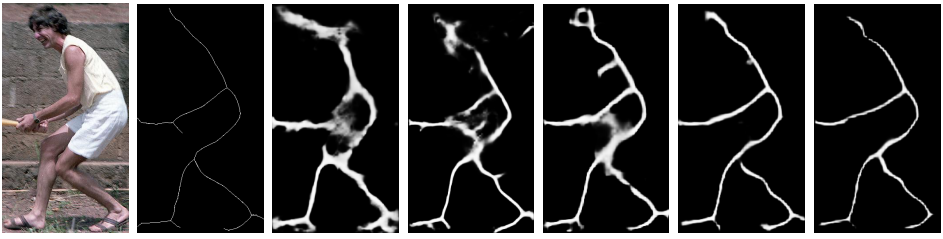

Figure 5: Qualitative comparison on the role of modules.

Data Augmentation. We employ the standard data augmentation following [26] in training stage: The original image is resized to 3 scales (0.8x, 1.0x, 1.2x), rotated with 4 angles (0°, 90°, 180°, 270°) and then flipped to 3 directions (none, left-to-right, up-to-down).

4.4 Comparison with the State-of-the-art

As indicated in Table 1, our method outperforms the current state-of-the-art by a decent margin on all datasets in terms of F-measure ODS. Compared to the recently proposed DeepFlux [35], our GeoSkeletonNet improves by 2.5%, 3.2% and 1.8% on the SK-LARGE, SK-SMALL and SYM-PASCAL datasets respectively, under a similar network design. The PR curve shown in Figure 4 also indicates a clear performance boost.

In Figure 3, we provide a qualitative comparison between our method and the current approaches. In Figure 3 (a), (b) and (c), our method significantly reduces the blurry effect in the previous HED-based methods [26, 40]. Meanwhile, our method avoid the occasional dis-continuities in DeepFlux with acceptable sacrifice of accurate localization (*e.g.* in (a), our predicted skeleton has a thicker junction near the knee while maintaining the whole leg complete). Figure 3 (d) reveals a failure case: Our method is not able to detect the skeleton of the screen, but still tries to reduce false positive predicted points as possible. In contrast, both FSDS and DeepFlux generate a noisy background.

4.5 Ablation Study

Role of Modules. We conduct an ablation analysis to understand the role of modules in performance contribution. Table 2 shows the F-measure ODS on SK-LARGE dataset under multiple module settings of the model. Average gradients (AvgGrad), resolution normalization (ResNorm) and geometrically weighted cross-entropy loss (GWC) greatly contribute to the final performance. Especially, the ResNorm brings the substantial 1.8% improvement, which reflects the difficulty of training on images with various scales.

Baseline	AvgGrad	ResNorm	GWC	WHD+PPL	F-measure ODS
✓					0.712
✓	✓				0.724
✓	✓	✓			0.741
✓	✓	✓	✓		0.753
✓	✓	✓		✓	0.746
✓	✓	✓	✓	✓	0.757

Table 2: Quantitative comparison on the role of modules.

In addition, Figure 5 compares the qualitative results of different module settings. Visually, AvgGrad and ResNorm slightly refine the results from baseline. On the contrary, GWC mitigates the blurry effect by a large margin, which shows the benefits of geometric awareness. Weighted Hausdorff distance (WHD) and patch-based point loss (PPL) further makes the predicted skeleton thinner and provides a better localization.

It is noteworthy that we do not separate the WHD and PPL in the performance analysis on the role of modules. The reason is that directly training on WHD is unstable: WHD is able to provide better localization in the predicted skeleton, but it sometimes generates unexpected disconnection or thickness after certain epochs and leads to a performance drop. Thus, we employ the PPL to stablize the WHD training and always evaluate the performance when both loss terms are turned on. Besides, we also do not include an inference speed comparison since AvgGrad, GWC and WHD+PPL do not affect the time of inference. Only ResNorm will bring a bit overhead, but it is neglectable.

Influence of Distance Hyper-parameter γ. We further analyze the influence of the distance hyper-parameter γ in geometrically weighted cross-entropy loss (GWC). When $\gamma \to 0$, the GWC reduces to the weighted cross-entropy loss (WC).

Table 3 shows that $\gamma = 0.5$ provides the best F-measure ODS in the GWC ablation setting (Baseline + AvgGrad + ResNorm + GWC). Thus, we set $\gamma = 0.5$ in all other experiments.

Distance Hyper-parameter	$\gamma = 0.125$	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 1.0$
F-measure ODS	0.746	0.751	0.753	0.745

Table 3: Influence of distance hyper-parameter γ.

5 Conclusion

In this paper, we have developed an end-to-end skeleton detection method that employs geometric awareness. Specifically, we devise a geometry-aware objective function to compute the global similarity between the predicted skeleton map and the ground truth in an end-to-end learning framework. A weighted Hausdorff distance (WHD) is adopted. In addition, we propose a patch-based point loss (PPL) to mitigate the instability in optimizing WHD and capture local features. Furthermore, we adapt the weighted cross-entropy into a geometric form, which significantly boosts the performance of skeleton detection. Evaluation on five standard skeleton detection benchmarks demonstrates the advantages of our proposed method, consistently outperforming the current state-of-the-art methods. In the future, we would like to explore the possibilities of the geometry-aware objective in wider fields, such as semantic segmentation and object detection.

Acknowledgement This work is supported by NSF IIS-1717431 and NSF IIS-1618477. We thank Intuitive Surgical and Northrop Grumman for the gift funds. The authors thank Kwonjoon Lee, Justin Lazarow, Yifan Xu and Sainan Liu for valuable discussions.

References

[1] Harry Blum. Biological shape and visual science (part i). *Journal of theoretical Biology*, 38(2):205–287, 1973.

[2] Eran Borenstein and Shimon Ullman. Class-specific, top-down segmentation. In *European conference on computer vision*, pages 109–122. Springer, 2002.

[3] Yuanqiang Cai, Weiqiang Wang, Haiqing Ren, and Ke Lu. Spn: short path network for scene text detection. *Neural Computing and Applications*, Feb 2019. ISSN 1433-3058. doi: 10.1007/s00521-019-04093-0. URL https://doi.org/10.1007/s00521-019-04093-0.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[6] M-P Dubuisson and Anil K Jain. A modified hausdorff distance for object matching. In *Proceedings of 12th international conference on pattern recognition*, volume 1, pages 566–568. IEEE, 1994.

[7] James H Elder and Richard M Goldberg. Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4):5, 2002.

[8] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[9] Yue Gao, Meng Wang, Rongrong Ji, Xindong Wu, and Qionghai Dai. 3-d object retrieval with hausdorff distance learning. *IEEE Transactions on industrial electronics*, 61(4):2088–2098, 2014.

[10] Jeong-Hun Jang and Ki-Sang Hong. A pseudo-distance map for the segmentation-free skeletonization of gray-scale images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 18–23. IEEE, 2001.

[11] Koteswar Rao Jerripothula, Jianfei Cai, Jiangbo Lu, and Junsong Yuan. Object co-skeletonization with co-segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3881–3889. IEEE, 2017.

[12] Oliver Jesorsky, Klaus J Kirchberg, and Robert W Frischholz. Robust face detection using the hausdorff distance. In *International conference on audio-and video-based biometric person authentication*, pages 90–95. Springer, 2001.

[13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[14] Wei Ke, Jie Chen, Jianbin Jiao, Guoying Zhao, and Qixiang Ye. Srn: side-output residual network for object symmetry detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1068–1076, 2017.

[15] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015.

[16] Alex Levinshtein, Cristian Sminchisescu, and Sven Dickinson. Multiscale symmetric part detection and grouping. *International journal of computer vision*, 104(2):117–134, 2013.

[17] Tony Lindeberg. Edge detection and ridge detection with automatic scale selection. *International journal of computer vision*, 30(2):117–156, 1998.

[18] Tony Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, 1998.

[19] Chang Liu, Wei Ke, Fei Qin, and Qixiang Ye. Linear span network for object skeleton detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 133–148, 2018.

[20] Tyng-Luh Liu, Davi Geiger, and Alan L Yuille. Segmenting by seeking the symmetry axis. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, volume 2, pages 994–998. IEEE, 1998.

[21] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai. Richer convolutional features for edge detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3000–3009, 2017.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[23] David Martin, Charless Fowlkes, Doron Tal, Jitendra Malik, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. 2001.

[24] Javier Ribera, David Güera, Yuhao Chen, and Edward Delp. Weighted hausdorff distance: A loss function for object localization. *arXiv preprint arXiv:1806.07564*, 2018.

[25] Wei Shen, Xiang Bai, Zihao Hu, and Zhijiang Zhang. Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images. *Pattern Recognition*, 52:306–316, 2016.

[26] Wei Shen, Kai Zhao, Yuan Jiang, Yan Wang, Zhijiang Zhang, and Xiang Bai. Object skeleton extraction in natural images by fusing scale-associated deep side outputs. In *CVPR*, 2016.

[27] Wei Shen, Kai Zhao, Yuan Jiang, Yan Wang, Xiang Bai, and Alan Yuille. Deepskeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images. *IEEE Transactions on Image Processing*, 26(11):5298–5311, 2017.

[28] Jamie Shotton, Andrew W Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images.

[29] Kaleem Siddiqi and Stephen Pizer. *Medial representations: mathematics, algorithms and applications*, volume 37. Springer Science & Business Media, 2008.

[30] Kaleem Siddiqi, Ali Shokoufandeh, Sven J Dickinson, and Steven W Zucker. Shock graphs and shape matching. *International Journal of Computer Vision*, 35(1):13–32, 1999.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] Stavros Tsogkas and Iasonas Kokkinos. Learning-based symmetry detection in natural images. In *European Conference on Computer Vision*, pages 41–54. Springer, 2012.

[33] Qingfu Wan, Wei Zhang, and Xiangyang Xue. Deepskeleton: Skeleton map for 3d human pose regression. *arXiv preprint arXiv:1711.10796*, 2017.

[34] Tie-Qiang Wang and Cheng-Lin Liu. Fully convolutional network based skeletonization for handwritten chinese characters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[35] Yukang Wang, Yongchao Xu, Stavros Tsogkas, Xiang Bai, Sven Dickinson, and Kaleem Siddiqi. Deepflux for skeletons in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[36] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[37] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1403, 2015.

[38] Zeyun Yu and Chandrajit Bajaj. A segmentation-free approach for skeletonization of gray-scale images via anisotropic vector diffusion. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[39] TY Zhang and Ching Y Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984.

[40] Kai Zhao, Wei Shen, Shanghua Gao, Dandan Li, and Ming-Ming Cheng. Hi-fi: hierarchical feature integration for skeleton detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1191–1197. AAAI Press, 2018.

[41] Song Chun Zhu and Alan L Yuille. Forms: a flexible object recognition and modelling system. *International journal of computer vision*, 20(3):187–212, 1996.