Keypoints as Dynamic Centroids: Human Semantic Segmentation using Centroid Representations

Niaz Ahmad, Jawad Khan, Youngmoon Lee
Department of Robotics
Hanyang University
{niazahamd89, jkhanbk1, youngmoonlee}@hanyang.ac.kr

***Abstract*—Existing proposals cannot leverage keypoint heatmaps with segmentation masks, calling for a joint representation of keypoints and semantic segmentation in end-to-end training. This paper proposes Keypoints as Dynamic Centroids (KDC), a new centroid-based representation for human semantic segmentation. KDC follows a bottom-up paradigm to generate Keypoint Feature Maps for both soft and hard keypoints. Then the high-resolution representation of keypoints is used as dynamic centroids in the embedding space to generate MaskCentroid to cluster the pixels to a particular human instance. The representation of human semantic segmentation is enabled by proposing a new SemanticSeg module that collects the feature of human keypoints and segmentation. Experimental results using MSCOCO and OCHuman benchmarks demonstrate the effectiveness and generalization ability of the proposed method on challenging scenarios such as occlusions, entangled limbs, and overlapped people in terms of both accuracy and runtime performance. The code has been made available at: https://github.com/RaiseLab/KDC.**

I. Introduction

Human-body semantic segmentation is widely used for human-computer interactions, robot-human interaction, robot scene understanding, pedestrians crossing for self-drive cars, real-time image/video analytics, and many more. The main goal of human body segmentation is to identify individuals and their activities from the 2D positioning of human joints and their body shape structure. There are two primary challenges for human representing semantically: (i) an unknown number of individuals are overlapped, occluded, or have entangled limbs, and (ii) computational complexity rapidly increases with the number of individuals. An image can have an undefined number of individuals at any location and distance. Human-to-human interactions, especially for socially engaged ones, incur complex spatial interference because of contacts, obstructions, and articulation of their limbs, making it difficult to associate body parts. In turn, computational cost increases prohibitively with the number of people in the image, calling for an efficient, scalable, and accurate human semantic segmentation model.

This paper presents KDC, a new centroid-based representation for the human semantic segmentation model. KDC uses PoseNet to detect individual keypoint in a bottom-up manner. Moreover, SegNet is designed to take the high confident keypoint as a dynamic centroid to associate the pixels to the right instance. Unlike top-down approaches [3, 6, 9, 10], KDC detects humans without requiring a box detector or incurring runtime complexity.

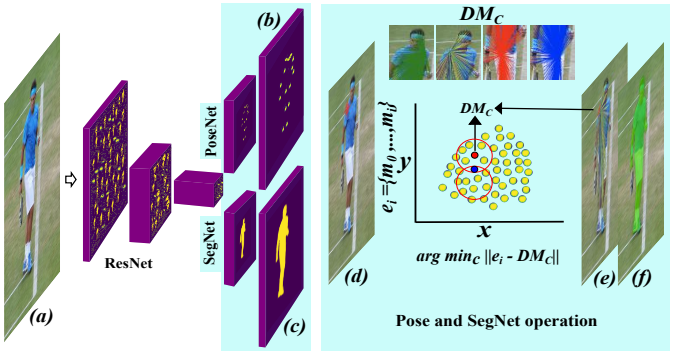

Fig. 1: KDC architecture. Both head Networks (PoseNet and SegNet) get high-level features from an input image (a) and generate keypoint feature map (a) and segmentation map(b). A highly confident keypoint (d) is used as a dynamic mask centroid DM_c (e) for the pixels in the embedding space to assign the pixel to the right instance. Finally, the SemanticSeg module uses the information from both networks for the final representation of instance-level segmentation (f).

KDC is not the very first method to leverage bottom-up approaches [14] to detect keypoints and perform segmentation task [1, 8, 16]. However, existing models [14] use human poses to refine pixel-wise clustering for segmentation and thus do not perform segmentation task well. Moreover, they suffer high overheads because of the extra computation of a person detector [8], scalability issues, for instance, segmentation [16], and model complexity of [1], which makes them unsuitable for crowd scenarios and real-time applications. Unlike existing models, KDC does not incur the high overheads associated with top-down approaches because of the person detector, nor the segmentation performance and scalability concerns associated with bottom-up approaches due to pixel-wise clustering without knowing the pixels core points.

KDC addresses the aforementioned challenges using two primary networks: PoseNet and SegNet (Figure 1). The PoseNet generates keypoint feature maps (KFM) from the input image (Figure 1a) using the keypoint disk that estimates the relative displacement between pairs of keypoints and improves the precision of long-range, occluded, and proximate keypoints (Figure 1b) to use them as a dynamic centroid for mask pixels. The SegNet performs pixel-level classification and produces a segmentation map (Figure 1c) using the dy-

namic high confident keypoint (Figure 1d) as a MaskCentroid (Figure 1e). The MaskCentroid defines the embedding space to associate pixels to the right instance by using the keypoints as centroids. At last, semantic segmentation (Figure 1f) is enabled utilizing the high-level features from PoseNet and SegNet.

We evaluated the performance of KDC using the MSCOCO [11], and OCHuman [16] benchmarks. To the best of our knowledge, this is an original effort using keypoints as dynamic centroids that could be changed in case of rapid occlusions in real-time. Compared with previous SOTA pipelines, KDC shows better robustness in detection errors and alleviates the complexities of occlusions, entangled limbs, and overlapped people.

II. Related Work

There are two primary approaches to identify objects semantically : (1) single-stage 1[5, 13, 2] and (2) multi-stage [8, 15]. The single-stage generates intermediate and distributed feature maps based on the input image. The InstanceFCN [5] produces instance-sensitive scoring maps and applies the assembly module to the output instance. This approach is based on repooling and other non-trivial computations (e.g., mask voting), which is unfavorable for real-time processing. YOLACT [2] runs a set of mask prototypes and uses coefficient masks to perform segmentation; however, this method is critical to obtain a high-resolution output. Multi-stage instance segmentation follows the detect-then-segment paradigm. It first performs box detection, and then pixels are classified to obtain the final mask in the box region. Mask R-CNN [8] is based on multi-stage instance segmentation that extends Faster R-CNN [15] by adding a branch for predicting segmentation masks for each Region of Interest. The method presented by [12] improves the accuracy of the Mask R-CNN by enriching the Feature Pyramid Network features.

SOTA developments have been made to use keypoints and perform human semantic segmentation. Pose2Seg [16] proposed human pose-based instance segmentation. This method separates instances based on the human pose, rather than the proposal region. It takes already generated pose as input that makes concerns an end-to-end training model. PersonLab [14] group keypoints by using greedy decoding. This method also reports a part-induced geometric embedding descriptor for human class instance segmentation. However, this approach fails to perform segmentation on highly entangled instances. A new Pose*Plus*Seg [1] played an important role in this regard; however, it compromises the performance of the model using static centroid and a couple of refined networks making it a complex structure model. We propose a simple, yet effective system to overcome the above complications KDC introduces KFM and uses a high confident keypoint as a dynamic clustering point that helps to cluster the mask pixels to a particular instance and generate a precise semantic mask even in high occluded cases.

III. Technical Approach

A. Keypoint Feature Maps

KDC generates KFM by employing the PoseNet, as illustrated in Figure 1b. In this stage, each individual keypoint

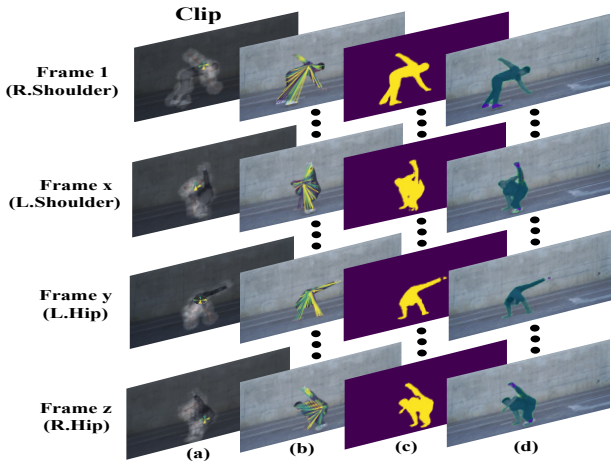

Fig. 2: (a) Indicates high confident keypoints (b) shows MaskCentroid a dynamic centroid based on the high keypoint confidence score (c) presents a precise segmentation map (d) represents human semantic segmentation.

is detected and concatenated for the output feature maps. Specifically, we adopt the residual-based network for our multi-person pose setting to produce KFM and calculate the high confident keypoint. We modeled our keypoint prediction as follows:

Let p_i represent the keypoint position in the image, where $i = \{1, \ldots, N\}$ are mapped to the 2D positions of the pixels. A keypoint disk $D_R(q) = \{p : \|p-q\| \leq R\}$ of radius R is focused at point q. Additionally, let $q_{j,k}$ be the 2D position of the j^{th} keypoint of the k^{th} person instance, where $j = \{1, \ldots, I\}$ and I is the number of individual keypoints in the image. For each known keypoint j, a binary classification approach is followed. Specifically, every predicted keypoint pixel p_i is binary classified such that $p_i = 1$ if $p_i \in D_R$ for each person keypoint j, otherwise $p_i = 0$. Thus, for every keypoint, there are independent dense binary classification tasks. To obtain the KFM for each keypoint j, we define a disk D_R of radius $R = 32$ (diameter = 64) independent of the keypoint scale. To equally weigh person keypoints in the classification loss, we choose a disk radius that does not scale according to the instance size. Note that R is constant for all experiments in this paper for optimal results. We train the model and compute the KFM loss based on the annotated image positions, back-propagating across the entire image, excluding the regions with individuals who are not fully annotated with keypoints (e.g., crowded areas and small individual segments).

B. MaskCentroid

Object segmenting is a practice of pixels classification challenge where to allocate pixels to the right instance. For this task, we use high confident keypoint 2a as dynamic centroid and called it MaskCentroid M_c as illustrated in Figure 2b to cluster the mask pixels with the defined centroid C_i inside each annotated person instance with 2D mask pixels, points from image position x_i to the position C_i of the corresponding instance. At each semantically identified human instance, the

pixels embedding $e(x_i)$ reflects a local approximation of the absolute location of each mask pixel of an individual to whom it corresponds, i.e., it depicts the expected 2D shape structure of the human body. Therefore, for each pixel, we learn pixels offset, pointing to the C_i. Every C_i is a high confident dynamic keypoint cluster the pixels to a particular human instance in the embedding space. The purpose of segmenting the human body is to assign a set of pixels $P_i = \{m_0, m_1, m_2, ..., m_i\}$ and its 2D embedding vectors $e(m_i)$, into a set of instances $I = \{N_0, N_1, N_2, ..., N_j\}$ to generate a 2D mask for each human instance as shown in Figure 2c. Pixels are clustered to their corresponding centroid:

$$C_i = \frac{1}{N} \sum_{m_i \in N_j} m_i. \tag{1}$$

This is attained by defining pixel offset vector v_i for each known pixel m_i, so that the resulting embedding $e_i = m_i + v_i$ points from its respective instance centroid. We penalize pixel offset loss by the L_1 loss function throughout model training, averaging and back-propagating at the image position x_i that corresponds to an instance of a specific individual entity:

$$L = \sum_{i=0}^{n} \|v_i - \hat{v}_i\|, \tag{2}$$

where $\hat{v}_i = C_i - m_i$ for $m_i \in N_j$. In order to cluster the pixels to their centroid, it is important to specify the positions of the instance centroids and to assign pixels to a particular instance centroid. We propose to use a gaussian function ϕ_j for each instance N_j, which converts the distance between a (spatial) pixel embedding $e_i = m_i + v_i$ and the instance centroid C_i into a probability of belonging to that instance:

$$\phi_j(e_i) = exp\left(-\frac{\|e_i - C_i\|^2}{2\sigma_j^2}\right). \tag{3}$$

The threshold for low and high probability is defined as 0.5, if the pixel embedding e_i is close to the instance centroid and is likely to belong to that instance, while a low probability means that the pixel is more likely to belong to the background (or another instance). More specifically, if $\phi_j(e_i) > 0.5$, than that pixel, at location x_i, will be assigned to instance N_j.

Dynamic Center of Attraction . Significant innovation has been made in SegNet over SOTA [1]. SOTA model used centroid as a fixed parameter to cluster the mask pixel and thus suffered inferior results if the centroid occludes in real-time cases. However, we can also let the network learn the optimal center of attraction by introducing the dynamic centroid. This can be done by defining the high confident keypoint as a learnable parameter that could be changed in case of rapid occlusions in real-time. By doing so, the network can influence the location of the center of attraction by changing the location of the embeddings:

$$\phi_j(e_i) = exp\left(-\frac{\|e_i - \frac{1}{|N_j|}\sum_{e_j \in N_j} e_j\|^2}{2\sigma_j^2}\right). \tag{4}$$

During inference, using the keypoints as a dynamic centroid for the mask pixels effectively addresses the challenging scenarios where more than 70% of the human body is occluded. Our experimental study analyzes the effectiveness of both Static MaskCentroid SM_c and Dynamic MaskCentroid DM_c on human instance segmentation (§V-A).

Instance-wise Gaussian Optimization (IGO). To precisely align the predicted semantic maps SegNet performs Gaussian smoothing [4] at the instance-level, i.e., *instance-wise Gaussian optimization*. We apply instance-wise smoothing to reduce the noise and retain useful information while producing individual semantic maps. We fix the σ ranging from 0.1 to 1 in order to obtain precise instance-level segmentation concerning the ground truth. We find that the σ value close to 0.1 produce a more precise segmentation mask when individuals are entangled and overlapped. Our ablation experiments demonstrate this observation and the effectiveness of instance-wise smoothing in(§V-B).

C. SemanticSeg Module

We introduce a new algorithm called SemanticSeg, which uses the high confident keypoint as a clustering point for human semantic segmentation, as illustrated in Figure 2a. The SemanticSeg module uses the high-level features generated from PoseNet and SegNet to produce the optimal semantic map, as shown in Figure 2b. Initially, keypoints and their 2D coordinates are cached in a priority queue. These keypoints are then used to detect body instances and gradually connect adjacent keypoints. High confident keypoints are calculated based on the confident score using the keypoint disk and the visible stability of keypoints. Further, KDC performs semantic level segmentation for all detected humans by clustering the pixels to its centroids defined for each human instance. Specifically, if the pixel m_i probability $P(m_i) > 0.5$, then that pixel at position x_i is assigned to the relevant human instance. We assume pixels with a probability threshold > 0.5 are close to the instance centroid and considers as a part of a particular instance. Otherwise, the pixels belong to another instance or background.

IV. Evaluation

We evaluate KDC using the standard benchmarks, COCO [11], and OCHuman [16] that focus on heavily occluded individuals and compare the computational cost and inference time with SOTA models. The model is trained end-to-end using the COCOPersons training set. Ablations are conducted on the COCO *val* set. The residual-based network ResNet-101 (RN-101) and ResNet-152 (RN-152) [7] are used for training and testing. The hyperparameters for training were: learning rate = $0.1 \times e^{-4}$, image size $= 401 \times 401$, batch size $= 4$, and Adam optimizer. We performed various transformations during model training, such as scale, flip, and rotate operations. and conducted synchronous training for 400 epochs on a single TITAN RTX using TensorFlow.

Results. Table I and Table II present the results of COCO Segmentation *val* and *test* sets. KDC achieved an mAP of 0.610 on the *val* set, and improved the AP by 0.192 compared with PersonLab [14](multi-scale), 0.055 AP compared with

Fig. 3: Example results from COCO and CrowdPose datasets.

Models	Backbone	AP	$AP^{.50}$	$AP^{.75}$	AP^{M}	AP^{L}	AR
PersonLab †	RN-101	0.382	0.661	0.397	0.476	0.592	0.162
PersonLab †	RN-152	0.387	0.667	0.406	0.483	0.595	0.163
PersonLab ‡	RN-101	0.414	0.684	0.447	0.492	0.621	0.170
PersonLab ‡	RN-152	0.418	0.688	0.455	0.497	0.621	0.170
Pose2Seg	RN-50-fpn	0.555	-	-	0.498	0.670	-
Pose*Plus*Seg	RN-152	0.563	0.701	0.557	0.509	0.683	0.701
KDC	RN-101	0.580	0.721	0.559	0.521	0.691	0.699
KDC	RN-152	0.610	0.764	0.579	0.567	0.724	0.706

TABLE I: Comparison on **COCO** Segmentation *val* set. † indicates single-scale testing. ‡ indicates multi-scale testing.

Pose2Seg [16], and 0.047 compared with Pose*Plus*Seg [1]. Furthermore, on the test set, KDC deliver top accuracy of 0.476 mAP and improved the AP by 0.105 over Mask-RCNN [8], 0.059 over PersonLab[14] (multi-scale), and 0.031 over Pose*Plus*Seg [1]. Table III shows segmentation performance compared with Pose2Seg [16] using the OCHuman *val* and *test* sets. Qualitative visual results from COCO and OCHuman datasets are presented in 3

Models	Backbone	AP	$AP^{.50}$	$AP^{.75}$	AP^{M}	AP^{L}	AR
Mask-RCNN	RNeXt-101	0.371	0.600	0.394	0.399	0.535	-
PersonLab †	RN-101	0.377	0.659	0.394	0.480	0.595	0.162
PersonLab †	RN-152	0.385	0.668	0.404	0.488	0.602	0.164
PersonLab ‡	RN-101	0.411	0.686	0.445	0.496	0.626	0.169
PersonLab ‡	RN-152	0.417	0.691	0.453	0.502	0.630	0.171
Pose*Plus*Seg	RN-152	0.445	0.794	0.471	0.524	0.651	0.677
KDC	RN-101	0.457	0.804	0.478	0.535	0.674	0.677
KDC	RN-152	0.476	0.818	0.487	0.546	0.678	0.682

TABLE II: Comparison on **COCO** Segmentation *test* set. † indicates single-scale testing. ‡ indicates multi-scale testing.

Models	Backbone	Val mAP	Test mAP
Pose2Seg	RN-50-fpn	0.544	0.552
KDC	RN-101	0.567	0.570
KDC	RN-152	0.583	0.596

TABLE III: Performance comparison using **OCHuman** segmentation *val* and *test* datasets.

Models	AP	$AP^{.50}$	$AP^{.75}$	AP^{M}	AP^{L}
PersonLab	0.418	0.688	0.455	0.497	0.621
Pose2Seg	0.555	-	-	0.498	0.670
Pose*Plus*Seg	0.563	0.701	0.557	0.509	0.683
KDC:					
ResNet101 (SM_c)	0.568	0.721	0.561	0.531	0.690
ResNet152 (SM_c)	0.570	0.734	0.565	0.547	0.698
ResNet101 (DM_c)	0.599	0.751	0.572	0.551	0.717
ResNet152 (DM_c)	0.610	0.764	0.579	0.567	0.724

TABLE IV: Comparison between existing models vs. KDC's component (SM_c and DM_c).

Inference Time. Figure 4 presents the average precision (x-axis) and inference time (y-axis) of KDC with SOTA competitors using an image size of 401×401 resolution. Mask-RCNN [8] and Pose*Plus*Seg [1] results are obtained using COCO. Pose2Seg [16] and KDC results are obtained on OCHuman datasets.

V. Ablation Experiments

A. Static vs. Dynamic MaskCentroids

We analyze the Static MaskCentroid SM_c and Dynamic MaskCentroid SM_c that play an important role to divide individuals semantically. We compared KDC's MaskCentroid with human segmentation models. Table IV shows the MaskCentroid accuracy trade-off compared with PersonLab [14], Pose2Seg [16], and Pose*Plus*Seg [1].

B. Impact of IGO

Finally, we examine the impact of *instance-wise Gaussian optimization* on human semantic segmentation task. We tested the sensitivity of σ ranging from 0.1 to 0.5 on human instance segmentation. Figure 5 shows the results with different σ values, where low σ provides a precise segmentation mask and performs better in crowded cases.

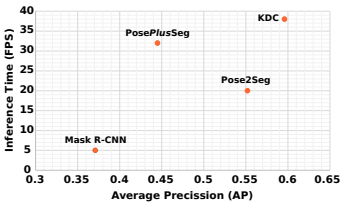

Fig. 4: Comparison of runtime performance. All the results are obtained from their papers.

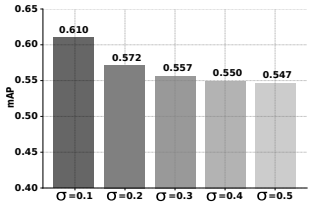

Fig. 5: Impact of instance-wise Gaussian optimization with varied σ.

VI. Conclusion

This paper rethinks the well-known human semantic segmentation problem in the context of challenging multi-person scenarios. KDC inaugurate a keypoint feature map using keypoint disk. In addition, a MaskCentroid is introduced that defines the keypoints as a dynamic centroid to cluster the pixels with the right instance in the embedding space even in the high occlusions. The effectiveness of KDC is evaluated using COCO and OCHuman challenging datasets. We conclude that KDC is a highly effective approach for the task of human semantic segmentation. Our in-depth evaluation has demonstrated its advantages in both accuracy and run-time performance over existing models.

ACKNOWLEDGMENTS

This work was supported in part by the National Research Foundation of Korea (NRF) grant 2022R1G1A1003531, 2022R1A4A3018824, RS-2023-00230593 and Institute of Information and Communications Technology Planning and Evaluation (IITP) grant IITP-2022-2020-0-01741, RS-2022-00155885 funded by the Korea government (MSIT).

REFERENCES

[1] Niaz Ahmad, Jawad Khan, Jeremy Yuhyun Kim, and Youngmoon Lee. Joint Human Pose Estimation and Instance Segmentation with PosePlusSeg. In *AAAI*, 2022.

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019.

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018.

[4] Moo K Chung. Gaussian kernel smoothing. *arXiv preprint arXiv:2007.09539*, 2020.

[5] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *ECCV*, 2016.

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[9] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *ICCV*, 2017.

[10] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[12] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018.

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

[14] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[16] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *CVPR*, 2019.