Dynamic Interaction Dilation for Interactive Human Parsing

Yutong Gao, Congyan Lang, Fayao Liu, Yuanzhouhan Cao, Lijuan Sun, Yunchao Wei

***Abstract*—Interactive segmentation pursues generating high-quality pixel-level predictions with a few user-provided clicks, which is gaining attention for its convenience in segmentation data annotation. Users are allowed to iteratively refine the prediction by adding clicks until the result is satisfactory. Existing interactive methods usually transform the clicks into a set of localization maps by Euclidian distance computation or RGB texture extraction to guide the segmentation, which makes the click transformation a core module in interactive segmentation networks. However, when adopted in human images where large poses, occlusions, and bad illuminations are prevailing, prior transformation methods tend to cause uncorrectable overlapping across localization maps, *i.e.*, one click corresponds to multiple transformed values at the same position in different localization map channels, which are difficult to form a good match among human parts and limit the interaction efficiency. Furthermore, the inappropriately transformed information is hard to be refined with the static transformation manner, *i.e.*, based on the fixed formulas / RGB textures, which is out of tune with the dynamically refined interaction process. Hence, we design a dynamic transformation scheme for interactive human parsing (IHP) named Dynamic Interaction Dilation Net (DID-Net), which serves as an initial attempt to break the limitations of static transformation while capturing long-range dependencies of clicks within each human part. Specifically, we construct a Dynamic Dilation Module (DD-Module) to dilate clicks radially in several directions assisted by human body edge detection. The continually refined edges guide to improve the dilation quality in each interaction iteration, thereby better fitting user intention. Furthermore, we propose an Adaptive Interaction Excitation Block (AIE-Block) to exploit potential semantic clues buried in the dilated clicks and emphasize semantic expression for each human part by feature recalibration. Our DID-Net achieves state-of-the-art performance on 3 public human parsing benchmarks.**

***Index Terms*—Human parsing, Interactive image segmentation, Semantic image segmentation.**

I. Introduction

Human parsing is a particular segmentation task, aiming to identify each semantic part, *i.e.*, head, arms, clothes, etc., which is one of the most critical problems in various human-centric applications, like human behavior and attribute analysis [1], [2], expression recognition [3], [4] to name a few. Excellent human parsing models [1], [2], [5]–[16] are often built on a large number of finely annotated pixel-level labels, more training data often leads to better parsing performance. However, the amount of public human parsing data is limited, and annotating more fine-grained human images is time-consuming and expensive. Therefore, Interactive Human Parsing (IHP) [17] is gaining attention to save the cost of data annotation by parsing with the guidance of a few user-provided clicks. Users can improve the parsing result by iteratively adding correction clicks until the prediction is satisfactory. In addition, the IHP has promising applications due to its capability to understand human provided information, such as marking human part occlusions [18] and reducing noisy annotations [19], [20] that are prevalent in the wild scenarios.

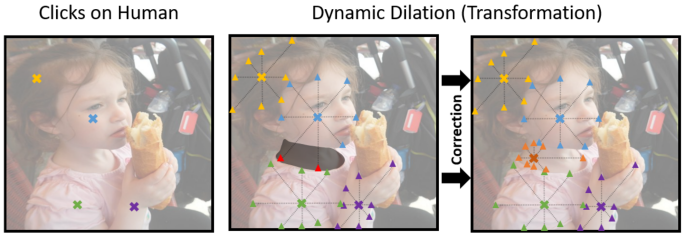


Fig. 1. Illustration of the click dilation transformation. The × indicates the clicks in human parts, and the △ indicates the dilation points of those clicks. Different colors represent clicks / dilation points in different classes. The shadow in the middle figure represents the mislabeled area that needs to be corrected, and the red △ in the shadow represents the inappropriate dilation points, which are dynamically fixed after one correction. Best view in color.

To make user clicks trainable with images, interactive models [17], [21]–[33] usually transform clicks into a set of localization maps and then feed them into the network together with the RGB channels. Hence, click transformation is a core module for interactive methods, a suitable transformation method can guide the interaction process more efficiently. Attempts have been made to explore effective transformation methods, *e.g.*, some [17], [21]–[23] leverage Euclidean / Gaussian functions to detect objects / boundaries, some [24], [34] use super-pixels to explore a reasonable way for transformation. Others [25]–[27] leverage extracted features to enhance the transformation expression on the basis of Euclidean distance computation. However, for the IHP task, the clicks are usually densely distributed in the closely connected human parts, and we argue that existing transformation methods tend to cause uncorrectable overlapping across localization maps, *i.e.*, one click corresponds to multiple transformed values at the

Yutong Gao and Congyan Lang *(corresponding author)* are with the Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China. Fayao Liu is a research scientist at Institute for Infocomm Research, A*STAR, Singapore. Yuanzhouhan Cao is currently a lecturer at the School of Computer Science and Information Technology, Beijing Jiaotong University, Beijing 100044, China. Lijuan Sun is currently a Postdoctor in the School of Economics and Management, Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China. Yunchao Wei is with the Institute of Information Science, Beijing Jiaotong University, Beijing, China, 10044. E-mail: wychao1987@gmail.com.

same position in different localization map channels, which are difficult to form a good match among different human parts and limit the interaction efficiency. Also, most prior works transform the clicks based on fixed formulas or RGB textures, keeping the basis of transformation static during the interaction process. And the inappropriately transformed information is hard to be refined in such static transformation manners which are out of tune with the dynamically refined interaction process.

To address the above issues, we design a dynamic transformation scheme for the IHP named Dynamic Interaction Dilation Net (**DID-Net**), which serves as a reliable transformation solution to fit in the dense-connected human parts and an initial attempt to break the limitations of static transformation. As mentioned by Ruan *et al.* [11] and Zeng *et al.* [16], one of the characteristics of human parsing is the closely connected human parts, which makes the edge identification of body parts particularly important. We hence utilize edge detection to construct a Dynamic Dilation Module (**DD-Module**) to dilate clicks radially in several directions, as shown in Fig. I, aiming at mining potential interaction cues and capturing long-range dependencies within each human part. Specifically, clicks in each direction are dilated to the category junction of different body parts according to the edge confidence, with one dilation point generated in each direction. The small number of dilation points effectively reduces the confusion guidance caused by the overlapping between different channels of the localization maps and thus leads to higher interaction efficiency. Furthermore, in each round of interaction, the new click guides optimize the edge confidence, thereby dynamically improving the dilation quality. The dilated clicks can be adjusted with feature optimization during interaction and continuously fit the user's intention. In particular, to reduce over-dilation / under-dilation caused by detected edge missing / redundant, we do not simply use a confidence threshold to decide where the dilation should stop, but propose a confidence accumulation method. We accumulate the edge confidence in each dilation direction, the dilation stops automatically when the accumulated confidence reaches a certain value, then we filter and retain those reliable dilations based on the edge confidence at the dilation point.

Meanwhile, to further exploit potential semantic clues contained in the dilated clicks, an efficient Adaptive Interaction Excitation Block (**AIE-Block**) is proposed. Specifically, we aggregate dilated click features to guide the recalibration of each human part features to emphasize semantic expression. The aggregated features are adaptively updated in each round of interaction with the newest clicks and dilation points to fit user provided cues.

We conduct various experiments to prove the effectiveness of all the designs we proposed on three human parsing benchmarks, and achieve state-of-the-art IHP performance. It is worth mentioning that our model can exceed state-of-the-art methods by a large margin using only the DD-Module.

In conclusion, the major contributions of this paper lie in the following aspects:

1) We initially explore the dynamic transformation manner for interactive segmentation to improve transformation quality and better fit user cues during the interaction. We demonstrate its advantages through extensive experiments.
2) We propose an effective dynamic transformation scheme for the IHP to capture long-range dependencies of clicks in each human part. We further exploit semantic clues contained in the dilated clicks to emphasize semantic expression by feature recalibration.
3) Our method achieves significant improvement on three human parsing benchmarks, and the performance outperforms state-of-the-art methods by a large margin. Compared with state-of-the-arts, our method can save 40% and 46.1% correction efforts to achieve 85% mIoU on LIP [6] and PASCAL-Person-Part [35] datasets, respectively.

II. Related works

A. Human Parsing

Human Parsing aims to identify each pixel in a human image into several human-related semantic categories, *e.g.*, body parts, and clothes. It has received considerable interests due to its huge application potential. Compared with other semantic segmentation tasks, human parsing is characterized by tighter connections and stronger structure relations between human parts. A lot of excellent models [1], [2], [5]–[16] are produced to produce more accurate human masks. Some works [11], [16] utilizes the edge identification as an important clue of the closely connected human parts to represent the body parts more expressively. Wang *et al.* [15] propose a hierarchical human parser that exploits the representational capacity of deep graph networks and also the hierarchical human structures. All these models are hungry for pixel-level annotation labels for training. However, the amount of existing publicly available human parsing data is limited. Meanwhile, it is costly to obtain more finely annotated human image data where large poses, appearance variations, occlusions, and bad illuminations are prevailing.

B. Interactive Segmentation

The goal of interactive segmentation is to segment images with the guidance of human interactions so that users can obtain satisfactory pixel-level labels at the cost of a small amount of annotations. It has promising applications due to its capability to understand human provided information, and can be used for fast pixel level data annotation, data denoising in noisy annotations [19], [20], and marking human part occlusions [18] to name a few.

Interactive models have strong plasticity to generate segmentation results, and users can obtain better segmentation masks by iteratively adding annotations, *e.g.*, clicks, bounding boxes. Therefore, interactive models pursue higher interaction efficiency, *i.e.*, achieving satisfying performance with fewer annotation efforts. Early interactive works [36]–[46] use low-level features such as textures and colors of images to model the relations between pixels, and predicts segmentation results by their local continuity. These methods are inefficient due to the lack of high-level semantic perceiving.

Deep learning based interactive models [21]–[31], [34], [47]–[54] have become popular in recent years. Compared with earlier interactive works, they can assist interactions more efficiently b y e xtracting h igh-level s emantic i nformation of images. Most of them are focus on the the instance-level segmentation setting.

To enable user clicks to be trained with images, Xu *et al.* [21] propose to transform clicks located in the instance (object) and background into a pair of localization maps by Euclidean distance calculation and provide a user click simulation strategy for training. Maninis [22] explores to detect instance edge by transforming extreme clicks, *i.e.*, left-most, right-most, top, bottom pixels, with Gaussian functions. Wang *et al.* [23] propose to further refine p redictions b y d ragging c licks to object boundaries. Chen *et al.* [34] and Majumder *et al.* [24] employ super-pixels to explore a reasonable transformation scheme to achieve better click guidance. Other methods [25]–[27] leverage extracted features to enhance the transformation expression based on the Euclidean distance calculation.

In particular, Gao *et al.* [17] use scaled Euclidean distance maps and explore several click simulation strategies for the IHP. They demonstrate the effectiveness of semantic-level interactive models which are better than instance-level ones on the IHP task. Compared with semantic-level interactive segmentation, most instance-level interactive segmentation methods require separate annotation for each semantic class during the annotation process, difficult to share semantic information of different classes and leads to inefficient i nteraction. In contrast, for semantic-level interactive segmentation, the clicks of different semantic classes can fully complement each other's semantic cues and effectively guide the data annotation of semantic segmentation tasks.

Though effective, most existing interactive methods are based on the static transform manner, *i.e.*, based on the fixed formulas / RGB textures, which brings a lot of uncorrectable noises and overlapping across localization maps. Furthermore, we argue that the static transformation manner is incompatible with the dynamically refined interaction process. In this paper, we hence design a sparse and dynamic click transformation scheme for the IHP task which can dynamically refine the inappropriately transformed area to better fit u ser cues.

III. Proposed Method

In this section, we first describe the basic interaction process of the IHP, and then introduce the overview framework of our DID-Net. Further, we describe the design of the DD-Module and the AIE-Block in detail in turn.

A. Interaction Process

We adopt the IHP pipeline proposed by Gao *et al.* [17], which is shown to be reasonable and efficient. The interaction process consists of two stages: the initialization stage and the correction stage. In the initialization stage, one click is added to the central area of each human part for the initialization mask. In the correction stage, clicks are iteratively added to the largest error area until a satisfactory result is generated.

B. Network Overview

Our DID-Net employs the ResNet101 [55] based DeepLab V3+ [56] as the backbone. As shown in Fig. 2(a), after an initial transformation of the clicks, the generated initial localization maps are concatenated with RGB images as the initial input to dilate clicks in the DD-Module for dynamic transformation. Then a set of dilation maps are generated and additionally concatenated with the initial input, and propagate together through the whole network for parsing predictions. Specifically, the DD-Module is composed of an edge detector and a click dilator, the edge detector is constructed to generate edge confidence maps by edge detection to provide the basis for the click dilation, and the click dilator generates dilation maps as the complement transformation of the initial localization maps. The structure of the edge detector is shown in Fig. 2(b). Furthermore, we add an AIE-Block after the feature extractor to recalibrate more expressive features based on the dilation maps for each human part.

C. Dynamic Dilation Module

Essentially, the process of user-model interaction is a dynamic process of cognitive unification, where the predictions of the model are continuously refined following the addition of the clicks. However, most extant interactive works [17], [21]–[25], [34] adopt static and dense click-transformation algorithms, result in inter-category transformation overlapping and are difficult to be corrected. We therefore design the DD-Module to build a sparse transformation algorithm that reduces transformation overlapping while dynamically matching user intent. Here, we elaborate on our click transformation method within the DD-module, including the dilation transformation form and the click dilation process.

Dilation Transformation Form. First, we introduce the transformation form of click dilation. Due to the fact that the human body parts are closely connected, the distribution of the clicks for IHP tasks is relatively dense, most existing transformation methods like Euclidean distance calculation [17], [21], [25], Gaussian function [22], [23] and super-pixels / object-proposals [24], [34] transform a click to a cluster of densely distributed values, tend to cause information overlapping across localization maps, *i.e.*, one click corresponds to multiple transformed values at the same position in different localization map channels, which are difficult to form a good match among the human parts.To address this, inspired by CE2P [11], we design a sparse transformation mode by leverage part edges as an important clue of the human body to assist the clicks dilate in several directions for long-range semantic dependencies capturing. Specifically, we dilate the clicks in eight uniformly distributed directions to where edges may exist according to the edge confidence map. Considering that detected edge missing / redundancy may lead to over-dilation / under-dilation, we do not simply use a confidence threshold to decide where the dilation should stop, but propose a confidence accumulation method. For each dilation direction, we accumulate the edge confidence and stop the dilation when the accumulated value reaches a preset threshold. Let e_n be the edge confidence value of the n-th pixel away from the

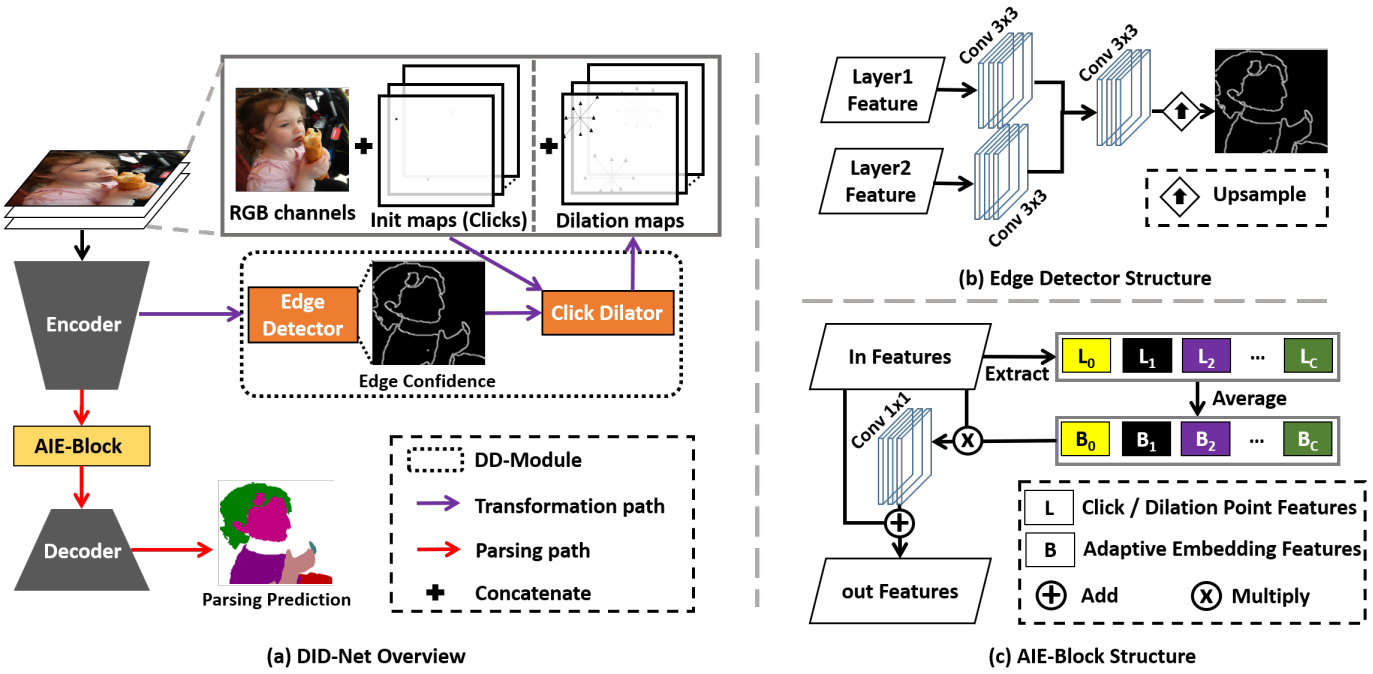


Fig. 2. (a) Pipeline of DID-Net. The $\rightarrow$ in purple, red and black indicate the forward propagation path during transformation, prediction, and both, respectively. "Add" indicates the addition between two features which have the same dimensions. "init maps" represents the initial localization maps. (b) Illustration of the edge detector in the DD-Module. "Layer1 Feature" and "Layer2 Feature" are the image features from the first and second ResNet101 [55] convolution blocks, respectively. (c) Illustration of our AIE-Block. "Extract" represents the click / dilation point features extraction. "C" is the total number of classes. "Add" and "Multiply" represent element-wise addition and multiplication between features. Best viewed in color.

click on the dilation direction. Then the end of the dilation can be formulated as,

$$\sum_{i=0}^{t} e_i \leq \theta < \sum_{i=0}^{t+1} e_i, \tag{1}$$

where θ is a predefined threshold hyperparameter for the accumulated value, the dilation ends at the t-th pixel in the dilation direction. Then, we decide whether the dilation ends at a near edge area by checking if e_t is "confident" enough, and we retain the dilation while generating a dilation point if $e_t > \delta$, where δ is an edge confidence threshold for dilation point generating. The semantic label of the dilation points is consistent with the click.

By digging the adjacency among different human parts, the dilation transformation form effectively reflect the long-range dependence within each parts. At the same time, the sparse transformation form based on several dilation points helps to reduce redundant transformed information and leads to a more efficient prediction guidance (refer to Section IV-D).

Click Dilation Process. Most of the existing interactive segmentation works [17], [21], [23], [24], [28], [29] transform the clicks based on fixed transformation formulas or RGB texture features, *e.g.*, Euclidean distance and super-pixel computation, to transform clicks in a static way. However, these static transformation are difficult to be corrected by subsequent correction clicks once semantic noises are generated in the localization maps, resulting in limited interaction efficiency. Therefore, in this section, we obtain the most reliable localization maps in each interaction round in a dynamic transformation manner based on the adjacency of body parts, *i.e.*, part edges. The edge optimization dynamically updates and refines the transformed information in the existing localization map.

As mentioned in Section III-B, to utilize the continue-refined adjacency correlation among the human parts, we divided our transformation process into 2 phases: initial transformation and dynamic transformation. We design the initial conversion as a static transformation process following most interactive segmentation methods [17], [21]–[25], [34], as the click updating, we employee the initial transformation as an aid to dynamically transform the clicks in the dynamic transformation phase. In the initial transformation phase, we simply transform each clicked pixel and generate a set of initial localization maps, to provide the updated edge confidence map according to the newest clicks. Then in the dynamic transformation phase, we utilize the latest edge confidence map to dilate the clicks in the click dilator and get several dilation points. We treat these dilation points as the continually updated guidance of clicks and encode them into a set of dilation maps to dynamically complement the semantic clues in the initial localization maps, and together guide the model predictions. The initial transformation and dynamic transformation phases share the encoding formula for clicks / dilation points. Let $M_c, c = 1, \ldots, C$ denotes the transformed maps for class c, *i.e.*, initial localization maps, dilation maps, where C is the number of semantic classes. $\mathcal{D}_c, c = 1, \ldots, C$ denotes the set of clicks / dilation points for class c. For each element

$m \in M_c$, the encoding formula can be expressed as,

$$m = \begin{cases} 255 & \exists d \in \mathcal{D}_c, \parallel d, m \parallel_2 = 0 \\ 0 & \forall d \in \mathcal{D}_c, \parallel d, m \parallel_2 \neq 0 \end{cases}, \tag{2}$$

where $\parallel *, * \parallel_2$ represents the Euclidean distance between two elements. We encode clicks and dilation points using the maximum transformation value of most prior works [17], [21]–[25], [34]. Differently, we abandon the transformation form of massively encoding non-click pixels through distance calculation / RGB texture, and only transform pixels where click / dilation points are located, aiming at reducing confusion guidance caused by transformation overlapping between localization maps.

We conduct virous experiments to prove the effectiveness and efficiency of our DD-Module. And make comparisons with several design choices of dilation strategies to attest our setting is more effective (refer to Section IV-D).

D. Adaptive Interaction Excitation Block

Based on the dynamic click dilation, we view the dilation points as continuously refined semantic dependencies of the clicks during interaction, which not only provide reliable transformation clues but also have the potential to motivate feature optimization in forward propagation process. Here we hence provide a simple and effective scheme to further exploit dynamic dilation points for adaptive feature optimization. Specifically, according to the dilation scheme in our DD-Module, the dilation points are usually distributed in the category proximity locations, which makes the clicks and dilation points of each category imply reliable semantic localization information. Therefore, we design an AIE-Block that leverage the clicks with their dilation points as semantic anchors, *i.e.*, the captured click-dependencies that represent different semantic categories, to adaptively and simultaneously recalibrate and optimize the features of each category.

As shown in Fig. 2(c), we extract the click / dilation point features within each human part for semantic expression enhancement. Due to inconsistent click numbers in different parts, it is hard to directly employ their features for recalibration. Therefore, average operations are performed semantically to obtain a set of dimension-consistent embedding features. During interaction, the semantic expression of these embedding features can be adaptively improved by the dynamically dilated clicks. Then we adopt the channel-wise feature interdependencies modeling [57] to recalibrate features of each human part by these embedding features to emphasize the semantic expression. Let $\mathcal{U}_c$ be the set of click & dilation point features of class c, F denote the image features input to the AIE-Block, then the feature recalibration in the AIE-Block can be expressed as,

$$\begin{cases} B_c & = \dfrac{1}{|\mathcal{U}_c|} \displaystyle\sum_{u \in \mathcal{U}_c} u \\ F'_c & = F \otimes B_c \end{cases}, c = 1, \ldots, C, \tag{3}$$

where $|\cdot|$ denotes the number of elements in the set. $\otimes$ is element-wise multiplication. B_c and F'_c are the embedding features and recalibrated features of class c respectively.

Finally, we fuse all the F'_c through a convolutional layer and form a residual structure with F to inject emphasized semantic expression, which can be formulated as,

$$\begin{cases} F_{fuse} & = \underset{c \in \{1,\ldots,C\}}{\text{Concat}} (F'_c) \\ R & = \text{sigmoid}(\text{Conv}(F_{fuse})) + F. \end{cases} \tag{4}$$

Here $Concat(*)$ represents the channel-wise concatenation. $Conv(*)$ is a convolutional layer with 1×1 kernel size, of which output dimension is consistent with F. $sigmoid(*)$ is the sigmoid activation function. R is the output of our AIE-Block.

We demonstrate the effectiveness of the AIE-Block through various experiments (refer to Section IV-D) and prove that dynamic dilation points possess potential for feature optimization.

E. Loss Function

The loss function of our DID-Net consists of a parsing loss and an edge perceiving loss, which is used to constrain the human parsing results and the edge confidence maps, respectively. Specifically, we adopt the standard Cross-Entropy loss for the parsing loss as in most human parsing works [11]–[16]. For the edge perceiving loss, we generate the edge label by implementing the edge detection method Canny [58] on the parsing labels. We adopt the balanced loss which has proven to perform well in boundary detection [59]. Let G denote the ground-truth semantic label of the human images, then the loss function of our DID-Net can be formulated as,

$$loss = \mathcal{L}_c(G, P) + \lambda \mathcal{L}_b(\text{Canny}(G), E), \tag{5}$$

where $\mathcal{L}_c(*, *)$ represents the standard Cross-Entropy for parsing loss. $\mathcal{L}_b(*, *)$ denotes the balanced loss for edge perceiving loss. $Canny(*)$ is the Canny edge detection operation. P and E represent the parsing prediction and the edge confidence map respectively. λ is a weighting hyperparameter.

IV. Experiments

A. Datasets

The **LIP** dataset [6] is a large-scale human parsing benchmark, containing 50462 human images with pixel-wise annotations on 19 semantic part labels.

The **PASCAL-Person-Part** dataset [35] provides pixel-wise labels for 6 human body parts in 3,535 annotated images.

The **CIHP** dataset [7] is a large-scale multi-human dataset with 38,280 diverse multi-human images, and labeled with 19 semantic human part categories.

B. Implementation Details

Click Simulation. As in most of the interactive segmentation works [17], [21], [25], [29], we simulate user clicks during training. We adopt the Central-Click simulation strategy proposed by [17] for locating human parts and using the iterative simulation method [31] to add extra clicks for robustness. Specifically, we feed the central clicks into our model in loop-one for initial prediction, and randomly pick 20 extra clicks

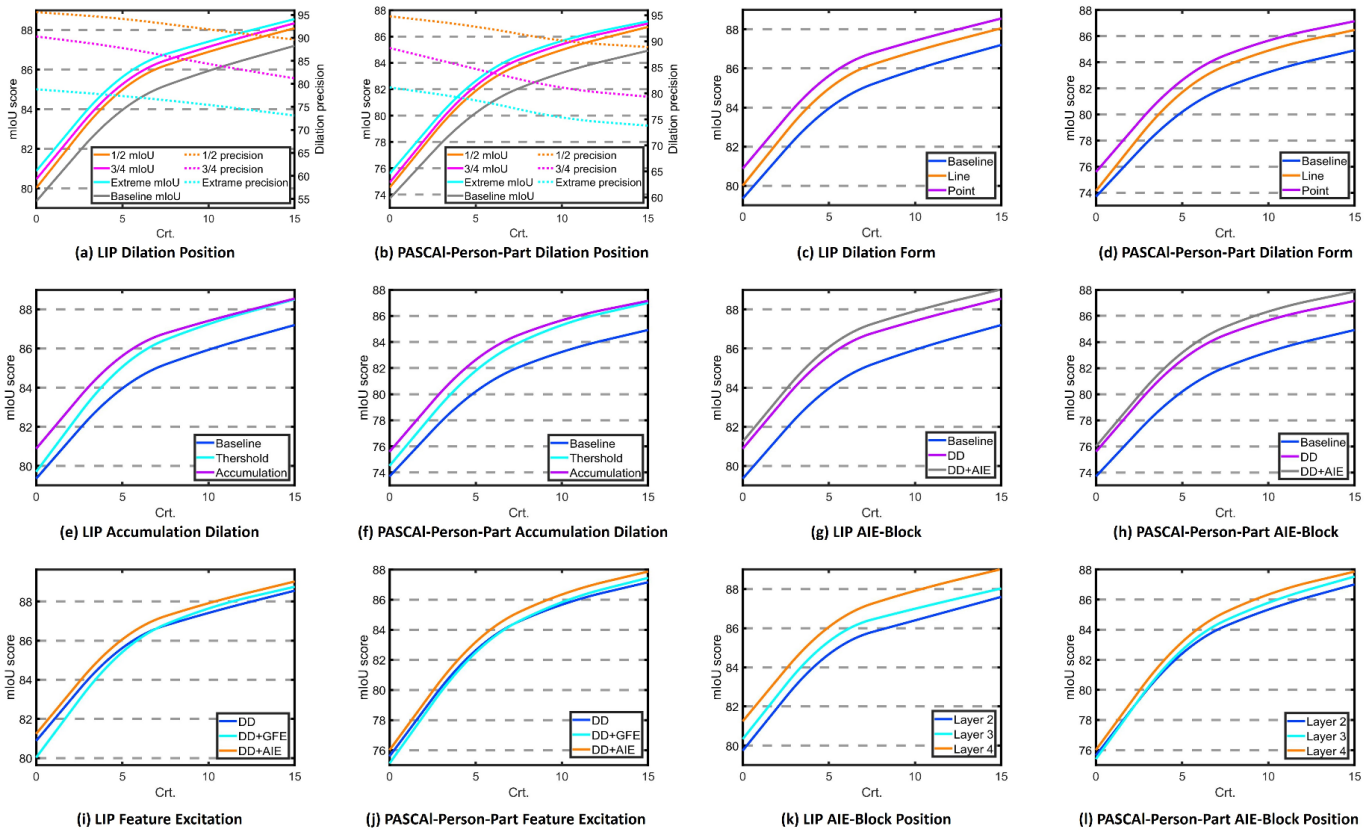


Fig. 3. Ablation experimental results on LIP and PASCAL-Person-Part datasets. "Baseline" represents our baseline model. (a)(b). Quantitative comparisons of using different dilation positions. "1/2", "3/4", and "Extreme" represents the variants of halfway position, three-quarters position, and extreme position, respectively. "mIoU" and "precision" represents the mIoU and dilation point precision score of the variants respectively. (c)(d). Quantitative comparisons between different dilation forms. "Line" and "Point" represent the line dilation form and point dilation form, respectively. (e)(f). Quantitative comparisons between different dilation ending. "Accumulation" and "Threshold" represents the dilations end according to accumulation value and threshold value, respectively. (g)(h). Effectiveness of the DD-Module and the AIE-Block. "DD" and "AIE" represent the model adopting the DD-Module and the AIE-Block, respectively. (i)(j). The AIE-Block *vs.* global feature excitation method. "GFE" represent the model adopt global feature excitation. (k)(l). Quantitative comparisons of different AIE-Block positions. "Layer 2", "Layer 3", "Layer 4" represent the AIE-Block setting after the second, the third and the fourth layer of ResNet101 encoder.

from the mislabeled pixels in the initial result. We then feed all the simulation clicks to our network to generate the final prediction for backpropagation.

Training Settings. We adopted ResNet101 [55] based DeepLab V3+ [56] as our backbone, and fine-tune the network pre-trained on MS COCO [60] with SGD optimizer. We set the learning rate, batch size, input size and training steps consistent within CM [17]. With C semantic categories contained in the dataset, the channel number of the first convolution layer is changed from 3 to $2\times C+3$. We set the hyperparameters θ, δ, and λ of our DID-Net to 4, 0.3, and 0.01, respectively. Our method is implemented with the Pytorch framework [61] and running on two NVIDIA TITAN RTX GPUs. The entire network is trained end-to-end on each benchmark.

C. Evaluation Metrics

Following most of the human parsing methods [1], [2], [5]–[15], we adopt mean intersection over union (mIoU) as the evaluation metric. During testing, we adopt the two-stage interaction process proposed by Gao *et al.* [17], and perform a certain number of corrections on each image to observe the change of mIoU. We adopt the "Average Click Number" ("Avg." for short) and the "Correction Click Number" ("Crt." for short) as evaluation metrics. "Avg." and "Crt." reflect the interaction cost to achieve a certain mIoU. Specifically, "Avg." is a semantic-level metric that computes the average click number on the cumulative category amount, and "Crt." is an image-level metric that records the correction efforts for each image. The less interaction cost of "Avg." and "Crt." to achieve a satisfied mIoU (80% for PASCAL-Person-Part, 85% for LIP, PASCAL-Person-Part and CIHP), the better the interaction efficiency. We compute the category precision of dilation points (dilation precision) in ablation experiments to observe its relationship to interaction efficiency. We also compute the F1-measure of the area within 5 pixels to part edges to observe how the dynamic transformation benefits the interaction efficiency. We adopt the model trained without the DD-Module / AIE-Block and only using the initial transformation as our baseline model.

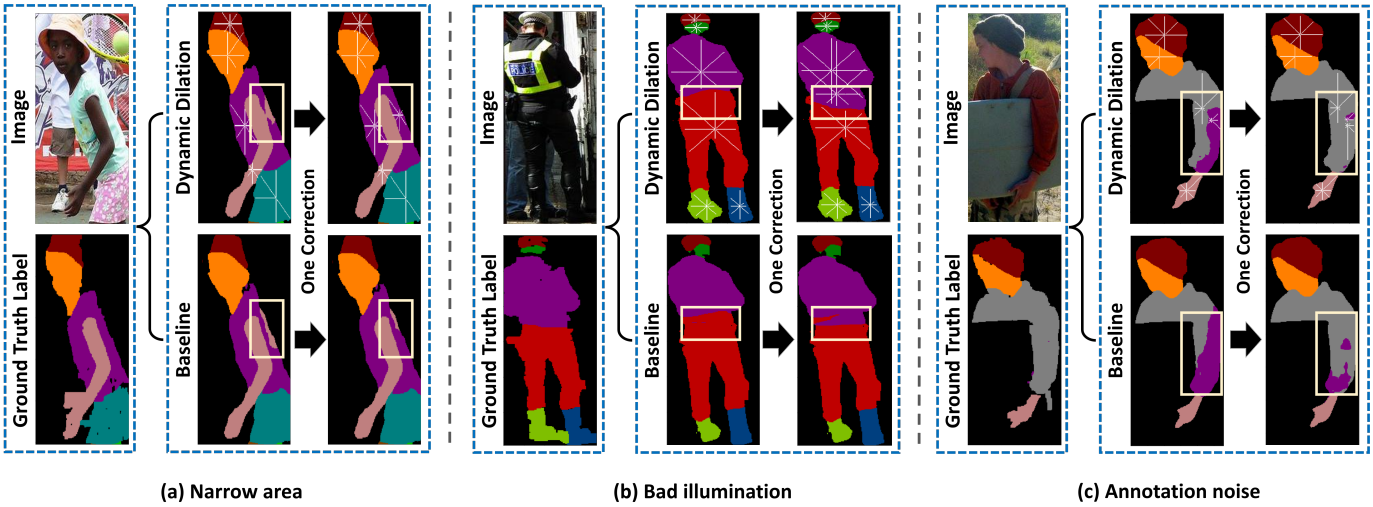


Fig. 4. Quantitative comparisons between the transformation with and without dynamic dilation on the LIP dataset. "Baseline" indicates the results generated by our baseline model. "Dynamic Dilation" indicates the results generated by our DID-Net. The white line is the dilation path, we generate a dilation point at the end of each dilation path. Invalid dilation points have no visible dilation path.

D. Ablation Study

We conduct various ablation experiments on LIP [6] and PASCAL-Person-Part [35] datasets to validate the effectiveness of our method.

Effects of Different Dilation Position. In the DD-Module, clicks are dilated under the constraints of the edge confidence map, one dilation point is generated in each direction. Assume a **dilation path** consist of the points between the click and where the dilation ends. We investigate the effects of different dilation positions, *i.e.*, where to set the dilation point, on the dilation path, We conduct three variants: "halfway position", "three-quarters position", and "extreme position", which set the dilation point halfway, three-quarters away, and farthest from the click on the dilation-path.

The quantitative comparisons can be seen in Fig. 3(a)(b). The variants of halfway position, three-quarters position, and extreme position improve the parsing mIoU by around 1% / 1.5%, 1.1% / 1.75%, and 1.6% / 2.2% on LIP / PASCAL-Person-Part dataset, respectively. The results show that the extreme position variant performs best and is more suitable for the dynamic transformation manner.

Interestingly, the extreme position variant has the lowest dilation precision while providing the best performance. Meanwhile, with correction clicks adding, the dilation precisions tend to decrease, while the interaction is still efficient. We can infer that the improvement of dilation points is more reflected in providing more reasonable context information among human parts, rather than accurately labeling the pixels. We adopt the extreme position in the rest of our paper.

Effects of Different Dilation Form. To explore the efficient dilation form, we conduct two variants named "point dilation" and "line dilation". Specifically, the point dilation variant only generates one dilation point on the dilation path while the line dilation variant generates a set of dilation points consisting of all the points on the dilation path.

As can be seen in Fig. 3(c)(d), the point dilation variant lead to a mIoU boost around 1.5% / 2.2% on LIP / PASCAL-Person-Part than the baseline, while the line dilation variant makes around 0.9% / 1.4% improvement. We can conclude that the point dilation is a better choice for the dilation form. This is primarily because most of the points on the dilation path are redundant for prediction, which causes the model to overfit with the redundant information during training and difficult to efficiently capture the long-range dependency within human parts. While the point dilation form can make the dilation points complementary to the interaction clicks. Hence we adopt the point dilation as our dilation form in the rest of our paper.

Effectiveness of Confidence Accumulation Dilation. To investigate the effects of the confidence accumulation dilation method, we here conduct a comparison with a threshold dilation method, *i.e.*, dilations stop once reaching the detected edge (edge confidence larger than 0.5). As shown in Fig. 3(e)(f), the accumulation dilation is more efficient, especially at the first several corrections. We analyze that the accumulation dilation tends to reduce inappropriate dilations, *i.e.*, over-dilation / under-dilation, when edge confidence quality is relatively low.

Effectiveness of the AIE-Block. To prove the effectiveness of the proposed AIE-Block, we conduct two variants by train with and without the AIE-Block on the basis of learning with the DD-Module.

The result is shown in Fig. 3(g)(h), the parsing mIoU boosted by the AIE-Block on the LIP / PASCAL-Person-Part is about 0.6% / 0.5%. Note that the 0.6% mIoU improvement is not trivial considering that it is very challenging to further improve the predictions produced by the interactive parsing model. This demonstrates that the adaptive feature recalibration in the AIE-Block can further benefit the interaction efficiency and improve the parsing quality effectively. The performance gains are similar on the two datasets which well prove that the AIE-Block is generally effective on different

benchmarks.

The AIE-Block *vs.* Global Feature Excitation. To compare the performance difference between the AIE-Block and the global feature excitation [57], we conduct a variant named "GFE" to apply the feature recalibration in Eq. 3 on the global features with the semantic agnostic setting instead of the click and dilation point features. As can be seen in Fig. 3(i)(j), the GFE variant performance is almost the same as using only the DD-Module, and the performance using the AIE-Block is ahead of GFE throughout the whole interaction process. The result indicates that our AIE-Block mainly benefits from the semantic representation of our dynamic transformation, demonstrating that our dynamic dilation transformation harbor the potential to better motivate the feature optimization .

Effects of the AIE-Block Follow Different Layers. To investigate the effects of the AIE-Block setting after different layers, we conduct three variants to set the AIE-Block after layer 2, 3, 4 of the ResNet101 backbone respectively. The results are shown in Fig. 3(k)(l), it can be seen that the AIE-Block added at the layer 4 shows better performance, indicating that our AIE-Block is more suitable for processing deeper features. We hence set the AIE-Block after the feature encoder in this paper.

Visual Illustration of Dynamic Dilation. In Fig.4, we visually illustrate how dynamic dilation based transformation differs from the baseline model. We set the clicks at similar positions for the two models. The comparison is based on three situations that could happen during the interaction process, which are clicking in a narrow area, dealing with bad illumination, and fixing annotation noise. From Fig.4, it can be seen that our DID-Net shows better efficiency in all three cases, and the inappropriate dilation points can be dynamically fixed by adding correction clicks.

E. Comparison with the State-of-the-Arts

We compare our DID-Net with the existing powerful standard interactive approaches, including IHP-DeepLab V3+ [21], IHP-DEXTR [22], ITIS [31], CA-learning [62] and CM [17]. When comparing with the ITIS, for each batch, we combine the CM and ITIS to randomly simulate 15 clicks in the central-click misprediction in loop-one, and get the final results in loop-two, and we compare the CA-learning based on the ITIS. For fair comparisons, we align the backbones of these approaches with our model, *i.e.*, DeepLab V3+. Here our model is trained with the optimal setting of the DD-Module and the AIE-Block on the three benchmarks, *i.e.*, LIP, PASCAL-Person-Part, and CIHP.

We conduct comparisons from two aspects. One is comparing the interaction cost, *i.e.*, the Crt. and the Avg., to achieve a certain mIoU, *i.e.*, 80% on PASCAL-Person-Part, 85% on LIP, CIHP, and PASCAL-Person-Part. The other one is comparing parsing performance with the same Avg..

As seen from Tab. I, the Avg. of our DID-Net to achieve the mIoU of 85% on LIP, PASCAL-Person-Part, CIHP, and 80% on PASCAL-Person-Part are 1.47, 3.34, 3.5, and 2.38 clicks, respectively. Compared with the powerful baseline CA-learning [62], our DID-Net saves correction efforts by 40% and 46.1% on LIP and PASCAL-Person-Part (85%) respectively. Our method easily achieves 85% mIoU on CIHP with the annotation effort of 8 Crt., which is hard for CA-learning.

As shown in Fig. 5, with the increase of Avg., our DID-Net performs the best among the state-of-the-arts. The parsing performance of our DID-Net outperforms the CA-learning [62] by a large margin on all three benchmarks. Specifically, the mIoU gains are around 1.1%, 2.6%, and 0.8% on LIP, PASCAL-Person-Part, and CIHP, respectively. Besides, it can be seen that our DID-Net trained with only the DD-Module is already outperforming state-of-the-arts, which proves the effectiveness and efficiency of our click transformation method for the IHP task. We put the comparison with the instance interactive segmentation separately in Fig. 6 to show the performance gap more clearly. These results demonstrate the superiority of our DID-Net over other interactive segmentation methods on interaction effectiveness for the IHP task.

F. Other Experimental Results

Testing Complexity. The computational complexity and memory usage of our proposed model is similar to other interactive segmentation models, *e.g.*, CM [17]. Specifically, 2,005 MB of GPU memory is consumed during testing, and the model inference time for each image is about 87 *ms*. Meanwhile, compared with standard human parsing models without interaction, *e.g.*, CE2P [11], our method can maintain the similar computational complexity and memory consumption.

Dynamic Dilation *vs.* Static Transformation. To prove the effectiveness of our dynamic transformation compare to the static transformation manner. We conduct the Euclidian distance map based transformation method [17], [21], [25], [28], [29] by setting the maximum encoding radius to 255 [21], 100, 16 [17], and 0 respectively. Plus the super-pixel level transformation method [24] named "Sup-p", forming five variants in total. The result can be seen in Fig. 7, the small-radius static transformations, *i.e.*, 0 and 16, exhibit better performance. The large-radius ones including the "Sup-p" are overloaded with noise due to the large overlap between its localization map channels, resulting in their performance approaching that of non-interactive semantic segmentation. In contrast, our dynamic transformation is ahead of the small-radius static transformation, which shows that our click-dilation transformation form possesses strong interactive-guidance capability to capture long-range semantic dependencies while avoiding confusion from overlapping transformation information.

The DD-Module *vs.* Edge Detector. To prove the effectiveness of the DD-Module is brought by the dynamic transformation rather than the edge perceiving loss in Eq. 5. We conduct a variant named "Baseline+Edge" by only train the model with the loss in Eq. 5 without the click dilation transformation. As can be seen in Tab. II, the Baseline+Edge variant perform similar to our baseline while the model trained with the DD-Module imporve a large performance gain. The results indicate that the mIoU benefit of the DD-Module is brought by the dynamic dilation transformation manner, and the role of edge perceiving loss is focused on assisting the click dynamic dilation.

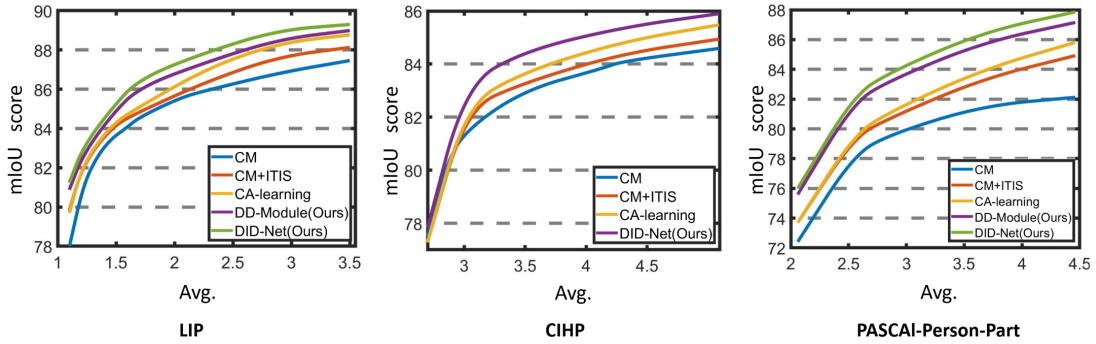


Fig. 5. Quantitative comparisons between DID-Net and other interactive models. "CM", "ITIS", and "CA-learning" represent the methods of [17], [31], and [62], respectively.

TABLE I
QUANTITATIVE COMPARISONS WITH STATE-OF-THE-ARTS METHODS ON THREE BENCHMARKS. BEST RESULTS ARE MARKED IN **BOLD**. "-" INDICATES THE MODEL IS HARD TO ACHIEVE THE PREDEFINED MIOU WITHIN 20 CRT.

	LIP (85%mIoU)		PASCAL-Person-Part (80%mIoU)		PASCAL-Person-Part (85%mIoU)		CIHP (85%mIoU)	
	Crt.	Avg.	Crt.	Avg.	Crt.	Avg.	Crt.	Avg.
CM(R) [17]	8	2.07	10	3.66	-	-	-	-
CM(N) [17]	9	2.19	15	4.46	-	-	-	-
CM(SPloss) [17]	7	1.95	6	3.02	-	-	-	-
CM+ITIS [31]	6	1.83	5	2.86	-	-	-	-
CA-learning [62]	5	1.71	4	2.70	13	4.14	-	-
DID-Net (Ours)	**3**	**1.47**	**2**	**2.38**	**8**	**3.34**	**8**	**3.50**

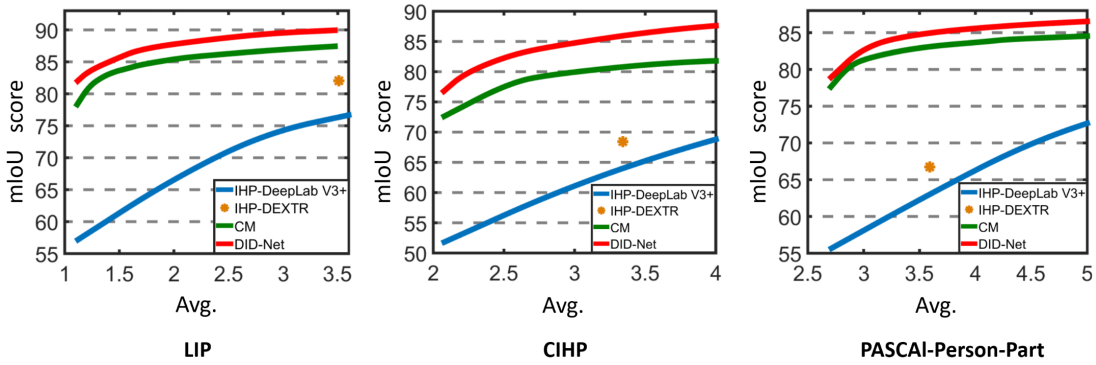


Fig. 6. Quantitative comparisons between DID-Net and other interactive models. "CM", "IHP-DeepLab V3+", and "IHP-DEXTR" represent the method of [17], [21], and [22], respectively. "DID-Net" represents our method.

TABLE II
MIOU PERFORMANCES OF DYNAMIC DILATION AND THE EDGE PERCEIVING LOSS ON LIP DATASET. "DILATION" REPRESENTS THE MODEL TRAINED WITH THE DD-MODULE. "BASELINE+EDGE" REPRESENTS THE MODEL TRAINED BY OUR LOSS FUNCTION EQ. 5 WITHOUT THE CLICK DILATION. "BASELINE" IS OUR BASELINE MODEL.

Crt.	0	5	10	15
Baseline	79.35	84.39	85.98	87.20
Baseline+Edge	79.28	84.23	86.28	87.33
Dilation	80.88	86.11	87.44	88.55

Effectiveness of Dilation Point Updating Process. During the evaluation, we dynamically update the dilation points in each round of the interaction process. To investigate the effectiveness of the dilation point updating process (DUP), we conduct comparisons between the variants with or without the DUP on LIP dataset. Here we respectively make comparisons based on the model trained with and without the AIE-Block. As can be seen in Tab. IV-F, using the DUP can boost the mIoU gains by around 0.5% on LIP without the AIE-Block, and can obtain 0.9% performance improvement when adopt the AIE-Block. It can be found that the performance improvement of the AIE-Block mainly comes from better dilation points, which further proves that the click-dependency information contained in the dynamically updated localization maps can be further exploited to improve interaction efficiency. The improvements brought by the DUP demonstrate the effectiveness of the dynamic transformation method. This value is not significant without the use of the AIE-Block, we analyze that the improvement obtained by the DD-Module comes mainly from the sparse transformation form and the click-dependency capturing. The few overlapping between channels

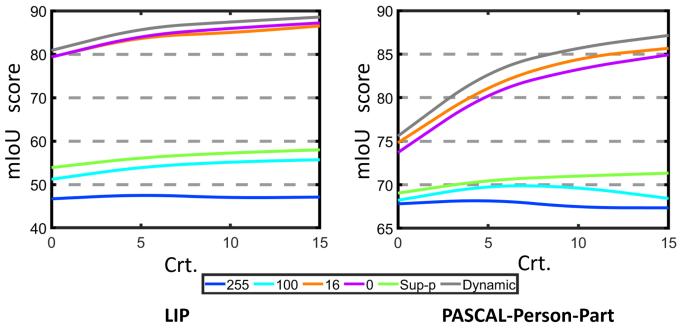


Fig. 7. Comparison results between various static transformation methods and the dynamic dilation method on LIP and PASCAL-Person-Part datasets. "255", "100", "16", and "0" represent the static Euclidian distance transformation with radius of 255, 100, 16, and 0, respectively. "Sup-p" represents the super-pixel based transformation. "Dynamic" represent our dynamic dilation transformation, *i.e.*, adopting the DD-Module.

TABLE III
QUANTITATIVE COMPARISONS FOR DUP ON LIP DATASET. "DUP" AND "N-DUP" INDICATES THE INTERACTION PROCESS WITH AND WITHOUT DYNAMIC DILATION POINT UPDATING RESPECTIVELY. "O" AND "AIE" INDICATES THE OUR MODEL WITH AND WITHOUT THE AIE-BLOCK RESPECTIVELY. "F1-MEASURE" INDICATES THE F1-MEASURE VALUE OF THE NEAR EDGE AREA (WITHIN 5 PIXELS TO EDGE). "D-PRECISION" INDICATES THE PRECISION OF THE DILATION POINTS.

Crt.	0	5	10	15	
N-DUP(O)	80.88	85.86	87.11	88.29	mIoU
DUP(O)	80.88	86.11	87.44	88.55	
N-DUP(AIE)	81.24	85.90	87.01	88.23	
DUP(AIE)	81.24	86.55	87.95	89.02	
N-DUP(O)	70.70	72.57	73.62	74.33	F1-measure
DUP(O)	70.70	73.12	74.56	75.66	
N-DUP(AIE)	71.61	72.27	73.22	74.36	
DUP(AIE)	71.61	74.22	75.08	76.19	
N-DUP(O)	78.82	76.16	73.1	69.92	D-precision
DUP(O)	78.82	77.41	75.48	73.12	
N-DUP(AIE)	79.57	76.38	72.72	70.17	
DUP(AIE)	79.57	78.02	76.28	73.98	

makes the noise in the localization map inherently low, which makes it difficult to further significantly reduce the noise in the localization map and obtain a greater improvement directly from the DUP.

To further investigate how the performance gains benefits from the DUP, we compute the F1-measure of the near edge area and the dilation point precision with different Crt. during the interaction process. As shown in Tab. IV-F, we can observe that the F1-measure is increasing with clicks added, which assists in the improvement of dilation precision. Thereby, the updated dilation points can guide the model to provide better parsing performance.

Interestingly, the performance shows that the interaction efficiency without DUP is still satisfactory, which performs better than the CA-learning [62] by around 0.7% mIoU.

In addition, we conclude the time-consuming of using the DUP, which slightly increase the average inference time about 16 *ms*, and does not bring much burden on the calculation.

Limitations. Our model is designed on the basis that the clicks provided by the user are correct, so our DID-Net has low robustness to the noise in the clicks, and the wrong labeling of the clicks will incorrectly guide the model to wrongly predict segmentation results. However, users can reduce the number of mislabeled clicks by getting familiar with the data semantics and reducing the impact of inappropriate clicks by revoking the mislabeled clicks or adding additional clicks. Also, how to improve the robustness of the model to click noise is a worthy research direction [19], [20]. In addition, our DID-Net is not applicable to most instance-level interactive segmentation methods. In the instance-level interactive segmentation task, due to the class-agnostic classification setting, the model produces only one object mask in each segmentation, which leads to the clicks is relatively sparse, hence the model needs relatively dense transformation information to better target object locations, and our sparse transformation manner is difficult to achieve excellent results based on this sparse click distribution. How to design a dynamic transformation strategy with dense transformation information is worth further investigation.

V. CONCLUSION

We propose a simple yet effective DID-Net for interactive human parsing as an initial attempt to transform clicks in a dynamic manner. A DD-Module is designed for capturing long-range dependencies of the clicks within each human part by dynamically dilating clicks in several directions. Then an AIE-Block is designed to further exploit potential semantic clues contained in the dilated clicks and enhance semantic expression by feature recalibration. Despite its simplicity, extensive experiments show the high interaction efficiency of our DID-Net across different datasets, demonstrating its superiority as an annotation tool for human image labeling.

VI. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under Grant (No.62072027), the National Key Research and Development Program of China (No.2022ZD0118502), the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funds (Grant No. A20H6b0151), and the Major Projects of National Natural Science Foundation of China (No.72293583).

REFERENCES

[1] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, and S. Yan, "Fashion parsing with video context," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 467–476.

[2] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1175–1186, 2016.

[3] H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei, "The elements of end-to-end deep face recognition: A survey of recent advances," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–42, 2022.

[4] D. Zeng, H. Liu, H. Lin, and S. Ge, "Talking face generation with expression-tailored generative adversarial network," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1716–1724.

[5] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 871–885, 2018.

[6] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1386–1394.

[7] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 770–785.

[8] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[9] L. Yang, Q. Song, Z. Wang, and M. Jiang, "Parsing r-cnn for instance-level human analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 364–373.

[10] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.

[11] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 4814–4821.

[12] T. Huang, Y. Xu, S. Bai, Y. Wang, and X. Bai, "Feature context learning for human parsing," *Science China Information Sciences*, vol. 62, no. 12, pp. 1–14, 2019.

[13] X. Zhang, Y. Chen, B. Zhu, J. Wang, and M. Tang, "Part-aware context network for human parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8971–8980.

[14] T. Li, Z. Liang, S. Zhao, J. Gong, and J. Shen, "Self-learning with rectification strategy for human parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9263–9272.

[15] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen, and L. Shao, "Hierarchical human parsing with typed part-relation reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8929–8939.

[16] D. Zeng, Y. Huang, Q. Bao, J. Zhang, C. Su, and W. Liu, "Neural architecture search for joint human parsing and pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 385–11 394.

[17] Y. Gao, L. Liang, C. Lang, S. Feng, Y. Li, and Y. Wei, "Clicking matters: Towards interactive human parsing," *IEEE Transactions on Multimedia*, 2022.

[18] S. Ge, C. Li, S. Zhao, and D. Zeng, "Occluded face recognition in the wild by identity-diversity inpainting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3387–3397, 2020.

[19] S. Li, T. Liu, J. Tan, D. Zeng, and S. Ge, "Trustable co-label learning from multiple noisy annotators," *IEEE Transactions on Multimedia*, 2021.

[20] S. Li, X. Xia, S. Ge, and T. Liu, "Selective-supervised contrastive learning with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 316–325.

[21] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang, "Deep interactive object selection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 373–381.

[22] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool, "Deep extreme cut: From extreme points to object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 616–625.

[23] Z. Wang, D. Acuna, H. Ling, A. Kar, and S. Fidler, "Object instance annotation with deep extreme level set evolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7500–7508.

[24] S. Majumder and A. Yao, "Content-aware multi-level guidance for interactive instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 602–11 611.

[25] W.-D. Jang and C.-S. Kim, "Interactive image segmentation via backpropagating refinement scheme," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5297–5306.

[26] X. Chen, Z. Zhao, F. Yu, Y. Zhang, and M. Duan, "Conditional diffusion for interactive segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7345–7354.

[27] C.-T. Lin, W.-C. Tu, C.-T. Liu, and S.-Y. Chien, "Interactive object segmentation with dynamic click transform," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2284–2288.

[28] K. Sofiiuk, I. Petrov, O. Barinova, and A. Konushin, "f-brs: Rethinking backpropagating refinement for interactive segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8623–8632.

[29] S. Zhang, J. H. Liew, Y. Wei, S. Wei, and Y. Zhao, "Interactive object segmentation with inside-outside guidance," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 234–12 244.

[30] Z. Lin, Z. Zhang, L.-Z. Chen, M.-M. Cheng, and S.-P. Lu, "Interactive image segmentation with first click attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 339–13 348.

[31] S. Mahadevan, P. Voigtlaender, and B. Leibe, "Iteratively trained interactive segmentation," *The British Machine Vision Conference*, 2018.

[32] T. Wang, Z. Ji, J. Yang, Q. Sun, and P. Fu, "Global manifold learning for interactive image segmentation," *IEEE Transactions on Multimedia*, vol. 23, pp. 3239–3249, 2020.

[33] K. Kim and S.-W. Jung, "Interactive image segmentation using semi-transparent wearable glasses," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 208–223, 2017.

[34] D.-J. Chen, J.-T. Chien, H.-T. Chen, and L.-W. Chang, "Tap and shoot segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[35] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1971–1978.

[36] C. Rother, V. Kolmogorov, and A. Blake, "" grabcut" interactive foreground extraction using iterated graph cuts," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[37] E. N. Mortensen and W. A. Barrett, "Intelligent scissors for image composition," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995, pp. 191–198.

[38] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[39] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[40] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, vol. 1. IEEE, 2001, pp. 105–112.

[41] V. Vezhnevets and V. Konouchine, "Growcut: Interactive multi-label nd image segmentation by cellular automata," in *proc. of Graphicon*, vol. 1, no. 4. Citeseer, 2005, pp. 150–156.

[42] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," *ACM Transactions on Graphics (ToG)*, vol. 23, no. 3, pp. 303–308, 2004.

[43] X. Bai and G. Sapiro, "Geodesic matting: A framework for fast interactive image and video segmentation and matting," *International journal of computer vision*, vol. 82, no. 2, pp. 113–132, 2009.

[44] A. Criminisi, T. Sharp, and A. Blake, "Geos: Geodesic image segmentation," in *European Conference on Computer Vision*. Springer, 2008, pp. 99–112.

[45] B. L. Price, B. Morse, and S. Cohen, "Geodesic graph cut for interactive image segmentation," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3161–3168.

[46] L. Grady, "Random walks for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006.

[47] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang., "Deep grabcut for object selection," *The British Machine Vision Conference*, 2017.

[48] J. Liew, Y. Wei, W. Xiong, S.-H. Ong, and J. Feng, "Regional interactive image segmentation networks," in *2017 IEEE international conference on computer vision (ICCV)*. IEEE Computer Society, 2017, pp. 2746–2754.

[49] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "Extreme clicking for efficient object annotation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4930–4939.

[50] H. Le, L. Mai, B. Price, S. Cohen, H. Jin, and F. Liu, "Interactive boundary prediction for object selection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 18–33.

[51] D. Acuna, H. Ling, A. Kar, and S. Fidler, "Efficient interactive annotation of segmentation datasets with polygon-rnn++," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 859–868.

[52] Z. Li, Q. Chen, and V. Koltun, "Interactive image segmentation with latent diversity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 577–585.

[53] M. Andriluka, J. R. Uijlings, and V. Ferrari, "Fluid annotation: a human-machine collaboration interface for full image annotation," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1957–1966.

[54] E. Agustsson, J. R. Uijlings, and V. Ferrari, "Interactive full image segmentation by considering all regions jointly," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 622–11 631.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[56] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[57] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[58] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[59] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[60] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[61] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[62] T. Kontogianni, M. Gygli, J. Uijlings, and V. Ferrari, "Continuous adaptation for interactive object segmentation by learning from corrections," in *European Conference on Computer Vision*. Springer, 2020, pp. 579–596.

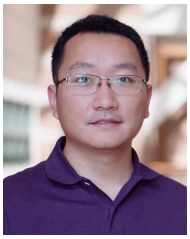
Yutong Gao received the MA. degree from National Taipei University of Technology, Taipei, China, in 2017. received the B.S. degree from Northeastern University, Liaoning, China. He is currently working toward the PhD degree in the School of Computer and Information Technology,Beijing Jiaotong University, Beijing, China. His current research interests include computer vision and machine learning. He is a student member of the IEEE.

Congyan Lang received the Ph.D. degree from the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, in 2006. She was a Visiting Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, from 2010 to 2011. From 2014 to 2015, she visited the Department of Computer Science, University of Rochester, Rochester, NY, USA, as a Visiting Researcher. She is currently a Professor with the School of Computer and Information Technology, Beijing Jiaotong University. Her current research interests include multimedia information retrieval and analysis, machine learning, and computer vision.

Fayao Liu Fayao Liu is a research scientist at Institute for Infocomm Research, A*STAR, Singapore. She received her PhD in computer science from University of Adelaide, Australia in Dec. 2015. She mainly works on machine learning and computer vision problems, with particular interests in self-supervised learning, few-shot learning and generative models.

Yuanzhouhan Cao Yuanzhouhan Cao is currently a lecturer at the School of Computer Science and Information Technology, Beijing Jiaotong University. He has obtained Master of Engineering and PhD of Computer Science from Sichuan University, China, and The University of Adelaide, Australia, in 2013 and 2017 respectively. He was a research scientist at Idiap Research Institute, Switzerland from 2017 to 2018, and a lecturer in the University of Adelaide from 2019 to 2020. His research interests are computer vision, machine learning and deep learning.

Lijuan Sun Lijuan Sun received the Ph.D. degree in the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, P.R. China, in 2021. She is currently a Postdoctor in the School of Economics and Management, Beijing University of Posts and Telecommunications. She is currently a key member of Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education. Her research interests include machine learning and computer vision.



Yunchao Wei is a Professor of Beijing Jiaotong University. He was a Senior Lecturer of Australian Artificial Intelligence Institute at the University of Technology Sydney from 2019 to 2021, a Postdoctoral Researcher in Beckman Institute at UIUC from 2017 to 2019, and a Postdoctoral Researcher of National University of Singapore from 2016 to 2017. He was selected as MIT TR35 China by MIT Technology Review in 2021, was named as one of the five top early-career researchers in Engineering and Computer Sciences in Australia by The Australian in 2020. He received the Discovery Early Career Researcher Award of the Australian Research Council in 2019, the 1st Prize in Science and Technology awarded by China Society of Image and Graphics (CSIG) in 2019.