Contour-Aware Equipotential Learning for Semantic Segmentation

Xu Yin, Dongbo Min, Yuchi Huo, Sung-Eui Yoon

***Abstract*—With increasing demands for high-quality semantic segmentation in the industry, hard-distinguishing semantic boundaries have posed a significant threat to existing solutions. Inspired by real-life experience, i.e., combining varied observations contributes to higher visual recognition confidence, we present the equipotential learning (EPL) method. This novel module transfers the predicted/ground-truth semantic labels to a self-defined potential domain to learn and infer decision boundaries along customized directions. The conversion to the potential domain is implemented via a lightweight differentiable anisotropic convolution without incurring any parameter overhead. Besides, the designed two loss functions, the point loss and the equipotential line loss implement anisotropic field regression and category-level contour learning, respectively, enhancing prediction consistencies in the inter/intra-class boundary areas. More importantly, EPL is agnostic to network architectures, and thus it can be plugged into most existing segmentation models. This paper is the first attempt to address the boundary segmentation problem with field regression and contour learning. Meaningful performance improvements on Pascal Voc 2012 and Cityscapes demonstrate that the proposed EPL module can benefit the off-the-shelf fully convolutional network models when recognizing semantic boundary areas. Besides, intensive comparisons and analysis show the favorable merits of EPL for distinguishing semantically-similar and irregular-shaped categories.**

***Index Terms*—Supervised Semantic Segmentation, Category-level contour learning, Semantic boundary refinement.**

I. Introduction

EXISTING deep semantic segmentation approaches [3], [22], [28], [29], [43] are usually trained with the cross-entropy (CE) loss for multi-class classification. In the training phase, this loss measures the mismatch between the areas determined by the probability estimation from the neural network and areas defined by the ground-truth semantic label. However, the existing benchmarks' inherent drawbacks may prevent CE from achieving better performance, especially in semantic boundary areas.

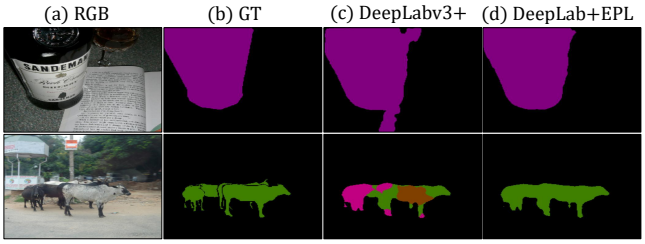


Fig. 1: Demonstration of the semantic boundary problem. (c) shows the prediction from DeepLabv3+ [7], and (d) shows the output when combined with EPL module. Both rows respectively demonstrate the segmentation difficulties in inter-class and intra-class boundary areas.

The first disadvantage relates to the label noise. It is studied [1], [35] that the misaligned ground-truth annotations with the real object edges [1], [35], [38] would mislead CE and results in low segmentation accuracy on the boundary regions. At the same time, the inherent inductive bias [10], [26] of convolutional neural networks (CNNs) is also a big barrier for CE to learn a clear semantic boundary.

In this work, we propose a novel method to help CE handle the boundary segmentation problem. We mainly identify the "semantic boundary area" into two types (shown in Fig. 1): the inter-class and intra-class. The inter-class case refers to transition areas between different categories; these categories (on the first row) either have similar visual characteristics (patterns/textures) or have strong semantic relationships. The intra-class case (on the second row) areas are often seen when segmenting multiple instances of the same category in a small area, particularly on objects with complicated contours.

To this end, we present a novel framework to address the semantic boundary segmentation problem. Our central idea comes from two aspects:

- People get a good visual understanding of objects in real life by changing the relative observation distance or varying the direction/perspectives (in Fig. 2 (a)). We conclude that better visual expression combines observations from various positions and perspectives. To this end, we propose a novel operator (anisotropic convolution) to expand the semantic labels and a loss function to refine the segmentation estimation in different directions.
- Objects (e.g., animals in Fig. 1) with similar geometric appearances are usually classified as the same categories.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C3002833), (No. 2021R1A4A1032582), and in part by Zhejiang Lab (121005-PI2101) (Corresponding authors: Yuchi Huo; Sung-eui Yoon.)

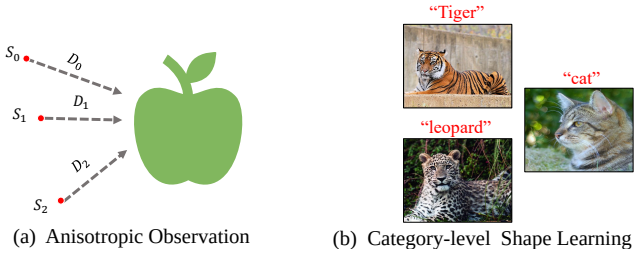
Xu Yin is with the School of Computing, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea (E-mail: yinofsgvr@kaist.ac.kr).

Dongbo Min is the Faculty of the Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, South Korea (E-mail: dbmin@ewha.ac.kr).

Yuchi Huo is with the State Key Lab of CAD and CG, Zhejiang University, China and Zhejiang Lab, China 310058 (E-mail: huo.yuchi.sc@gmail.com).

Sung-eui Yoon is with the Faculty of School of Computing, Korea Advanced Institute of Science and Technology, Deajeon 34141, South Korea (E-mail: sungeui@gmail.com).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes the source code and more experiment results. This material is 2066 kb in size.



(a) Anisotropic Observation (b) Category-level Shape Learning

Fig. 2: (a) This work takes inspiration from the daily visual observation, i.e., changing the relative distance D and the perspective S contributes the better object-contour recognition. (b) The category-level contour information is an important cue for image classification. For example, people would categorize tigers and leopards into cat species by external contours, even though they have different textures.

In real life, people keep those characteristics in mind and use them as empirical evidence to recognize new species (in Fig. 2 (b)). In this work, we suggest learning category-level contours to achieve this effect.

Specifically, we propose the equipotential learning (EPL) module, which refines the segmentation predictions in a new domain, termed the potential domain. Following the deep segmentation tradition [3], [22], [28], we define the semantic segmentation task as a pixel-wise classification problem in the conventional probability domain and then propose to convert the probability field into the potential field with anisotropic convolution to obtain visual observation from multiple perspectives. We implement two loss functions for refining the segmentation prediction in the potential field, the point loss and equipotential line loss, respectively performing anisotropic regression and category-level contour learning.

In summary, our contributions are as follows:

- We design an anisotropic convolution, a novel operator that converts the deep semantic segmentation problem to the self-defined potential domain, aiming to optimize neural network predictions from different directions. To the best of our knowledge, it is the first time to use this idea to deal with the boundary segmentation problem.
- In the potential domain, we build a point loss, requiring the segmentation predictions to fit the real image content anisotropically, enforcing the consistency between predicted pixels and their nearby decision boundaries.
- We present an equipotential line loss to learn each category's contour. This loss specifically learns the edge regions and optimizes the corresponding predictions for better boundary segmentation performances.
- Experimental results on Pascal Voc 2012 and Cityscapes demonstrate that our method can help current segmentation solutions [16], [31], [41], [42] refine their predictions on the semantic boundary areas.

II. Related Work

FCN methods for semantic segmentation. FCN refers [22] to networks adopting the convolutional layer throughout the architecture. With more similar methods [3], [6], [28] proposed, FCN has become the standard practice in the image segmentation community. Also, methods like conditional random field (CRF) [6] and point-based sampling [13], [18], [23] are put to enforce FCN models' segmentation ability on decision boundaries. However, building specified operators brings extra computational costs. Also, these studies have relatively weak performance in learning category-level characteristics. In this work, EPL module maps each category to an independent channel, learning category-level characteristics without increasing the parameter size.

Distance field regression. The distance field (DF) is well-known in the computer vision and graphics community. The value of a point in the field is defined as the distance to the nearest boundary, enabling high-quality feature representation. Audebert *et al.* [2] design a multi-task model for FCN, which requires feature maps to fit a DF, apart from estimating probabilities. Recently, Xue *et al.* [34] suggested networks to fit the signed distance field (SDF) directly and then use a smoothed Heaviside function to turn the distance prediction into probabilistic predictions. With incorporated information from DF, one could effectively regulate the layout of segmentation results. However, the primary issue of field-based studies [24], [32]–[34] is that DF itself does not carry any category identity information, which may mislead the neural network when learning multiple categories jointly. We borrow the concept of "distance field" but map category-level image contents to independent spaces and then learn the contour information in the potential domain to address this issue.

Boundary supervision: Many loss functions are specifically designed for calculating boundary loss. For instance, Discriminative Feature Network (DFN) [36] employs an edge detection step for feature maps and tries to match the edge map [20] by a sigmoid loss. In the medical image segmentation community, Dice loss [30] is widely used to solve the class-imbalance problem in the boundary region. Another loss proposed in [17] re-calculates the distance field's metric in an integral regional way and achieves great success over the binary organ segmentation task. Our work uses the equipotential lines in the potential domain as the boundary supervision and learns all categories' contours with the proposed line loss.

III. Methodology

This section first introduces the anisotropic convolution, an operator that convolves the semantic segmentation problem from the probability domain to the potential domain (Probability $\rightarrow$ Potential). Secondly, we elaborate on fitting anisotropic observation and performing the category-level contour learning in the potential domain. Finally, we plug EPL into FCN models to improve their boundary segmentation performance.

A. Preliminaries

Before going deep into the anisotropic convolution, we clarify two important concepts used throughout the paper.

- **Field.** We use "field" to represent the basic unit for domain conversion. The predicted probability fields refer

to probability estimations from FCN, and the ground-truth probability fields are the one-hot encoding result of label annotations. Similarly, we name the conversion results in the potential domain, the predicted/ground-truth potential fields.

- **Potential energy.** We define the potential energy as the pixels' numerical value in potential fields. The anisotropic convolution converts the pixel-level probability estimations to potential energies in different directions by performing domain conversion in the training phase.

B. Anisotropic Convolution for Domain Conversion

To implement the domain conversion, we introduce the anisotropic convolution (AC), a differentiable convolutional operator that proceeds in specific directions. Using probability fields as the input, AC extends the image content to get its anisotropic semantic extensions.

A general AC operator consists of a filtering kernel W and anisotropic splitter S, corresponding to the variables of "relative distance" and "perspective" in the visual observation process. We let X and Y ($Y \in [0, 1]$) stand for input images and their ground-truth probability fields. In the supervised K-class semantic segmentation task, we train network f with the parameter θ. $\hat{Y} = \{\hat{y_1}, \hat{y_2}, ...\hat{y_K}\}$ denotes the category-level probability estimation, where:

$$\hat{Y} = f_\theta(X), \tag{1}$$

For any point $\hat{y}^p \in \hat{Y}$ ($p \in P$ is the spatial coordinate set of Y and $\hat{Y}$), we get its energy $E(\hat{y}^p)$ in the potential domain by performing the conversion on its $w \times w$ neighborhood space $V^{\hat{p}}$. Formally, we express this process as:

$$E(\hat{y}^p) = V^{\hat{p}} * (W \circ S), \tag{2}$$

where $V^{\hat{p}}$ has the same kernel size as W, and $*$ and $\circ$ denote the convolution and Hadamard product, respectively. Also, we apply the same conversion on all possible $y^p \in Y$ to get the ground-truth $E(y^p)$.

In real life, we usually adjust the perspective to get different views of an object since anisotropic observations help us better understand the object. In our case, the changeable perspective is realized by letting the splitter contain different direction vectors. For instance, the splitter S in Fig. 3 consists of four elements ($S = \{s_1, s_2, s_3, s_4\}$), denoting the directions of up, down, left, and right, respectively. In the experiment section, we test three splitters A, B and C (shown in the bottom of Fig. 3) to explore the effect of S in semantic segmentation, containing 4, 4, and 8 directions, respectively. To mitigate the training difficulty and reduce the computational overhead, we set the filtering kernel W with a box pattern [27] and maintain both weights of W and S unchanged in the training phase. AC does not increase the parameter budget of θ in the full conversion process.

Fig. 3 presents a domain conversion example, where a 7×7 probability field $E(\hat{Y})$ is converted to four potential fields in different directions using 5×5 AC operator. We think of each potential field as an observation for the input semantic part (shown in orange) in a direction.

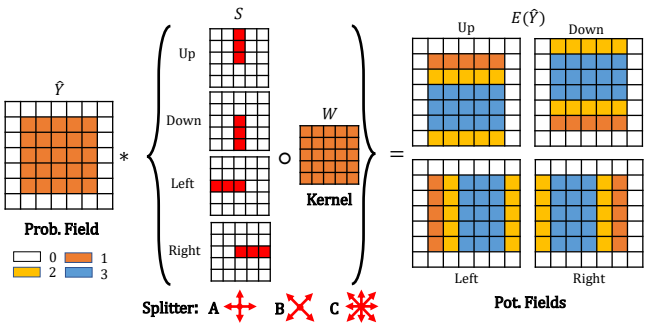


Fig. 3: Example of the domain conversion with 5×5 anisotropic convolution (AC). Here, AC includes four directions (referring to splitter A), which expands the content contained in the probability field (**Prob. Field**) in specified directions. The potential energy of the resulting potential fields (**Pot. Fields**) ranges from 0 to 3, and we label their distribution with different colors. In experiments, we test all three splitters (A, B, and C) in ablation studies to see their effectiveness.

Generally, domain conversion has two benefits:

- **Visual:** In Fig. 3, we observe that the category-level semantic is extended in preset directions. We think of each potential field as a view in a specific direction. In the training phase, after converting the ground-truth probability fields to potential fields, one can get the anisotropic observations of the real image content after integrating the context information in all potential fields.
- **Physical:** In standard deep segmentation practices, the neural network can only get the supervision information from the semantic labels. AC spreads the image context anisotropically and maps the sparse contour label to a dense annotation in the potential fields. $E(Y)$ provides more comprehensive and stronger supervision for input X than Y, especially in the boundary areas. Besides, domain conversion enables the network f to explore a broader solution space since we can simultaneously optimize f in the probability and potential domains.

Besides, AC operator enables users flexibly to change the "observation distance" and "perspective" variables by adjusting W and S, based on the property of object instances (refer to Sec. IV for details).

C. Loss functions

This section presents the point loss and the equipotential line loss that enforce the optimization in the potential domain. The first term implements the anisotropic field regression throughout the potential domain, while the second loss aims to precisely learn each category's contour.

Point Loss for anisotropic field regression: We propose a point loss to fit the ground-truth potential fields $E(Y)$ in all directions to learn the contextual information in the potential domain. This loss regresses the fields in the potential domain globally; therefore, the information learned from the potential domain would correct prediction errors in the probability domain.

Also, the potential energy integrates information from the neighborhood space (of the same size as W) for better optimization when the loss function regresses points in the semantic boundary. In other words, the computed gradient in the potential domain on y^p will push the main segmentation network to refine y^p's involved neighborhood predictions in the probability domain through backpropagation, enhancing the prediction consistency among pixels.

Formally, we compute the point loss L_{point} by averaging the error in each direction $s \in S$ as:

$$L_{point} = \frac{1}{|S|} \sum_{s \in S}^{|S|} \sum_{p \in P} ||E_s(y^p) - E_s(\hat{y}^p)||, \tag{3}$$

In the experiment section, we set the point loss with the $L1$ and $L2$ norm and then test their effectiveness.

In conclusion, the potential fields denote the anisotropic observations for the image content; when L_{point} regresses the prediction to fit the ground-truth potential fields, the segmentation network will look for an anisotropically-stable state and therefore achieve a globally-semantic balance.

Equipotential Line Loss for category-level contour learning: Although L_{point} helps improve the prediction consistency, it also brings a risk of blur boundaries because of the predicted in-between values, i.e., close to 0.5, in the potential fields (see the blur part in Fig. 6). It would result in an intra-class indistinction problem in the category channels. Besides, the point loss designed for global regression lacks the effects of contour learning, thus failing to fully utilize the semantic boundary information in the potential domain. To solve this issue and enable contour learning, we specifically present an equipotential line loss to strengthen the optimization in the object boundary-related region.

In the ground-truth potential fields converted by the $w \times w$ AC, we define the equipotential line as a set of points having an equal energy value in the range of $[1, \lfloor \frac{w}{2} \rfloor]$. After domain conversion, the yielding ground-truth $E(Y)$ always follows a discrete distribution in $[0, \frac{w}{2}]$. By contrast, the predicted $E(\hat{Y})$'s distribution has a continuous shape since it takes the probability estimation $\hat{y}$ as input. Once matched with the ground-truth equipotential lines, the network predicts concrete category-level contours. Therefore, we suggest $E(\hat{Y})$ to specifically learn the equipotential lines that carry affluent contour information.

Similarly, we provide an example in Fig. 4 , showing how to learn a single category's contour with L_{line}. With the 7×7 AC operating in direction s, we get three equipotential lines (marked with different colors). As observed in Fig. 4, all equipotential lines are closely located around the dog's real edge and can be used to depict its contour. We generalize this example to the general $w \times w$ AC, quantify the line loss between the predicted and ground truth equipotential lines in all directions, and learn the category-level contours.

Loss formulation: Before elaborating on L_{line}'s formulation, we need to instantiate the ground-truth/predicted equipotential line regions (denoted with L and $\hat{L}$) in $E(Y)$ and $E(\hat{Y})$.

Formally, we index $E(Y)$ in the ascending order and composite L by iteratively assigning $E(Y)$'s points to equipotential lines according to their energy values. When generalizing to the category-level case, for each category $i \in K$, its specified equipotential line region $l_{i,s}$ in direction $s \in S$ is represented as: $l_{i,s} = \{l^1_{i,s}, ..., l^{\lfloor \frac{w}{2} \rfloor}_{i,s}\}$; the superscript denotes the energy.

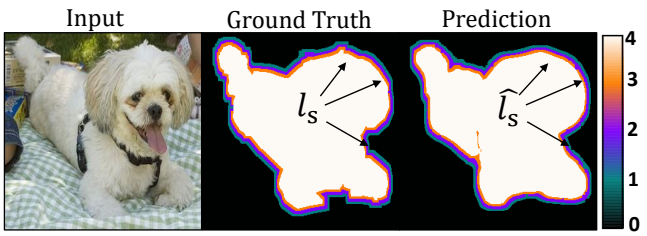


Fig. 4: Example of the equipotential line loss for category-level contour learning with 7×7 AC. We take the "dog" category as the example and visualize its ground-truth and predicted equipotential lines (in the middle and right figure) in direction s, termed l_s and $\hat{l_s}$, and mark the energy distribution with the colors in the bar. To learn the dog's contour, L_{line} optimizes the mismatch region between l_s and $\hat{l_s}$, in any $s \in S$, in the range $[1, \lfloor \frac{w}{2} \rfloor]$ ($\lfloor \frac{w}{2} \rfloor = 3$ in this example).

In order to fit the continuous prediction with the discrete ground truth, we assume that each line in $\hat{l}_{i,s}$ is composed of the same number of points as its ground truth in $l_{i,s}$. Therefore, we reorder and index $E(\hat{Y})$ to let $\hat{l}_{i,s}$ be an equal-count counterpart of $l_{i,s}$. It means that for any integer $\mu \in [1, \lfloor \frac{w}{2} \rfloor]$, we have:

$$|l^{\mu}_{i,s}| = |\hat{l}^{\mu}_{i,s}|, \tag{4}$$

where $|\cdot|$ denotes the number of points in the individual line.

To formulate the loss, we get inspiration from dice loss [30] and optimize the equipotential line loss L_{line} in all directions by enlarging the intersection between each $\langle l_{i,s}, \hat{l}_{i,s} \rangle$ pair.

Algorithm 1 presents the implementation details of L_{line}. To compute the intersection part, we specifically apply an exponential activation with factor μ to punish large mismatches and then measure the overlapped area between $l_{i,s}$ and $\hat{l}_{i,s}$. Additionally, we use a constant value C to normalize the equipotential dice coefficient (EDC) within $[0, 1]$ as dice loss. It is clear to see that L_{line} decreases when the corresponding lines in $l_{i,s}$ and $\hat{l}_{i,s}$ match with each other.

Intuitively, the equipotential line loss achieves the goal of contour learning by enforcing a strong geometric constraint on each category's edge area. L_{line} punishes the predicted in-between values in $E(\hat{Y})$ and optimizes the predictions in semantic boundary areas from the distribution perspective. Compared with L_{point}, L_{line} concentrates on optimizing the edge part determined by W, making the segmentation network better contour-representation ability. Also, L_{line} integrates the line-level misalignment information of $\langle L, \hat{L} \rangle$ in different directions and, therefore, can more accurately localize the mismatch boundary, even though it happens in a small region. These two characteristics effectively help address the intra-class problem. Besides, our L_{line} is more concise than other semantic boundary refinement studies [17], [34] and introduces no additional parameters in the training or inference phase.

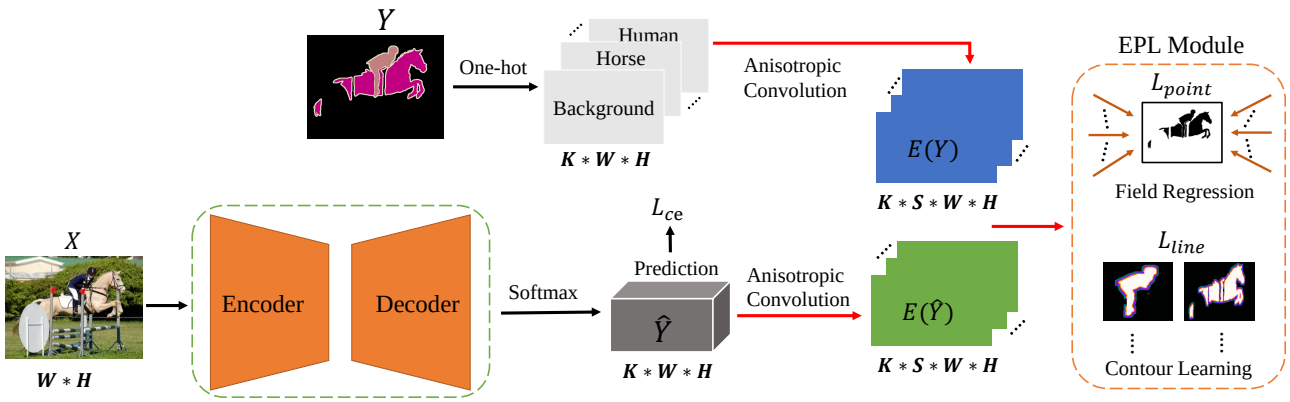


Fig. 5: Assemble FCN model with EPL for K-class semantic segmentation. After proceeding with the anisotropic convolution, along any direction in S, the point loss L_{point} and the equipotential line loss L_{line} respectively enable the anisotropic field regression and the category-level contour learning.

Algorithm 1 Equipotential Line Loss

Input: the ground-truth/predicted line region $L, \hat{L}$, and the normalization factor μ.

Output: L_{line}

Initialized L_{line} = 0;
for each category i in K **do**
  for each direction s in S **do**
    $l_{i,s}, \hat{l}_{i,s} \leftarrow L[i][s]$, $\hat{L}[i][s]$; ▷ Category-level unit
    for $\tau \leftarrow 1$ to $\lfloor \frac{w}{2} \rfloor$ **do**
      $d_{i,s} \leftarrow e^{-(l_{i,s}-\tau)^{\mu}}$; ▷ Exponential activation
      $\hat{d}_{i,s} \leftarrow e^{-(\hat{l}_{i,s}-\tau)^{\mu}}$;
      Represent the intersection area:
        $IoU_{i,s} \leftarrow ||d_{i,s} \cdot \hat{d}_{i,s}||_1$;
      Measure the mismatch area:
        $C \leftarrow \frac{||d_{i,s}||_1}{||d_{i,s} \cdot \hat{d}_{i,s}||_1}$;
        $EDC^{\mu}_{i,s} \leftarrow \frac{2 \cdot C \cdot IoU_{i,s}}{||d_{i,s}||_1 + ||\hat{d}_{i,s}||_1}$ ▷ Coefficient
        $L_{line} \leftarrow L_{line} + (1 - EDC^{i}_{s})$
  end for
end for
return $L_{line}/|S|$

D. Applications of EPL in FCN

The full EPL module can be assembled to most FCN models (as illustrated in Fig. 5) to achieve better segmentation result. In the training phase, we follow the common practice to compute the cross-entropy loss (termed L_{ce}) in the probability domain. Next, we use AC to convert Y and $\hat{Y}$ to the potential domain, where L_{point} and L_{line} are then employed for further optimization. Empirically, we use λ_1 and λ_2 as balance weights and express the final loss term as:

$$Loss = L_{ce} + \lambda_1 L_{point} + \lambda_2 L_{line}. \quad (5)$$

In the inference stage, EPL is discarded and incurs no extra computational overhead.

IV. Evaluation

In this section, we perform two sets of experiments. In the ablation study, we add EPL module on PSPNet [41] to reveal the effectiveness of L_{point} and L_{line}. Besides, we discuss the impact of anisotropic evolution (AC) when adopting different splitters. Moreover, we compare two loss functions with related studies to qualitatively evaluate their efficiency. Finally, we report the overall performance gains achieved with EPL on other baseline models.

A. Setup

- **Datasets:** All experiments are conducted on two segmentation benchmarks: Pascal Voc 2012 and Cityscapes. The former dataset includes 21 classes, with 10,582 and 1,449 images for training and validation, while the latter contains 19 categories, including 2,979 training (fine-annotated) and 500 validation images.
- **Baseline Models:** We deploy EPL module on five FCN models: PSPNet [41], PSANet [42], DeepLab v3+ [7], CCNet [16] and GSCNN [31]. We adopt the reliable PyTorch implementations [1][2][3] to reproduce the baselines and achieve strong performances. The ablation experiments are conducted on PSPNet with resnet50 [14] while other baselines' performances are reported in the main result.
- **Data augmentation and experimental details:** We strictly follow the experiment settings in the adopted implementations without changing any other hyperparameters except the loss weight λ_1, λ_2, and μ. The involved data augmentation operations include random scale (in $[0.5, 2.0]$), random rotation (degree within $[-10, 10]$), Gaussian blur, horizontal flipping, and random crop. In the ablation part, we conduct all experiments on the small-size images with crop sizes of 256×256 and 256×512 on Pascal Voc 2012 [12] and Cityscapes [11] (batch=12). As for practical experiments, we train all

[1]https://github.com/hszhao/semseg
[2]https://github.com/NVIDIA/semantic-segmentation
[3]https://github.com/jfzhang95/pytorch-deeplab-xception

Dataset	Method	Kernel	mIoU(%)
Pascal Voc	Baseline	-	73.52
	+Point^1	7	74.07
		9	73.85
		11	**74.08**$_{(+0.56)}$
	+Point^2	7	74.09
		9	**74.42**$_{(+0.90)}$
		11	74.27
Cityscapes	Baseline	-	71.66
	+Point^1	9	70.44
		11	70.96
		13	**72.71**$_{(+1.05)}$
	+Point^2	9	**72.39**$_{(+0.73)}$
		11	70.93
		13	71.05

TABLE I: **Point loss** performance on Pascal Voc 2012 and Cityscapes validation set. The superscript of "+Point" denotes the norm type.

models in a batch of 16 with reported image size in Table IX and VIII. Note that all experiments are conducted on 4×NVIDIA RTX Titan.

- **Evaluation protocol:** All segmentation performances are evaluated on the validation set of benchmarks. We present qualitative evaluations on each component of EPL and reveal its effect in multiple-scale inference results, in the range of [0.5, 0.75, 1.0, 1.25, 1.5, 1.75]. Except **for** the mean Intersection-over-union (mIoU), we specially employ **F-Measure** [23], [25] and **Trimap IoU** [6], [9] to quantize the models' segmentation performance in semantic boundary areas.

B. Ablation Study

We use **Point** and **Line** to denote the point loss and the equipotential line loss, respectively. In this part, we mainly apply the splitter A (shown in Fig. 3) and set it with different kernel sizes to evaluate the effectiveness of both loss functions. To make a fair comparison, we empirically set μ (Eq. 4), λ_1, λ_2 (Eq. 5), as 10, 0.1, and 0.01, respectively. Discussions of other splitters (B and C) and choices of hyperparameters are reported in Appendix B.

Ablation for the Point loss. We add L_{point} to the baseline network to regress the potential fields. To verify its effect, we respectively set L_{point} (Eq. 3) of $L1$ and $L2$ norm (marked with the superscript) and report their results in Table I. One can see that we achieve considerable improvements over both datasets. We do not observe obvious performance differences between the two norms and therefore set L_{point} of $L2$ norm in the later experiments.

In Fig. 6, we see that the prediction of the dog is considerably refined by L_{point}. However, after visualizing one potential field (the second row), we still observe blurs between the paws, indicating the in-between predictions.

Ablation for the equipotential line loss. Table II reports the performance of L_{line} on both datasets. After learning the category-specific contours, we observe that PSPNet achieves up to 0.63%/1.60% mIoU increasements on Pascal Voc 2012/Cityscapes. Similarly, we visualize the contour learning

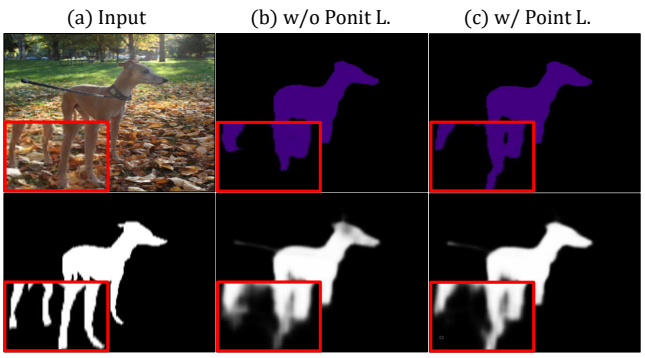


Fig. 6: Example of the point loss on Pascal Voc 2012. In the second row, we visualize the potential field.

Dataset	Method	Kernel	mIoU(%)
Pascal Voc	Baseline	-	73.52
	+Line	7	**74.15**$_{(+0.63)}$
		9	73.47
		11	74.05
Cityscapes	Baseline	-	71.66
	+Line	9	71.78
		11	**73.26**$_{(+1.60)}$
		13	72.11

TABLE II: **The equipotential line loss** performance on Pascal Voc 2012 and Cityscapes validation set.

effects in Fig. 8 and observe that L_{line} learned the subtle and complicated shape features of the bicycle category.

Boundary Segmentation evaluation: To furtherly verify the learning effect of both losses, we introduce two boundary-quality measures, F-measure [25], and Trimap [6], [23], to evaluate segmentation results in the semantic boundary area. Both metrics measure the matching level between the prediction and the ground truth in a narrow band region from the ground-truth semantic boundary given a pixel width. We evaluate the segmentation results multiple times with different pixel widths and present the comparison with/without employing L_{line} in Fig. 7; one can see that both L_{point} and L_{line} improve the segmentation capability in the boundary areas. Besides, we see that L_{line} achieves more performance enhancements than L_{point} on the areas near the edge (pixel width < 10), indicating L_{line} 's effectiveness on contour learning.

Effects of anisotropic convolution: In this part, we ablate the AC operator with the standard convolution (**SC**) to verify its efficiency. Empirically, SC destroys the ground-truth image content and distorts its probability distributions. In Table III, we replace AC with SC and optimize the potential fields with L_{point} and L_{line} similarly. We observe that SC degrades the baseline performance on both datasets, confirming the advantages of AC and the feasibility of our "anisotropic observation" assumption.

Effects of directional splitters: The choice of the splitters has a crucial influence on EPL. We test the other two splitters B and C (reported in Appendix, Table XII & XIII) and observe that splitter A performs the best among all three candidates and can achieve consistent improvements on both datasets, indicating that the best semantic observation comes from the

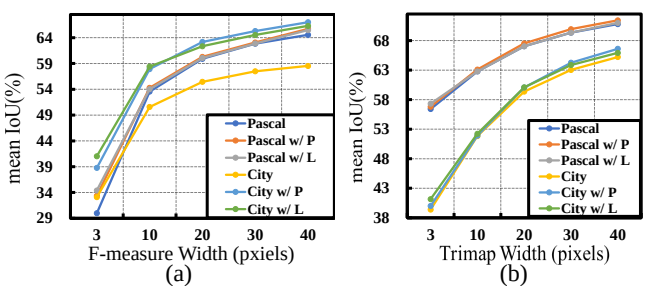


Fig. 7: (a)(b) plot how F-measure and Trimap mIoU change with the pixel width of the ground-truth boundary region on Pascal Voc 2012 (**Pascal**) and Cityscapes (**City**) before and after employing L_{line} (**w/L.**) and (**w/P.**).

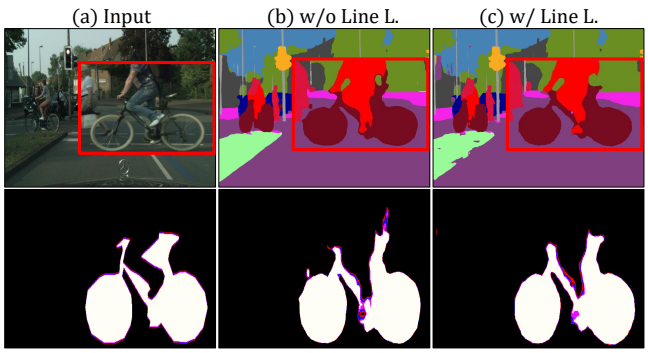


Fig. 8: Example results of the equipotential line loss on Cityscapes. In the second row, we compare the visualized potential fields of the red cropped bicycle area.

integrated result in the up, down, right, and left directions.

Effects of Kernel size: In Table I & II, we see that both losses' performances do not necessarily increase with AC's kernel size. We see W's size as the range of semantic area that aims to optimize. In ablation experiments (Table I, II, XII and XIII), we tested multiple kernels and found that the large ones do not work well with small object instances because the decision boundary area becomes negligible in these cases. Therefore, we use the optimal range of kernel sizes found in the ablation part, then scale AC in proportion to fit other network output resolutions in later experiments.

Effects of the exponential activation: To implement L_{line}, we apply an exponential action factor μ to measure the overlapped part between the ground-truth L and predicted $\hat{L}$ (see line 6, 7 in Algorithm 1) where μ must be even. Here, we test the effect of μ with five even values $2, 4, 10, 16, 20$, and then experiment L_{line} ($\lambda_1 = 0, \lambda_2 = 0.2$) with PSPNet on two datasets. Table IV presents PSPNet's performance when setting μ with different values. We observe that the best segmentation performances on Pascal Voc 2012 and Cityscapes are achieved when $\mu = 10, 2$. Therefore, we apply both values in later experiments.

C. Comparison with related works

The section compares L_{point} and L_{line} with their related works. Note that none of the compared losses has been adapted to the multi-class segmentation problem in previous studies.

Method	mIoU(P.%)	mIoU(C.%)
PSPNet(baseline)	73.52	71.66
PSPNet+SC+L_{point}	71.08(-2.44)	68.37(-3.29)
PSPNet+SC+L_{line}	72.17(-1.35)	70.93(-0.73)
PSPNet+AC+L_{point}	**74.84**(+1.32)	**72.39**(+0.73)
PSPNet+AC+L_{line}	**74.71**(+1.19)	**73.26**(+1.60)

TABLE III: Compare results of AC with standard convolution (**SC**) on Pascal Voc 2012 (**P.**) and Cityscapes (**C.**).

μ	mIoU(P.%)	mIoU(C.%)
Baseline	73.52	71.66
2	73.12	**72.37**
4	73.73	71.04
10	**74.71**	71.47
16	73.85	65.51
20	73.60	66.51

TABLE IV: Effects of μ on segmentation results on Pascal Voc 2012 (**P.**) and Cityscapes (**C.**)

Point loss vs. boundary loss: A principal property of the potential fields is that points' potential energy increase with their spatial distance to the semantic boundary, conforming to the character of the distance fields. We consider the boundary loss [17] related to our method and then assemble it to PSPNet (see Appendix B). To make a fair comparison, we test the boundary loss L_{bd} and L_{point} with five loss weights $\{0.05, 0.10, 0.20, 0.25, 0.50\}$, and then report both losses' best segmentation performances on two datasets. In Table VII, we see L_{point} outperforms L_{bd} because the boundary loss is not applicable for the multi-class task and is less informative than L_{point} that conducts field regression in all directions (see Table XIV in Appendix).

Equipotential line loss vs. dice loss: This part compares the L_{line} with the dice loss (L_{dice}) that has a similar optimization principle. We see L_{dice} as an auxiliary for image segmentation and adopt the same evaluation protocol as experiments with L_{point} (see Appendix B and Table XV for implementation details and more comparisons). In the end, we report the comparison results in Table VII and observe that L_{line} performs better than L_{dice} [30]. This comparison result proves the effect of anisotropic convolution and indicates that contour learning can benefit semantic segmentation.

D. Main Results

After loss balance, we apply the full EPL on five FCN baselines: PSPNet [41], PSANet [42], DeepLab V3+ [7], CCNet [16] and GSCNN [31]. Besides, we conduct intensive experiments with variable backbones: ResNet-50/101 [14], MobileNet [15], DRN [37] and WRNet38 [40]. Note that all models are trained in a batch of 16 with reported image size in Table VIII & IX.

Performance on Pascal Voc 2012: In Table VIII, we compare baseline models' performances with and without using EPL. Models with EPL consistently outperform their corresponding baseline models with considerable mIoU increasements. In the best case, we improve the mIoU of DeepLab v3+ (with ResNet101) up to 1.62%. Fig. 9 visualizes

Method	Backbone	mIoU	road	swalk	build	wall	fence	pole	tligh.	tsign	veg	terr.	sky	pers.	rider	car	truck	bus	train	mcyc	bcyc
PSANet	ResNet-101	78.01	98.1	85.1	92.5	54.1	60.9	64.7	70.8	78.6	92.6	65.5	94.6	82.9	63.2	95.0	74.9	88.0	77.6	65.1	78.0
+EPL	ResNet-101	**79.46**	96.8	83.7	92.9	**56.7**	**62.7**	**68.8**	**75.5**	**81.6**	93.3	66.0	94.7	**85.2**	**66.9**	95.6	70.3	**90.3**	76.7	**71.5**	**80.5**
PSPNet	ResNet-101	78.35	98.4	86.3	92.9	55.3	63.1	65.9	73.1	79.8	93.0	66.0	94.9	83.8	64.8	95.1	74.0	85.8	75.8	71.1	79.3
+EPL	ResNet-101	**79.44**	96.7	83.5	**94.1**	**56.3**	62.4	**69.4**	73.7	**82.1**	92.7	**67.4**	94.2	**86.4**	**66.3**	95.7	72.3	**88.5**	**78.1**	**72.3**	80.2
DeepLab	WResNet-38	79.38	98.2	85.7	92.7	60.1	62.9	66.6	70.4	79.2	92.7	66.6	94.6	82.7	64.2	95.2	80.5	89.6	80.9	67.1	78.3
+EPL	WResNet-38	**80.34**	97.5	84.3	92.6	**61.2**	**64.1**	66.5	70.7	80.1	90.5	**68.1**	94.7	**84.5**	65.1	94.9	**81.6**	**90.7**	**82.1**	67.4	**79.5**

TABLE V: Category-level mIoU Comparison on Cityscapes validation set.

Method	Backbone	mIoU	bgr	aero	bicy	bird	boat	bottle	bus	car	cat	chair	cow	dt.	dog	horse	motor	pers.	pott	sheep	sofa	train	tv.
PSANet	ResNet-101	79.25	94.8	91.2	43.7	89.8	76.0	80.6	94.4	88.4	93.3	44.5	88.6	56.0	89.4	87.5	85.8	88.2	65.0	92.2	50.7	86.7	77.4
+EPL	ResNet-101	**80.33**	95.1	92.0	44.4	90.5	**77.9**	**82.1**	94.5	**90.1**	**94.4**	43.8	89.0	**59.2**	90.1	88.1	**88.0**	88.9	65.0	91.9	**53.9**	**89.3**	**78.9**
PSPNet	ResNet-101	79.50	94.9	90.8	44.2	90.0	74.0	81.0	95.3	90.2	94.2	42.8	87.9	57.0	89.5	87.2	89.8	88.3	65.7	91.3	47.3	88.7	79.3
+EPL	ResNet-101	**80.46**	**95.1**	92.4	44.7	88.9	**75.7**	81.7	95.8	**91.5**	94.6	43.5	**89.1**	**59.4**	**90.5**	**88.3**	90.4	88.4	64.7	**93.4**	**49.0**	**91.5**	76.9
DeepLab	ResNet-101	79.15	93.9	89.7	42.3	90.4	69.0	81.1	93.0	91.0	93.5	41.8	90.2	62.1	90.8	88.6	86.4	86.8	67.0	87.5	50.3	87.5	79.2
+EPL	ResNet-101	**80.77**	94.2	89.8	**43.5**	90.6	**72.9**	81.8	93.4	90.3	93.9	**47.2**	**92.3**	**64.7**	**91.8**	89.2	**88.7**	87.7	**70.4**	**89.9**	**59.8**	87.8	77.3

TABLE VI: Category-level mIoU comparison on Pascal Voc 2012 Validation set.

Dataset	Method	mIoU(%)
Pascal Voc	$+L_{ce}$	73.52
	$+L_{ce}+L_{bd}$	74.55
	$+L_{ce}+L_{point}$	**74.84**(+1.32)
	$+L_{ce}+L_{dice}$	74.45
	$+L_{ce}+L_{line}$	**74.71**(+1.19)
Cityscapes	$+L_{ce}$	71.66
	$+L_{ce}+L_{bd}$	71.81
	$+L_{ce}+L_{point}$	**72.40**(+0.74)
	$+L_{ce}+L_{dice}$	70.30
	$+L_{ce}+L_{line}$	**73.26**(+1.60)

TABLE VII: Comparisons with the boundary loss [17] (L_{bd}) and dice loss [30] (L_{dice}) on two datasets' validation set.

Model	Size	Backbone	Method	mIoU(%)
PSANet	465×465	ResNet-50	Baseline	77.85
			+EPL	**79.08**(+1.23)
		ResNet-101	Baseline	79.25
			+EPL	**80.33**(+1.08)
PSPNet	473×473	ResNet-50	Baseline	78.02
			+EPL	**78.84**(+0.82)
		ResNet-101	Baseline	79.50
			+EPL	**80.46**(+0.96)
DeepLab	513×513	MoblieNet	Baseline	71.49
			red+EPL	**72.65**(+1.16)
		DRN-54	Baseline	79.58
			+EPL	**80.63**(+1.05)
		ResNet-101	Baseline	79.15
			+EPL	**80.77**(+1.62)

TABLE VIII: Overall results on Pascal Voc validation set.

Model	Size	Backbone	Method	mIoU(%)
PSANet	560×560	ResNet-50	Baseline	76.69
			+EPL	**77.61**(+0.92)
		ResNet-101	Baseline	78.01
			+EPL	**79.46**(+1.45)
PSPNet	560×560	ResNet-50	Baseline	77.34
			+EPL	**78.49**(+1.15)
		ResNet-101	Baseline	78.35
			+EPL	**79.44**(+1.09)
DeepLab	560×560	WRNet-38	Baseline	79.38
			+EPL	**80.34**(+0.96)
CCNet	560×560	WRNet-38	Baseline	77.73
			+EPL	**78.81**(+1.08)
GSCNN	560×560	WRNet-38	Baseline	80.67
			+EPL	**81.78**(+1.11)

TABLE IX: Overall results on Cityscapes validation set.

the segmentation results, and we see that EPL can greatly help existing FCN-based segmentation models solve the challenges from the inter-class (the first and the second row) or the intra-class regions (the third row).

Performance on Cityscapes: We report the result on cityscapes in Table IX. Once again, we achieve substantial performance gains over all three FCN baseline models when we employ EPL, regardless of the backbone type. Similarly, visual comparisons are exhibited in Fig. 10.

Category-level evaluation. We show the category-level mIoU comparison on both datasets before and after adopting EPL in Table V & VI. On Cityscapes, we observe that EPL significantly improves baselines' ability to distinguish category pairs that have similar semantics or strong semantic relationships, such as {"person", "rider,"}, {"rider", "bicycle"}, and {"traffic sign," "traffic light"}. On Pascal Voc 2012, EPL enhances baseline models' ability to segment categories with complicated shapes, such as the cow, dog, and dining table.

E. Comparison with state-of-the-art methods

In this section, we compare the proposed EPL module with the existing boundary segmentation approaches [18], [19], [39] in Cityscapes. In Table VIII and IX, we observe that EPL enhances all segmentation baselines by over 1% regardless of the backbone models. Compared (in Table X) with other boundary segmentation studies [4], [18], [19], [39], EPL is slightly worse than InverseForm (**InF**) yet competitive against the other studies. Besides, we compare the properties of each method in Table XI and observe that EPL is the only approach that does not utilize edge maps for the network training and brings no increased parameter size and inference cost.

V. Conclusion

This paper addresses the semantic boundary segmentation problem with anisotropic field regression and category-level contour learning. With the proposed EPL (equipotential learning) module, we transfer the original probability estimation problem to the self-defined potential domain with the

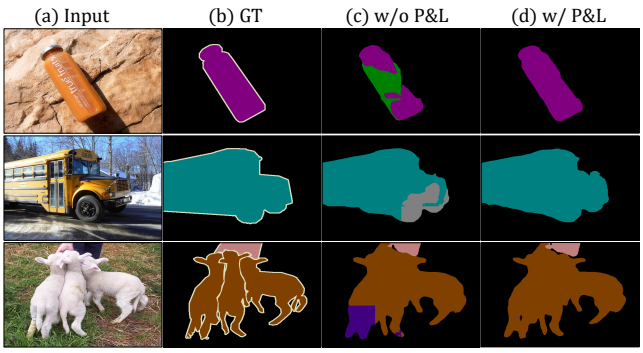


Fig. 9: Qualitative comparison for segmentation results on Pascal Voc 2012 validation set.

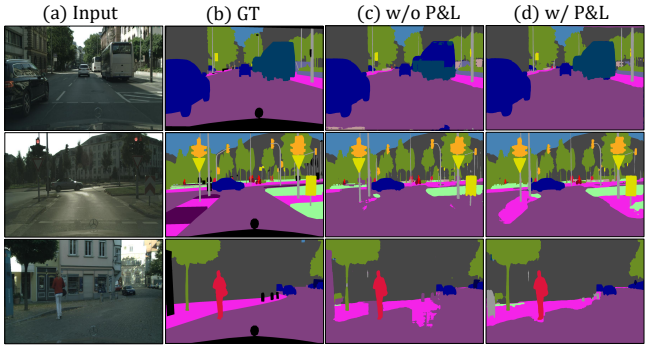


Fig. 10: Qualitative comparison for segmentation results on Cityscapes validation set.

Method	Backbone	mIoU(%)
PSPNet	ResNet-50	77.34
PSPNet w/ EPL	ResNet-50	**78.49**(+1.15)
PSANet	ResNet-101	78.01
PSANet w/ EPL	ResNet-101	**79.46**(+1.45)
DeepLabv3	ResNet-101	77.80
DeepLabv3 w/ PR [18]	ResNet-101	78.40(+1.20)
DeepLabv3	ResNet-50	79.18
DeepLabv3 w/ IABL [32]	ResNet-50	79.94(+0.76)
DeepLabv3+	WRNet-38	79.50
DeepLabv3+ w/ SFix [39]	WRNet-38	80.30(+0.80)
DeepLabv3+	ASPP [7]	81.30
DeepLabv3+ w/ DSN [19]	ASPP [7]	82.40(+1.10)
DeepLabv3+ (our Impl.)	WRNet-38	79.38
DeepLabv3+ w/ EPL	WRNet-38	**80.34**(+0.96)
GSCNN	WRNet-38	81.0
GSCNN w/ InF [4]	WRNet-38	82.60(+1.50)
GSCNN (our Impl.)	WRNet-38	80.67
GSCNN w/ EPL	WRNet-38	**81.78**(+1.11)

TABLE X: Comparisons with SOTA studies on Cityscapes. Note that our result (**our Impl.**) is experimented with 560×560 images and hence results in a small mIoU gap.

anisotropic convolution. Besides, we sequentially introduced the point loss to fit the image content along different directions from variable distances and the equipotential line loss to enforce the category-level contour learning. Experiments on Pascal Voc 2012, Cityscapes show that the designed EPL, serving as an add-on method, can significantly enhance existing

Method	No param. overhead	No edge super.	No infer. overhead
SFix [39]	-	-	-
DSN [19]	-	-	-
PR [18]	-	✓	-
InF [4]	-	-	✓
IABL [32]	-	✓	✓
EPL	✓	✓	✓

TABLE XI: Comparisons of the SOTA boundary segmentation methods. We evaluate these studies' properties from three perspectives: whether using edge maps for boundary supervision (**super.**) and increasing parameters (**param.**) and inference (**infer.**) overhead.

FCN models' performance on the semantic boundary regions. Compared with other studies [4], [18], [19], [23], [39], our approach does not introduce additional supervision information and adds no parameters and inference cost. We believe that EPL can be generalized and benefit other segmentation tasks, like point cloud [5], [21], instance, and panoptic segmentation [8].

APPENDIX

In this part, we report more comparison results of our approach in ablation experiments. In the end, we present more quantitative results to support our paper. This section presents more experiment details and comparison results for the ablation experiment in Section IV-B.

A. Comparisons results of different splitters:

We experiment AC with with three different splitters, A, B and C (shown in Fig. 3), that contain different directional vectors. Next, we compare the segmentation performance of the point loss and the equipotential line loss on PSPNet when using different splitters. As reported detailed comparison results in Table XII & XIII, we see that L_{point} and L_{line} perform best when using the splitter A.

B. Comparisons with related works

In Section IV-C, we compared the point loss with the boundary loss, the equipotential line loss, and the dice loss. Here, we report the detailed results of each comparison pair.

Implementation details: To apply the boundary loss to the segmentation network, we follow the practice of [17], firstly computing the ground-truth boundary-distance map of the true semantic labels and then integrating the boundary loss with the cross-entropy loss. Also, we adopt the same implementation protocol when comparing the dice loss with our equipotential line loss.

Detailed comparison results: For each comparison pair, we test 5 different loss weights and report the experiment details and conditions where the performances are achieved. We respectively set the splitter, kernel size, and loss norm of 'A', 7 and 'A', 11 when experimenting on Pascal Voc 2012 and cityscapes. Besides, the point loss is formulated as the L_2 norm consistently.

Dataset	Method	Kernel	mIoU (%)		
Pascal Voc	Baseline	-	73.52		
			A	B	C
	+Point^1	7	74.07	74.24	73.80
		9	73.85	74.36	73.83
		11	**74.08**	**74.55**	**74.35**
	+Point^2	7	74.09	**74.42**	**74.63**
		9	**74.42**	73.60	71.21
		11	74.21	72.96	74.29
Cityscapes	Baseline	-	71.66		
			A	B	C
	+Point^1	9	70.44	**72.07**	61.37
		11	70.96	55.04	67.01
		13	**72.71**	66.61	**72.22**
	+Point^2	9	**72.39**	71.50	68.25
		11	70.93	**73.25**	**73.07**
		13	71.02	64.72	71.78

TABLE XII: Detailed performance comparisons of the Point loss on Pascal Voc 2012 and Cityscapes validation set. The superscript of "+Point" denotes the norm type, and A, B, C columns indicate that performances are obtained using the corresponding splitter.

Dataset	Method	Kernel	mIoU (%)		
Pascal Voc	Baseline	-	73.52		
			A	B	C
	+Line	7	**74.15**	73.50	**73.94**
		9	73.47	73.33	73.73
		11	74.05	**73.74**	73.62
Cityscapes	Baseline	-	71.66		
			A	B	C
	+Line	9	71.78	**72.03**	71.19
		11	**73.26**	66.80	**71.35**
		13	72.11	60.42	66.96

TABLE XIII: Detailed performance comparison of the equipotential line loss on Pascal Voc 2012 and Cityscapes validation set. A, B, C denote the splitter type.

Dataset	weight	mIoU(B.%)	mIoU(P.%)
Pascal Voc	Baseline	73.52	
	0.05	72.88	73.31
	0.10	**74.55**	**74.84**
	0.20	73.59	72.92
	0.25	72.17	64.44
	0.50	43.95	73.61
Cityscapes	Baseline	71.66	
	0.05	**71.81**	65.92
	0.10	68.75	70.93
	0.20	71.78	**72.40**
	0.25	69.65	66.11
	0.50	66.86	69.04

TABLE XIV: Quantitative comparison between the boundary loss (**B.**) and our point loss (**P.**).

Dataset	weight	mIoU(D.%)	mIoU(L.%)
Pascal Voc	Baseline	73.52	
	0.05	72.88	73.49
	0.10	74.07	73.68
	0.20	**74.45**	**74.71**
	0.25	74.36	68.22
	0.50	73.45	72.83
Cityscapes	Baseline	71.66	
	0.05	57.49	56.72
	0.10	57.89	68.58
	0.20	61.30	71.47
	0.25	60.54	70.87
	0.50	**70.30**	**73.26**

TABLE XV: Quantitative comparison between the dice loss (**D.**) and our equipotential line loss (**L.**)

REFERENCES

[1] D. Acuna, A. Kar, and S. Fidler, "Devil is in the edges: Learning semantic boundaries from noisy annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 11 075–11 083.

[2] N. Audebert, A. Boulch, B. Le Saux, and S. Lefèvre, "Distance transform regression for spatially-aware deep semantic segmentation," *Comput. Vis. and Image Underst.*, vol. 189, p. 102809, 2019.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[4] S. Borse, Y. Wang, Y. Zhang, and F. Porikli, "Inverseform: A loss function for structured boundary-aware segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Virtual/Online, 2021, pp. 5901–5911.

[5] C. Chen, S. Qian, Q. Fang, and C. Xu, "Hapgn: Hierarchical attentive pooling graph network for point cloud segmentation," *IEEE Trans. Multimedia*, vol. 23, pp. 2335–2346, 2021.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, and K. e. a. Murphy, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal.*, vol. 40, no. 4, pp. 834–848, 2017.

[7] L.-C. Chen, Y. Zhu, and G. e. a. Papandreou, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 801–818.

[8] B. Cheng, M. D. Collins, Y. Zhu, and T. e. a. Liu, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Virtual/Online, 2020, pp. 12 475–12 485.

[9] B. Cheng, R. Girshick, P. Dollár, and A. C. e. a. Berg, "Boundary iou: Improving object-centric image segmentation evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Virtual/Online, 2021, pp. 15 334–15 342.

[10] N. Cohen and A. Shashua, "Inductive bias of deep convolutional networks through pooling geometry," 2017. [Online]. Available: https://arxiv.org/abs/1605.06743

[11] M. Cordts, M. Omran, S. Ramos, and T. e. a. Rehfeld, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, Nevada, 2016, pp. 3213–3223.

[12] M. Everingham and J. Winn, "The pascal visual object classes challenge 2012 (voc2012) development kit," *Pattern Anal., Statist. Modelling and Computat. Learning, Tech. Rep*, vol. 8, 2011.

[13] Z. Gu, L. Niu, H. Zhao, and L. Zhang, "Hard pixel mining for depth privileged semantic segmentation," *IEEE Trans. Multimedia*, vol. 23, pp. 3738–3751, 2021.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, Nevada, 2016, pp. 770–778.

[15] A. G. Howard, M. Zhu, B. Chen, and D. K. et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: http://arxiv.org/abs/1704.04861

[16] Z. Huang, X. Wang, L. Huang, and C. e. a. Huang, "Ccnet: Criss-cross attention for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Shimon Ullman, 2019, pp. 603–612.

[17] H. Kervadec, J. Bouchtiba, C. Desrosiers, and E. e. a. Granger, "Boundary loss for highly unbalanced segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Shenzhen, China, 2019, pp. 285–296.

[18] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, Washington, 2020, pp. 9799–9808.

[19] X. Li, X. Li, L. Zhang, and G. e. a. Cheng, "Improving semantic segmentation via decoupled body and edge supervision," in *Proc. Eur. Conf. Comput. Vis.* Virtual/Online: Springer, 2020, pp. 435–452.

[20] T.-Y. Lin, P. Goyal, R. Girshick, and K. e. a. He, "Focal loss for dense

object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Tomaso Poggio, 2017, pp. 2980–2988.
[21] H. Liu, Y. Guo, Y. Ma, and Y. e. a. Lei, "Semantic context encoding for accurate 3d point cloud segmentation," *IEEE Trans. Multimedia*, vol. 23, pp. 2045–2055, 2021.
[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, Massachusetts, 2015, pp. 3431–3440.
[23] D. Marin, Z. He, P. Vajda, and P. e. a. Chatterjee, "Efficient segmentation: Learning downsampling near semantic boundaries," in *Proc. IEEE Int. Conf. Comput. Vis.*, Shimon Ullman, 2019, pp. 2131–2141.
[24] P. Naylor, M. Laé, F. Reyal, and T. Walter, "Segmentation of nuclei in histopathology images by deep regression of the distance map," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 448–459, 2018.
[25] F. Perazzi, J. Pont-Tuset, B. McWilliams, and L. e. a. Van Gool, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, Nevada, 2016, pp. 724–732.
[26] N. Rahaman *et al.*, "On the spectral bias of neural networks," in *Int. Conf. on Machine Learning*. Long Beach, US: PMLR, 2019, pp. 5301–5310.
[27] R. Rau and J. H. McClellan, "Efficient approximation of gaussian filters," *IEEE Trans. Image Process.*, vol. 45, no. 2, pp. 468–471, 1997.
[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*. Munich, Germany: Springer, 2015, pp. 234–241.
[29] Q. Song, J. Li, C. Li, H. Guo, and R. Huang, "Fully attentional network for semantic segmentation," *arXiv preprint arXiv:2112.04108*, 2021.
[30] C. H. Sudre, W. Li, T. Vercauteren, and S. e. a. Ourselin, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in med. image anal. and multimod. learning for clin. dec. supp.* Springer, 2017, pp. 240–248.
[31] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Shimon Ullman, 2019, pp. 5229–5238.
[32] C. Wang, Y. Zhang, M. Cui, J. Liu, P. Ren, Y. Yang, X. Xie, X. Hua, H. Bao, and W. Xu, "Active boundary loss for semantic segmentation," 2021. [Online]. Available: arXivpreprintarXiv:2102.02696
[33] E. Xie, P. Sun, X. Song, and W. Wang, "Polarmask: Single shot instance segmentation with polar representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Virtual/Online, 2020, pp. 12 193–12 202.
[34] Y. Xue, H. Tang, Z. Qiao, and G. e. a. Gong, "Shape-aware organ segmentation by predicting signed distance maps," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, Hilton New York Midtown, 2020, pp. 12 565–12 572.
[35] S. Ye, D. Chen, S. Han, and J. Liao, "Learning with noisy labels for robust point cloud segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Virtual/Online, 2021, pp. 6443–6452.
[36] C. Yu, J. Wang, C. Peng, and C. e. a. Gao, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, Utah, 2018, pp. 1857–1866.
[37] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, Hawaii, 2017, pp. 472–480.
[38] Z. Yu, W. Liu, Y. Zou, C. Feng, S. Ramalingam, B. Kumar, and J. Kautz, "Simultaneous edge alignment and learning," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 388–404.
[39] Y. Yuan, J. Xie, X. Chen, and J. Wang, "Segfix: Model-agnostic boundary refinement for segmentation," in *Proc. Eur. Conf. Comput. Vis.* Virtual/Online: Springer, 2020, pp. 489–506.
[40] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016. [Online]. Available: http://arxiv.org/abs/1605.07146
[41] H. Zhao, J. Shi, X. Qi, and X. e. a. Wang, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, Hawaii, 2017, pp. 2881–2890.
[42] H. Zhao, Y. Zhang, S. Liu, and J. e. a. Shi, "Psanet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 267–283.
[43] L. Zhou, C. Gong, Z. Liu, and K. Fu, "Sal: Selection and attention losses for weakly supervised semantic segmentation," *IEEE Trans. Multimedia*, vol. 23, pp. 1035–1048, 2020.