NEURAL INFORMATION PROCESSING SYSTEMS

Diffusion-Driven Two-Stage Active Learning for Low-Budget Semantic Segmentation

Jeongin Kim[1] Wonho Bae[2] YouLee Han[1] Giyeong Oh[3] Youngjae Yu[4] Danica J. Sutherland[2,5] Junhyug Noh[1]

[1]EWHA [2]UBC [3]YONSEI [4]SNU [5]Amii

Poster ID: 118018 | Schedule: Wednesday, Dec 11th | Exhibit Hall C,D,E

Background

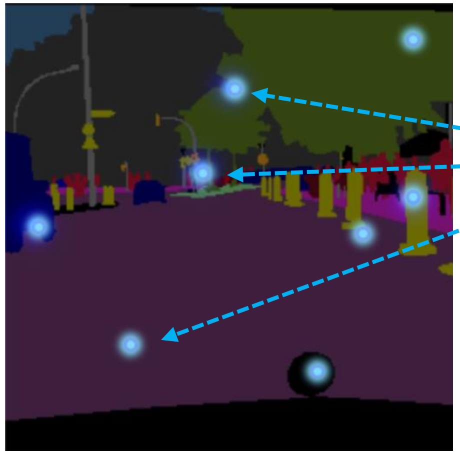
- Problem: fully-supervised semantic segmentation requires dense pixel labeling.
    - Manual annotation is extremely costly and time-consuming
      → Need to reduce labeling costs while maintaining strong performance

- Active Learning (AL) alleviates this by selecting informative samples for annotation
    - Prior AL studies query samples at different granularities:
        - Image-level (i.e., subset of images)
        - Region-level (e.g., super-pixels)
        - Pixel-level (e.g., identical pixel budget per image)

Background

- Our Approach: Defining a New Practical Setting
  → Select an extremely small number of pixels from a unified candidate pool across images

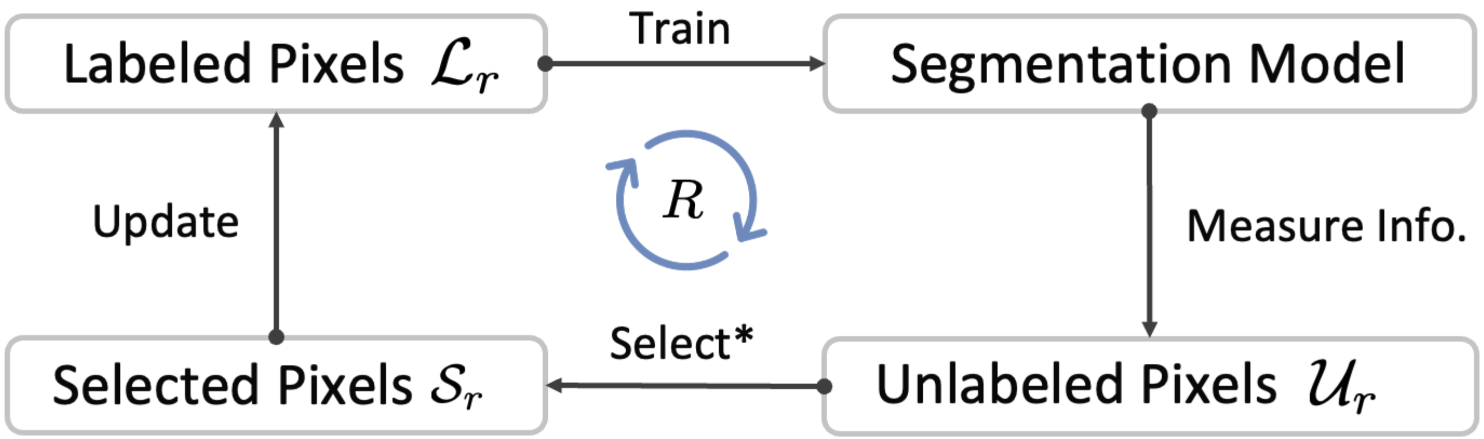
- A Key Question:



Which pixels are the most informative to label to improve segmentation performance?

Problem Setup: Active Learning

- At each active learning round,
  the model selects b informative pixels from unlabeled pixels for annotation
- Selected pixels are added to the labeled set
- The segmentation model is retrained iteratively



*Selection based on Acquisition Function

Problem Setup: Budget Setting

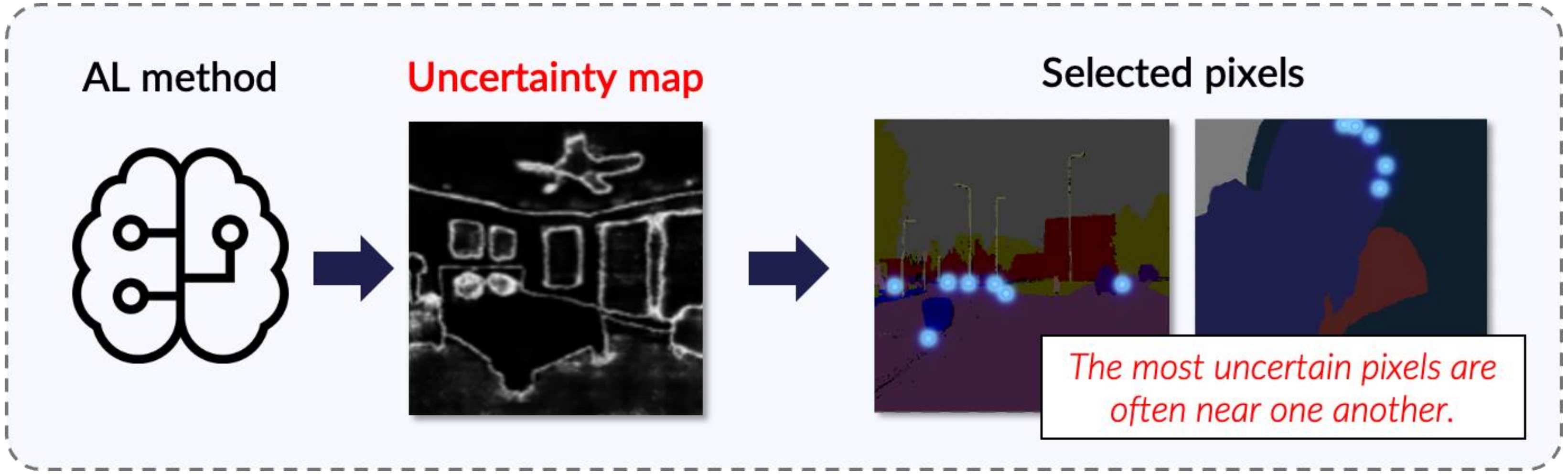
- Total Budget (B): We fix the total annotation budget $B = N$
  Where N is the total number of images in the dataset
  → Averages to only one labeled pixel per image

- AL Process: We split this budget evenly into 10 AL rounds
  → Yields a budget of $b = 0.1N$ pixels per round

- In total, only 0.0015% of all pixels are annotated after 10 rounds
  → *An extreme low-budget regime*

Motivation

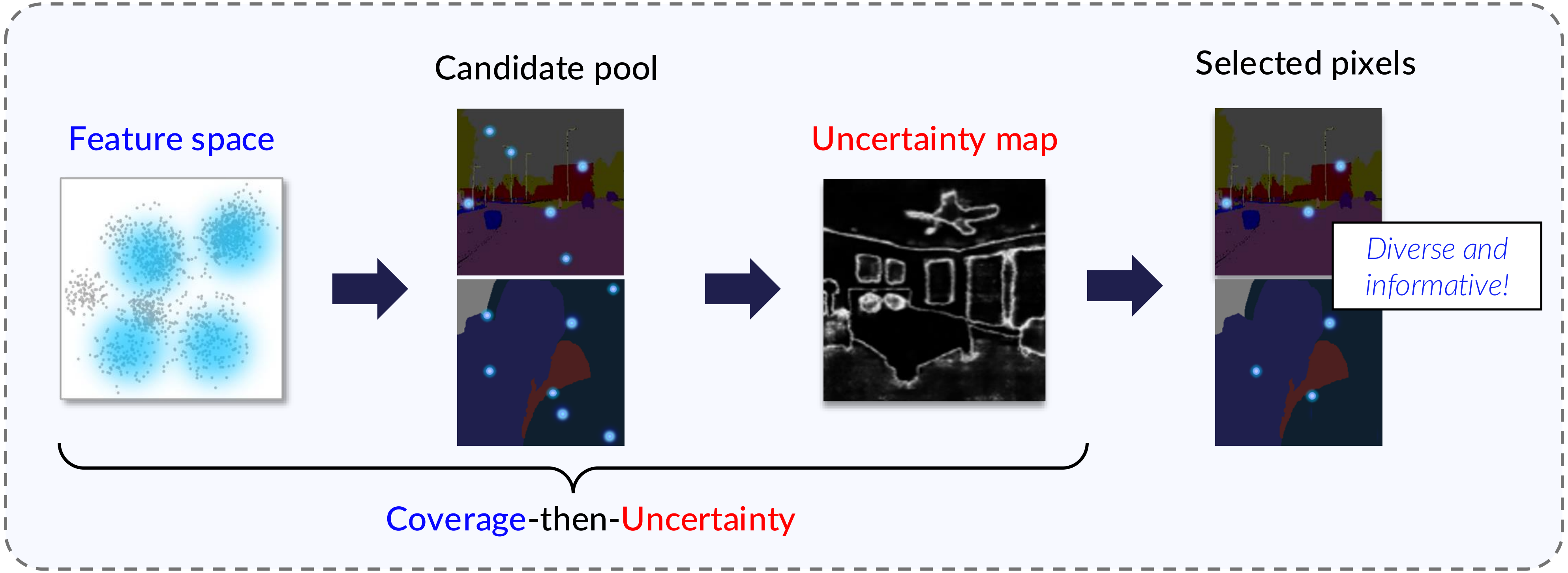
- Existing pixel-level active learning studies [1, 2]
  → Apply uncertainty-based acquisition functions
  → But they tend to select redundant pixels
  as highly uncertain pixels are clustered along local regions (e.g., object boundaries)



[1] Gyungin Shin, Weidi Xie, and Samuel Albanie. “All You Need Are a Few Pixels: Semantic Segmentation With PixelPick.” ICCVW 2021.
[2] Sima Didari, Wenjun Hu, Jae Oh Woo, Heng Hao, Hankyu Moon, and Seungjai Min. “Bayesian Active Learning for Semantic Segmentation.” arXiv preprint 2024.

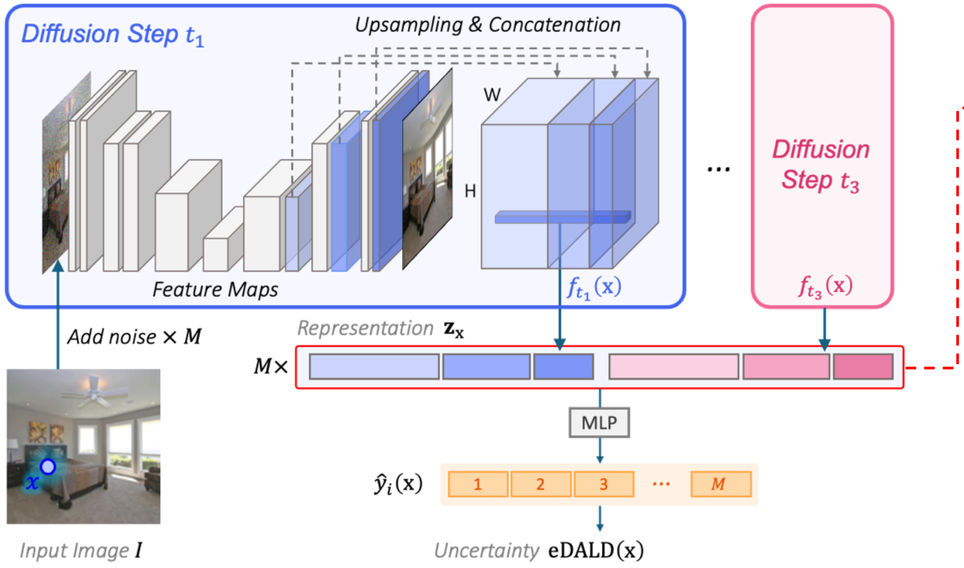
Our Solution

- How do we alleviate this redundancy? We introduce a two-stage sampling strategy
  1. Selects a representative candidate pool
  2. Narrows down the most uncertain samples



Diffusion Representations for Segmentation

- Building on LEDM [3], we extract multi-scale features from a pre-trained diffusion model
  → Utilizing pre-trained features is essential in low-budget regimes
  as sparse labels provide insufficient supervision for learning effective representations
- For a given pixel x, we obtain a multi-scale representation $\mathrm{z_x}$, which is fed to lightweight segmentation head $s_\theta: \mathbb{R}^D \rightarrow \mathbb{R}^C$



$$\mathbf{z}_\mathrm{x} := \left[f_{t_1,l_1}(\mathrm{x}); \ldots; f_{t_1,l_L}(\mathrm{x}); \ldots; f_{t_T,l_1}(\mathrm{x}); \ldots; f_{t_T,l_L}(\mathrm{x}) \right] \in \mathbb{R}^D$$

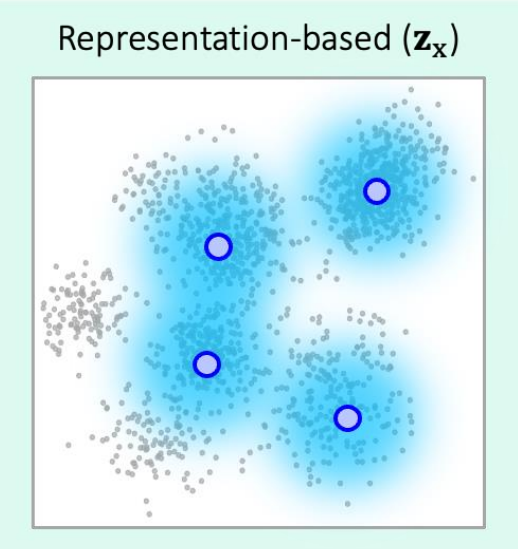
$$D = \sum_{i=1}^{T} \sum_{j=1}^{L} D_{t_i,l_j}$$

The stochastic nature of diffusion enables us to capture uncertainty from variations in multi-timestep features

[3] Dmitry Baranchuk, et al. "Label-Efficient Semantic Segmentation with Diffusion Models." ICLR 2022.

Two-stage eDALD

1st stage: Representation-based candidate selection



MaxHerding-based Greedy selection [4]

$$\tilde{\mathrm{x}}^* \in \underset{\tilde{\mathrm{x}} \in I, \tilde{\mathrm{x}} \in \mathcal{U}}{\operatorname{argmax}} \hat{\mathrm{C}}_k(\mathcal{L} \cup \{\tilde{\mathrm{x}}\})$$

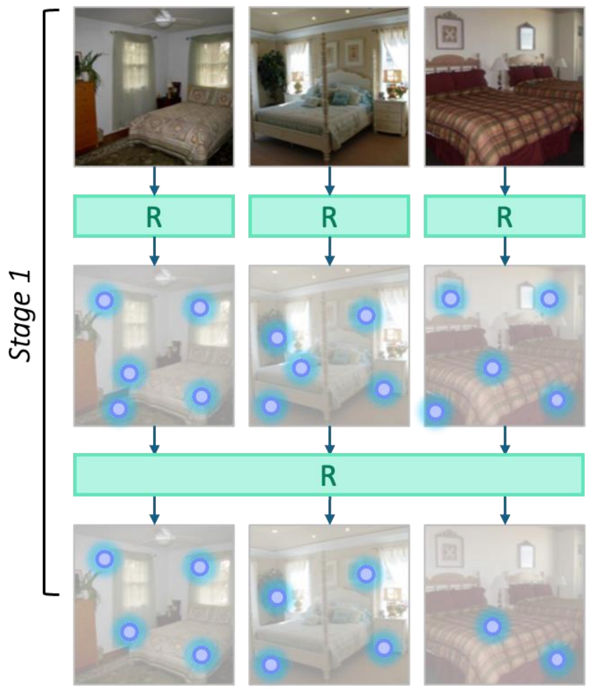
where $\hat{\mathrm{C}}_k(\mathcal{L} \cup \{\tilde{\mathrm{x}}\}) := \frac{1}{|\mathcal{U}|} \sum_{\mathrm{x} \in \mathcal{U}} \left[\max_{\mathrm{x}' \in \mathcal{L} \cup \{\tilde{x}\}} k(\mathrm{x}, \mathrm{x}') \right]$

$$k(\mathrm{x}, \mathrm{x}') = \exp\left(- \frac{\|\mathbf{z}_{\mathrm{x}} - \mathbf{z}_{\mathrm{x}'}\|_2^2}{\sigma^2} \right)$$ (RBF Kernel)

[4] Wonho Bae, Junhyug Noh, and Danica J Sutherland. “Generalized Coverage for More Robust Low-Budget Active Learning.” ECCV 2024.

Two-stage eDALD

1^{st} stage: Representation-based candidate selection



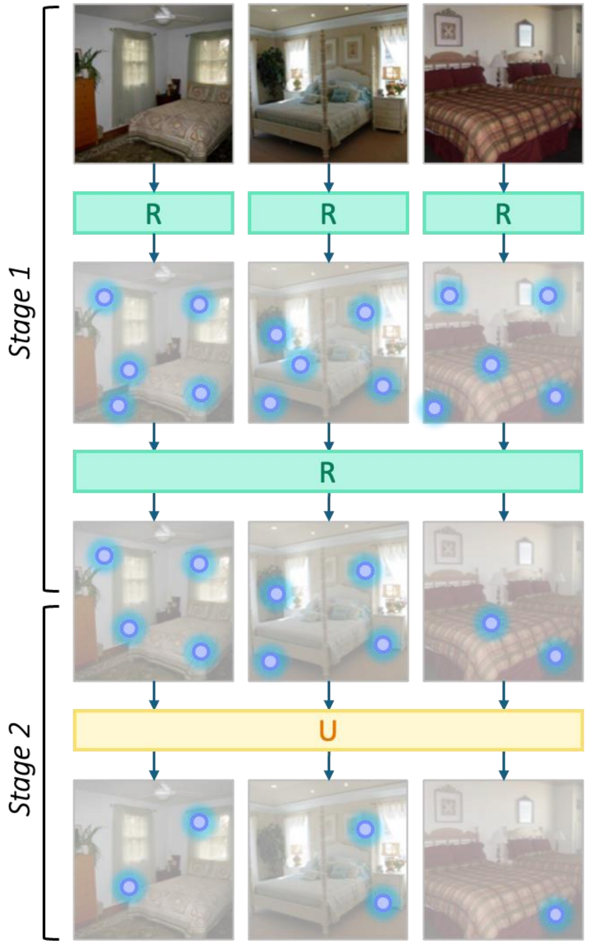
Two-step MaxHerding

I. **Local Herding**
Identify K representative pixels within each image
$\rightarrow$ Form initial pool $\mathcal{M}_0$

II. **Global Herding**
Apply MaxHerding across $\mathcal{M}_0$ to select $\mathcal{M}$ diverse candidates

Two-stage eDALD

2nd stage : Uncertainty-based selection



Diffusion-based Active Learning by Disagreement (DALD)

$$\mathrm{x}^* \in \underset{\mathrm{x} \in \mathcal{U}}{\operatorname{argmax}} \; \mathrm{I}(\hat{Y}; Z \mid \mathrm{x}, s_\theta, f)$$

$$\mathrm{I}(\hat{Y}; Z \mid \mathrm{x}, s_\theta, f) = \underbrace{\mathrm{H}(\hat{Y} \mid \mathrm{x}, s_\theta, f)}_{\text{Unconditional entropy}} - \mathop{\mathbb{E}}_{\mathbf{z} \sim q(\cdot \mid \mathrm{x})} \Big[\underbrace{\mathrm{H}(\hat{Y} \mid Z = \mathbf{z}, \mathrm{x}, s_\theta)}_{\text{Conditional entropy}} \Big]$$

Inspired by BALD [5]

→ Measures model disagreement (mutual information, MI)

→ Across stochastic forward passes

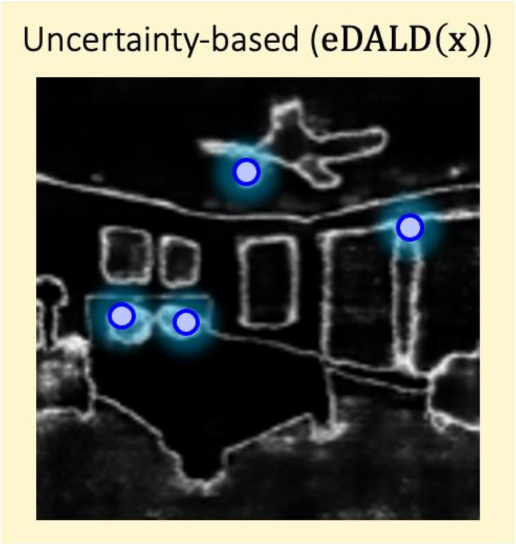
→ Multiple noisy inputs → varied features → higher MI

→ stronger epistemic uncertainty

[5] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. “Deep Bayesian Active Learning with Image Data.” ICML 2017.

Two-stage eDALD

2nd stage : Uncertainty-based selection



Entropy-Augmented DALD (eDALD)

To further account for aleatoric uncertainty and overall predictive noise, we introduce an Entropy term.

$$\mathrm{x}^* = \underset{\mathrm{x}\in\mathcal{U}}{\mathrm{argmax}}\ \mathrm{eDALD}(\mathrm{x}),$$

where $\mathrm{eDALD}(\mathrm{x}) = \mathrm{I}\big(\hat{Y}; Z \mid \mathrm{x}, s_\theta, f\big) + \mathrm{H}\big(\hat{Y} \mid \mathbf{z}^{(0)}, \mathrm{x}\big)$

Datasets

- Evaluate the proposed metho on **four** standard semantic segmentation benchmarks

	CamVid	ADE-Bed	Cityscapes	Pascal-Context
Sample				
# of class	11	30	19	33
Scene	Urban driving	Bedroom	Street-scene	Everyday scenes
Example	road, building, car	bed, lamp, pillow	road, pedestrian, car	person, dog, chair

Experiments

- Effect of representation-first filtering on uncertainty sampling

Uncertainty	UC Only	Herding $\rightarrow$ UC	Gain (pp)	Gain (%)
Entropy	25.26 ± 0.36	30.77 ± 0.44	$+5.51$	$+21.81$
Margin	31.27 ± 1.10	32.77 ± 0.75	$+1.50$	$+4.80$
BALD	24.59 ± 0.97	22.79 ± 0.89	-1.80	-7.32
DALD	23.81 ± 3.60	21.05 ± 1.05	-2.76	-11.59
PowerBALD	30.03 ± 0.76	31.57 ± 0.79	$+1.54$	$+5.13$
PowerDALD	31.30 ± 1.22	32.00 ± 0.66	$+0.70$	$+2.24$
eBALD (Entropy + BALD)	25.96 ± 1.92	32.12 ± 0.40	$+6.16$	$+23.73$
eDALD (Entropy + DALD)	25.14 ± 0.57	36.12 ± 0.24	$+10.98$	$+43.68$

- MaxHerding-based filtering exhibits the best synergy with eDALD
  → Reflects the complementary role of the two signals:

  1. **Disagreement**: Identifies perturbation-sensitivity that entropy may overlook
  2. **Entropy**: Recovers low-confidence areas that disagreement misses

Experiments

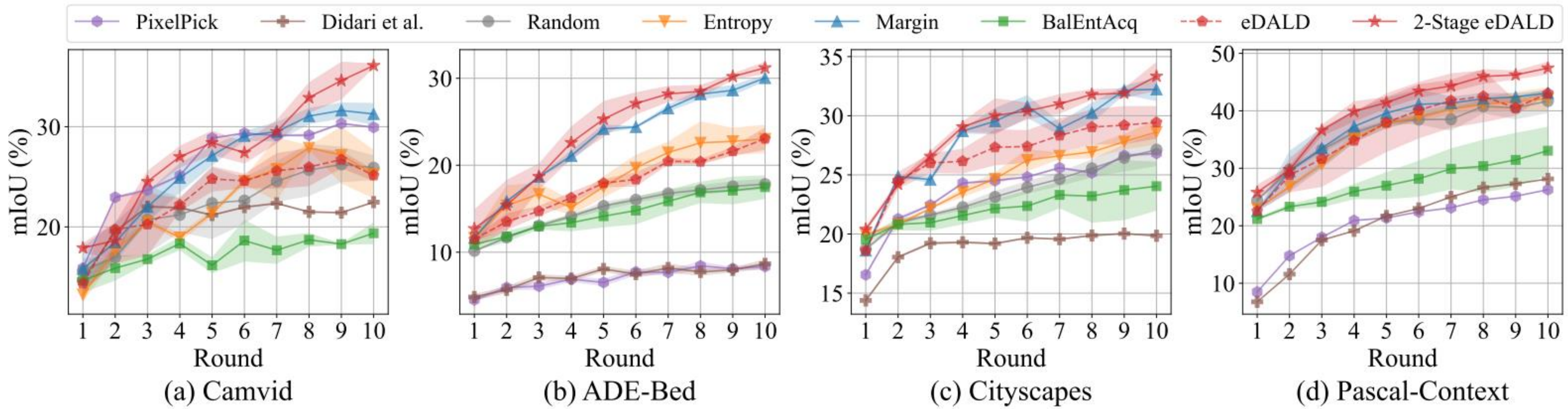
- Performance comparison with baselines

Backbone	Method	CamVid	ADE-Bed	Cityscapes	Pascal-C	Avg
DeepLabV3	PixelPick [1]	29.93 ± 0.12	8.35 ± 0.41	26.82 ± 0.14	26.28 ± 0.09	22.85
	Didari et al. [2]	22.47 ± 0.10	8.66 ± 0.53	19.85 ± 0.07	28.15 ± 0.11	19.78
DDPM	Random	25.91 ± 1.23	17.83 ± 0.62	27.13 ± 1.38	41.70 ± 2.08	28.14
	Entropy	25.26 ± 0.36	23.02 ± 1.64	28.62 ± 1.05	42.09 ± 1.99	29.74
	Margin	31.27 ± 1.10	30.03 ± 0.37	32.23 ± 1.21	45.11 ± 2.45	34.66
	BalEntAcq	19.37 ± 1.10	17.48 ± 1.36	24.04 ± 2.07	33.06 ± 4.18	23.49
	eDALD	25.14 ± 0.57	23.06 ± 1.29	29.44 ± 1.38	43.05 ± 0.12	30.17
	2-Stage eDALD	$\mathbf{36.12 \pm 0.24}$	$\mathbf{31.12 \pm 0.20}$	$\mathbf{33.34 \pm 0.78}$	$\mathbf{47.98 \pm 0.41}$	**37.14**

- DDPM backbone significantly outperforms DeepLabV3 backbone
- Our 2-Stage eDALD achieves the best average performance, consistently outperforming all baselines

Experiments

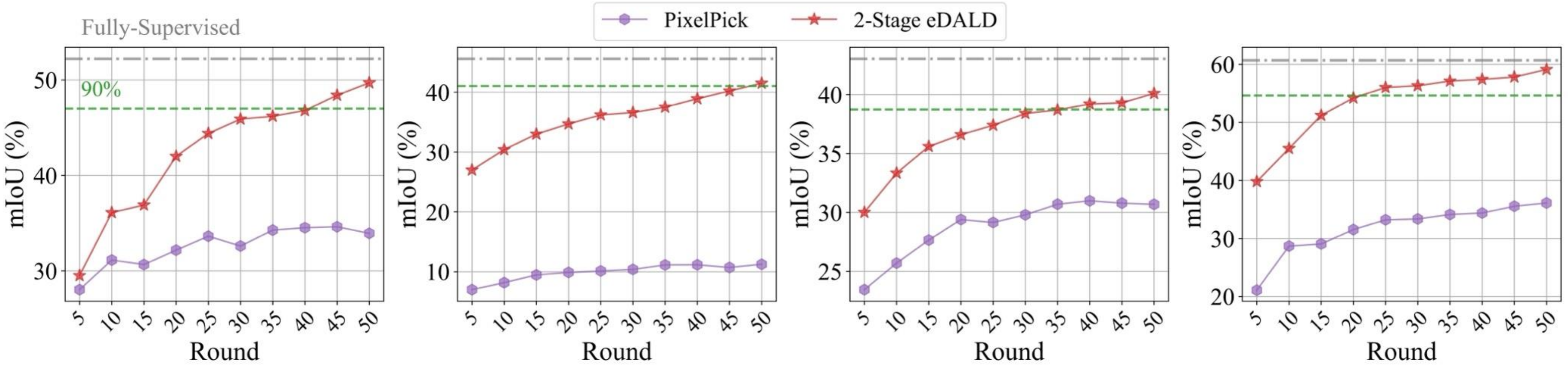
- Round-wise learning curves



- Our 2-Stage eDALD consistently achieves the highest final-round performance across all datasets
- Uncertainty-based single-stage methods show only slow, gradual improvements, leading to smaller overall gains

Experiments

- Convergence to fully-supervised performance



(a) Camvid (b) ADE-Bed (c) Cityscapes (d) Pascal-Context

- More rounds progression when 2-stage eDALD (coverage → uncertainty) is applied only in the low-budget phase
- Reaches 90% of fully-supervised mIoU using only 0.003–0.007% of pixels

Conclusion

- We proposed 2-stage eDALD, a novel strategy for low-budget active learning:
  1. **Stage 1** ensures diverse coverage using diffusion-based representations
  2. **Stage 2** refines selection by eDALD, our novel uncertainty metric

- Our metric (eDALD) is a confidence-aware disagreement-based metric that effectively captures both epistemic and aleatoric uncertainty

- Key Finding: We found that combining representation diversity with diffusion-based uncertainty enables effective learning even under extremely low annotation budgets

Thanks for listening !

For more information, please refer to our paper and code

Paper

Code

Poster ID: 118018 | Schedule: Wednesday, Dec 11th | Exhibit Hall C,D,E