A. Societal impact

Our method enables a segmentation model trained on a synthetic dataset more generalizable when deploying in a real-world dataset with the limited annotation cost. Thus, our work may have a positive impact on communities to reduce the cost of annotating the out-of-domain data, which is economic and environmental friendliness. We carry out experiments on benchmark datasets and do not notice any societal issues. It does not involve sensitive attributes.

B. Sensitivity to different hyper-parameters

We conduct detailed experiments about the sensitivity to different hyper-parameters related to our method. If not specifically mentioned, all the experiments below are carried out based on DeepLab-v2 with the backbone ResNet-101 on GTAV $\rightarrow$ Cityscapes using Ours (RA). When we investigate the sensitivity to a specific hyperparameter, other parameters are fixed to the default values, i.e., τ=0.05, α_1=0.1 and α_2=1.0 for all experiments.

B.1. Effect of the negative threshold τ

In Table 7, we show the results of Ours (RA) with varying τ, i.e., $\tau \in \{0.01, 0.02, 0.05, 0.08, 0.10, 0.20\}$. Ours (RA) achieves consistent results within a suitable insecure threshold (≤ 0.10), but will have a performance drop with a large value of τ like 0.20.

B.2. Effect of the consistency regularization loss weight α_1

In Table 8, we show the results of Ours (RA) with varying α_1, i.e., $\alpha_1 \in \{0.0, 0.05, 0.1, 0.2, 0.5, 1.0\}$. Note that when α_1=0.0, the model is trained without any consistency constraint on source data. As we can see, the best performance is achieved at α_1=0.1. A smaller or larger value of α_1 will either induce a weaker or stronger constraint.

B.3. Effect of the negative learning loss weight α_2

In Table 9, we show the results of Ours (RA) with varying α_2, i.e., $\alpha_2 \in \{0.0, 0.1, 0.5, 1.0, 1.5, 2.0\}$. Note that when α_2=0.0, the model is trained without negative learning loss on target data. The performance is stable varying α_2, which signifies the equal magnitude between negative learning loss and supervised learning loss.

C. Comparison with a simple edge detector

In § 4.2, since we set k=1 that is relative small for RA, the selected regions via Ours (RA) indicate that the edge pixels are mostly favored for labeling. This makes sense due to the criterion favoring regions which have high spatial entropy. Therefore, we further compare our RIPU with a simple Canny algorithm [3] + prediction uncertainty (ENT [52]) to select uncertain pixels from pre-detected edge and the results are reported in Table 10. The performance drops 1.4 mIoU (under 2.2% budget) and 8.6 mIoU (under 40 pixels budget).

For a better understanding of the section procedure, we illustrate the selected regions for annotating from the Canny + ENT and Our methods. From Fig. 7, we can clearly see that the edges inside an object will be selected by Canny + ENT, which is unhelpful. But the region impurity is actually low (low spatial entropy) and these regions would not be selected in our method.

Table 7. Effect of the negative threshold τ.

τ	0.01	0.02	**0.05**	0.08	0.10	0.20
mIoU	68.81	69.54	**69.62**	69.17	68.94	68.13

Table 8. Effect of the consistency regularization loss weight α_1.

α_1	0.0	0.05	**0.1**	0.2	0.5	1.0
mIoU	69.22	69.45	**69.62**	69.61	69.58	69.39

Table 9. Effect of the negative learning loss weight α_2.

α_2	0.0	0.1	0.5	**1.0**	1.5	2.0
mIoU	69.04	69.27	69.44	**69.62**	69.36	69.16

Table 10. Comparison with edge detector on GTAV $\rightarrow$ Cityscape.

Method	Budget	mIoU	Budget	mIoU
Canny + ENT	40 pixels	56.9	2.2%	68.2
Ours	PA, 40 pixels	**65.5**	RA, 2.2%	**69.6**

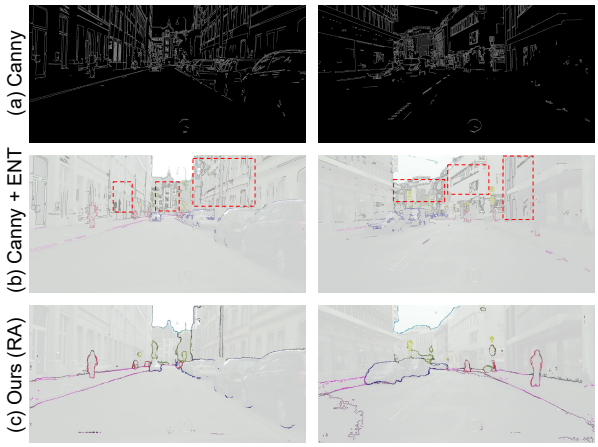

Figure 7. **Visualization of queried pixels to annotate (2.2%) on GTAV $\rightarrow$ Cityscapes.** (a) Canny: the edges of target images detected by Canny algorithm; (b) Canny + ENT: a simple Canny edge detector with uncertainty sampling as a baseline. The similarities are that both methods favor the edge pixels to annotate, the differences are that Canny + ENT may pick out pixels inside objects, such as windows in a building while the proposed method is capable of avoiding this issue, demonstrating the benefits of diverse region selection with both impurity and uncertainty.

Table 11. Experiments on different active selection methods on GTAV → Cityscapes. Best results are shown in **bold**.

Method	Budget	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	mIoU
RAND	40 pixels	94.2	**69.9**	85.2	44.4	40.6	33.1	41.7	49.2	85.6	51.5	88.3	62.6	37.5	87.6	61.6	62.3	49.9	41.8	59.0	60.3
ENT [52]	40 pixels	93.1	55.9	85.4	35.6	30.6	30.0	28.3	39.0	86.8	45.9	88.5	65.5	32.0	88.0	55.9	54.7	27.7	39.4	62.8	55.0
SCONF [10]	40 pixels	92.1	56.2	86.2	38.4	36.2	37.8	41.4	48.1	87.4	46.8	87.8	67.1	39.1	88.6	57.5	56.6	45.7	47.5	63.0	59.1
Ours (PA)	40 pixels	**95.6**	69.6	**88.0**	**47.3**	**45.1**	**37.8**	**45.9**	**56.5**	**88.2**	**54.2**	**89.0**	**69.7**	**45.4**	**90.9**	**67.0**	**69.9**	**54.1**	**52.4**	**65.8**	**64.9**
RAND	2.2%	95.3	72.7	86.8	45.3	43.7	38.4	45.2	53.1	87.2	54.3	90.0	65.6	42.5	59.3	67.8	67.1	**59.2**	45.0	63.2	63.8
ENT [52]	2.2%	94.8	71.1	87.3	52.3	46.1	38.7	47.2	56.3	87.9	55.2	89.3	69.5	47.9	90.5	**74.8**	71.2	58.6	52.7	66.5	66.2
SCONF [10]	2.2%	94.9	69.9	88.1	52.0	50.0	40.4	49.7	59.4	88.1	55.8	89.7	71.1	49.9	90.7	71.6	69.7	52.5	53.1	67.4	66.5
Ours (RA)	2.2%	**96.9**	**76.2**	**89.9**	**55.5**	**52.4**	**44.6**	**54.5**	**63.8**	**89.9**	**57.0**	**92.1**	**73.1**	**52.7**	**92.3**	71.9	**72.7**	41.0	**55.8**	**70.1**	**68.5**

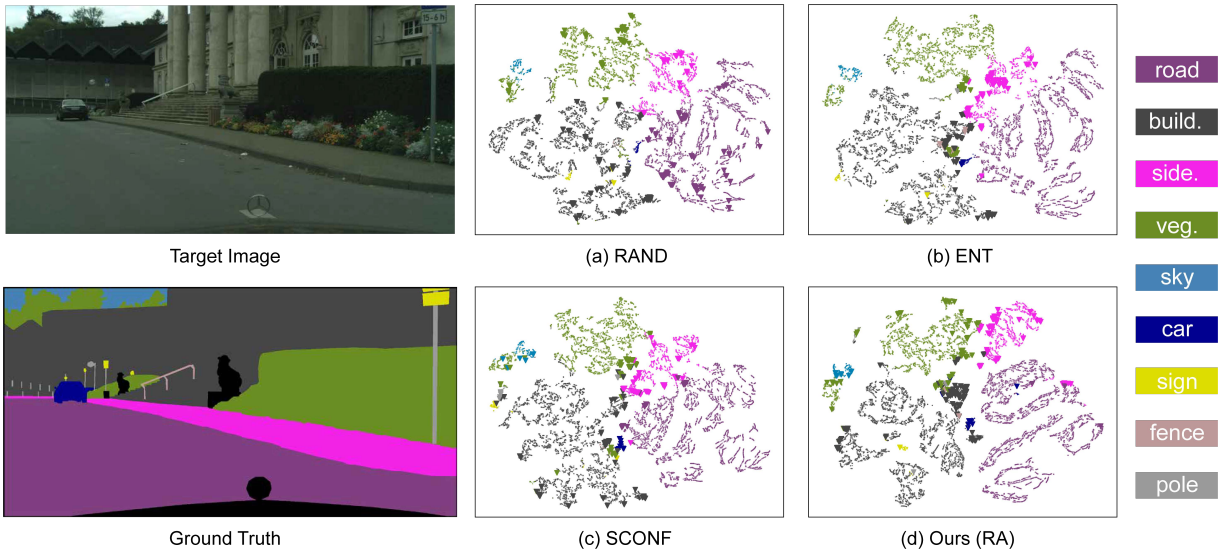

Figure 8. t-SNE visualization [68] of queried regions from Cityscapes training set on the task GTAV → Cityscapes. Compared to RAND, ENT, SCONF, Ours (RA) is able to select the most diverse and uncertain regions of an image. Large triangles ▼ denote the selected pixels and small points • are the remaining pixels in a target image. Please zoom in to see the details.

D. Comparison of different active selection methods

D.1. Details about comparison methods and performance per class

The performance of active learning depends principally upon depends the selection strategies. In the main paper (Table 4), we compare Ours (RA) and Ours (PA) with other common selection methods such as Random selection (RAND), entropy (ENT) [52] and softmax confidence (SCONF) [10] on GTAV → Cityscapes. All methods do not use additional loss $\mathcal{L}_{cr}^{s}$ or $\mathcal{L}_{nl}^{t}$.

Rand selection (RAND): pixels and regions are randomly sampled with equal probability form each target image.

Entropy (ENT) [52]: pixels with the highest prediction entropy, i.e., $-\sum_{c=1}^{C} \mathbf{P}_t^{(i,j,c)} \log \mathbf{P}_t^{(i,j,c)}$ are sampled for PA. And regions with the highest average prediction entropy of all pixels in a region are sampled for RA.

Softmax confidence (SCONF) [10]: query the most unsure pixels by the softmax confidence score for PA

$$1 - \max_{c \in \{1,\cdots,C\}} \mathbf{P}_t^{(i,j,c)},$$

where larger value indicates less confident. And for RA, select regions with the largest average score of all pixels in a region.

In Table 11, we extend the results of Table 4 by adding the per class IoU for each method. Indeed, our methods select more regions or pixels belonging to the majority classes than baseline methods. Note that Ours (RA) works specially well for rare object categories, such as "fence", "pole", "sign" or "rider", among others, which is a side effect of directly optimizing for per class IoU and mean IoU.

D.2. t-SNE visualization

In Fig. 8, we illustrate the sampling behavior of Ours (RA) with different selection strategies via t-SNE visualization [68]. We visualize the feature representations of pixels sampled via RAND, ENT, SCONF and Ours (RA) (large,

Table 12. **Quantitative evaluation on GTAV → Cityscapes.** Results are based on DeepLab-v2 with ResNet-101 architecture. SF indicates whether the method supports source-free adaptation. Best results are shown in **bold**.

Method	SF	Budget	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	mIoU
URMA [59]	✓	-	92.3	55.2	81.6	30.8	18.8	37.1	17.7	12.1	84.2	35.9	83.8	57.7	24.1	81.7	27.5	44.3	6.9	24.1	40.4	45.1
LD [82]	✓	-	91.6	53.2	80.6	36.6	14.2	26.4	31.6	22.7	83.1	42.1	79.3	57.3	26.6	82.1	41.0	50.1	0.3	25.9	19.5	45.5
SFDA (w/ cPAE) [27]	✓	-	91.7	53.4	86.1	37.6	32.1	37.4	38.2	35.6	86.7	48.5	89.9	62.6	34.3	87.2	51.0	50.8	4.2	42.7	53.9	53.4
Ours (RA)	✓	2.2%	95.9	**76.2**	88.4	45.4	47.8	42.1	53.0	**62.8**	88.6	56.6	91.4	72.1	52.2	91.2	59.5	74.2	55.0	54.4	68.3	67.1
Ours (RA)	✗	2.2%	**96.5**	74.1	**89.7**	**53.1**	**51.0**	**43.8**	**53.4**	62.2	**90.0**	**57.6**	**92.6**	**73.0**	**53.0**	**92.8**	**73.8**	**78.5**	**62.0**	**55.6**	**70.0**	**69.6**

Table 13. **Quantitative evaluation on SYNTHIA → Cityscapes.** Results are based on DeepLab-v2 with ResNet-101 architecture. We report the mIoUs in terms of 13 classes (excluding the "wall", "fence", and "pole") and 16 classes. Best results are shown in **bold**.

Method	SF	Budget	road	side.	buil.	wall*	fence*	pole*	light	sign	veg.	sky	pers.	rider	car	bus	motor	bike	mIoU	mIoU*
URMA [59]	✓	-	59.3	24.6	77.0	14.0	1.8	31.5	18.3	32.0	83.1	80.4	46.3	17.8	76.7	17.0	18.5	34.6	39.6	45.0
LD [82]	✓	-	77.1	33.4	79.4	5.8	0.5	23.7	5.2	13.0	81.8	78.3	56.1	21.6	80.3	49.6	28.0	48.1	42.6	50.1
SFDA (w/ cPAE) [27]	✓	-	90.5	50.0	81.6	13.3	2.8	34.7	25.7	33.1	83.8	89.2	66.0	34.9	85.3	53.4	46.1	46.6	52.0	60.1
Ours (RA)	✓	2.2%	96.6	75.9	89.0	**49.2**	46.6	40.2	48.4	60.7	89.8	91.7	70.0	48.8	92.3	**81.1**	51.0	67.8	68.7	74.1
Ours (RA)	✗	2.2%	**96.8**	**76.6**	**89.6**	45.0	**47.7**	**45.0**	**53.0**	**62.5**	**90.6**	**92.7**	**73.0**	**52.9**	**93.1**	80.5	**52.4**	**70.1**	**70.1**	**75.7**

triangles, ▼) along with the remaining pixels (small, points, •) in a target image. We clearly observe that the RAND baseline performs average sampling regions in Fig. 8(a), which can waste the annotation budget on labeling redundant areas within objects, such as "road" and "building". Fig. 8(b) and Fig. 8(c) show that ENT and SCONF chose most uncertain regions, but seldom chose the infrequent object categories such as "sign", "fence" and "pole" which can be selected via Ours (RA) in Fig. 8(d). Across all strategies, we find that our method samples regions that are diverse (often present in a region with much more object categories) and uncertain (often present in a cluster of unpredictable object boundaries).

In short, Ours (RA) is the best option among various possible selection strategies regarding both the performance gain (Table 11) and visual presentation (Fig. 8).

E. Extension of RIPU to source-free scenario

Active domain adaptation, which achieves enormous performance gains at the expense of annotating a few target samples, has attracted a surge of interest due to its utility. Considering the data sensitivity and security issues, we further evaluate the generalization of our RIPU to a challenging scenario called source-free domain adaptation (SFDA), where only a source domain pre-trained model and unlabeled target data are accessible to conduct adaptation [27]. In SFDA extension, we start from a source domain pre-trained model, then we optimize the model with $\mathcal{L}_{CE}^{t}$ of active samples and $\mathcal{L}_{nl}^{t}$ of target data, without utilizing the source domain. We adopt the DeepLab-v2 [5] with the backbone ResNet-101 and carry out experiment on both two popular domain adaptation benchmarks, i.e., GTAV → Cityscapes and SYNTHIA → Cityscapes, with the annotation budget of 2.2% regions per image. With respect to training procedure, we keep in line with details of the main paper, such as learning rate schedule, batch size, maximum iteration, and input size etc.

As shown in Table 12 and Table 13, with little workload to manually annotate active regions in a target image, Ours (RA) achieves significant improvements compared to existing SFDA approaches [27, 59, 82], in detail, 13.7∼22.0 mIoU on GTAV → Cityscapes and 16.7∼29.1 mIoU on SYNTHIA → Cityscapes. These results suggest that our method better facilitates the performance on SFDA. In addition, we can observe a slight performance degradation without source data participating during the training process. In a nutshell, RIPU can be well generalized to SFDA and shows great potential for further exploration of performance increases.

F. Additional qualitative results

We follow the same conventions as Fig. 3 and Fig. 4 of the main paper, and present additional results for qualitative comparisons under various settings, including GTAV → Cityscapes (Fig. 9 and Fig. 11), SYNTHIA → Cityscapes (Fig. 10 and Fig. 12).

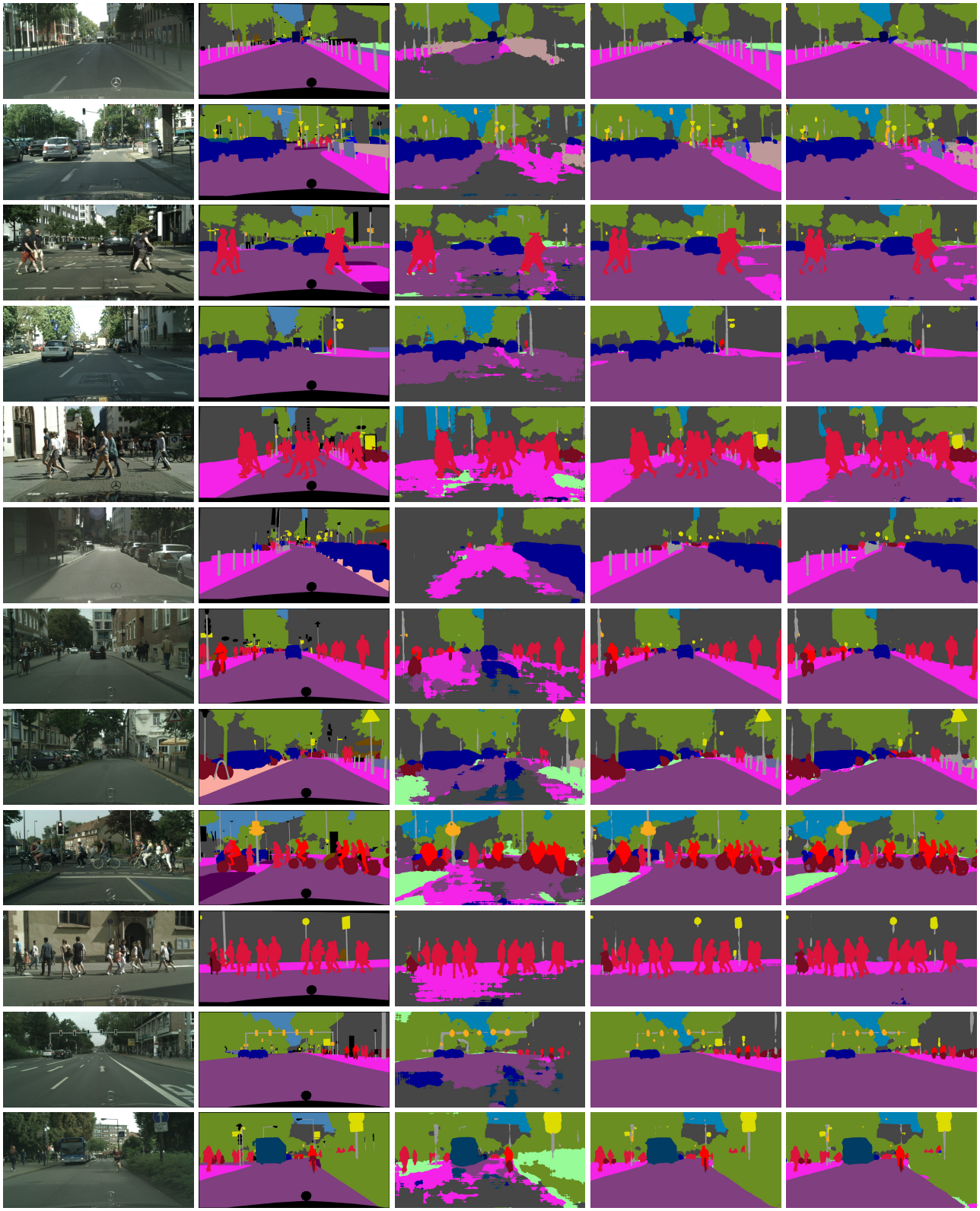

Figure 9. **Visualization of segmentation results on GTAV → Cityscapes.** From left to right: original target image, ground-truth label, result predicted by Source Only model, result predicted by Ours (RA), and result predicted by Ours (PA) are shown one by one.

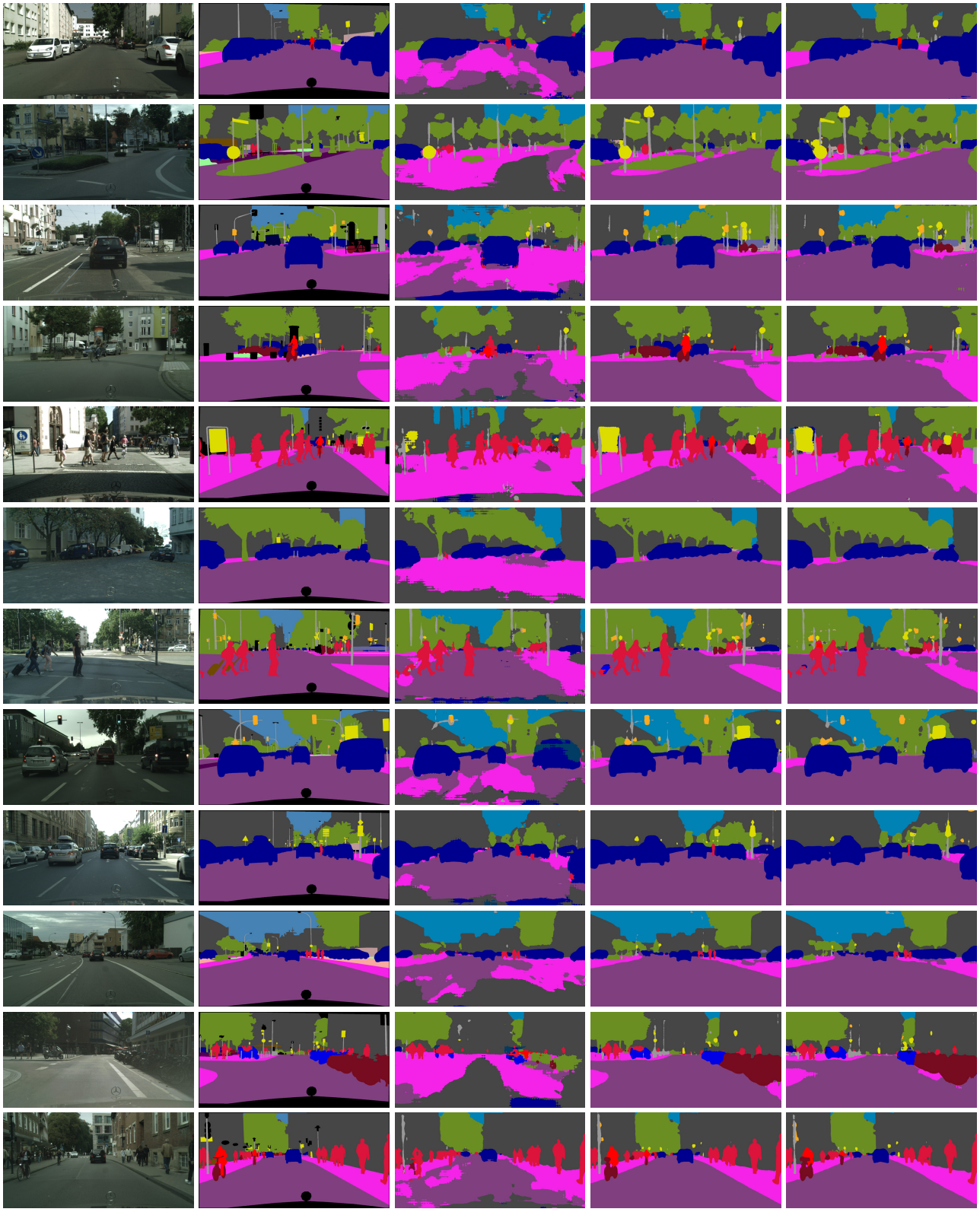

Figure 10. **Visualization of segmentation results on SYNTHIA → Cityscapes.** From left to right: original target image, ground-truth label, result predicted by Source Only model, result predicted by Ours (RA), and result predicted by Ours (PA) are shown one by one.

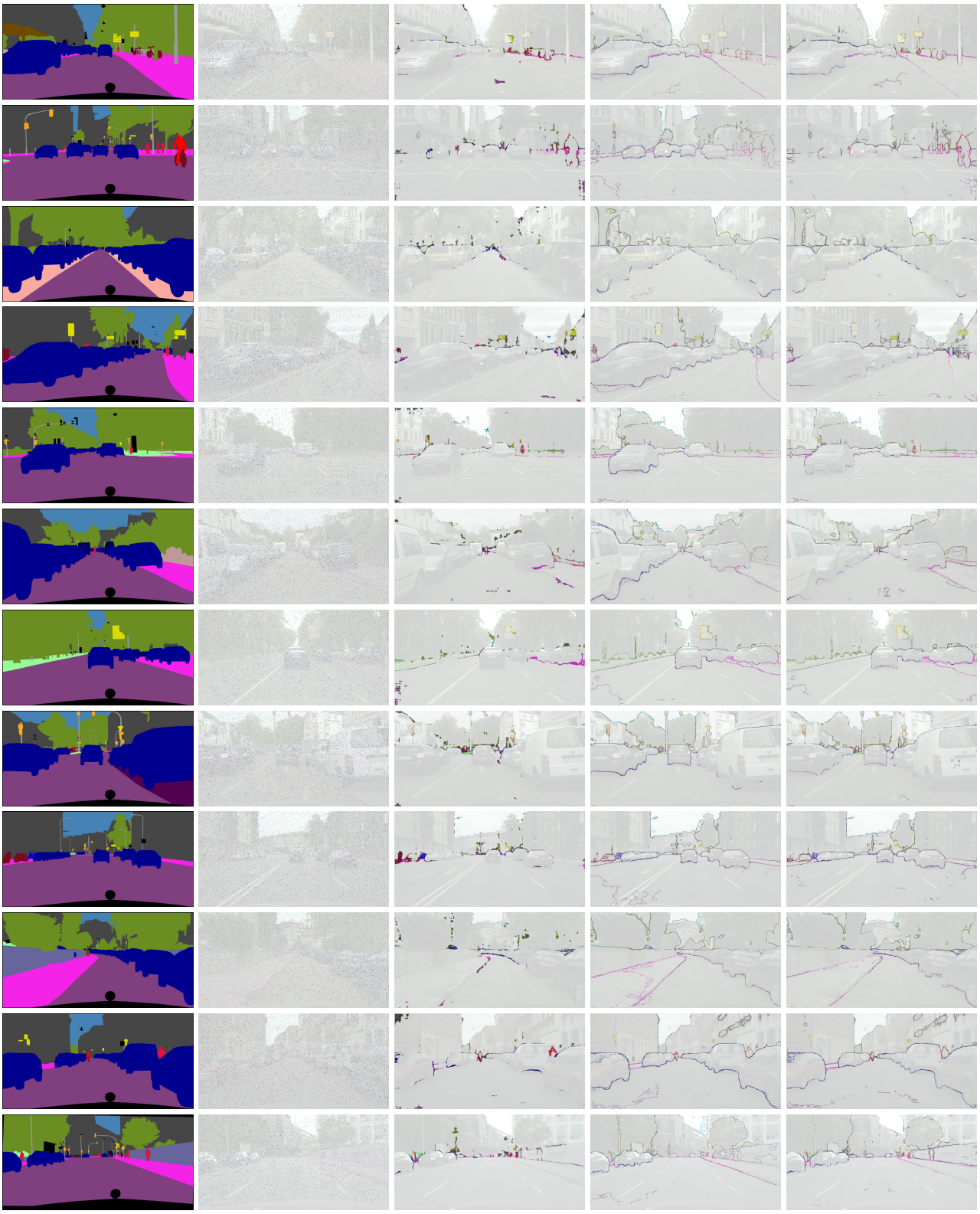

Figure 11. **Visualization of queried regions to annotate (2.2%) on GTAV → Cityscapes.** Compared to RAND, Uncertainty Only, and Impurity Only, Ours (RA) is able to select the most diverse and uncertain regions of an image. Please zoom in to see the details.

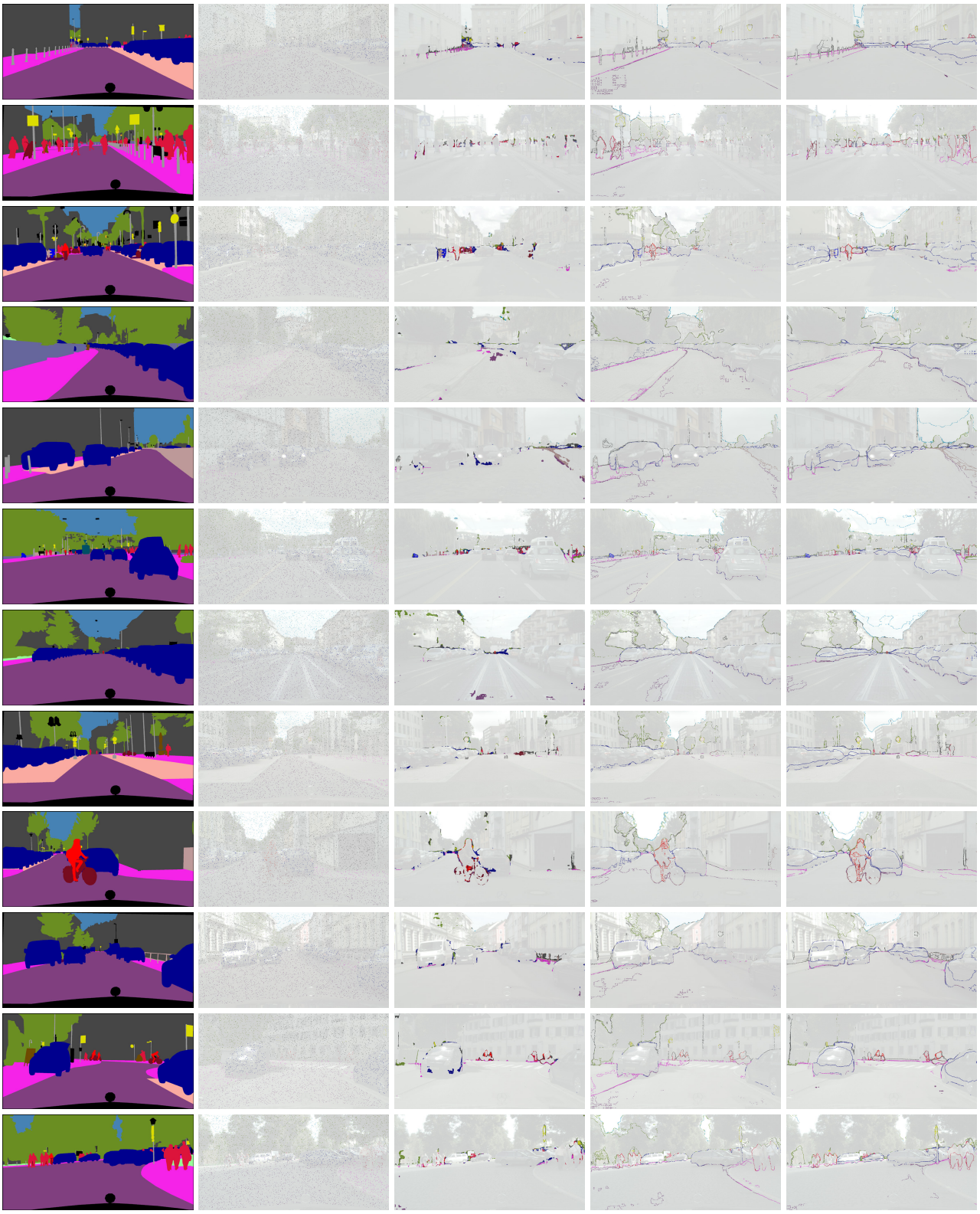

Figure 12. **Visualization of queried regions to annotate (2.2%) on SYNTHIA → Cityscapes.** Compared to RAND, Uncertainty Only, and Impurity Only, Ours (RA) is able to select the most diverse and uncertain regions of an image. Please zoom in to see the details.