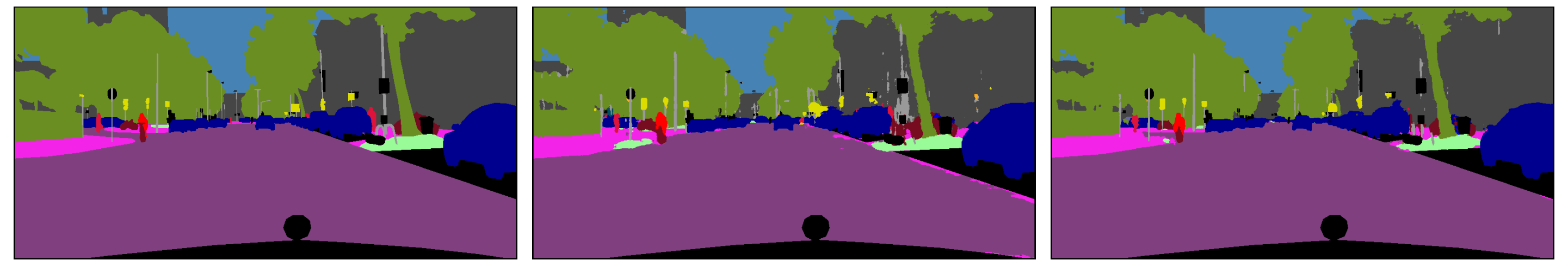
I Bet You Are Wrong:
Gambling Adversarial Networks for Structured Semantic Segmentation

Laurens Samson[1,2], Nanne van Noord[2], Olaf Booij[1], Michael Hofmann[1], Efstratios Gavves[2] and Mohsen Ghafoorian[1]
TomTom[1]/University of Amsterdam[2], The Netherlands

Overview: Structured Prediction

Deep neural networks have obtained significant success in semantic segmentation. Despite these achievements, convolutional neural networks do not possess a mechanism to capture high-level structures, such as continuity and neighbouring consistency. The disability to capture high-level qualities stems from the fact that the network addresses the problem as a pixel-wise classification task.



Labels — Pixel level loss — Gambling nets

Since pixel level losses are not inherently able to enforce these high-level qualities, other methods often address this with:

- **Hand-crafted post-processing**: ad-hoc and domain specific, often computationally expensive.
- **Conditional random fields**: partial coverage of consistencies, extra computational burden at inference time.
- **Additional engineered loss terms**: ad-hoc and domain specific, often tricky to formulate differentiable loss terms.
- **Adversarial training**: not easy to train, loss of uncertainty notion, value-based discrimination.

Conventional Adversarial Semantic Segmentation

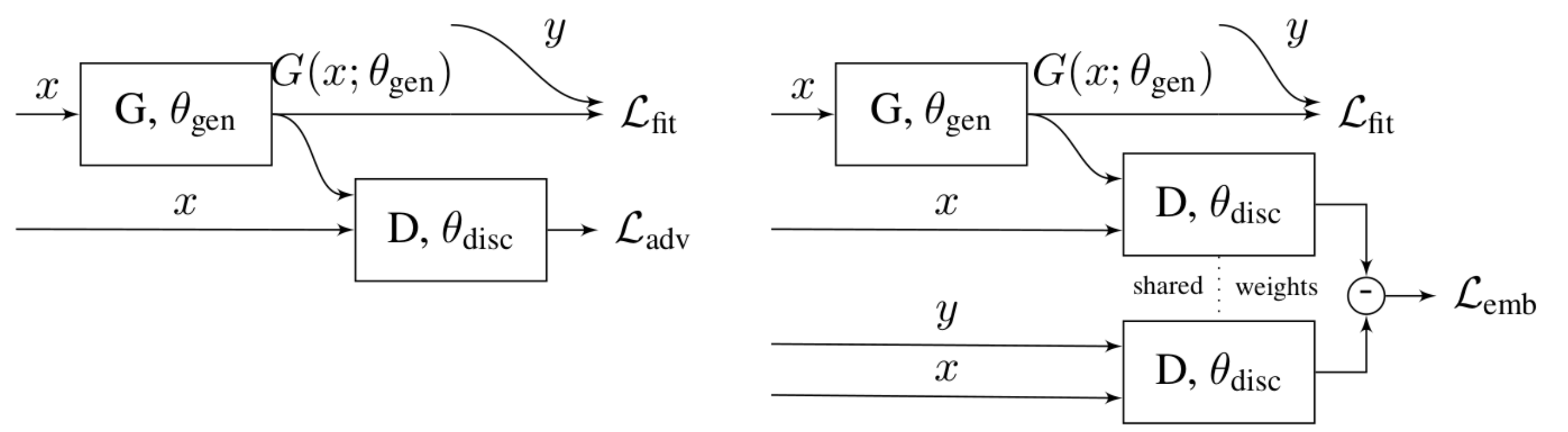


Figure: Schematic of adversarial semantic segmentation. Left: conditional GAN, right: EL-GAN.

Adversarial training [1, 2, 3] loss terms:

$$\mathcal{L}_{gen}(x,y;\theta_{gen},\theta_{disc}) = \mathcal{L}_{ce}(G(x;\theta_{gen}),y) + \mathcal{L}_{adv}(D(x,G(x,\theta_{gen});\theta_{disc}),1), \quad (1)$$

$$\mathcal{L}_{disc}(x,y;\theta_{gen},\theta_{disc}) = \mathcal{L}_{adv}(D(x,y;\theta_{disc}),1) + \mathcal{L}_{adv}(D(x,G(x;\theta_{gen});\theta_{disc}),0). \quad (2)$$

Gambling Adversarial Networks

Despite of success in previous work with adversarial training in semantic segmentation, we foresee two potential issues in the current setup:

1. **Value-based discrimination**: The discriminator learns to distinguish the real from the fake distribution by evaluating the numerical values. The values in the ground-truth are either zeroes or ones (one-hot vector), whereas the values of the generated predictions range between zero and one (softmax vector).
2. **Generator faking certainty**: The generator tries to mimic the one-hot vectors in order to fool the discriminator, which leads to the loss of the ability to express uncertainty in the predictions of the generator.

Similar to conventional adversarial training, in gambling adversarial networks a mini-max game is played between two players, the segmenter and the gambler.

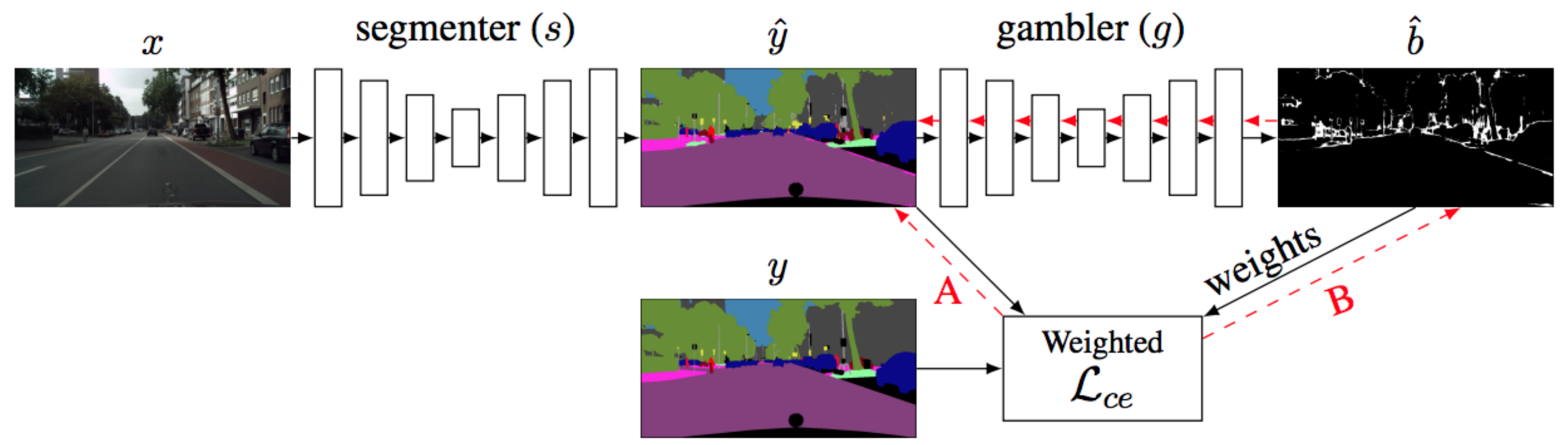


Figure: An overview of gambling adversarial networks. The solid black arrows indicate the forward pass. The red dashed arrows represent the two gradient flows of the weighted cross-entropy loss. Gradient flow A provides pixel-level feedback independent of other pixel predictions. Gradient flow B, going through the gambler network, enables feedback reflecting the inter-pixel and structural consistency.

1. **The segmenter**: The segmenter aims to generate images such that the gambler has no obvious clues to localize incorrect predictions.

$$\mathcal{L}_s(x,y;\theta_s,\theta_g) = \mathcal{L}_{ce}(s(x;\theta_s),y) - \mathcal{L}_g(x,y;\theta_s,\theta_g). \quad (3)$$

2. **The gambler**. Given a limited investment budget, the gambler predicts an image-sized betting map, where high bets indicate pixels that are likely incorrectly classified, given the contextual prediction clues around it.

$$\mathcal{L}_g(x,y;\theta_s,\theta_g) = -\frac{1}{wh}\sum_{i,j}^{w,h} g(x,s(x;\theta_s);\theta_g)_{i,j}\mathcal{L}_{ce}(s(x;\theta_s)_{i,j},y_{i,j}). \quad (4)$$

The budget of the gambler is limited by changing the betting maps into a smoothed probability distribution:

$$g(x,\hat{y};\theta_g)_{i,j} = \frac{g_\sigma(x,\hat{y};\theta_g)_{i,j}+\beta}{\sum_{k,l}^{w,h} g_\sigma(x,\hat{y};\theta_g)_{k,l}+\beta}. \quad (5)$$

Experiments and Results

Recovered uncertainty representation:

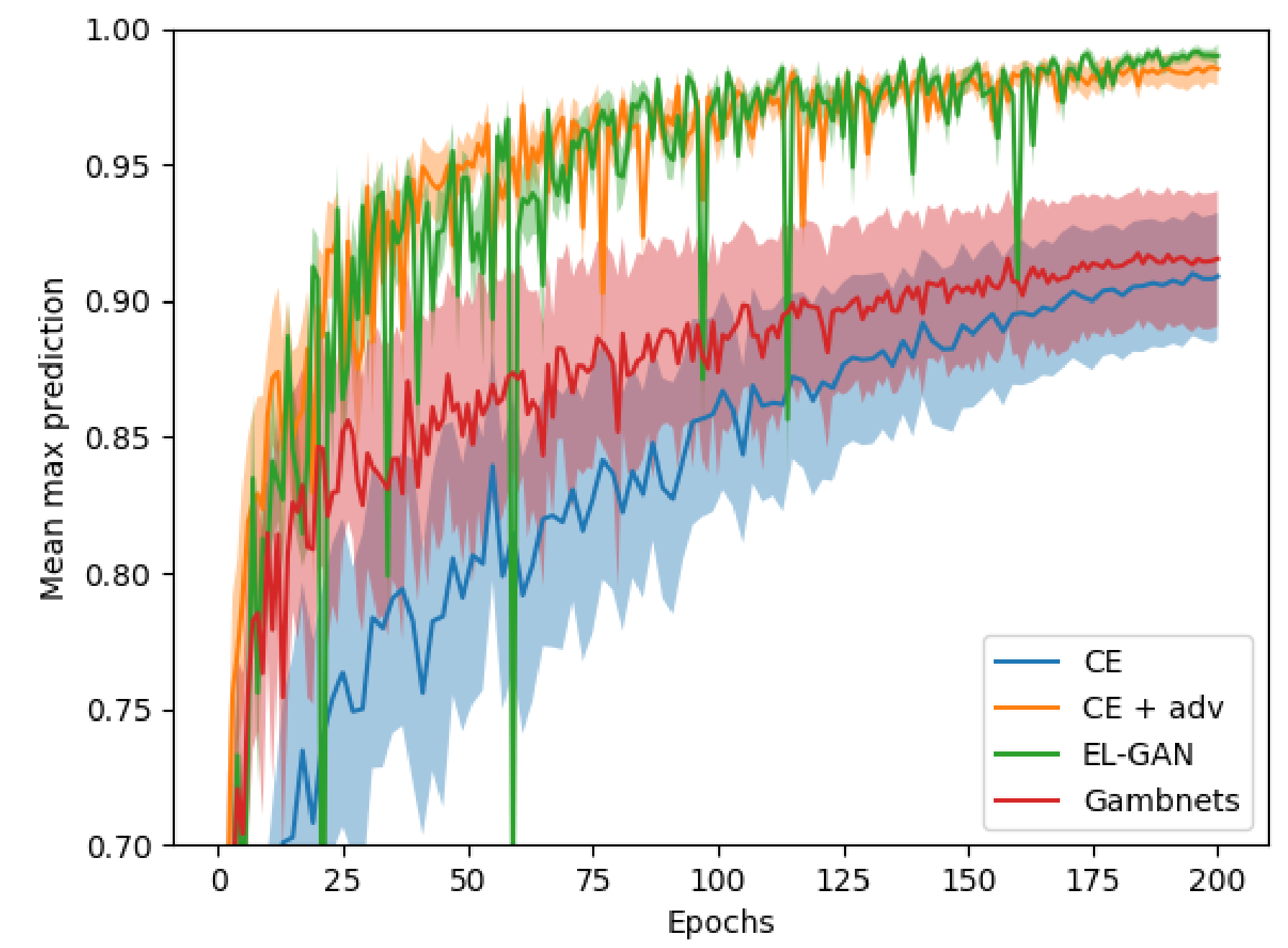


Figure: Mean maximum class-likelihoods (mean confidence) over time on the Cityscapes validation set. Solid central curves and the surrounding shaded area represent the mean and standard deviation respectively.

Used metrics for structured semantic segmentation:

- Modified Hausdorff (Chamfer) distance:

$$\mathrm{d}_H(X,Y) = \frac{1}{2}\sum\left\{\frac{1}{|X|}\sum_{x\in X}\inf_{y\in Y}\mathrm{d}(x,y), \frac{1}{|Y|}\sum_{y\in Y}\inf_{x\in X}\mathrm{d}(x,y)\right\}, \quad (6)$$

- BF-score:

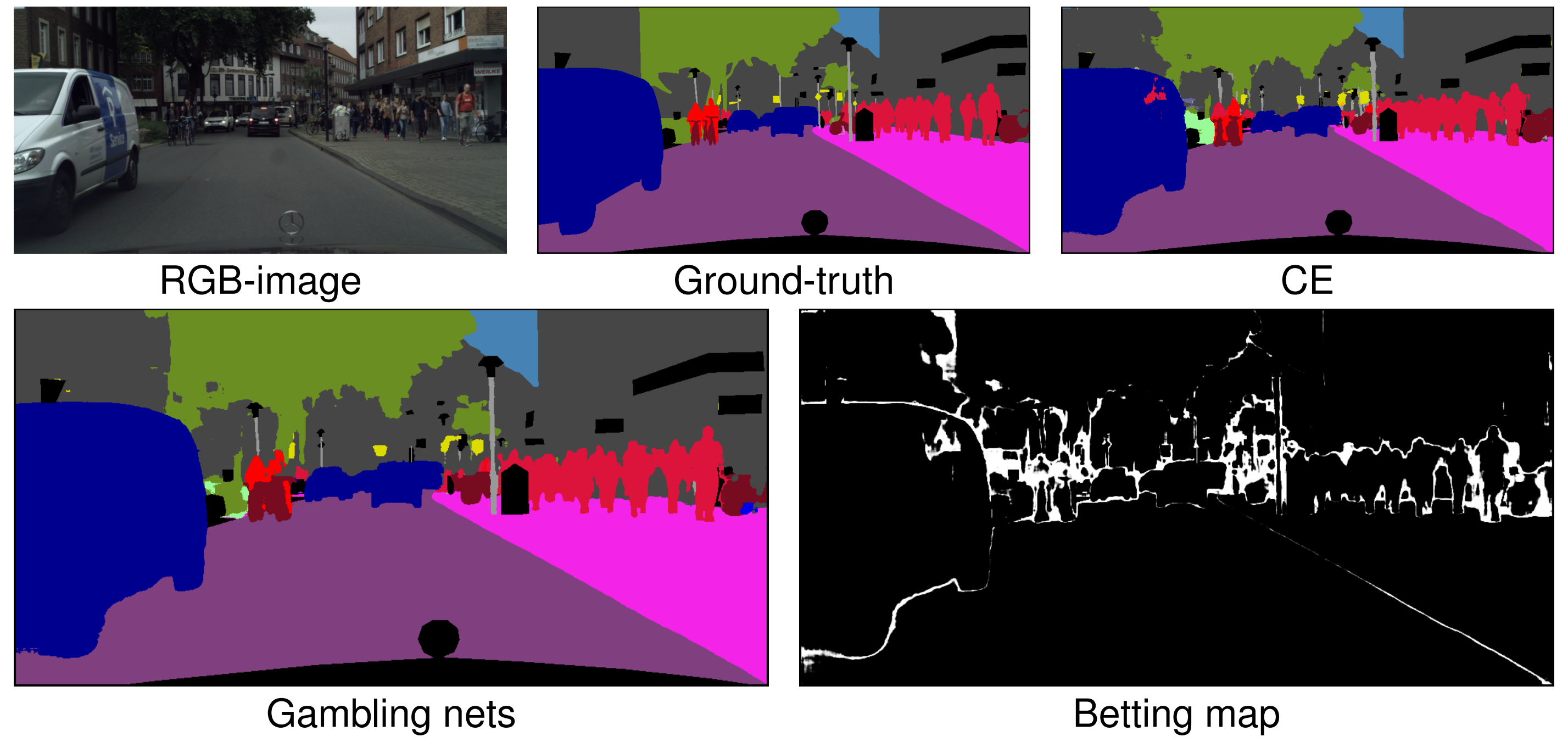
$$d_\theta(X,Y) = \frac{1}{|X|}\sum_{x\in X}[[\inf_{y\in Y} d(x,y) < \theta]] \quad (7)$$

$$BF(X,Y) = \frac{2d_\theta(X,Y)d_\theta(Y,X)}{d_\theta(X,Y)+d_\theta(Y,X)}, \quad (8)$$

where X and Y are the boundaries of corresponding classes for prediction and ground-truth.

Method	Mean IoU	BF-score	Hausdorff
CE	52.7	49.0	36.8
Focal loss [4]	56.2	55.3	30.2
CE + adv [2]	56.3	57.3	31.3
EL-GAN [3]	55.4	54.2	31.6
Gambling nets	**57.9**	**58.5**	**27.6**

Table: Results on Cityscapes with a U-Net based architecture as segmentation network.



RGB-image — Ground-truth — CE — Gambling nets — Betting map

Figure: Qualitative results on Cityscapes with a U-Net based architecture.

Method	road	swalk	build	wall	fence	pole	tlight	sign	veg.	ter.	sky	pers	rider	car	truck	bus	train	mbike	bike	mean
CE	95.2	68.4	84.4	26.0	30.9	43.0	38.9	51.3	87.2	50.3	91.5	59.0	32.6	85.5	22.8	43.2	**19.2**	15.4	57.4	57.2
Focal loss [4]	96.0	71.3	87.1	32.2	34.9	48.6	47.6	57.8	88.9	54.2	92.7	62.9	33.5	87.2	28.5	**47.5**	18.3	19.3	60.0	56.2
CE + adv [2]	95.9	72.7	83.5	28.9	35.2	49.8	47.8	59.3	89.0	54.8	92.3	66.4	38.4	87.2	27.8	41.4	15.3	20.3	62.5	56.3
EL-GAN [3]	96.1	71.1	86.8	**33.5**	37.0	48.7	46.6	57.3	88.9	53.6	92.9	62.6	34.4	87.1	26.0	38.3	16.3	17.8	58.9	55.4
Gambling	**96.3**	**73.0**	**87.6**	33.4	**39.1**	**52.9**	**51.3**	**61.9**	**89.7**	**55.8**	**93.1**	**68.1**	**38.9**	**88.7**	**30.3**	40.2	11.5	**24.8**	**63.2**	**57.9**

Table: IoU per class on the validation set of Cityscapes with a U-Net based architecture as segmentation network.

Table: Results on Cityscapes with a PSPNet as segmentation network.

Method	Mean IoU	BF-score	Hausdorff
CE	72.4	69.0	19.4
Focal loss [4]	71.5	67.4	21.2
CE + adv [2]	68.0	67.0	20.9
EL-GAN [3]	71.3	67.0	21.2
Gambling nets	**73.1**	**70.1**	**18.7**

Table: Results on Camvid with a PSPNet as segmentation network.

Method	Mean IoU	BF-score	Hausdorff
CE	72.5	71.8	17.9
Focal loss [4]	70.8	71.4	17.7
CE + adv [2]	**72.7**	72.7	17.1
EL-GAN [3]	70.1	69.6	19.1
Gambling nets	72.1	**73.8**	**16.0**

Method	road	swalk	build	wall	fence	pole	tlight	sign	veg.	ter.	sky	pers	rider	car	truck	bus	train	mbike	bike
CE	84.8	69.0	77.3	15.6	13.7	66.4	31.3	53.7	82.3	28.7	82.0	47.5	29.2	76.0	8.3	12.2	2.6	8.9	44.1
Focal loss [4]	87.2	72.4	80.7	19.7	16.0	71.0	40.1	62.3	86.1	**35.8**	84.8	51.6	32.0	79.4	9.2	18.0	4.3	12.1	50.5
CE + adv [2]	82.6	72.3	79.8	16.2	16.2	72.1	43.6	65.7	86.2	34.5	83.3	54.8	34.4	78.8	8.7	17.5	4.4	14.0	52.0
EL-GAN [3]	86.9	72.3	79.9	19.3	16.4	70.7	38.2	63.4	85.5	32.7	84.0	51.2	32.7	78.1	9.5	16.8	**4.8**	8.8	47.0
Gambling	**87.4**	**74.3**	**81.3**	**20.7**	**18.6**	**74.0**	**45.7**	**67.8**	**87.2**	35.4	**85.4**	**57.0**	**38.8**	**80.0**	**11.2**	**19.3**	4.4	**15.6**	**52.9**

Table: BF-score per class on the validation set of Cityscapes with U-Net based architecture as segmentation network.

References

[1] P. Isola *et al.*, “Image-to-image Translation with cGANs,” in *CVPR*, 2017.

[2] P. Luc *et al.*, “Semantic Segmentation using Adversarial Networks,” in *NIPS-W*, 2016.

[3] M. Ghafoorian *et al.*, “EL-GAN: embedding loss driven GANs for lane detection,” in *ECCV*, 2018.

[4] T.-Y. Lin *et al.*, “Focal loss for dense object detection,” in *ICCV*, 2017.

Our Paper: https://arxiv.org/abs/1908.02711