Gated-SCNN

Gated Shape CNNs for Semantic Segmentation

Contents

- Motivation
- Contribution
- Architecture
- Loss function and regularizer
- result

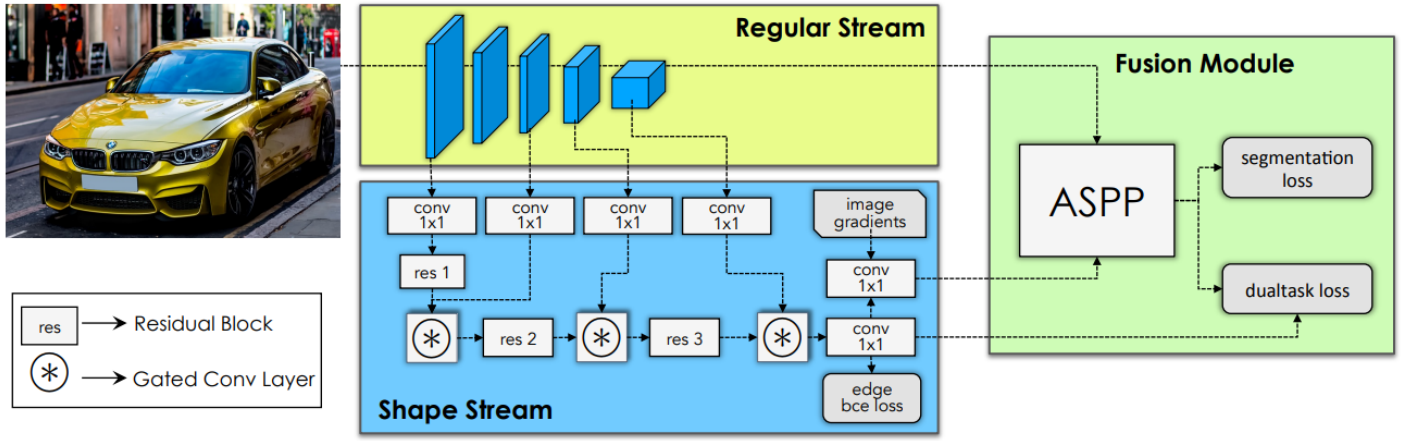
Motivation

- Processing the color, shape, and texture information all together may not be ideal.

Contribution

- Proposing Gated-SCNN(GSCNN), a new two-stream CNN architecture for semantic segmentation that wires shape into a separate parallel stream
- Using new dual task regularizer loss function

Architecture



apply GCL, we first obtain an attention map $\alpha_t \in \mathbb{R}^{H \times W}$ by concatenating r_t and s_t followed by a normalized 1×1 convolutional layer $C_{1\times1}$ which in turn is followed by a sigmoid function σ :

$$\alpha_t = \sigma(C_{1\times1}(s_t||r_t)), \tag{1}$$

where $||$ denotes concatenation of feature maps. Given the attention map α_t, GCL is applied on s_t as an element-wise product $\odot$ with attention map α followed by a residual connection and channel-wise weighting with kernel w_t. At each pixel (i, j), GCL $\circledast$ is computed as

$$\begin{aligned} \hat{s}_t^{(i,j)} &= (s_t \circledast w_t)_{(i,j)} \\ &= ((s_{t_{(i,j)}} \odot \alpha_{t_{(i,j)}}) + s_{t_{(i,j)}})^T w_t. \end{aligned} \tag{2}$$

Loss function and regularizer

$\zeta \in R^{H \times W}$ be a potential that represents whether a particular pixel belongs to a semantic boundary in the input image I. It is computed by taking a spatial derivative on segmentation output as follows:

$$\zeta = \frac{1}{\sqrt{2}} ||\nabla (G * \arg\max_k p(y^k | r, s))|| \tag{4}$$

where G denotes Gaussian filter. If we assume $\hat{\zeta}$ is a GT binary mask computed in the same way from the GT semantic labels $\hat{f}$, we can write the following loss function:

$$\mathcal{L}_{reg_{\rightarrow}}^{\theta\,\phi,\gamma} = \lambda_3 \sum_{p^+} |\zeta(p^+) - \hat{\zeta}(p^+)| \tag{5}$$

where p^+ contains the set of all non-zero pixel coordinates in both ζ and $\hat{\zeta}$. Intuitively, we want to ensure that boundary pixels are penalized when there is a mismatch with GT boundaries, and to avoid non-boundary pixels to dominate the loss function. Note that the above regularization loss function exploits the duality between boundary prediction and semantic segmentation in the boundary space.

Loss function and regularizer

Similarly, we can use the boundary prediction from the shape stream $s \in \mathbb{R}^{H \times W}$ to ensure consistency between the binary boundary prediction s and the predicted semantics $p(y|r, s)$:

$$\mathcal{L}_{reg_{\leftarrow}}^{\theta\,\phi,\gamma} = \lambda_4 \sum_{k,p} \mathbb{1}_{s_p} [\hat{y}_p^k \log p(y_p^k | r, s)], \tag{6}$$

where p and k runs over all image pixels and semantic classes, respectively. $\mathbb{1}_s = \{1 : s > thrs\}$ corresponds to the indicator function and *thrs* is a confidence threshold, we use 0.8 in our experiments. The total dual task regularizer loss function can be written as:

$$\mathcal{L}^{\theta\,\phi,\gamma} = \mathcal{L}_{reg_{\rightarrow}}^{\theta\,\phi,\gamma} + \mathcal{L}_{reg_{\leftarrow}}^{\theta\,\phi,\gamma} \tag{7}$$

Here, λ_3 and λ_4 are two hyper-parameters that control the weighting of the regularizer.

result

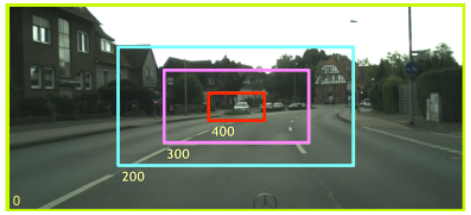


Figure 3: Illustration of the crops used for the distance-based evaluation.

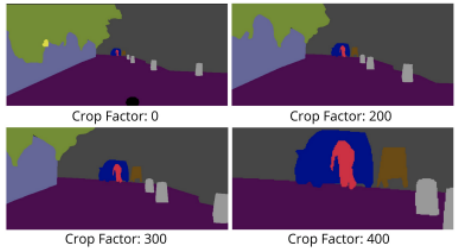


Figure 4: Predictions at diff. crop factors.

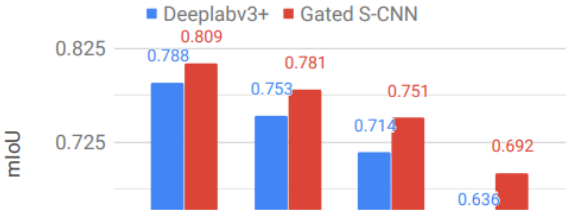


Figure 5: **Distance-based evaluation**: Comparison of mIoU at different crop factors.

Method	road	s.walk	build.	wall	fence	pole	t-light	t-sign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mean
LRR [18]	97.7	79.9	90.7	44.4	48.6	58.6	68.2	72.0	92.5	69.3	94.7	81.6	60.0	94.0	43.6	56.8	47.2	54.8	69.7	69.7
DeepLabV2 [9]	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.8	70.4
Piecewise [32]	98.0	82.6	90.6	44.0	50.7	51.1	65.0	71.7	92.0	72.0	94.1	81.5	61.1	94.3	61.1	65.1	53.8	61.6	70.6	71.6
PSP-Net [58]	98.2	85.8	92.8	57.5	65.9	62.6	71.8	80.7	92.4	64.5	94.8	82.1	61.5	95.1	78.6	88.3	77.9	68.1	78.0	78.8
DeepLabV3+ [11]	98.2	84.9	92.7	**57.3**	62.1	65.2	68.6	78.9	92.7	63.5	95.3	82.3	62.8	95.4	**85.3**	89.1	80.9	64.6	77.3	78.8
Ours (GSCNN)	**98.3**	**86.3**	**93.3**	55.8	**64.0**	**70.8**	**75.9**	**83.1**	**93.0**	**65.1**	**95.2**	**85.3**	**67.9**	**96.0**	80.8	**91.2**	**83.3**	**69.6**	**80.4**	**80.8**

Table 1: Comparison in terms of IoU vs state-of-the-art baselines on the Cityscapes val set.

Thrs	Method	road	s.walk	build.	wall	fence	pole	t-light	t-sign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mean
12px	DeepLabV3+	**92.3**	80.4	87.2	59.6	53.7	83.8	75.2	81.2	90.2	60.8	90.4	76.6	78.7	91.6	**81.0**	87.1	92.6	**81.8**	78.0	80.1
	Ours	92.2	**81.7**	**87.9**	**59.6**	**54.3**	**87.1**	**82.3**	**84.4**	**90.9**	**61.1**	**91.9**	**80.4**	**82.8**	**92.6**	78.5	**90.0**	**94.6**	79.1	**82.2**	**81.8**
9px	DeepLabV3+	91.2	78.3	84.8	58.1	52.4	82.1	73.7	79.5	87.9	59.4	89.5	74.7	76.8	90.0	**80.5**	86.6	92.5	**81.0**	75.4	78.7
	Ours	**91.3**	**80.1**	**86.0**	**58.5**	**52.9**	**86.1**	**81.5**	**83.3**	**89.0**	**59.8**	**91.1**	**79.1**	**81.5**	**91.5**	78.1	**89.7**	**94.4**	78.5	**80.4**	**80.7**
5px	DeepLabV3+	88.1	72.6	78.1	55.0	49.1	77.9	69.0	74.7	81.0	55.8	86.4	69.0	71.9	85.4	**79.4**	85.4	92.1	**79.4**	68.4	74.7
	Ours	**88.7**	**75.3**	**80.9**	**55.9**	**49.9**	**83.6**	**78.6**	**80.4**	**83.4**	**56.6**	**88.4**	**75.4**	**77.8**	**88.3**	77.0	**88.9**	**94.2**	76.9	**75.1**	**77.6**
3px	DeepLabV3+	83.7	65.1	69.7	52.2	46.2	72.0	62.8	67.7	71.8	52.0	80.9	61.5	66.4	78.8	**78.2**	83.9	91.7	**77.9**	60.9	69.7
	Ours	**85.0**	**68.8**	**74.1**	**53.3**	**47.0**	**79.6**	**74.3**	**76.2**	**75.3**	**53.1**	**83.5**	**69.8**	**73.1**	**83.4**	75.8	**88.0**	**93.9**	75.1	**68.5**	**73.6**

Table 2: Comparison vs baselines at different thresholds in terms of boundary F-score on the Cityscapes val set.