LEARNABLE CONTEXTUAL REGULARIZATION FOR SEMANTIC SEGMENTATION OF INDOOR SCENE IMAGES

Jun Chu[1],Xu Xiao[1,2], Gaofeng Meng[2], Lingfeng Wang[2] and Chunhong Pan[2]

1.Software College of Nanchang Hangkong University

2.National Laboratory of Pattern Recognition

Outline

I. ABSTRACT
II. PROPOSED MODEL
III. EXPERIMENTS
IV. CONCLUSIONS

ABSTRACT

Semantic segmentation of **indoor scene images** has a wide range of applications.

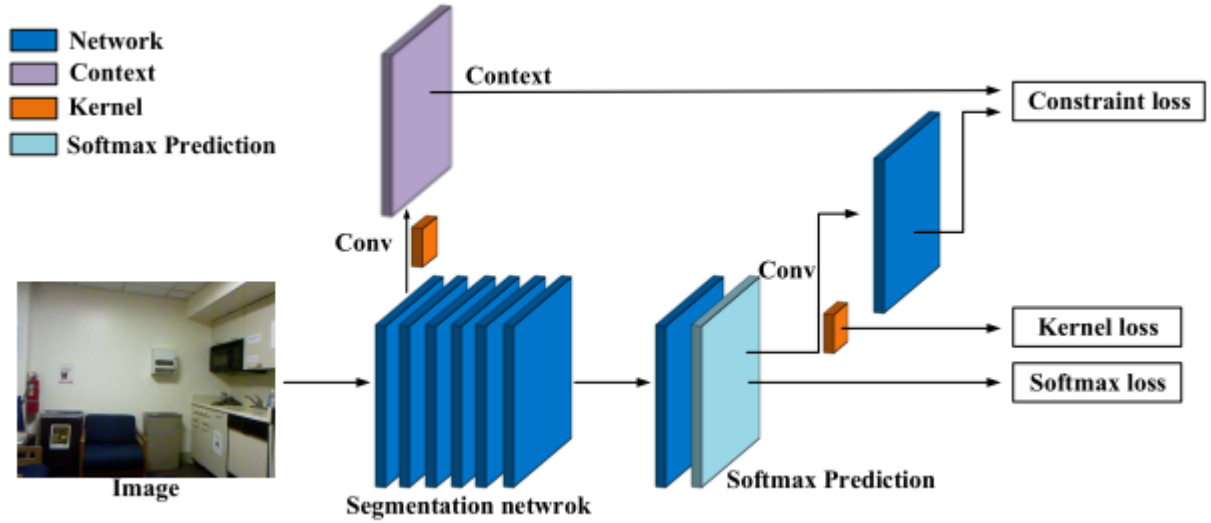
- **Challenges:** Due to a large number of classes and uneven distribution in indoor scenes, mislabels are often made when facing small objects or boundary regions.

- **Motivation:** Contextual information may benefit for segmentation results, but has not yet been exploited sufficiently. We propose a learnable contextual regularization model for enhancing the semantic segmentation results of color indoor scene images.

ABSTRACT

- Our main contributions:
  - Our model, derived from the inherent contextual regularization on the indoor scene objects, benefits much from the learnable constraint layers bridging the lower layers and the higher layers in the deep convolutional network.
  - The constraint layers are further integrated with a weighted L1-norm based contextual regularization between the neighboring pixels of RGB values to improve the segmentation results.
  - This regularization model is combined with a deep convolutional segmentation network without significantly increasing the number of additional parameters.

PROPOSED MODEL



- Our proposed model trained end-to-end to optimize the output semantic segmentation quality.
- The contextual regulation is modeled as a loss layer with some convolutional kernels to constrain the softmax predictions. The kernel of the contextual regulation layer is learnable, enabling the layer more robust and flexible to different types of indoor scenes.
- We will learn the context relationship both from the lowest layer and the highest segmentation output layer by means of weight sharing.

Contextual Regularization

An image can be considered as a combination of image patches in orders. We derive a contextual regularization from an input RGB image and use it to constrain the segmentation results. A weighting function W :

$$W(i, j)(f(i) - f(j)) \quad (1)$$

- i and j are two neighboring pixels, and f is the last output that will be regularized.
- $W \in [0,1]$, for $W(i, j) = 0$,the contextual constraint between $f(i)$ and $f(j)$ will be terminated.
- W plays a switch role to control whether the constraint between i and j will be canceled.

➢Contextual Regularization

how to choose a reasonable $W(i,j)$:

- if two neighboring pixels in original image have similar RGB value, weight function will put significant constraints on the two pixels.
- the more difference between the neighboring pixels the smaller value of the $W(i,j)$ is.

$$W(i,j) = e^{-\|I(i)-I(j)\|^2/2\sigma^2} \quad (2)$$

Integrating the weighted contextual constraints in the whole image domain leads to the following:

$$W_j(i) = e^{-\left(\sum_c \left|(D_j \otimes I^c)_i\right|^2\right)/2\sigma^2} \quad (3)$$

Learnable Constraint Loss Layer

The learnable constraint loss function evident from Eq. (3), we employ L 1 -norm which is more robust to outliers than L 2 -norm and the boundary effect is better. The discrete form of Eq. (3) as below:

$$\sum_{j\in\omega_i}\sum_{i\in \mathrm{I}}\omega_{ij}\left|(D_j \otimes f)\right|$$

more compactly

$$\sum_{j\in\omega}\left\|W_j \,\mathrm{o}(D_j \otimes f)\right\|_1$$

Learnable Constraint Loss Layer

The expression of weight maps is as follows:

$$W_j(i) = e^{-\left(\sum_c \left|(D_j \otimes I^c)_i\right|^2\right)/2\sigma^2}$$

with N-dimensional input *f = (f 1 · · · f N)* in our layer we can then formulate the constraint loss objective as:

$$l_c = \lambda \sum_{n=1}^{N} \sum_{j \in \omega} \left\| W_j \, o(D_j \otimes f) \right\|_1$$

where the constraint loss l_c is the accumulation of all pixels of the N-maps and λ is the super parameter.

Optimization Of Our Model

➢**Global Optimization :**

softmax loss in semantic segmentation considers pixel-wise loss for the right classified while ignores the wrong classified neighboring pixels that share similar colour values.

We introduce the contextual constraint loss, which can better utilize the local relationship of neighboring pixels and improve the initial network segmentation.

$$L = l_s + l_c$$

l_s is softmax loss, and l_c is our contextual regularization loss

Optimization Of Our Model

- **Local Optimization:**

need control the kernels,to learn the right context regularization features corresponding to the segmentation task. Let l_k be the convolution kernel loss:

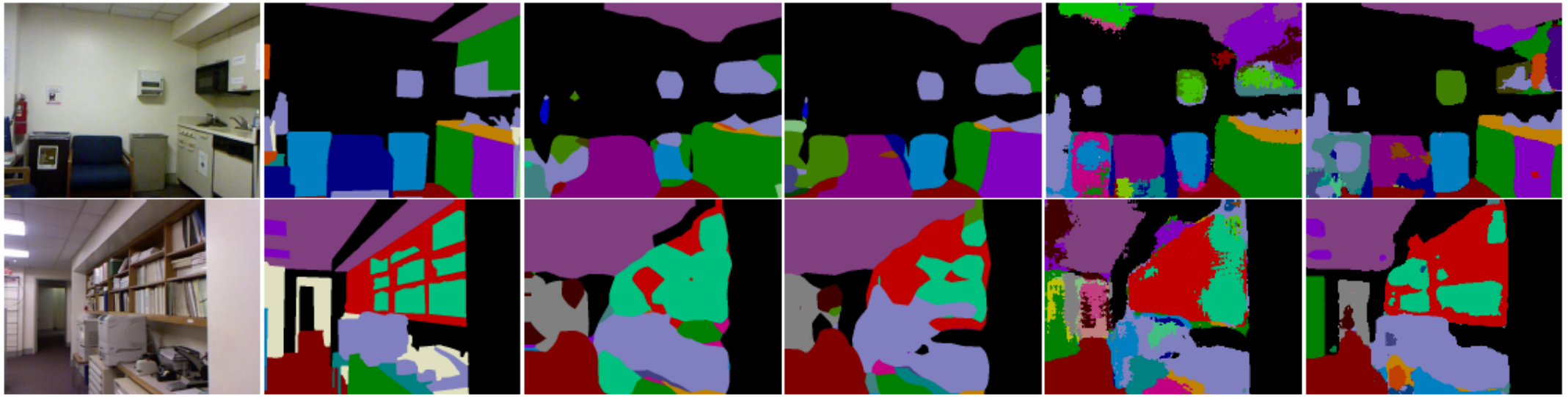
$$l_k = \sum_{i \in \varpi} \left\{ \alpha \left(\sum d_i^2 - 1 \right)^2 + \beta \sum d_i^2 \right\}$$

EXPERIMENTS

➢ **Dataset and Metrics**

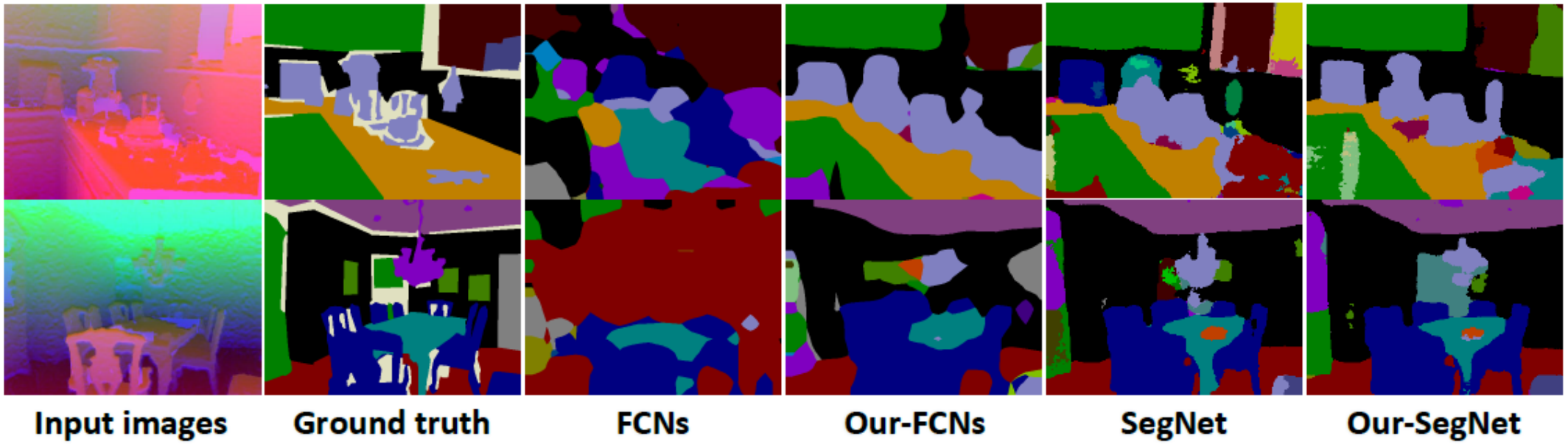
- NYUDv2 dataset :1,449 RGB-D， 795 training image， 654 testing images.
- 40-class and 4-class segmentation task
- Pixel-acc: $\sum_i n_{ii} / \sum_i t_i$
- Mean-acc: $(1/n_{cl})\sum_i n_{ii}/t_i$
- Mean IoU: $(1/n_{cl})\sum_i n_{ii}/(t_i + \sum_j n_{ji} - n_{ii})$
- Frequency weighted IoU: $(\sum_k t_k)^{-1}\sum_i t_i n_{ii}/(t_i + \sum_j n_{ji} - n_{ii})$

Visualization



Examples of semantic segmentation results on the NYUDV2 dataset RGB images. Respectively, Our-FCNs and Our-SegNet are the results by our method based on FCNs and SegNet. Compared with original FCNs, our method can distinguish small objects from large-sized ones. By adding contextual regularization, objects in the adjacent area can also be correctly classified, resulting in more accurate boundaries.

Visualization



Examples of semantic segmentation results on the NYUDV2 dataset HHA images. HHA images [18] encoder the depth image of three channels (horizontal disparity, height above ground, and the angle the pixel’s local surface normal). Compared with original SegNet, local constrains are used in our method to achieve more consistent classification results in adjacent regions which proves that our method is also effective for HHA images.

Comparisons evaluation metrics

	Pixel-acc	Mean-acc	Mean IoU	Fw IoU
	61.8	44.7	31.6	46.0
FCNS(32S-HHA)	58.3	44.8	31.7	46.3
Gupta et al	60.3	-	28.6	47
Our(32s-RGB)	**62.5**	**46.3**	**33.6**	**49.9**
Our(32s-HHA)	60.7	45.3	32.0	47.9

FCNs-32 40classes

	Pixel-acc	Mean-acc
Couprie et al	64.5	63.5
Khan et al	69.2	65.6
Stuckler et al	70.9	67.0
Muller et al	72.3	71.9
Gupta et al	78	-
Our(32s-RGB)	**81.1**	**80.3**

FCNs-32 4classes

	Pixel-acc	Mean-acc	Mean IoU	Fw IoU
SegNet-RGB	46.8	22.2	14.2	33.4
Our (SegNet-RGB)	55.1	**32.1**	**22.3**	39.1
SegNet-HHA	54.1	30.5	21.0	38.5
Our (SegNet-HHA)	**55.6**	31.7	21.8	**39.9**

SegNet 40classes

CONCLUSIONS

- We proposed a learnable contextual regularization for semantic segmentation of indoor scene images. This regularization term can be flexibly used in many segmentation networks, such as FCNs and SegNet. We consider not only the low-level information but also the upper-level information.
- We construct a set of learnable weight matrixes from the low-level that can impose additional contextual constraints of all the pixels on the segmentation network output, not limited to the label pixels. Our method helps models adapt to different segmentation tasks.