2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

RelationNet: Learning Deep-Aligned Representation for Semantic Image Segmentation

Yueqing Zhuang, Li Tao, Fan Yang, Cong Ma, Ziwei Zhang, Huizhu Jia*, Xiaodong Xie
National Engineering Laboratory for Video Technology, Dept. EECS, Peking University,
No.5 Yiheyuan Road, Beijing 100871, Beijing, China
{zhuangyq, chntaoli, fyang.eecs, cong-reeshard.ma, zhangziw, hzjia, donxie}@pku.edu.cn

***Abstract*—Semantic image segmentation, which assigns labels in pixel level, plays a central role in image understanding. Recent approaches have attempted to harness the capabilities of deep learning. However, one central problem of these methods is that deep convolutional neural network gives little consideration to the correlation among pixels. To handle this issue, in this paper, we propose a novel deep neural network named RelationNet, which utilizes CNN and RNN to aggregate context information. Besides, a spatial correlation loss is applied to train RelationNet to align features of spatial pixels belonging to same category. Importantly, since it is expensive to obtain pixel-wise annotations, we exploit a new training method to combine the coarsely and finely labeled data. Experiments show the detailed improvements of each proposal. Experimental results demonstrate the effectiveness of our proposed method to the problem of semantic image segmentation, which obtains state-of-the-art performance on the Cityscapes benchmark and Pascal Context dataset.**

I. Introduction

Semantic image segmentation is about labeling each pixel in the image with the class of its enclosing object or region. It attracts increasing attentions rapidly in computer vision and pattern recognition research community due to its importance for automatic driving, remote sensing and medical image processing. IAt's important to use semantic segmentation to estimate the precise boundary rather than using object detection to obtain coarse bounding box which only delineates rough location of an object.

To address this task, in the previous decades, traditional methods depend on hand-crafted features combined with classifiers. Structured prediction technique [1][2] and context information embedding[3] have achieved substantial improvements. Recently, deep learning network has been widely used for semantic segmentation which achieves promising performance and has become the dominant solution. FCN (Fully Convolutional Network) [4] replaces fully connected layers with convolutional layers, and is adopted by state-of-the-art image semantic segmentation methods. These deep learning based methods mainly contain two steps, feature extraction and pixel-wise classification.

Designing a discriminative feature representation of a pixel is the key challenge in pixel-wise labeling problem. FCN framework makes progress with development of a more discriminative feature representation from VGG [5][6], ResNet [7][8][9] to DenseNet [10]. Multi-Scale technology is another solution to learn better feature representation, and can be roughly classified into three categories: image pyramid, encoder-decoder networks and extra module for multi-scale feature. Typical works [11][12][13] use image pyramid to extract and merge features from different scales, as the small scale image contains the context information while larger scale image includes the details. Since higher layers of CNN have larger scale of receptive field than lower layers, the encoder-decoder structure [14] learns multi-scale representation by combining the feature maps from different layers. Extra modules on the top of the original feature extraction network [6][7] use scale-aware operation (eg. dilated convolution or grid pooling) to get multi-scale feature embedding.

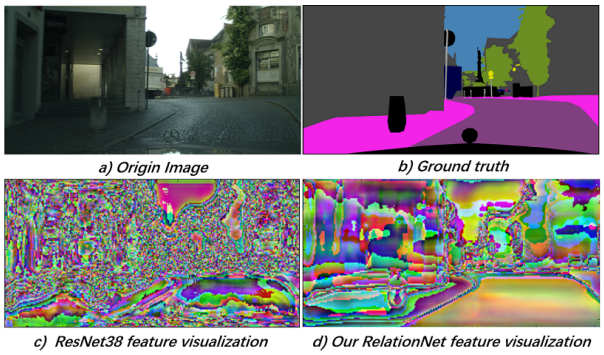

Fig. 1. Visualization of feature embedding computed densely from input images. Two different network used to extract the feature before the classification subnet (compressed to the three dimensions by the PCA for visualization), our RelationNet with SCL has more consistency inside objects and the sharper boundary near contours in feature space.

Pixel-wise classification usually uses structural prediction to refine the result according to image edges, appearance and spatial consistency. The pioneer work [6] uses CRF (Conditional Random Field) as post processing to refine the results. Following methods [16][17] incorporate CRF to FCN which are jointly trained while other work [13] uses deep convolution neural network to estimate the CRF. Besides, domain transform [18] is used to combine edge detection with semantic image segmentation.

FCN considers image semantic segmentation as a pixel-level classification problem, while it ignores the relation of pixel-wise features belonging to the same object and discrim-

* means the corresponding author. Huizhu Jia is with Peking University, also with Cooperative Medianet Innovation Center and Beida Information Research

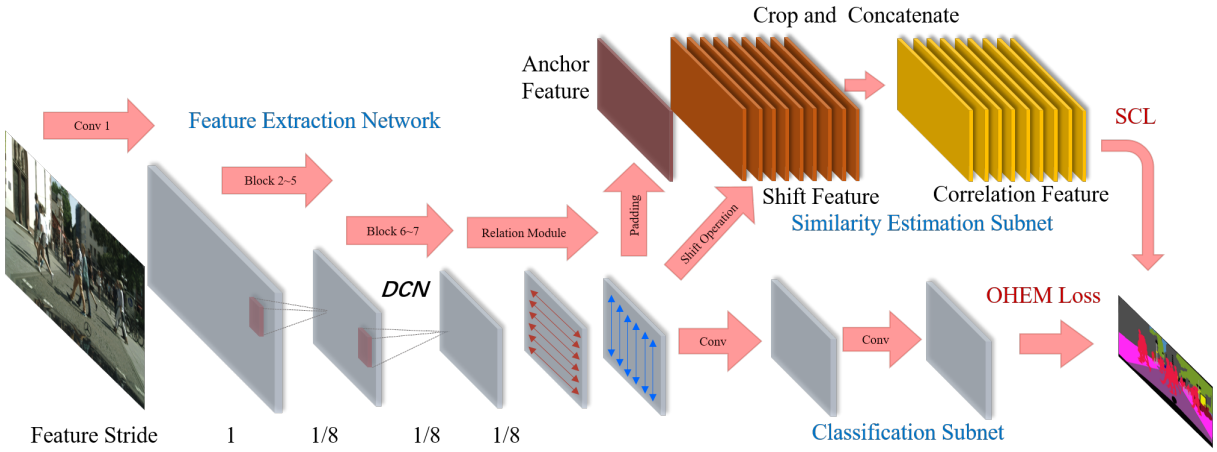

Fig. 2. Visualization of our proposed architecture. The architecture for feature extraction is ResNet38 [9], but we replace the second convolution in the block 6 and 7 with Deformable Convolution [15]. Next, feature maps are fed into relation module to aggregate features. In the end, classification subnet aims at pixel-wise classification when similarity estimation subnet pays attention to the relatedness of spatial pixels.

ination of features from different object (Fig. 1c). Different from the structural classification that considers relations in the final result, in this paper, we propose a Relation Module which consider the relation with high level features. Moreover, similarity estimation subnet with spatial correlation loss is proposed to make the learned features more discriminative (Fig. 1d).

Experimental results demonstrate that our approach achieves state-of-the-art performance on the Cityscapes [19] and Pascal Context [20] dataset with mean IoU 82.4% and 48.4% without CRF. Our main contributions are three holds:

- Relation Module is proposed to aggregate the feature representation inside the objects.
- Similarity Estimation Subnet with SCL (Spatial Correlation Loss) is developed to learn the relatedness of adjacent features.
- We exploit a better training method called *Alternating Training* for combining finely annotated data with coarsely annotated data.

II. Methods

As mentioned above, FCN is a straight architecture which doesn't consider about relatedness of adjacent features. To solve this problem, we propose a method based on network architecture and supervision. The method is illustrated in Fig. 2, which can be divided into three parts for architecture and two parts for losses. RelationNet contains feature extraction, feature aggregation and similarity estimation, while loss includes classification loss and spatial correlation loss. In testing phase, the similarity estimation subnet is removed, which is only used to supervise the similarity of adjacent high-level features in the training proceed.

A. RelationNet

1) Network Architecture For Feature Exaction: Similar to the work in [9], the proposed method uses ResNet38 to learn feature representation. As demonstrated in [21], dilated convolutions can maintain internal representations for high resolution which are reducted by spatial pooling operation. Moreover, deformable convolution [15] can learn suitable dilations in order to fit the scale of objects.

Different from [9], we replace pooling layers (before Block 2, 3 and 4) with increasing stride of corresponding convolutional layers to 2 and substitute the second convolution in Block 5 with dilation of 2. The second convolutions in Block 6, 7 are replaced with deformable convolution [15]. Detailed experiments are shown in section III-A.

2) Relation Module For Feature Aggregation: In FCN, a spatial channel of feature map means the embedding of corresponding pixel. Spatial adjacent features have more overlap in receptive field. Therefore, adjacent features assigned to the same label should have more similarity than those assigned to the different label. With the observation that FCN can't distinguish features from each other even when their spatial space is adjacent (Fig. 1 c), we implement Relation Module to aggregate features as below:

$$fo_{m,n} = \overrightarrow{Cell}(fi_{m,n}, fi_{m+\triangle w, n+\triangle h}) \tag{1}$$

Where fi, fo are the feature maps of input and output respectively. There are four directions GRU (Fig. 1 c) left-right bidirectional GRU is followed by up-down bidirectional GRU. This setting makes network consider more about the relationship among adjacent features. Due to lesser GPU memory occupation and faster convergence of GRU (Gated Recurrent Unit) than RNN (Recurrent Neural Network) and LSTM (Long Short-Term Memory), We choose GRU as the element of our Relation Module.

3) Similarity Estimation Subnet For Comparison: The features assigned to the same label should have more similarity than the features assigned to different label in feature space. Feature alignment means features belonging to the same label should be aggregated. To solve feature-aligned problem, we propose a similarity estimation subnet, which is supervised by SCL (Spatial Correlation Loss) to classify a pair of features

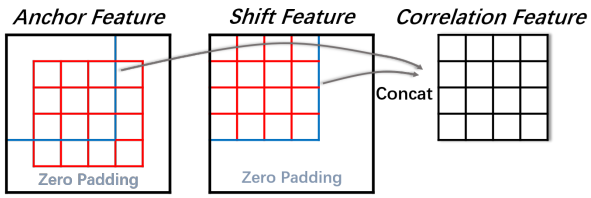

Fig. 3. Visualization of shift feature operation. For example, there are 4×4 feature map, and first, we pad *zero* in the border which called *anchor feature*, and then the feature shift in the 6×6, and finally we get *correlation feature* with the operation of cropping and concatenating *anchor feature* and *shift feature*.

which belong to the same object or not (i.e. binary classification).

Fig. 3 shows our shift feature operation to align feature. We first apply padding operation on the feature map out of Relation Module and get *anchor feature*, then shift *anchor feature* with offset $\triangle w$, $\triangle h$ to get *shift feature*. With operation of cropping and concatenating aligned features of adjacent pixels, we attain spatial *correlation feature*. Then *correlation feature* is fed into similarity estimation subnet which is supervised by SCL in Section. II-B2.

B. Training objective

Our final loss formulates is as follows:

$$\mathcal{L} = \mathcal{L}_{ohem} + \lambda \mathcal{L}_{scl} \tag{2}$$

Fig. 4 shows the difference between two terms of the loss. The $\mathcal{L}_{ohem}$ dominates in total loss which pays attention to judging class label, while the $\mathcal{L}_{scl}$ acts as auxiliary loss for predicting the similarity of adjacent features. λ is weight coefficient used to balance these two losses.

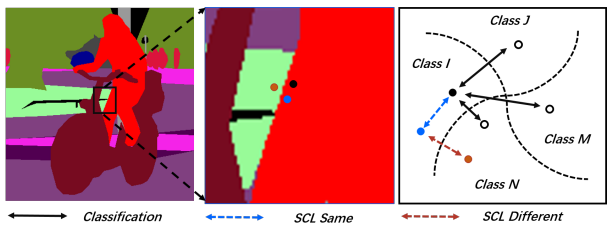

Fig. 4. Illustration of the goals with different losses. The OHEM aims at distinguishing which class the feature should be (Black point and arrow), while the SCL predicts the similarity so that pull the distance of the same (Green color) and push distance of the different(Gray color).

1) Online Hard Example Mining(OHEM): Unbalanced samples are usually present in datasets, especially in sematic image segmentation dataset, causing the preference on training networks and less improvement on the hard examples of semantic segmentation. To solve this problem, we adopt OHEM [9] from [22] which can learn hard examples at stage of training the network, loss function is formulated as below:

$$\mathcal{L}_{ohem} = \frac{1}{\sum_i^N \sum_j^K \mathcal{I}\{y_i=j \ and \ p_{ij}<t\}} * \sum_i^N \sum_j^K \mathcal{I}\{y_i = j \ and \ p_{ij} < t\} log p_{ij} \tag{3}$$

Let K be the number of category c_j in label space. For simplicity, suppose that we convert the image into a one-dimensional pixel array and there are N pixels we should do prediction, and i is the mark number identifying the pixel. And p_{ij} is the probability of $pixel_i$ assigned to the category c_j. For the ground truth, $P(X, Y)$ is set 1 where X and Y belong to the same category. y_i is the target label of $pixel_i$. Comparing with Cross-Entropy Loss, OHEM would sample the pixel-wise loss according to threshold t so that the network will pay more attention to the hard examples at the training stage.

2) Spatial Correlation Loss (SCL): The loss of vanilla classification (same as OHEM) doesn't consider about the consistency inside instances in the spatial space. Inspired by Center Loss [23], which consider the intra-class variations, we propose SCL to considers intra and inter relation among adjacent pixels (Fig. 4). Due to the importance of feature alignment mentioned in Section II-A3, SCL is attached to *similarity estimation subnet* to predict the similarity. We formulate SCL for a pixel in the spatial space as below:

$$\mathcal{L}(X_{mn}, X_{m+\triangle w, n+\triangle h}) = \begin{cases} \alpha log(1 - P(X_{mn}, X_{m+\triangle w, n+\triangle h})) & if \ y_{mn} \neq y_{m+\triangle w, n+\triangle h} \\ \beta log P(X_{mn}, X_{m+\triangle w, n+\triangle h}) & if \ y_{mn} = y_{m+\triangle w, 0+\triangle h} \end{cases} \tag{4}$$

in which

$$\alpha = \frac{|y_{mn} = y_{m+\triangle x, n+\triangle y}|}{|y_{mn} = y_{m+\triangle w, n+\triangle h}| + |y_{mn} \neq y_{m+\triangle w, n+\triangle h}|}$$
$$\beta = \frac{|y_{mn} \neq y_{m+\triangle w, n+\triangle h}|}{|y_{mn} = y_{m+\triangle w, n+\triangle h}| + |y_{mn} \neq y_{m+\triangle w, n+\triangle h}|} \tag{5}$$

$|y_{mn} = y_{m+\triangle w, n+\triangle h}|$ and $|y_{mn} \neq y_{m+\triangle w, n+\triangle h}|$ mean the number of positive sample set and negative sample set respectively. α, β are the weights for unbalanced data. $P(X_{mn}, X_{m+\triangle w, n+\triangle h})$ is prediction probability of the spatial relevant pixels belonging to the same category. In our SCL, the $pixel_{m,n}$, should be compared to the nearby $pixel_{m+\triangle w, n+\triangle h}$, where $\triangle w$, $\triangle h$ range from [-1,0,1]. Therefore, our SCL builds as follows:

$$\mathcal{L}_{scl} = \frac{1}{N} \sum_{m=1}^{W} \sum_{n=1}^{H} \sum_{\triangle w=-1}^{1} \sum_{\triangle h=-1}^{1} \mathcal{L}(X_{mn}, X_{m+\triangle w, n+\triangle h}) \tag{6}$$

As in the equation 6, where $N = WH|\triangle w||\triangle h|$ acts as the term of normalization, a pixel is compared with the nearby 9 pixels including itself. Relation module and similarity estimation subnet with SCL bring about the discrimination near contour and the consistency inside instances (Fig. 1).

C. Unsampling Strategy

The output stride of RelationNet is $1/8$, as traditional interpolations like bilinear interpolations would be inaccurate for small objects and contours between two categories. Therefore,

4-steps testing is used to get prediction which is unsampled to origin size of image(as Fig. 5), which will suppress inaccuracy by interpolation because coordinate value of feature maps is the embedding of pixel locating at the center of respective field.

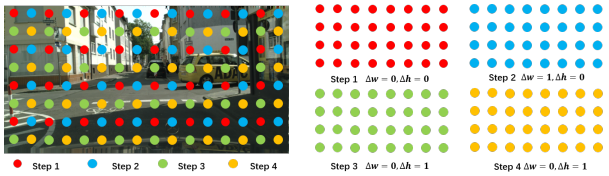

Fig. 5. Illustration of the multi-steps test. Each color means the result of one step testing, and the point in the image means the center of receptive field. We shift the image for testing in order to get the result of each pixel instead of using interpolations.

III. Experiments

Cityscapes [19] contains 5000 high quality pixel-level finely annotated images (Fig. 7 a) and 19998 coarsely annotated images (Fig. 7 b), collected from 50 cities, which is used for auto-driver. Images are divided into 2975, 500, 1525 images for training, validation, and testing respectively. This dataset contains 19 categories which occur in auto-driver at most often. For ablation experiments, we train network with fine training data or both fine and coarse data (marked with †), and test on the fine validation set. For final result, we train RelationNet with fine training, validation data as well as coarse data (marked with ‡), and test on the server [1].

We report metrics as below: 1) pixel accuracy, which is the percentage of correctly labeled pixels. 2) mean value of class-wise pixel accuracies, 3) mean IoU score, which is the mean value of class-wise intersection-over-union scores.

TABLE I
Ablation experimental results on Cityscapes validation. † means network is trained with adding coarse data.

Network	CE	OHEM	mIoU(%)	mAcc(%)	Acc(%)
		Loss Choice			
ResNet101	✓		70.14	78.93	95.04
		✓	72.12	80.43	95.29
	✓		73.64	82.16	95.27
		✓	77.39	84.47	95.99
		Architecture Choices			
ResNet38	+DCN		78.40	85.23	96.12
		+Relation	78.43	86.20	96.05
	+DCN	+Relation	79.17	85.99	96.17
		Alternative Training			
	Stage1		79.88	86.83	96.26
	Stage1 + Multi-Steps		80.92	87.54	96.53
	Stage2† + Multi-Steps		81.16	88.24	96.46
	Stage3 + Multi-Steps		**81.76**	**88.69**	**96.77**

A. Architecture Choices

To evaluate RelationNet, we conduct experiments with several setting as Tab. I, which includes the replacement of convolution, the effectiveness of our proposed Relation Module, and the combination of the DCN (Deformable Convolution Networks)[15] and Relation Module. As listed in Tab. I, ResNet with DCN works better than the traditional convolution, which yields results 78.40%/85.23% in terms of mIoU and mAcc, surpassing our ResNet-38 with OHEM by 1.01%/0.76%. Also, our Relation Module attached to the traditional convolution network gets the results 78.43%/86.20%, exceeding the ResNet-38 with OHEM by 1.04%/1.73%. As the DCN can be seen as the attention mechanism, Relation module with DCN gets the improvement of 1.78%/1.52% than the ResNet-38 with OHEM. All results mentioned above indicate that our relation module pays an important role on powerful feature representation.

B. Loss Choices

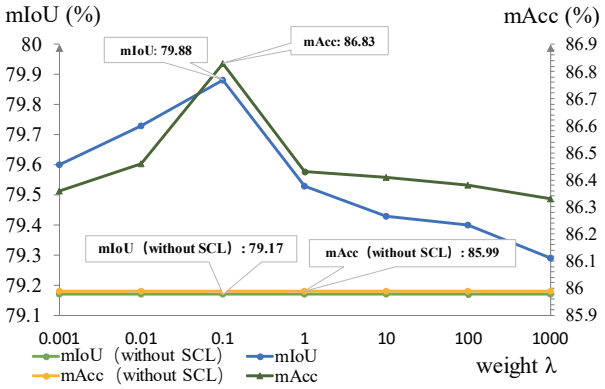

Fig. 6. Quantitative analysis of weight λ in terms of mIoU and mAcc. The results are obtained by single scale without multi-steps test.

As Tab. I shows, comparing OHEM with cross entropy loss, we find the effectiveness of the OHEM that it can improve the results because of unbalance samples. Experientially, we set threshold t 0.6 to sample hard examples. We experiment with setting SCL weight λ between 0.001 and 1000, Fig. 6 shows the results comparing to RelationNet without SCL. Empirically, $\lambda = 0.1$ yields the best performance. The introduced similarity estimation subnet with SCL helps to optimize the learning process while not affecting learning in the master branch. The improvements $0.71\%/0.84\%$ in terms of mIoU/mAcc make us believe that deep networks will benefit from the proposed similarity estimation subnet with SCL.

C. Training Methods

TABLE II
Results on Cityscapes val show that transfering from different pre-trained models may lead to different local minimums.

Network	Finetune	mIoU(%)	mAcc(%)	Acc(%)
ResNet101-Relation	cityscapes	72.23	80.34	95.28
ResNet101-Relation	ImageNet [24]	**73.70**	**81.19**	**95.46**
ResNet38-Relation	cityscapes	76.06	83.32	95.84
ResNet38-Relation	ImageNet [24]	**78.43**	**86.20**	**96.05**

1) The impact of the Pre-trained Model: As irrelevance of features in a trained CNN (eg. Fig. 1 c), directly training from a pre-trained model from Cityscapes may cause Relation

[1] https://www.cityscapes-dataset.com/benchmarks/

Module converges improperly. As Tab. II shows, experiments in ResNet38 and ResNet101 prove our hypothesis. Using the pre-trained model on the Cityscapes, Relation Module may fall into the local minimum. Only with pre-trained model from the ImageNet, RelationNet can converge properly when the network of feature extraction and the network of feature aggregation are trained at the same time, this process is also quite different from [18] which uses GRU to smooth the result.

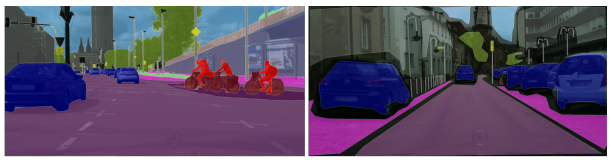

Fig. 7. Illustration of the data label in the Cityscapes dataset. The coarsely labeled data lose the boundary information.

2) Alternating Training: As coarse data is roughly labeled main parts of objects (eg. Fig. 7), while the contour between objects is not labeled so that it loses the information of contour, directly applying SCL in coarse data may cause the incorrectness of aligning features. Inspired by the idea of [25] which uses *Alternating Training* strategy to train RCNN and Region Proposed Network, we adopt *Alternating Training* to train the semantic image segmentation.

First, we train RelationNet with only fine data using SCL and OHEM (*Stage 1*), then train the network with coarse data but not with SCL (*Stage 2*). Finally, we fine-tune RelationNet with only fine data using SCL and OHEM (*Stage 3*). In this way, we can merge a large range of coarsely labeled data which contains main part with a little of finely labeled data that includes boundary information together. Results on Tab. I show the effectiveness of our proposed training methods with an improvement of 0.6%. It's remarkable that our proposed training method for combining coarse data and fine data can economize the expensive pixel-level labeling time.

D. Implementation Details

We set the *batchsize* to 8 during training for all experiments and use pre-trained model from ImageNet[24]. The learning rate sets 5×10^{-4} for the first 35 epochs and learning rate goes down linearly from the 5×10^{-4} to 5×10^{-6} for the last 25 epochs. For data arguments, we randomly flip and resize images ranging from 0.55 to 1.3 and randomly crop it to (512, 520). For *Alternating Training*, *Stage1, Stage2, Stage3* are trained with $60, 30, 15$ epoches respectively.

E. Experimental Results

1) Cityscapes: Statistics in Tab. III show that our proposed method outperforms other methods with notable advantage. For equal comparison, we use the fine training data (2975 images) and coarse data to train our final model and our method yields 80.8% mIoU. Fig. 8 shows our comparison with our baseline, which indicates that our experimental results have less noise and more consistency inside an object. With

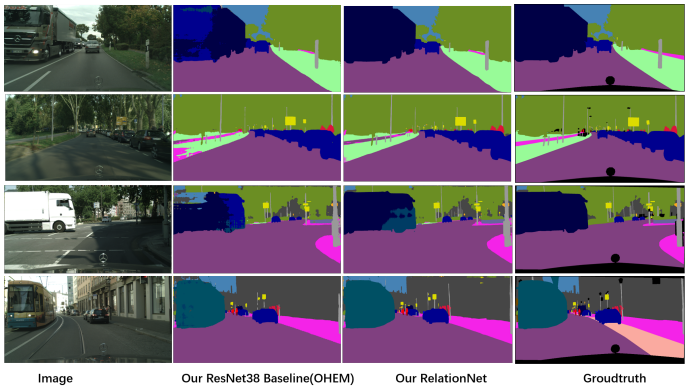

Fig. 8. Results of RelationNet on cityscapes obtained by single scale and single model, which are compared with our baseline(OHEM).

TABLE III
RESULTS ON CITYSCAPES TESTING SET. †,‡ MEANS TRAINING WITH BOTH THE COARSE AND FINE TRAINING OR TOTAL FINE DATA. IIOU IS SPECIAL METRIC IN CITYSCAPES DATASETS.

Method	IoU cla.(%)	iIoU cla.(%)	IoU cat.(%)	iIoU cat.(%)
CRF-RNN [16]	62.5	34.4	82.7	66.0
FCN [4]	65.3	41.7	85.7	70.1
LRR [12]	69.7	48.0	88.2	74.7
DeepLabv2_CRF [6]	70.4	42.6	86.4	67.7
Piecewise [13]	71.6	51.7	87.3	74.1
Global-Local-Refinement [26]	77.3	53.4	90.0	76.8
TuSimple [27]	77.6	53.6	90.1	75.2
SAC_multiple [28]	78.1	55.2	90.6	78.3
PSPNet [7]	78.4	56.7	90.6	78.6
ResNet38 [9]	78.4	59.1	90.9	81.1
Our RelationNet	**79.3**	**60.7**	**91.2**	**81.6**
LRR [12] ‡	71.8	47.9	88.4	73.9
Segmodel [29] ‡	79.2	56.4	90.4	77.0
TuSimple_Coarse [27] ‡	80.1	56.9	90.7	77.8
Netwarp [30] ‡	80.5	59.5	91.0	79.8
ResNet38 [9] ‡	80.6	57.8	91.0	79.1
PSPNet [7] ‡	81.2	59.6	91.5	79.2
Our RelationNet ‡	**82.4**	**61.9**	**91.8**	**81.4**

adding fine validation dataset, we achieve 82.4% mIoU over the benchmark[2]. Results on Tab. I show the improvements 1.05% in terms of mIoU on Cityscapes val.

2) Pascal Context: This dataset [20] consists of 4998 images for training and another 5105 images for validation. Pixels either belong to background category or 59 semantic categories, including *bag, food, sign, ceiling, ground, and snow*. Since the test set is not available, here we directly test our result on the validation set. As shown in Tab. IV, our method again performs the best scores in three evaluation metric.

IV. CONCLUSION

We propose an effective Relation Module for feature aggregation, use spatial correlation loss to get better feature representation for each pixel, and exploit the effectiveness of each proposal and *Alternating Training* strategy. Our experiments suggest that our RelationNet joint trained with SCL and OHEM gets powerful performances. Moreover, *Alternating Training* would save expensive time-costing for labeling data.

[2]https://www.cityscapes-dataset.com/benchmarks/

TABLE IV
RESULTS ON PASCAL CONTEXT[20] VAL SET WITH 5105 IMAGES. THE BASELINE IS OUR RESNET38 WITHOUT OHEM AND SCL.

Method	Acc(%)	mAcc(%)	mIoU(%)
FCN-8s [4]	65.9	46.5	35.1
BoxSup [31]	-	-	40.5
Context [32]	71.5	53.9	43.3
VeryDeep [21]	72.9	54.8	44.5
DeepLab_v2 [6]	-	-	45.7
ResNet38 [9]	75.0	58.1	48.1
Our BaseLine	73.2	52.2	43.8
Our RelationNet	**75.2**	**58.9**	**48.4**

We get 82.4% and 48.4% mIoU which is state-of-the-art results in the Cityscapes and Pascal Context dataset.

ACKNOWLEGEMENT

This work is partially supported by National Key Research and Development Program of China under contract No. 2016YFB0402001, the Major National Scientific Instrument and Equipment Development Project of China under contract No. 2013YQ030967, National Science Foundation of China under contract No. 61602011 and NVIDIA NVAIL program.

REFERENCES

[1] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical crfs for object class image segmentation," in *IEEE International Conference on Computer Vision*, 2009, pp. 739–746.

[2] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[3] C. Rui, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *European Conference on Computer Vision*, 2012, pp. 430–443.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[6] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: http://arxiv.org/abs/1606.00915

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[9] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *CoRR*, vol. abs/1611.10080, 2016. [Online]. Available: http://arxiv.org/abs/1611.10080

[10] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 2261–2269.

[11] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, Aug 2013.

[12] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 519–534.

[13] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[15] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 764–773.

[16] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1529–1537.

[17] V. Jampani, M. Kiefel, and P. V. Gehler, "Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks," in *Computer Vision and Pattern Recognition*, 2016, pp. 4452–4461.

[18] L. C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4545–4554.

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[20] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[21] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[22] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.

[23] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, *A Discriminative Feature Learning Approach for Deep Face Recognition*. Springer International Publishing, 2016.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[26] M. L. J. L. S. Y. Rui Zhang, Sheng Tang, "Global-residual and local-boundary refinement networks for rectifying scene parsing predictions," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 3427–3433. [Online]. Available: https://doi.org/10.24963/ijcai.2017/479

[27] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," *arXiv preprint arXiv:1702.08502*, 2017.

[28] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan, "Scale-adaptive convolutions for scene parsing," in *Proc. 26th Int. Conf. Comput. Vis.*, 2017, pp. 2031–2039.

[29] F. Shen, R. Gan, S. Yan, and G. Zeng, "Semantic segmentation via structured patch prediction, context crf and guidance crf," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1953–1961.

[30] R. Gadde, V. Jampani, and P. V. Gehler, "Semantic video cnns through representation warping," *CoRR, abs/1708.03088*, 2017.

[31] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1635–1643.

[32] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Exploring context with deep structured models for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.