The progress of the reading plan:

05 / 50

Paper Information

Paper Title :

ExFuse: Enhancing Feature Fusion for Semantic Segmentation

Conference :

ECCV 2018

Authors and Institutions

Authors

- Zhenli Zhang 1
- Xiangyu Zhang 2
- Chao Peng 2
- Dazhi Cheng 3
- Jian Sun 2

Institutions

- 1 Fudan University
- 2 Megvii Inc. (Face++)
- 3 Beijing Institute of Technology

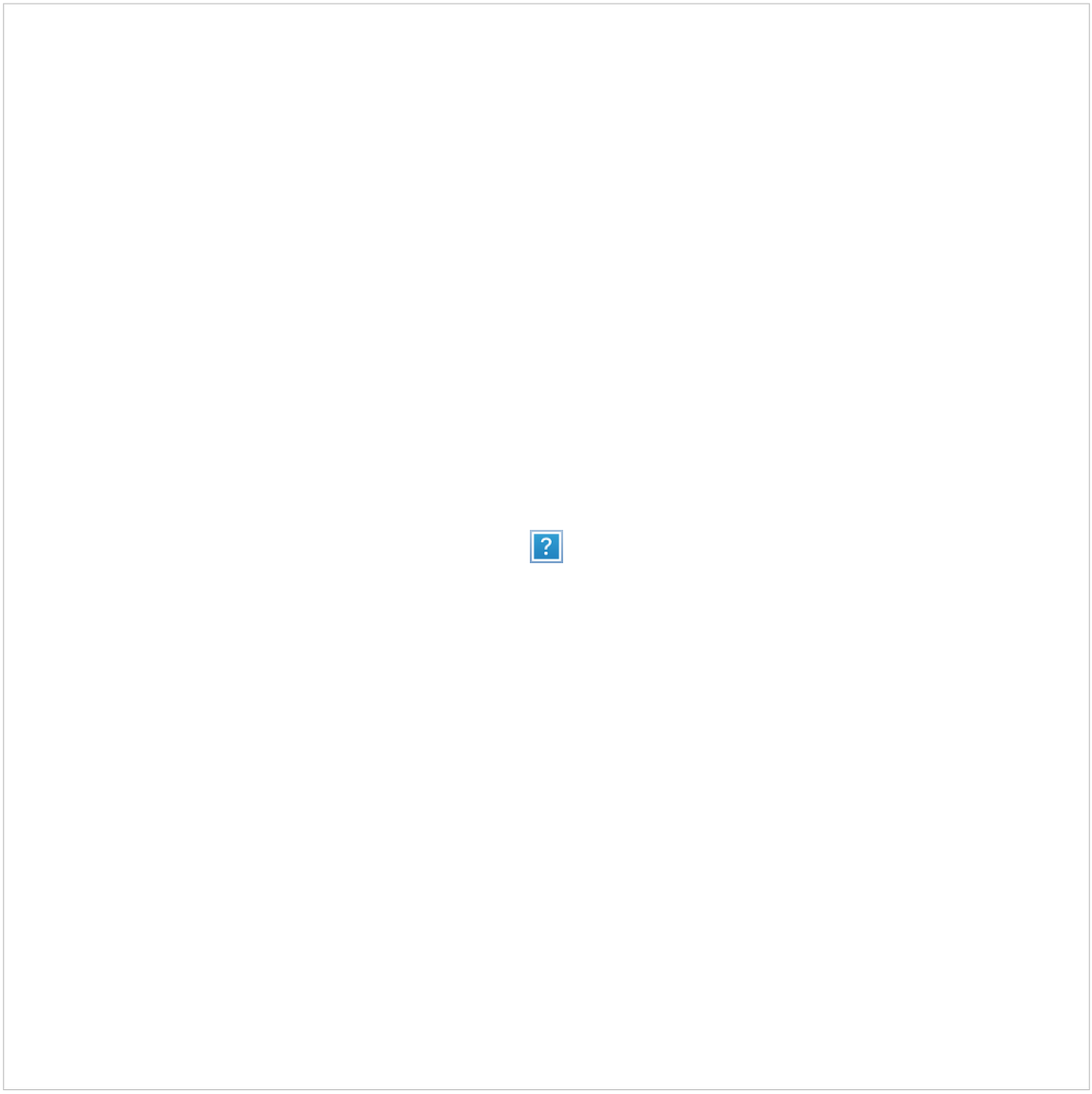
Official Codes

NO

Official Instruction (Chinese)

ECCV 2018 | 旷视科技提出ExFuse——优化解决语义分割特征融合问题

Network Structure



Note

Summary in one sentence.

Enhance the feature fusing method of U-Net structure semantic segmentation methods by introducing more semantic concepts into low-level features and by embedding more spatial information into high-level features.

Key Words

Feature Fusing

Five questions about this paper:

1. [Problem Definition / Motivation] What problem is this paper trying to

solve?

A lot of methods fuse low-level but high-resolution features and high-level low-resolution features don't considering the large semantic or resolution gap between low and high level features.

If low level feature does't have any high level semantic information, then it will contain many noises and become a burden when fusing with high level feature, and vice versa.

In other words, feature fusion could be enhanced **by introducing more semantic concepts into low-level features or by embedding more spatial information into high-level features**.

2. [Contribution / Method] What's new in this paper? / How does this paper solve the above problems?

1. introduce more semantic information into low-level features
    - layer rearrangement (**LR**): decrease the mount of layers between the low and high level feature.
    - semantic supervision (**SS**): Also called auxiliary supervisions. At training process, take account of the loss of all the auxiliary branches to force the low level features to contain more semantic informantion.

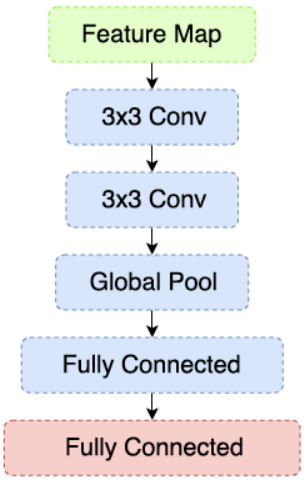


Fig. 3. Details of *Semantic Supervision (SS)* component in our pipeline.

    - **[THE MOST INTERESTING] semantic embedding branch (SEB)**: Using high level feature to refine the low level feature map, discarding some noises.

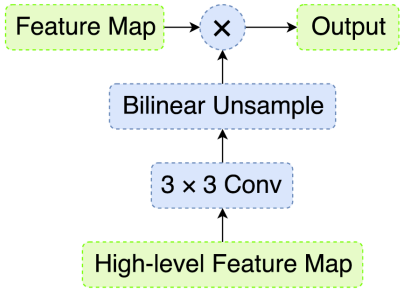


Fig. 4. Design of the *Semantic Embedding Branch* in Fig 2. The "×" sign means element-wise multiplication. If there are more than one groups of high-level features, the component outputs the production of each feature map after upsampling.

2. embed more spatial information into high-level features

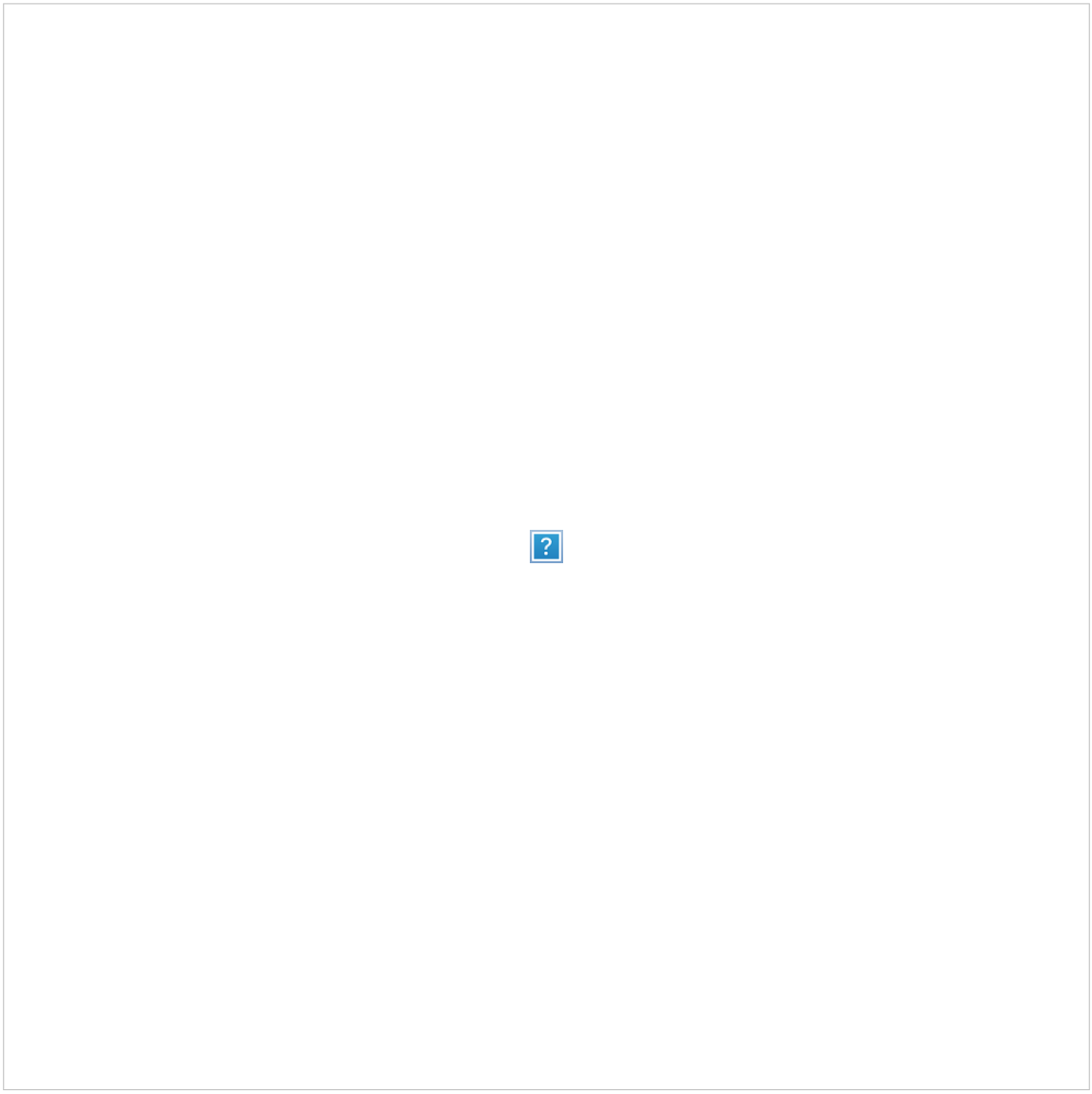
    - explicit channel resolution embedding (**ECRE**) : Adopting the method of ESPCN, which is a extraordinary idea for sub-pixel super resolution. But the auxiliary supervision here, combining with the sub-pixel upsampling is necessary.

Index	Method	mIoU (%)
1	Baseline	78.3
2	Deconv + Supervised	78.2
3	Sub-pixel Upsample Only	77.6
4	ECRE (Fig 5)	**78.8**

Table 4. Ablation study on the design of *Explicit Channel Resolution Embedding, (ECRE)*. The baseline model is in Table 3 (#3)

    - densely adjacent prediction (**DAP**)

Here is the ablation experiments of the methods above.



Some comparison conclusion:

Module	Index compared	Improvement(%)
SS - semantic supervision	1, 2	1.5
LR - layer rearrangement	2, 3	0.8
ECRE - explicit channel resolution embedding	3, 4	0.5
ECRE	6, 7	0.4
SEB - semantic embedding branch	3, 5	0.7
DAP - densely adjacent prediction	5, 6	0.6

3. Details about the experiment

3.1 Which Datasets are used?

- PASCAL VOC 2012
- Microsoft COCO dataset
- SBD

3.2 How is the experiment set up?

Employing Microsoft COCO dataset to pretrain the model.

3.3 What's the evaluation metric?

mIoU

3.4 (Optional）How to divide training data and test data?

- In stage-1, we mix up all images in COCO, SBD and standard PASCAL VOC 2012

images, resulting in 109892 images for training in total.

- In stage-2, we utilize SBD and PASCAL VOC 2012 training images.
- Finally for stage-3, we only employ standard PASCAL VOC 2012 training set.

3.5 (Optional）What is the ranking of the experiment results?

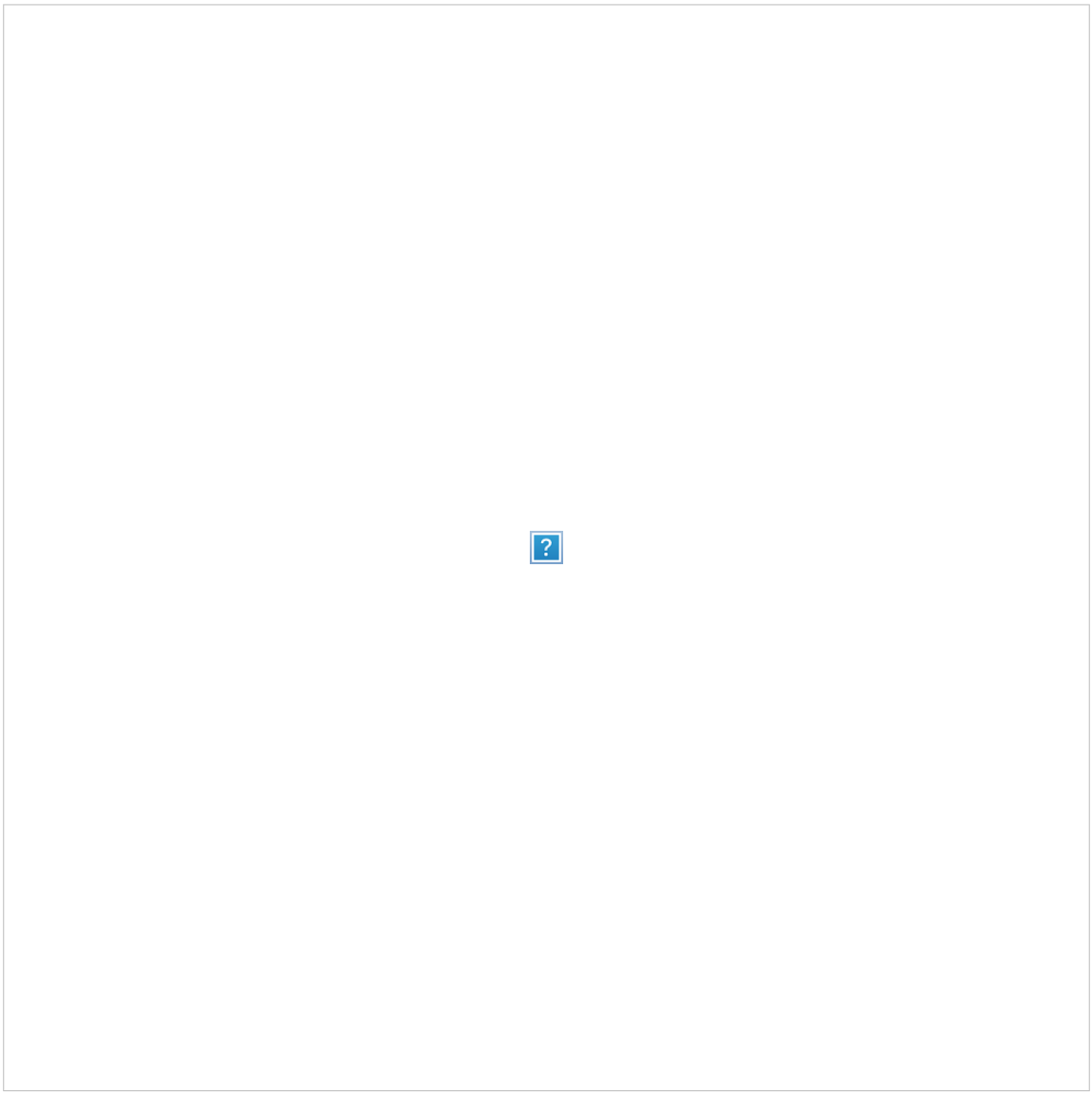
Method	mIOU
Tusimple [9]	83.1
Large_Kernel_Matters [8]	83.6
Multipath_RefineNet [11]	84.2
ResNet_38_MS_COCO [39]	84.9
PSPNet [10]	85.4
DeepLabv3 [5]	85.7
SDN [40]	86.6
DeepLabv3+ (Xception) [37]	87.8
ExFuse_ResNet101 (ours)	**86.2**
ExFuse_ResNeXt131 (ours)	**87.9**

Table 8. Performance on PASCAL VOC 2012 test set

Actually the ExFuse utilized more datas than Deeplabv3+ to train. And SBD isn't used in Deeplabv3 too.

So actually the comparision is not completed and unfair.

Here’s a table made by myself. No fairly comparision is common.



4. Advantages (self-summary rather than the author’s)

5. Disadvantages (self-summary rather than the author’s)

5.1

Comparing ExFuse with the atrous convolution, which can keep the resolution while going deeper layers, is important.

5.2 The writing is too tricky.

For example,

> feature maps close to semantic supervisions (e.g. classication loss) tend to encode more

Here the semantic supervisions actually means the high level layers.

> To make low-level features (res-2 or res-3) ‘closer’ to the supervisions, one straight-forward approach is to arrange more layers in the early stages rather than the latter.

Actually, here the author changed the structure of ResNeXt network, to dicrease the amount of building blocks from {3; 4; 23; 3} to {8; 8; 9; 8}, which means the amount of hiden layers is less. So naturally more the low-level layers are closer to the high level layers.

I suggest that neat and clear representations should be used in writing, or misleading may be happen.

5.3 About the "Semantic Embedding Branch"

> However, if the low-level feature contains little semantic information, it is insucient to recover the semantic resolution.

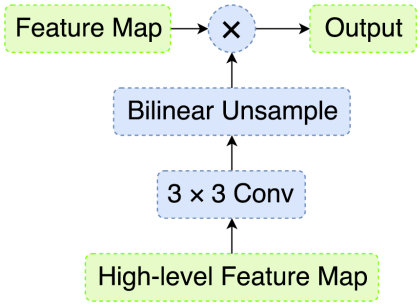


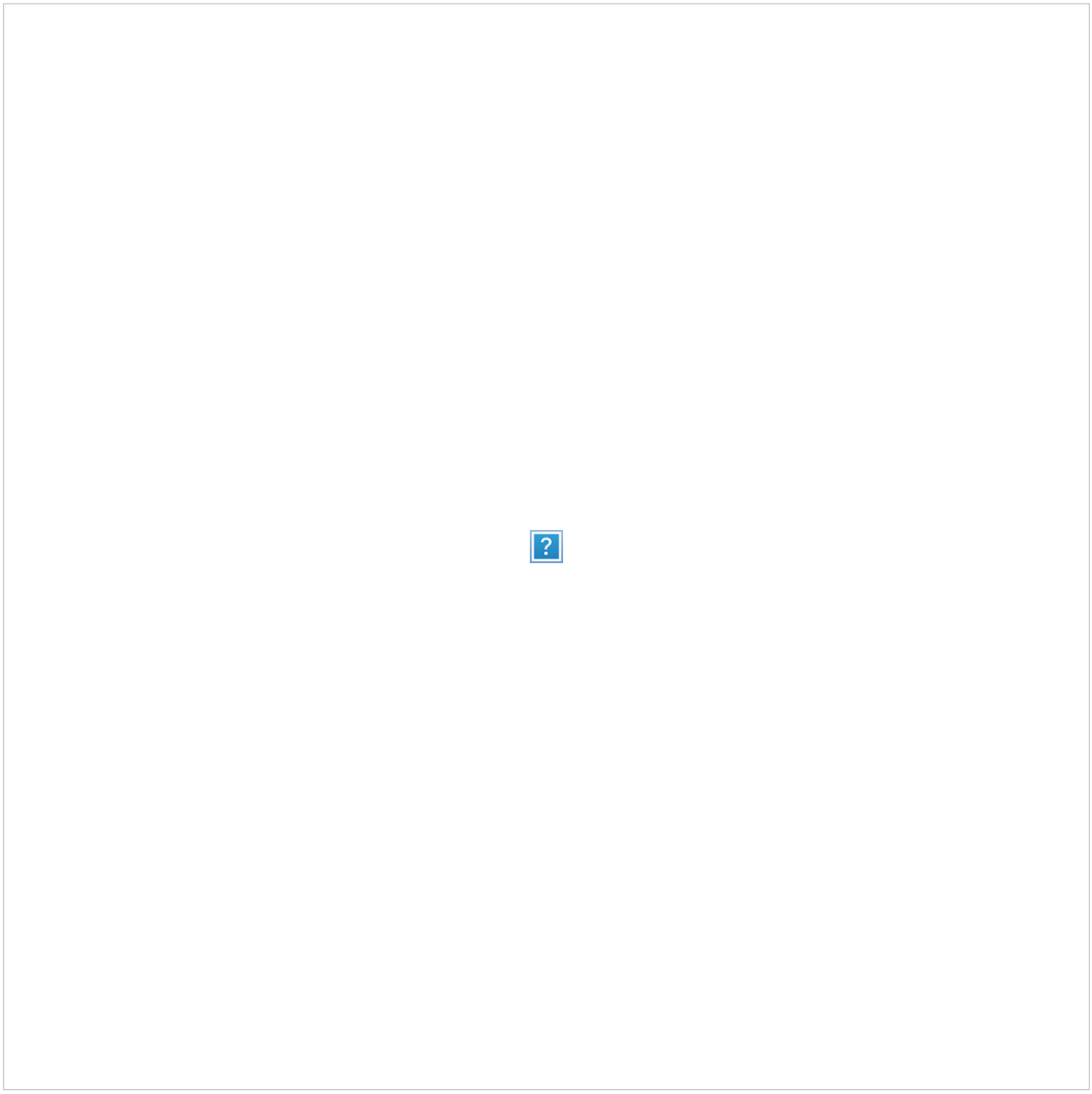
Fig. 4. Design of the *Semantic Embedding Branch* in Fig 2. The “×” sign means element-wise multiplication. If there are more than one groups of high-level features, the component outputs the production of each feature map after upsampling.

I don’t agree with the author. The low level feature has only very few semantic infomation, and most of it is the struture information. The semantic information, should be in the charge of high level information. That’s why we fuse the low and high level features.

So, the low level feature here, is not becoming richer of information. Actually it’s becoming poorer.

Because when we look at the struture of the "Semantic Embedding Branch", we can find that the low level feature map is multiplying element-wisely with the upsampled high level feature. And the high level feature contains many semantic informations. The information of whether two pixels are belonging to the same class is included in it. We can regard it as an attention map or a mask, the value it including implies each position in the low level feature is the boundary or not.

After the multiplication, the noises in the low level feature are decreased, by preserving the boundaries and discarding those inplace noisy values. Just look the following image.



Refinement is actually what's happening here.

So may we can design a structure to refine the low level feature to decrease its noise specifically.

6. What's more?

This ExFuse paper reveals importance of refining the low and high level features for the gap between them. So a iteration structure to enhance the low and high level features recurrently maybe helpful.