Noname manuscript No.
(will be inserted by the editor)

HSNet: Hierarchical Semantics Network for Scene Parsing

Xin Tan[1,2] · **Jiachen Xu**[1] · **Ying Cao**[2]† · **Ke Xu**[2] · **Lizhuang Ma**[1] · **Rynson W.H. Lau**[2]†

Received: date / Accepted: date

Abstract Scene parsing is one of the fundamental tasks in computer vision. Humans tend to perceive a scene in a hierarchical manner, i.e., first identifying the coarse category (e.g., vehicle) of a group of objects and then the fine category (e.g., bicycle, truck or car) of each of them. Despite recent tremendous progress on scene parsing, such a hierarchical semantics prior (HSP) has not been explicitly exploited. In this paper, we aim to introduce the HSP into scene parsing, by proposing a hierarchical semantics network (HSNet). Our key contribution is a bidirectional cross-level feature matching framework, which enables us to learn multi-level, hierarchy-aware features via forward feature transfer and backward feature regularization. In the forward stage, we train a coarse-to-fine module to learn fine-category features that explicitly encode hierarchical semantics information. In the backward stage, we introduce a fine-to-coarse module to collapse fine-category features to coarse-category features that are used to regularize the feature learning of our network. Experimental results on Cityscapes and Pascal Context show that our method achieves state-of-the-art performances. Our visualization also shows that our learned features capture semantic hierarchy favorably.

Keywords Hierarchical semantics · scene parsing · cross-level feature · bidirectional network

Xin Tan
E-mail: tanxin2017@sjtu.edu.cn

Jiachen Xu
E-mail: xujiachen@sjtu.edu.cn

Ying Cao
E-mail: caoying59@gmail.com

Ke Xu
E-mail: kkangwing@gmail.com

Lizhuang Ma
E-mail: ma-lz@cs.sjtu.edu.cn

Rynson W.H. Lau
E-mail: Rynson.Lau@cityu.edu.hk

† corresponding authors
[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Department of Computer Science, City University of Hong Kong, HKSAR, China

1 Introduction

Scene parsing aims to assign pixel-level semantic labels to an input image, which is important for autonomous driving. Given an input image as shown in Figure 1, it may not be easy to accurately and efficiently distinguish the sidewalk from the road, due to their similar locations (at the bottom of the image) and surface orientations (flat). We may need to inspect them carefully in order to tell their subtle differences. On the other hand, buildings are easier to be separated from the road and sidewalk given their obvious difference in position and orientation. For such visual recognition and segmentation tasks to be more efficient, the human visual perception system processes an observed scene in an semantically hierarchical way [19], based on our prior knowledge of the visual world. That is, we first group objects with similar semantical and structural properties into coarse semantic categories, and then further discriminate them into different fine categories.

Objects in urban street scenes have the hierarchical relations. With the Cityscapes dataset [6] as an example, as shown at the top of Figure 1, the object categories of an outdoor scene can be organized by a semantics hierarchy. Based on this observation, we propose to introduce semantic hierarchical knowledge as a hierarchical semantics prior (**HSP**) to scene parsing to naturally mimic the behaviors of the human visual perception system. While there are already some

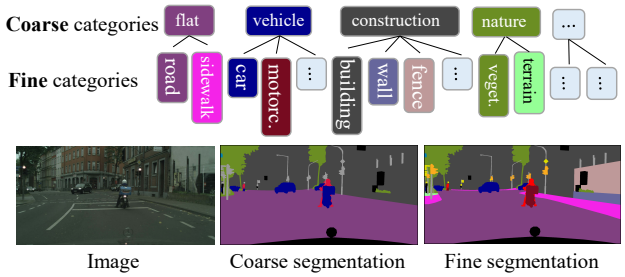


Fig. 1: An example of hierarchical semantics for a real-world scene. At the top is a hierarchy of semantic categories where multiple fine categories may belong to one coarse category. At the bottom is a street scene (left), together with its segmentation based on coarse (middle) and fine (right) categories.

existing efforts to incorporate hierarchical semantics knowledge into image classification models [20, 30], to the best of our knowledge, no study has been conducted so far to consider this knowledge explicitly in scene parsing.

There is a main challenge in this paper. How to model the human prior, i.e., HSP, into our network? In this paper, we take a step towards modeling HSP explicitly in scene parsing, by proposing a hierarchical semantics network (HSNet). In our proposed HSNet, we construct fine-category features (or coarse-category features) based on HSP with coarse-category features (or fine-category features), respectively. Our key contribution is a novel bidirectional cross-level feature matching strategy, which allows us to learn multi-scale features capturing semantic hierarchy knowledge. Our framework is composed of a forward feature transfer stage and a backward feature regularization stage. In the forward stage, we train a coarse-to-fine module to transform the coarse-category features of the network into fine-category features through explicitly integrating the HSP. The output fine-category features inherently characterize the pre-defined hierarchy of semantic categories.

In the backward stage, we introduce a fine-to-coarse module to collapse the fine-category features of the network back to coarse-category features, again based on the HSP. The output coarse-category features are used to regularize the feature learning of our network, which encourages alignment between the learned features at different category levels of the two stages. Our coarse-category features are different from the coarse features in previous work. For example, coarse features in AFNet [47] are with unclear boundaries in fine-level semantics. However, our coarse-category features aim to predict clear boundaries in coarse-level semantic information, and it may be not easy to train the model in shallow layers. Hence, we design the bidirectional network to enforce the deep layers optimize the coarse-category features in the backward stage.

In summary, the main contributions of this paper include:

- To the best of our knowledge, we are the first to investigate how to exploit hierarchical semantics knowledge under the context of scene parsing.
- We propose a novel bidirectional cross-level feature matching framework for learning powerful multi-scale, hierarchy-aware features, which are effective in boosting the performance of scene parsing without extra inference cost to distill HSP.
- We have conducted experiments to show that the hierarchical semantics prior can help significantly improve the performance of scene parsing, yielding state-of-the-art performances on the Cityscapes and Pascal Context dataset.

2 Related Works

2.1 Scene Parsing

In recent years, scene parsing has made significant progress due to the advance of FCNs [25, 32, 42, 52]. Recent efforts along this direction have been devoted to making use of human prior knowledge to learn discriminative features and guide the design of deep networks. Since lower layers of networks tend to capture low-level features (i.e., edges, structures) while higher layers tend to capture high-level features (i.e., semantic information), a lot of works attempt to combine the features from different layers to obtain more powerful multi-scale representations [23, 27, 50]. In addition, as not all regions are equally important for inferring the segmentation, numerous attention-based mechanisms are proposed to guide models to focus on more informative regions [10, 15, 31, 47]. Further, some works also claim that contextual information is very helpful to scene parsing, and propose different kinds of context aggregation approaches, such as PSP [51], ASPP [41], NL [34], RCCA [15] and object-contextual representations [46] . Some works observe that the object shape is also important for improving segmentation performance, and exploit boundary information [7], shape information [29], and depth information [18] to guide model learning. Some works solve the scene parsing by using the graph model to learn the object relations. For example, DGCNet [49] takes a dual graph convolutional network to learn global context of the input feature by modelling two orthogonal graphs in a single framework. CDGCNet [14] designs a class-wise dynamic graph convolution (CDGC) module to adaptively propagate information. In addition, Ji et al. [16] propose an encoder-decoder with cascaded CRFs for scene parsing. Wang et al. [35] proposes the intra-class feature variation distillation for scene parsing. Liu et al. [24]

unifies dependency reasoning at three semantic levels for parsing the images.

In this paper, our aim is to introduce hierarchical semantics to scene parsing, which has not been explored by any existing works.

2.2 Hierarchical Semantics

Semantics are useful in many tasks [9, 28, 38, 39, 53]. A recent study has shown that image classification models can be benefited by considering hierarchical semantics information (i.e., a hierarchy of classes) through network design [2]. Some prior works on large-scale image recognition have also taken hierarchical semantics into consideration. For example, HD-CNN [40] is a hierarchical deep network by embedding deep CNNs into a two-level category hierarchy that helps distinguish difficult classes using multiple classifiers. In [1], a tree-structured network with several branches is proposed, where each branch serves a specific subset of categories. In [30], the Multi Categorical-Level Networks are proposed to extract more discriminating features that contain human-categorization knowledge. In [20], a hierarchical classification framework is proposed to recognize unseen objects.

In fact, many popular large scale datasets for scene parsing are organised by semantic hierarchies. For example, one of the most popular scene parsing datasets, Cityscapes [6], contains 19 fine categories grouped into 7 coarse categories. Although the hierarchical semantics have been used in image classification, their effect upon scene parsing is unknown and still unexplored. In this paper, we show that the hierarchical semantics prior (HSP) is a promising information to use for scene parsing. We propose an unified framework for hierarchy-aware feature learning, and demonstrate its superiority over existing methods on Cityscapes.

3 Proposed Method

3.1 Network Overview

Figure 2 illustrates our proposed network. Our segmentation network (the middle blue stream) consists of a coarse convolution neural network (CoarseCNN), a fine convolution neural network (FineCNN), and a softmax layer. Given an input image $I \in R^{H\times W\times 3}$, our model first uses CoarseCNN to extract a coarse-category feature representation $h_c \in R^{H\times W\times M}$, where each channel corresponds to one of the M coarse semantic categories. The h_c is then fed into FineCNN to obtain a fine-category feature representation $h_f \in R^{H\times W\times N}$, where each channel denotes one of the N fine categories. Finally, the h_f is processed by a softmax layer to predict an output segmentation map. To train the segmentation stream, we apply a cross entropy loss $\mathcal{L}^f_{\text{seg}}$ between S_f and the ground truth $\overline{S}_f$. Further, we also use h_c to generate a side output S_c, and supervise it with a cross entropy loss $\mathcal{L}^c_{\text{seg}}$ where the target is the ground truth coarse semantic map $\overline{S}_c$.

Our key contribution is a bidirectional cross-level feature matching framework to enable our model to learn multi-scale, hierarchy-aware features, by explicitly taking into account the HSP. In the *forward* matching stage, we train a coarse-to-fine module (C2F) to transform h_c into a fine-category feature representation $\widetilde{h}_f$ that explicitly encodes semantic hierarchy information, by leveraging the pre-defined HSP. The inputs to C2F include coarse-category features h_c, a coarse scene parsing map $\overline{S}_c$, and a sub-category occurrence vector v, where $v = (N_1, N_2, \ldots, N_M)^T$ is a M-dimensional vector and N_c represents the number of fine categories under coarse category c. The learned HSP will then be transferred into the mainstream segmentation network by matching $\widetilde{h}_f$ and h_f in a L_1 sense: $\mathcal{L}_{\text{forward}} = \|\widetilde{h}_f - h_f\|_1$. In the *backward* matching stage, we use a fine-to-coarse module (F2C) to collapse h_f to $\widetilde{h}_c$ based on v and regularize the learning of h_c by matching h_c to $\widetilde{h}_c$: $\mathcal{L}_{\text{backward}} = \|\widetilde{h}_c - h_c\|_1$. Such a backward matching will encourage h_c to adapt to the hierarchical structure in h_f and, therefore, ease the learning of the hierarchy-aware features.

Our framework is learned by minimizing a total loss: $w^f_{\text{seg}}\mathcal{L}^f_{\text{seg}} + w^c_{\text{seg}}\mathcal{L}^c_{\text{seg}} + w_{\text{forward}}\mathcal{L}_{\text{forward}} + w_{\text{backward}}\mathcal{L}_{\text{backward}}$, where w^f_{seg}, w^c_{seg}, w_{forward} and w_{backward} are balancing weights. It is worth noting that, unlike the strategies used in classification methods [40] that require hierarchical semantics structure to guide the inference of different sub-networks during inference, we only use C2F and F2C in the training stage, and our mainstream segmentation network will be used as a standalone module during inference. The $\overline{S}_c$ is obtained by changing categorical labels in the fine scene parsing map (e.g., car and motorcycle) into their coarse label (e.g., vehicle) according to a pre-defined HSP.

3.2 C2F Module

To learn a fine-category feature representation that explicitly encodes HSP, we propose a C2F module as shown in Figure 3. The inputs to this module are a coarse-category feature representation $h_c \in R^{H\times W\times M}$, a coarse segmentation map $\overline{S}_c$, and a sub-category occurrence vector $v = \{N_1, N_2, \ldots, N_M\}$. For each coarse category $c \in \{1, 2, \ldots, M\}$, we introduce a $k\times k$ convolutional layer with N_c filters to map h_c to a representation $\widetilde{h}_f \in R^{H\times W\times N_c}$, which captures the features of its sub-categories. To make the resulting representation as discriminative as possible, we apply a softmax operation to normalize the activations at each spatial location and render them to be more mu-

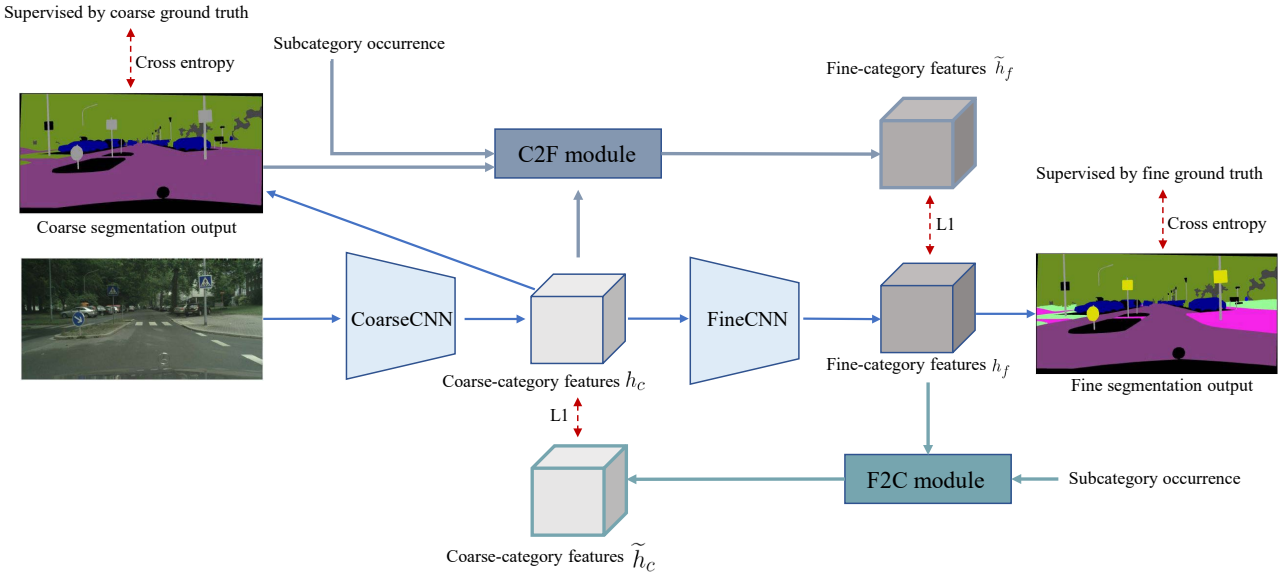


Fig. 2: The framework of our proposed network. The input image is sent to the CoarseCNN to generate the coarse-category features h_c, which is supervised by the coarse ground-truth. The h_c is then sent to the FineCNN to obtain the fine-category features h_f, which is supervised by the fine ground-truth. The h_c is also sent to C2F module to obtain $\widetilde{h}_f$ and the h_f is also sent to F2C module to obtain $\widetilde{h}_c$. Meanwhile, $\widetilde{h}_f$is learning h_f and $\widetilde{h}_c$is learning h_c.

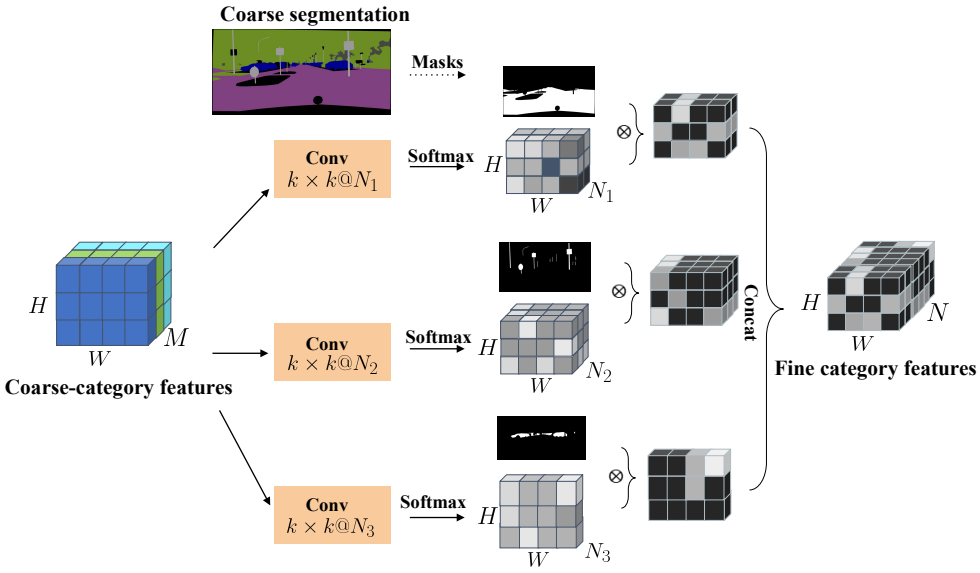


Fig. 3: Illustration of C2F module.

tually exclusive (i.e., amplify the difference among them). In addition, ideally, the representation for each coarse category should have zero activations at pixels that do not belong to the category. Therefore, we construct a binary spatial mask $\mathcal{M}_c \in \{0,1\}^{H \times W}$ for each coarse category c from the coarse segmentation S_c, so that $\mathcal{M}_c(x,y)$ is 1 if pixel (x,y) belongs to c and 0 otherwise. The $\mathcal{M}_c$ is then multiplied with $\widetilde{h}_c$ element-wise to zero out the values of those pixels that do not belong to c. Finally, the output fine-category feature representation $\widetilde{h}_f$ is formed by concatenating the resulting representations $\{\widetilde{h}_c\}_{c=1}^{M}$ in channel-wise.

The $\widetilde{h}_f$ is akin to a subdivision of h_c into fine-level categories, and is inherently aware of semantic hierarchy. This is because, for each spatial location on h_c, only the channels that belong to the same coarse category have non-zero activations. In addition, due to the use of softmax operations, the channel activations at each location can be interpreted as class probabilities, which enable a hierarchical classifi-

cation. More specifically, we may first look at the non-zero channels to identify which coarse category a pixel belongs to (coarse-level classification), and then discriminate among different fine categories using the values of the non-zero channels (fine-level classification).

3.3 F2C Module

F2C module can be regarded as a reverse process of the C2F module. It transforms fine-category features back to coarse-category features. However, since it is intended to regularize the learning of h_c, no learnable parameters are involved in this module. Intuitively, if there exists alignment between features across different semantic levels, it should be possible to abstract a fine-category feature representation back to a coarse-category one using the HSP. The F2C module is designed to enforce this property during learning. As shown in Figure 4, the F2C module first decomposes the fine-category feature representation h_f into a set of disjoint feature representations, each for a coarse category, according to a subcategory occurrence vector v. It then aggregates each resulting representation in channel-wise, producing a set of single-channel representations that are finally concatenated to output a coarse-category representation $\widetilde{h}_c$.

We adopt max pooling as the aggregation operation since it can preserve more detailed information, as compared with average pooling [3]. We have empirically found that max pooling is able to select discriminative features to allow our model to capture fine-grained patterns and improve its generalizability to complex scenes.

3.4 Discussion

Although hierarchy strategy is explored in some previous works [47, 51] and we both train the model in a coarse-to-fine manner. We would like to highlight our difference. In previous works, they learned the coarse features with noises and fine-tuned the features in the fining stage by cleaning the details, e.g., boundaries. They considered both coarse and fine features in fine-category. By contrast, our coarse features are already clear in details, but just in coarse-category, compared with the final outputs in fine-category. It indicates that our method is modeling the HSP to predict the segmentation.

In addition, it is worth pointing out that our method has the same number of parameters, compared with the backbone networks during test. That is to say, the C2F and F2C modules are only used for training. Hence,our method is also a one-stage method as the testing network is the same as its backbone network. It indicates that our method is also a effective method for distilling HSP information without any extra inference cost.

4 Experiments

In this section, we first compare our methods with state-of-the-art methods. We then conduct ablation studies to evaluate the bidirectional strategy and the effectiveness of C2F and F2C modules. Finally, we visualize the features learned by our model.

4.1 Experimental Setup

Implementation Details. We implement our model using Pytorch framework. Our model is trained on two Nvidia Titan RTX GPU cards. The initial learning rate is set to 0.01 and we decrease the learning rate using a polynomial policy with a decay rate of 0.9. The model is trained for $40,000$ iterations. We use random scaling (from 0.75 to 2.0) and random horizontal flipping to augment the training data. We take stage 1 and stage 2 of Resnet-101 [13] as the CoarseCNN, stage 3 and stage 4 of Resnet-101 and the pyramid pooling module [51] as the FineCNN. We adopt the pyramid pooling module [51] before the output to capture long-range context. In the experiment, the channel number for the coarse-category feature representations h_c and $\widetilde{h}_c$ is 7, and the channel number for the fine-category feature representations h_f and $\widetilde{h}_f$ is 19. Note that C2F and F2C are only used during training. The loss balance weight w^f_{seg},w^c_{seg}, w_{forward} and w_{backward} are all empirically set to 1 since these loss values are in the same order of magnitude. To get the optimal performance, we also take the multi-scale inference scheme and left-right flipping in test time, as in most scene parsing methods [7, 15, 41, 45, 51]. As all other compared methods, we use mean IoU (mIoU) to evaluate segmentation performance.

Datasets. We conduct our experiments on two popular datasets, Cityscapes [6] and Pascal Context [26].

- Cityscapes dataset is an urban street scene parsing dataset, which contains $5,000$ real-world images with pixel-level fine annotations. We use $2,975$ images for training, 500 images for validation and $1,525$ for testing. All images are of resolution 1024×2048. Cityscapes contains 19 fine categories under 7 coarse categories and their belonging relationship is given as Table 1. Hence, the subcategory occurrence vector of Cityscapes is $v = (2, 3, 3, 2, 1, 2, 6)^T$.
- Pascal Context is a set of additional annotations for PASCAL VOC 2010. Pascal Context dataset includes real-world $4,996$ images for training and $5,104$ images for validation. There are 59 fine categories. We group them following the rules of Cityscapes and add three coarse categories, which are furniture, animals and clothes. Hence, the subcategory occurrence vector of Cityscapes is $v = (3, 1, 8, 9, 9, 8, 1, 1, 10, 6, 3)^T$.

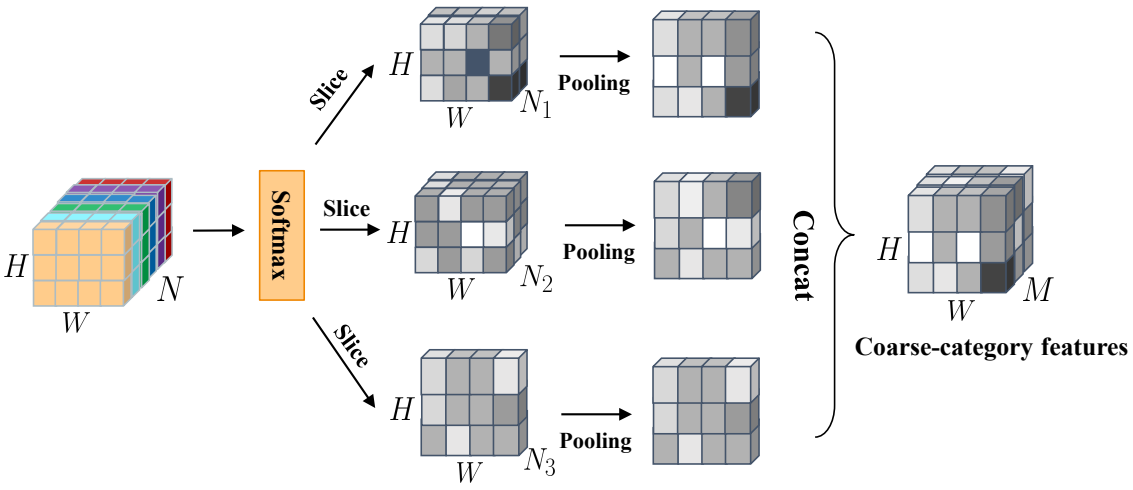


Fig. 4: Illustration of F2C module.

Coarse	Fine	HSP
flat	road sidewalk	2
construction	building wall fence	3
object	pole traffic sign traffic light	3
nature	vegetation terrain	2
sky	sky	1
human	person rider	2
vehicle	car truck bus train motorcycle bicycle	6

Table 1: The hierarchical semantics prior of Cityscapes.

Methods	Venue & Year	mIoU(%)
AED [33]	TVC 2020	72.5
PSPNet [51]	CVPR 2017	78.4
ACFNet [47]	ICCV 2019	81.8
ResNet-38 [37]	PR 2019	78.4
DGCNet [49]	BMVC 2019	80.8
ANN [55]	ICCV 2019	81.3
CPNet [43]	ECCV 2020	81.3
CCNet [15]	ICCV 2019	81.4
SpyGR [21]	CVPR 2020	81.6
BFP [7]	ICCV 2019	81.4
CDGCNet [14]	ECCV 2020	82.0
DependencyNet [24]	CVPR 2021	81.9
DAN-SAM [17]	TVC 2021	81.5
Ours		**82.3**

Table 2: Comparison with the state-of-the-art methods in terms of mIoU on the Cityscapes *test* set. The best results are in **bold**.

4.2 Comparison with State-of-the-art

Cityscapes. We compare our model with 13 state-of-the-art methods over Cityscapes dataset. As shown in Table 2, we can see that our method outperforms all the other methods on mIoU. Our model only uses the simple segmentation stream without the C2F and F2C modules during test time, but most of the other models need to use carefully designed networks in test time. This implies that our proposed model can effectively learn HSP, which is indeed helpful to scene parsing.

We choose to use PSPNet as a baseline since our segmentation stream has a similar architecture (using the pyramid pooling module that is the key contribution of PSPNet) and a similar number of parameters to PSPNet. Our model has a performance gain of 3.9% over PSPNet. In Figure 5, we also present the visual comparison with the baseline model as in many existing scene parsing works [15, 41, 44, 50]. The top three rows show that our model is able to discriminate similar fine categories, i.e., bus, truck and train in top two rows, vegetation and terrain in the third row. The fourth row shows that PSPNet may produce some noise in the segmentation (e.g., by mis-recognizing part of the sky as train.), while our model is more robust, giving a clearer output. In the last row, we can see that our model is able to recall some easily overlooked categories (e.g., a small part of the vegetation outlined). These qualitative and quantitative results show that the proposed model can obtain promising performance in scene parsing with the guidance of hierarchical semantics.

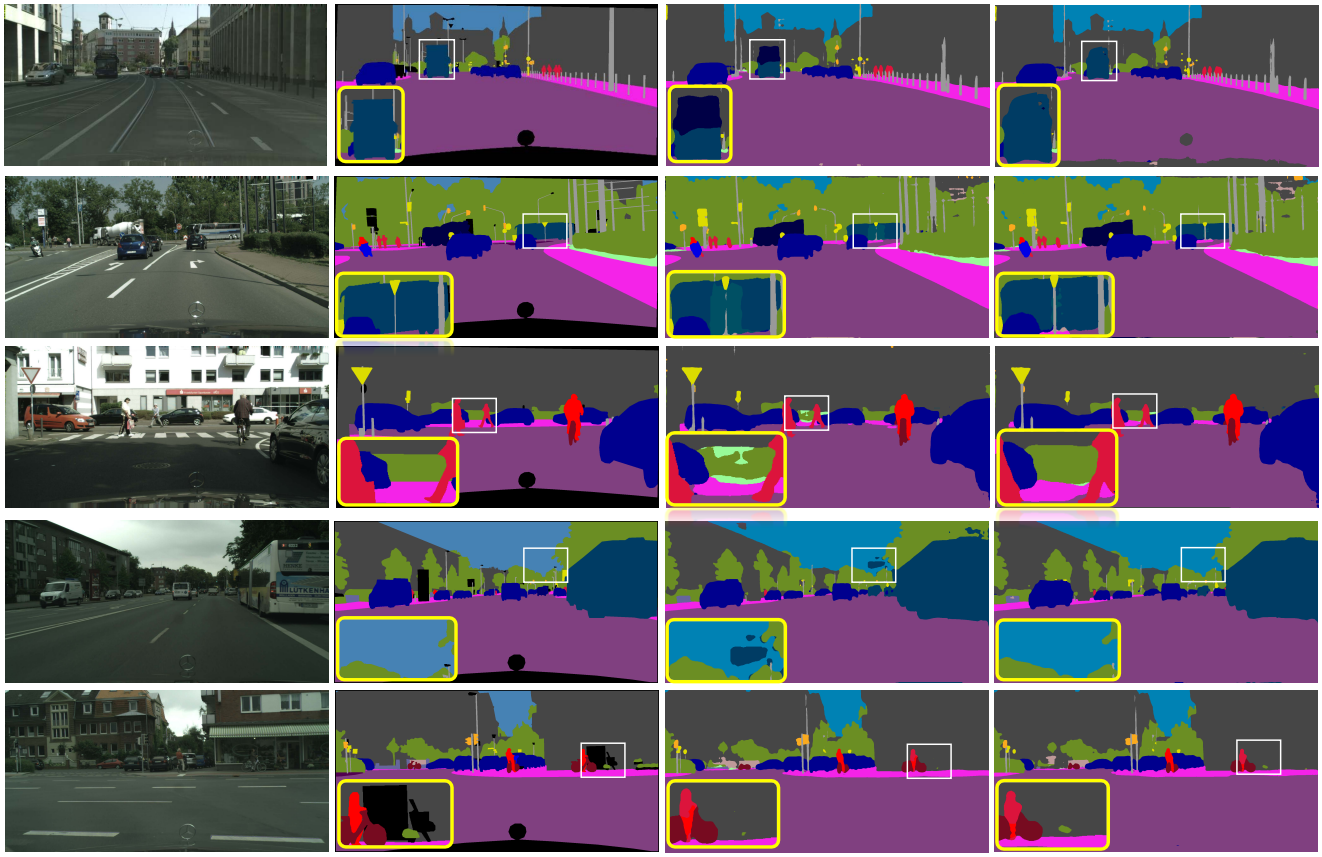


Fig. 5: Visual comparison of our model with the baseline. Our advantages are highlighted by white boxes, which are magnified at the bottom left corner of each image.

Pascal Context. We have conducted our method with two different models. One uses the pyramid pooling module [51] at the end of FineCNN (Ours-PSPNet), another replaces the pyramid pooling module to criss-cross attention module [15] (Ours-CCNet). The results are shown as Table 3. It can be seen that our method can improve different backbones and our method achieve the best results among the compared methods.

Methods	Venue & Year	mIoU(%)
Deeplab-v2 [4]	TPAMI 2017	45.7
RefineNet [23]	CVPR 2017	47.3
PSPNet [51]	CVPR 2017	47.8
MSCI [22]	ECCV 2018	50.3
EncNet [48]	CVPR 2018	51.7
DANet [11]	CVPR 2019	52.6
SVCNet [8]	CVPR 2019	53.2
DMNet [12]	ICCV 2019	54.4
CCNet [15]	ICCV 2019	52.1
SpyGR [21]	CVPR 2020	52.8
CPNet [43]	ECCV 2020	53.9
CFNet [36]	TMM 2021	52.4
EncNet+JPU [5]	CVPR 2021	52.8
Ours-PSPNet		51.2
Ours-CCNet		**54.9**

Table 3: Comparison with the state-of-the-art methods in terms of mIoU on the Pascal Context *val* set. The best results are in **bold**.

4.3 Ablation Study

To evaluate our network design, we have conducted four ablation studies over Cityscapes dataset. 1) We first study the effect of our bidirectional cross-level feature matching strategy, and then 2) analyze the effect of the proposed C2F and F2C modules, and 3) analyze the difference between L1 and L2 loss functions, finally 4) analyze the effect of loss weights. For simplicity, we use a batch size of two and single-scale predictions in this study.

The effect of our bidirectional cross-level feature matching strategy. To analyze the effect of the backward

stage, we explore two alternatives: 1) directly remove the L1 loss between h_c and $\widetilde{h}_c$ ($h_c \leftrightarrow \emptyset$); 2) remove the L1 loss and force $\widetilde{h}_c$ (rather than h_c) to match with the ground truth coarse-category segmentation ($(h_c, \widetilde{h}_c) \leftrightarrow (\emptyset, \overline{S}_c)$). Similarly, to analyze the effect of the forward stage, we 1) apply no supervision to h_f ($h_f \leftrightarrow \emptyset$), or 2) apply no supervision to h_f and use a fine-category segmentation loss on $\widetilde{h}_f$ ($(h_f, \widetilde{h}_f) \leftrightarrow (\emptyset, \overline{S}_f)$). Table 4 shows the performances of different settings. We can see that our bidirectional cross-level feature matching strategy can achieve the best performance while the models trained with only one-directional feature matching or without any feature matching perform worse.

Supervision	FWD	BWD	mIoU
$(h_c, h_f) \leftrightarrow (\emptyset, \emptyset)$	×	×	74.3
$(h_c, h_f, \widetilde{h}_c, \widetilde{h}_f) \leftrightarrow (\emptyset, \emptyset, \overline{S}_c, \overline{S}_f)$	×	×	75.1
$h_f \leftrightarrow \emptyset$	×	✓	75.1
$(h_f, \widetilde{h}_f) \leftrightarrow (\emptyset, \overline{S}_f)$	×	✓	75.3
$h_c \leftrightarrow \emptyset$	✓	×	75.7
$(h_c, \widetilde{h}_c) \leftrightarrow (\emptyset, \overline{S}_c)$	✓	×	75.8
Ours	✓	✓	**76.2**

Table 4: Results of the ablation study on our bidirectional cross-level feature matching strategy.

The effect of the C2F and F2C modules. To analyse the effect of these modules, we replace the C2F module by a 1×1 stride-1 convolutional layer with 19 filters (i.e., w/o C2F), and the F2C module by a 1×1 stride-1 convolutional layer with 7 filters (i.e., w/o F2C). Table 5 reports the performances with and without the C2F/F2C modules. We can see that the model without both C2F and F2C modules performs worst, and our model with both modules achieves the best performance.

Figure 6 presents visual comparison of different variants of our model. In the top row, our model with both modules can correctly segment the bike and its rider, while other variants misdetect part of the bike or rider as car. In the bottom row, our model can successfully detect the complete shape of the sidewalk, while other variants have difficulty in distinguishing between the sidewalk and road in some local region. These suggest that C2F and F2C are crucial to the performance of our model.

The difference between L1 and L2 loss functions. To analyse the difference of L1 and L2 loss functions, which are used between h_c and $\widetilde{h}_c$, and h_f and $\widetilde{h}_f$. Table 6 reports the different performances with L1 and L2 loss functions between h_c and $\widetilde{h}_c$, and h_f and $\widetilde{h}_f$.

It can be seen that L1 loss function performs better than L2 function. It is because L2 loss is more sensitive in the small details [54]. However, the difference between L1 and L2 loss functions is not very large, which indicates that our strategy is robust.

Method	mIoU
w/o C2F and F2C	74.5
w/o F2C	76.0
w/o C2F	75.5
Ours	**76.2**

Table 5: Performance comparison of our model with or without the C2F/F2C modules.

Method	L2	L1 (Ours)
mIoU	75.8	**76.2**

Table 6: Performance comparison of L1 and L2 loss functions, between h_c and $\widetilde{h}_c$, and h_f and $\widetilde{h}_f$.

The effect of loss weights. In this work, we have four loss functions with four weights, which are w_{forward}, w_{backward}, w^f_{seg} and w^c_{seg}, respectively. We have conducted 9 different loss weights as Table 7 shows. It can be seen that our model with four same loss weights can achieve the best results.

w_{forward}	w_{backward}	w^f_{seg}	w^c_{seg}	mIoU
1	1	2	2	74.9
1	1	5	5	72.9
2	2	1	1	73.4
5	5	1	1	67.9
1	2	1	2	74.2
2	1	2	1	74.8
1	5	1	5	74.0
5	1	5	1	68.2
1	1	1	1	**76.2**

Table 7: Performance comparison of different loss weights.

4.4 Visualization of Learned Features

We would like to further visualize the learned features of our model to investigate how well our model captures hierarchical semantics. To do this, we feed input images into the network, and visualize some channels in the coarse-category feature representation h_c along with their corresponding channels in the fine-category feature representation h_f. Figure 7 shows our results (Ours). We have two observations here. First, different channels emerge to detect different high-level semantic categories. For example, in the first row of Figure 7, the "flat" channel in h_c strongly responds to flat

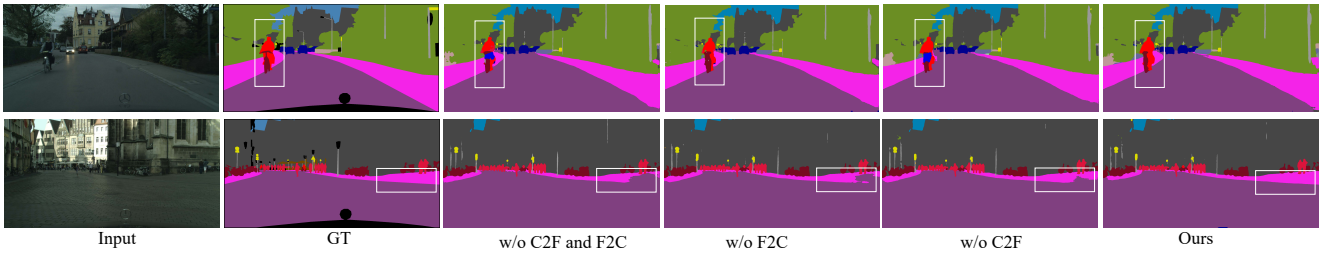


Fig. 6: Visual comparison of results with or without the C2F and F2C modules. Our advantages are highlighted by white boxes.

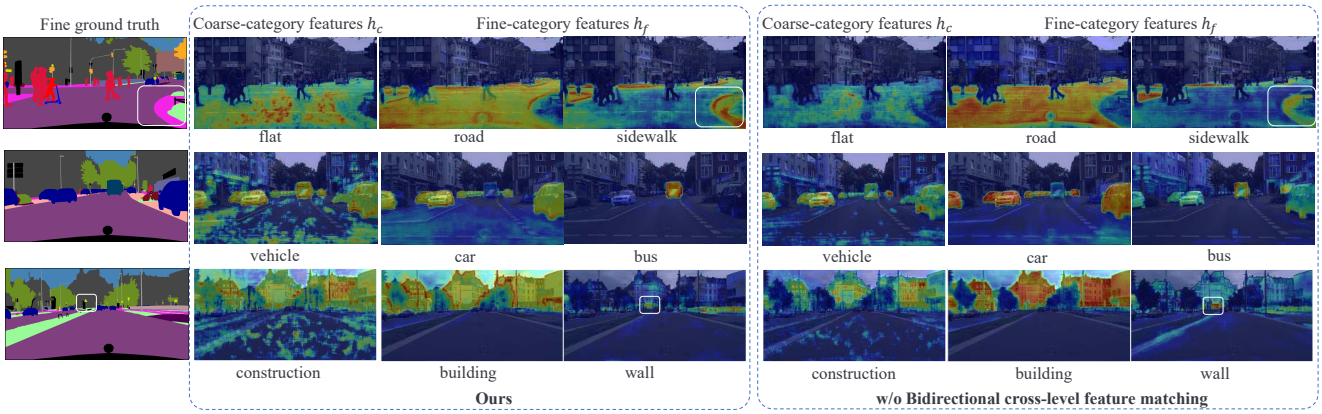


Fig. 7: Visualization of our learned features. For each channel in our coarse-category feature representation (e.g., flat), we show its corresponding channels in the fine-category feature representation (e.g., road and sidewalk). For comparison, we also show the visualized features of the model without bidirectional cross-level feature matching.

surfaces in the image, while the "road" and "sidewalk" channels are mainly activated by road and sidewalk regions, respectively. Second, our model learns an alignment between coarse- and fine-category features. For example, in the second row of Figure 7, the "vehicle" channel can be disentangled into the "car" and "bus" channels well. These suggest that our learned features are discriminative and can explicitly encode the hierarchical semantic prior (HSP).

To explore the effect of our bidirectional cross-level feature matching strategy upon the feature learning, in Figure 7, we also visualize the features of the model trained without bidirectional feature matching (i.e., removing F2C and C2F modules from our model). As can be seen, bidirectional feature matching allows our model to learn more discriminative features. For example, in the first row, our "sidewalk" channel has higher responses at the sidewalk region outlined by the white box. In addition, without bidirectional feature matching, the learned features are not disentangled well, particularly at fine category level. For example, in the second row, the "bus" channel also responds to cars, and in the third row, the "wall" channel generates high responses at a region belonging to "building".

4.5 Our Limitations

Our work does have a drawback. Our method requires the HSP prior, which defines the semantic categories in two level. Hence, our method can achieve competitive results on the dataset with a large number of semantic labels. However, it may not be seen the obvious advantages when the dataset has only a few semantic labels. Fortunately, most dataset contain many semantic labels since the real world is complicated.

5 Conclusion

In this paper, we have introduced the hierarchical semantics prior (HSP) to scene parsing. This prior, although common, has not be considered in scene parsing. The key contribution of our paper is to propose a bidirectional cross-level feature matching framework, which enables learning discriminative, multi-scale features that encode HSP. Experiments show that our proposed model achieves state-of-the-art performances on the Cityscapes and Pascal Context dataset. Our feature visualization reveals that a hierarchy of semantic categories automatically emerges in our learned features. We believe that this is the first work to bring hierarchical

semantics knowledge into scene parsing, and demonstrate that such knowledge is valuable for scene parsing. We envision that our work has potential to inspire follow-up works to further explore along this direction.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2019YFC1521104), National Natural Science Foundation of China (72192821, 61972157), Shanghai Municipal Science, Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200, 22YF1420300), a General Research Fund from RGC of Hong Kong (RGC Ref.: 11205620), and a Strategic Research Grant from City University of Hong Kong (Ref.: 7005674). Xin Tan is also supported by the Postgraduate Studentship (by Mainland Schemes) from City University of Hong Kong.

Compliance with ethical standards

The authors declare that they have no conflict of interest.

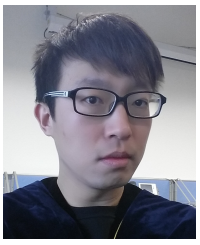
References

1. Karim Ahmed, Mohammad Haris Baig, and Lorenzo Torresani. Network of experts for large-scale image categorization. In *ECCV*, 2016. 3
2. Alsallakh Bilal, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Do convolutional neural networks learn class hierarchy? *TVCG*, 24(1):152–162, 2017. 3
3. Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *ICML*, pages 111–118, 2010. 5
4. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE TPAMI*, 2018. 7
5. Tse-Wei Chen, Deyu Wang, Wei Tao, Dongchao Wen, Lingxiao Yin, Tadayuki Ito, Kinya Osa, and Masami Kato. Cassod-net: Cascaded and separable structures of dilated convolution for embedded vision systems and applications. In *CVPR*, pages 3182–3190, 2021. 7
6. Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 3, 5
7. Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *ICCV*, 2019. 2, 5, 6
8. Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *CVPR*, pages 8885–8894, 2019. 7
9. Chaojie Fan, Yong Peng, Shuangling Peng, Honghao Zhang, Yuankai Wu, and Sam Kwong. Detection of train driver fatigue and distraction based on forehead eeg: A time-series ensemble learning method. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–11, 2021. 3
10. Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Huchuan Lu. Dual attention network for scene segmentation. *CVPR*, 2019. 2
11. Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019. 7
12. Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *ICCV*, pages 3562–3572, 2019. 7
13. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015. 5
14. Hanzhe Hu, Deyi Ji, Weihao Gan, Shuai Bai, Wei Wu, and Junjie Yan. Class-wise dynamic graph convolution for semantic segmentation. In *ECCV*, 2020. 2, 6
15. Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. *ICCV*, 2019. 2, 5, 6, 7
16. Jian Ji, Rui Shi, Sitong Li, Peng Chen, and Qiguang Miao. Encoder-decoder with cascaded crfs for semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 2
17. Min Jiang, Fuhao Zhai, and Jun Kong. Sparse attention module for optimizing semantic segmentation performance combined with a multi-task feature extraction network. *The Visual Computer*, pages 1–16, 2021. 6
18. Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson WH Lau, and Thomas S Huang. Geometry-aware distillation for indoor semantic segmentation. In *CVPR*, pages 2869–2878, 2019. 2
19. Norbert Kruger, Peter Janssen, Sinan Kalkan, Markus Lappe, Ales Leonardis, Justus Piater, Antonio J Rodriguez-Sanchez, and Laurenz Wiskott. Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *TPAMI*, 35(8):1847–1871, 2012. 1
20. Kibok Lee, Kimin Lee, Kyle Min, Yuting Zhang, Jinwoo Shin, and Honglak Lee. Hierarchical novelty detection for visual object recognition. In *CVPR*, 2018. 2, 3
21. Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *CVPR*, pages 8950–8959, 2020. 6, 7
22. Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *ECCV*, 2018. 7
23. Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, pages 1925–1934, 2017. 2, 7
24. Mingyuan Liu, Dan Schonfeld, and Wei Tang. Exploit visual dependency relations for semantic segmentation. In *CVPR*, pages 9726–9735, 2021. 2, 6
25. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2
26. Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 5
27. Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters – improve semantic segmentation by global convolutional network. In *CVPR*, pages 1743 – 1751, 2017. 2
28. Yong Peng, Yating Lin, Chaojie Fan, Qian Xu, Diya Xu, Shengen Yi, Honghao Zhang, and Kui Wang. Passenger overall comfort in high-speed railway environments based on eeg: Assessment and degradation mechanism. *Building and Environment*, 210:108711, 2022. 3
29. Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *ICCV*, 2019. 2

30. Youssef Tamaazousti, Hervé Le Borgne, and Céline Hudelot. Mucale-net: Multi categorical-level networks to generate more discriminating features. In *CVPR*, 2017. 2, 3
31. Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson W. H. Lau. Night-time scene parsing with a large real dataset. *IEEE Transactions on Image Processing*, 30:9085–9098, 2021. 2
32. Dan Wang, Guoqing Hu, and Chengzhi Lyu. Frnet: an end-to-end feature refinement neural network for medical image segmentation. *The Visual Computer*, 37:1101–1112, 2021. 2
33. Kang Wang, Jinfu Yang, Shuai Yuan, and Mingai Li. A lightweight network with attention decoder for real-time semantic segmentation. *The Visual Computer*, pages 1–11, 2021. 6
34. Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2
35. Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *ECCV*, pages 346–362, 2020. 2
36. Tianyi Wu, Sheng Tang, Rui Zhang, and Guodong Guo. Consensus feature network for scene parsing. *IEEE Transactions on Multimedia*, 2021. 7
37. Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 2019. 6
38. Ke Xu, Xin Tian, Xin Yang, Baocai Yin, and Rynson W. H. Lau. Intensity-aware single-image deraining with semantic and color regularization. *IEEE TIP*, 2021. 3
39. Ke Xu, Xin Wang, Xin Yang, Shengfeng He, Qiang Zhang, Baocai Yin, Xiaopeng Wei, and Rynson WH Lau. Efficient image super-resolution integration. *The Visual Computer*, 2018. 3
40. Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *ICCV*, pages 2740–2748, 2015. 3
41. Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018. 2, 5, 6
42. Xiang Ye, Heng Wang, and Yong Li. Image content-dependent steerable kernels. *The Visual Computer*, pages 1–12, 2021. 2
43. Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *CVPR*, pages 12416–12425, 2020. 6, 7
44. Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018. 6
45. Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 2018. 5
46. Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2019. 2
47. Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In *ICCV*, 2019. 2, 5, 6
48. Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, pages 7151–7160, 2018. 7
49. Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip HS Torr. Dual graph convolutional network for semantic segmentation. In *BMVC*, 2019. 2, 6
50. Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *ECCV*, 2018. 2, 6
51. Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2, 5, 6, 7
52. Chuanxia Zheng, Jianhua Wang, Weihai Chen, and Xingming Wu. Multi-class indoor semantic segmentation with deep structured model. *The Visual Computer*, 34(5):735–747, 2018. 2
53. Xiaoyang Zheng, Xin Tan, Jie Zhou, Lizhuang Ma, and Rynson W. H. Lau. Weakly-supervised saliency detection via salient object subitizing. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4370–4380, 2021. 3
54. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 8
55. Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *ICCV*, 2019. 6

Xin Tan is now a Ph.D. candidate at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. He is also a joint-Ph.D. student at the Department of Computer Science, City University of Hong Kong since 2019. He received his B.Eng. degree in Automation from the Chongqing University, China in 2017. His research interests lie in computer vision and deep learning, in particular, scene parsing and saliency detection. He serves as a program committee member/reviewer for CVPR, ICCV, AAAI, IJCAI and International Journal of Computer Vision (IJCV).

Jiachen Xu received his B.Eng. degree in Computer Science and Technology from the Central South University, China in 2019. He is now a master student at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. His research interests cover 3D point cloud and image segmentation.



Ying Cao received the Ph.D. degree in computer science from the City University of Hong Kong in 2014, and the M.Sc. and B.Eng. degrees in software engineering from Northeastern University, China, in 2010 and 2008, respectively. He was a Postdoctoral Fellow in the Department of Computer Science at the City University of Hong Kong from 2015 to 2016. His primary research interest lies in data-driven computational manga and graphic design.

Ke Xu is currently with the Department of Computer Science at City University of Hong Kong. He obtains the dual Ph.D. degrees from Dalian University of Technology and City University of Hong Kong. He has served as a program committee member/reviewer for several CV and AI

conferences and journals, including CVPR, ICCV, NeurIPS, ICLR, ICML, IJCV, TCSVT. His research interests include deep learning, image enhancement and editing.

Lizhuang Ma received his B.S. and Ph.D. degrees from the Zhejiang University, China in 1985 and 1991, respectively. He is now a Distinguished Professor, at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China and the School of Computer Science and Technology, East China Normal University, China. He was a Visiting Professor at the Frounhofer IGD, Darmstadt, Germany in 1998, and a Visiting Professor at the Center for Advanced Media Technology, Nanyang Technological University, Singapore from 1999 to 2000. His research interests include computer vision, computer aided geometric design, computer graphics, scientific data visualization, computer animation, digital media technology, and theory and applications for computer graphics, CAD/CAM. He serves as the reviewer of IEEE TPAMI, IEEE TIP, IEEE TMM, CVPR, AAAI etc.

Rynson W.H. Lau received the Ph.D. degree from the University of Cambridge. He was with the Faculty of Durham University. He is currently with the City University of Hong Kong. His research interests include computer graphics, image processing, and computer vision. He has also served in the committee of a number of conferences, including the Program Co-Chair of the ACM VRST 2004, the ACM MTDL 2009, the IEEE U-Media 2010, and the Conference Co-chair of CASA 2005, the ACM VRST 2005, the ACM MDI 2009, and the ACM VRST 2014. He has served as a Guest Editor of a number of journal special issues, including ACM Trans. on Internet Technology, IEEE Trans. on Multimedia, IEEE Trans. on Visualization and Computer Graphics, and IEEE Computer Graphics and Applications. He serves on the editorial board of International Journal of Computer Vision.