# Co-occurrent Features in Semantic Segmentation

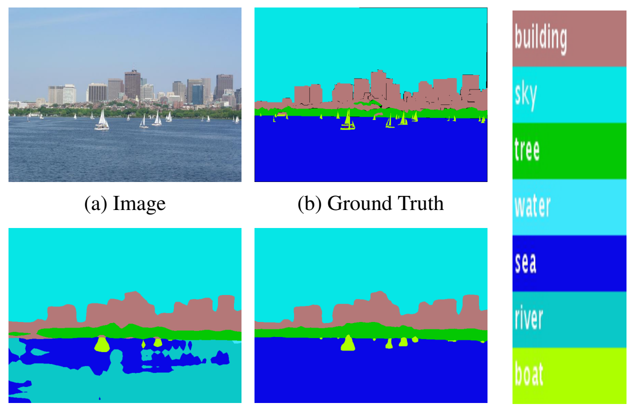
Hang Zhang [1] Han Zhang [2] Chenguang Wang [1] Junyuan Xie [1]

[1]Amazon Web Services [2]Google Brain

{hzaws,chgwang,junyuanx}@amazon.com, zhanghan@google.com

## Abstract

*Recent work has achieved great success in utilizing global contextual information for semantic segmentation, including increasing the receptive field and aggregating pyramid feature representations. In this paper, we go beyond global context and explore the fine-grained representation using co-occurrent features by introducing Co-occurrent Feature Model, which predicts the distribution of co-occurrent features for a given target. To leverage the semantic context in the co-occurrent features, we build an Aggregated Co-occurrent Feature (ACF) Module by aggregating the probability of the co-occurrent feature within the co-occurrent context. ACF Module learns a fine-grained spatial invariant representation to capture co-occurrent context information across the scene. Our approach significantly improves the segmentation results using FCN and achieves superior performance 54.0% mIoU on Pascal Context, 87.2% mIoU on Pascal VOC 2012 and 44.89% mIoU on ADE20K datasets. The source code and complete system will be publicly available upon publication*[1].




(a) Image (b) Ground Truth (c) FCN (baseline) (d) CFNet (ours) (e) legend

Figure 1: Some object categories are difficult to distinguish based on local appearance and scene context. In this example, *water, river* and *sea* are visually similar and all fit this scene context. Human can utilize the presence of the *boat* to make the prediction, as it typically co-occurs with the *sea*. Motivated by this, we introduce Aggregated Co-occurrent Feature Module to relook at the relations with all the co-occurrent features before making the predictions. (More visual examples in Figure 2)

## 1. Introduction

Semantic segmentation provides per-pixel label of object categories for the given image, which is a challenging task requiring accurate prediction of the object category, location and shape. Successful approaches are usually based on Fully Convolutional Network (FCN) [31], with a Deep Convolutional Neural Network (CNN) [22, 23] as the base network. Recent work achieves great success in leveraging contextual information, including enlarging receptive field size with pyramid-based representations [6,29,53] and learning category specific scaling factors using context embedding [51].

Despite the success in incorporating global contextual information, in some challenging scenarios, a rough holistic global context might not be enough for the classification of ambiguous objects in the scene. In addition, we observe natural scenes usually have reasonable and coherent composition of objects. The presence of one object, even in a spatially disjoint region, can be compelling evidence of the existence of the other. The co-occurrence property among objects can improve the robustness of the recognition system and help resolve the ambiguity of object labels against noises such as occlusion and variations in pose and illumination. For example, as shown in Figure 1, *sea, river* and *water* are very similar in appearance and the global context as a city scene is not able to disambiguate these three as they can all exist near a city. But object co-occurrence asserts that *sea* is more likely to appear when *boat* is around.

[1]Links can be found at http://hangzh.com/

Moreover, co-occurrence does not only exist between objects and it can also be generalized to different parts of an object. As shown in the $1^{st}$ row of Figure 2, an *armchair* is composed of armrests, legs, back and seat. It is difficult to resolve the ambiguity between a *chair* with an *armchair* without noticing the co-occurring armrest parts. In general, co-occurrent features play an important role in recognizing the class labels of image pixels. Therefore, a powerful approach directly capturing the co-occurrent features and utilizing their dependencies is desirable for semantic segmentation.

Existing approaches are not capable to capture the dependencies between co-occurrent features due to their fixed spatial structure. The baseline FCN [31] has a relatively local receptive field and fails to utilize co-occurrent features in distant portions of the image. Recent work simply enlarges the receptive field by utilizing the multi-scale feature representations using pyramid pooling method [17, 53] or different atrous rates of convolutions [6]. So the same pooling or atrous convolution operation is applied everywhere in the feature map. However, the distribution of crucial features for the recognition of different image regions varies tremendously. Instead of having fixed spatial connection, the network should be able to capture co-occurrent features across different relative locations, in a spatial invariant manner.

As the first contribution of this work, the feature co-occurrence is modeled as a probability distribution over the feature space conditioned on a given target feature, which we refer to as *Co-occurrent Feature Model (CFM)*. The CFM learns an inherent co-occurrence representation, where the similarities between features measure how likely the features would co-occur with the target in the same image. We therefore define a probability distribution conditioned on target feature using Softmax of the similarities between the target and co-occurrent features across the space, which inherits the spatial invariant nature. Moreover, we expect the co-occurrent features also capture the scene context. However, we find that the limitation in expressiveness of the Softmax distribution is a bottleneck for modeling the context information. For this, we propose a contextual prior as a conditional probability on the scene context. The CFM is then defined as a mixture of Softmaxes distribution with the contextual prior. With the proposed CFM, we build *Aggregated Co-occurrent Feature (ACF) Module* to integrate the context-aware information within the co-occurrent features, which allows the network to recap the whole scene before making individual predictions (overview in Figure 3).

The second contribution of this paper is constructing *Co-occurrent Feature Network* (CFNet), the state-of-the-art semantic segmentation architecture. With the proposed ACF Module, we build CFNet with pre-trained ResNet [18] as the base network. The proposed CFNet with ResNet-101 base network achieves state-of-the-art results 54.0% mIoU on Pascal Context [33], 87.2% mIoU on Pascal VOC 2012 [12] and 44.89% mIoU on ADE20K [56].

## 2. Co-occurrent Features

We refer to the features co-occurring with the target feature within the same input image/featuremap as *co-occurrent features*. In this section, we first introduce the *Co-occurrent Features Model* to capture the distribution of the co-occurrent features for a given target. We further introduce *Aggregated Co-occurrent Feature Module* to aggregate the contextual information of co-occurrent features across the scene as the output target feature representation.

### 2.1. Co-occurrent Feature Model

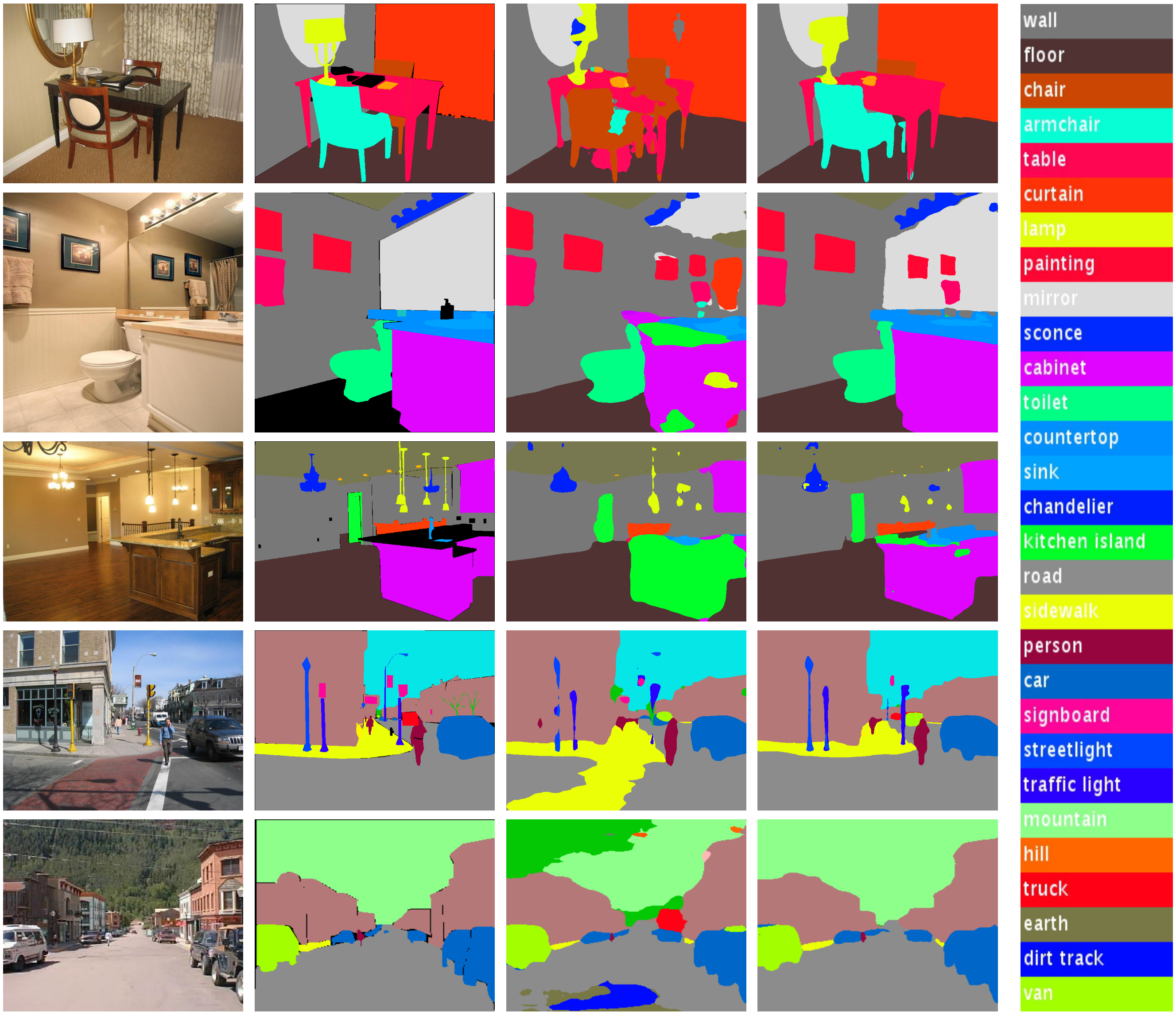
We tackle the feature co-occurrences as a probabilistic problem instead of predicting their presences, since the co-occurrent features for a given target are usually not deterministic. We build a *Co-occurrent Feature Model*, which learns an inherent representation via measuring the similarity between the co-occurrent feature and the target feature, indicating how likely they would co-occur[2]. Then the probability distribution of the co-occurrent features conditioned on target feature can be defined using Softmax of the similarities across the space. Consider the input CNN featuremap as $N$ number of channel-dimensional features $X = \{x_1, ...x_N\}$, and $x_i$ for $i \in \{1, ...N\}$ is the input feature at location $i$. The probability of the co-occurrent feature $x_c$ for a given target feature $x_t$ is:

$$p(x_c|x_t) = \frac{e^{s(x_c,x_t)}}{\sum_{i=1}^{N} e^{s(x_i,x_t)}}, \tag{1}$$

where $s(x_c, x_t)$ is the similarity between the co-occurrent feature $x_c$ and the target feature $x_t$. A natural parameterization for the similarity function $s$ is using dot product similarity $s(x_c, x_t) = u_{x_c}^{\top} v_{x_t}$, where $v_{x_t}$ and $u_{x_c}$ are the target and co-occurrent vector representations for feature $x_t$ and $x_c$. The vector representations are given by $u_{x_c} = \Phi_c(x_c)$ and $v_{x_t} = \Phi_t(x_t)$, where $\Phi_c$ & $\Phi_t$ are the learnable transformations using feed-forward networks. The proposed model in Eq. 1 is in the same spirit with the skip-gram model proposed in [32], which is used to capture the co-occurrent word representation.

**Contextual Prior.** We find it is difficult to model the co-occurrent features only using the target information without knowing the whole scene, because the distribution of the co-occurrent features for the same target varies in different context. For example, we may expect chairs or tables

[2]Inspired by the distributed hypothesis [16]: *the target feature representations are modeled to predict well co-occurrent features in its context.*




(a) Image (b) Ground Truth (c) FCN (baseline) (d) CFNet (ours) (e) Legend

Figure 2: Some challenging category labels are difficult to distinguish even using global semantic context, which requires understanding the fine-grained details in the co-occurrent features. In the $1^{st}$ example, it is hard to know whether the *chair* is a *armchair* without noticing co-occurring arm. For the $2^{nd}$ example, the baseline FCN fails to predict *mirror* parts that are far from *sconce*. Similarly, FCN fails to utilize the spatial layout to distinguish the *cabinet* with *kitchen island*. The proposed CFNet relooks at the relations with the co-occurrent features before classifying each pixel, which successfully segments the above mentioned objects and also distinguishes the *road* from *sidewalk* and *dirt track*, segments the *mountain* as a whole part in the last two examples. (Visual examples from ADE20K dataset [56].)

co-occurring with a human in the indoor scene, but expect vehicles and buildings instead in the outdoor scene. Recent study also shows that the Softmax-based models do not have enough capacity for high-rank problems [49]. We can hypothesize that predicting co-occurrent features is a high-rank problem in images, which we can show with empirical observations. If the co-occurrent features are low-rank, we could use finite number of bases to represent all possible co-occurrent features by a weighted combinations of these bases. However, this contradicts with our common sense about the varieties of the real-world images. Therefore, predicting co-occurrent features is a high-rank problem.

To tackle the above issues, we propose to model the scene context as *contextual prior*. Inspired by Yang *et*

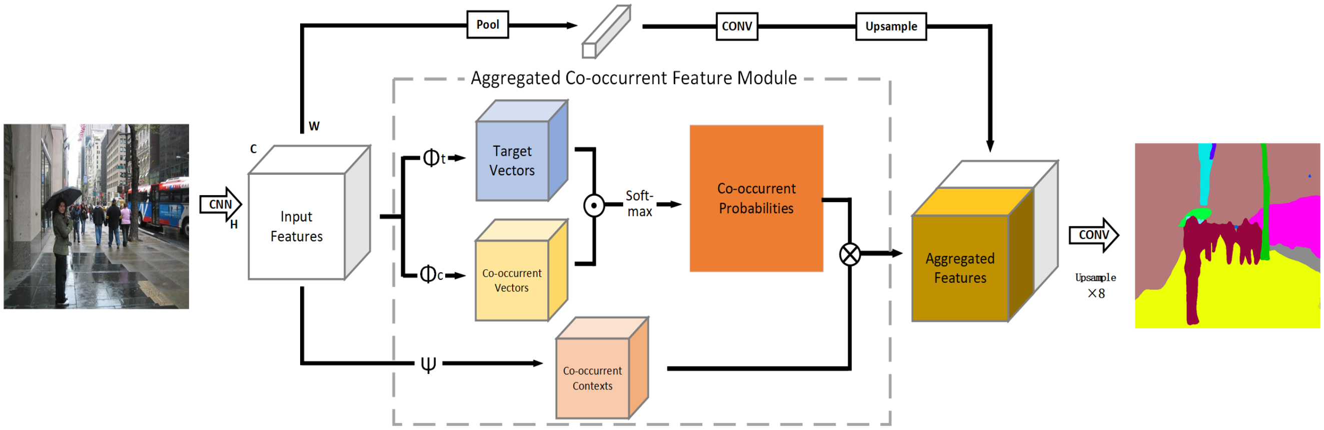



Figure 3: Overview of the proposed CFNet. Given an input image, the convolutional featuremaps are extracted using pre-trained CNN. Then the featuremaps are transformed into target and co-occurrent vector representations using the feed-forward networks $\Phi_t$ and $\Phi_c$. The Co-occurrent Feature Model estimates the probability of the co-occurrent features based on the pair-wise similarities in the vector spaces. The ACF Module aggregates the co-occurrent context captured using transformation $\Psi$ with the co-occurrent probability. Another branch captures the global feature using global average pooling followed by a convolution. Then the upsampled featuremaps are concatenated with ACF model output along the channel dimension. Finally, the concatenated feature representation is fed into the last convolutional layer to make the per-pixel predictions. ($\odot$ represents pairwise dot-product similarity, $\otimes$ means aggregation operation)

*al.* [49], the contextual prior is defined as a Mixture of Softmaxes (MoS) to learn a prior distribution for the co-occurrent features conditioned on the contextual information. The MoS formulates the co-occurrent probability of $x_c$ for target $x_t$ as:

$$p(x_c|x_t) = \sum_{k=1}^{K} \pi^k \frac{e^{s^k(x_c,x_t)}}{\sum_{i=1}^{N} e^{s^k(x_i,x_t)}}, \tag{2}$$

where $\pi^k$ is the prior or mixture weight of the $k$-th component, and $s^k(x_c, x_t)$ is the similarity in $k$-th component for $k \in \{1, ...K\}$. The vector representations $u_{x_c}$ and $v_{x_t}$ are chunked into $K$ sub-components, and the similarity of each component is given by $s^k(x_c, x_t) = u_{x_c,k}^\top v_{x_t,k}$. The prior of each mixture is conditional on the contextual information, which can be parameterized as $\pi^k = \frac{exp(w_k^\top \bar{v_x})}{\sum_{k'=1}^{K} exp(w_{k'}^\top \bar{v_x})}$, where $\bar{v_x} = \sum_i^N \frac{v_{x_i}}{N}$ captures the contextual information and $w_k$ is a learnable vector. The MoS allows the co-occurrent features under different semantic context can have different priors.

### 2.2. Aggregated Co-occurrent Feature Module

To utilize the co-occurrent features, we build *Aggregated Co-occurrent Feature Module* (ACF), which aggregates the co-occurrent contexts with their co-occurrent probabilities across the spatial locations in a self-attention [42] or non-local [3] manner:

$$z_t = \sum_{c=1}^{N} p(x_c|x_t) \cdot \psi_c, \tag{3}$$

where $z_t$ is the aggregated feature output for target $t$, $p(x_c|x_t)$ is the co-occurrent probability given by Equation 2 and $\psi_c$ is the co-occurrent context at location $c$. The co-occurrent context is given by $\psi_c = \Psi(x_c)$, and $\Psi$ is a learnable transformation using feed-forward network. The ACF Module captures the co-occurrent feature distributions, and aggregates the contextual information with the co-occurrent probabilities.

**Dropout and Multi-head Ensembles.** Model combination almost always improves the performance for machine learning algorithms. Dropout [40] randomly drops units during the training, so that it learns "thinned" networks and averages the logits during the inference. Dropout can avoid the network adapting too much on the training data for over-fitting. We apply dropout [40] on the co-occurrent features and expect the network to make correct predictions even if some of the concurrent features are missing, so that the network can generalize from limited patterns appeared in the training set. Another model combination we explore is using "multi-head" [42], which concatenates the features of module outputs using different weights to build a in-network ensemble. We adapt the multi-head strategy to further improve the model capacity.

**Global Pooling Feature.** Global pooling (GP) feature is commonly used in modern semantic segmentation approaches [6, 29, 53], which provides a global receptive field as a strong cue to distinguish category in confusing areas. The GP feature is captured by a global average pooling, followed by a 1×1 convolution, and then attached to each feature location. We extend the proposed Co-occurrent Feature Module with a global pooling feature branch to leverage the global context, as shown in Figure 3.

## 3. Co-occurrent Feature Network

With proposed Co-occurrent Feature Module, we build *Co-occurrent Feature Network* (CFNet) as shown in Figure 3. We use pre-trained ResNet [18] as the base network and apply dilated network strategy to Res-4 and Res-5[3] of ResNet, resulting stride-8 models. The proposed Aggregated Co-occurrent Feature Module and global pooling feature branch are added on top of the base network. ACF Module considers the input convolutional featuremap with the shape of $C \times H \times W$ as a set of $C$-dimensional features $X = \{x_1, ..x_N\}$, where $N = H * W$ is total number of features. The input features are transformed into vector spaces using $\Phi_t$ and $\Phi_c$. We use shared weights for the target transformation $\Phi_t$ and the co-occurrent transformation $\Phi_c$, and apply 3×3 average pooling with stride of 2 before the co-occurrent transformation $\Phi_c$ to reduce the computation. The probabilities of co-occurrent features are predicted based on the similarities in the vector space. Then the proposed ACF Module aggregates the co-occurrent contexts with the co-occurrent probabilities. The Co-occurrent Feature Model learns the co-occurrent feature distribution by learning the transformation functions $\Phi_t$ and $\Phi_c$ in vector spaces.

In another branch, a global pooling feature is concatenated to the output featuremap after convolution and upsampling. Finally, the last convolution predicts the per-pixel prediction of the object categories. We upsample the prediction featuremap by 8 times to make its size equal to input image size to calculate the segmentation loss. Since the proposed ACF Module is differentiable and can be learned with the rest of the network, it is compatible with existing FCN based algorithms.

### 3.1. Relation to Other Methods

**Semantic Segmentation.** CNN based method has achieved remarkable success in semantic segmentation and scene parsing. Early work classifies each individual patches/regions for generating segmentation masks [13, 15]. FCN [31] first replaces the fully connected layers of pre-trained network with the convolution layers for semantic segmentation. The adaption of CNN for image classification suffers from the loss of spatial resolution. DeconvNet [34] and SegNet [2] learn a decoder to recover the information from downsampled features. Applying Atrous/Dilated convolution in pre-trained network produces larger featuremap [5, 50]. UNet [38] concatenates the lower level features to the featuremap as skip connections. Prior work also adapts Dense CRF as post-processing to refine the FCN prediction boundaries [5, 8]. CRF-FCN allows end-to-end learning of CRF with FCN [55]. Recent work refines segmentation boundaries [26, 55] and increases spatial resolution [36]. Hwang *et al.* [20] and Ke *et al.* [21] also exploit label co-occurrence and structural label dependencies. These work mainly focus on the network training regularization, while our approach improves the network representation by directly model the feature co-occurrence.

[3]We refer to the network stages with the original strides of 16 and 32 as Res-4 and Res-5.

| **Network** | $224^2$ **center** | | $320^2$ **center** | |
|---|---|---|---|---|
| | top-1 | top-5 | top-1 | top-5 |
| ResNet-50 | 78.55 | 94.17 | 79.33 | 94.64 |
| ResNet-101 | 80.24 | 95.12 | 80.63 | 95.50 |

Table 1: Imagenet [10] pretraining for the base networks. The top-1 and top-5 accuracy (%) on validation set use center crop on image size of 224×224 and 320×320.

**Context Aggregation.** Pioneering work demonstrates that combining global features with local patches can improve the segmentation results [37, 39, 41]. ParseNet [29] proposes to concatenate a global pooling feature with original featuremap to capture global context and increase the receptive field size. Pyramid Pooling Module (PPM) [17, 53] concatenates the global pooling features from a multi-scale pyramid. Atrous Spatial Pyramid Pooling (ASPP) [6] uses a set of different atrous rate convolutions to capture pyramid feature representations with different receptive field sizes.

These methods have the predefined spatial connections. Consider the convolution operation as a matrix multiplication $y = Wx$, where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ are the flatten input and output and $W \in \mathbb{R}^{mn}$ is a transform matrix depending on the convolution kernel [11]. Matrix $W$ has $k_h \cdot k_w$ non-zero elements in each row for the convolution kernel with the shape of $k_h \times k_w$. Combining different atrous rate of convolutions as in ASPP [6] is adding the non-zero entries to each row, but the overall spatial connections are still sparse and the representation is spatial sensitive. The proposed ACF Module can also be formulated as $y = Wx$ and $W$ is the co-occurrent probability. Comparing to existing methods, the ACF Module captures the context across the whole scene with spatial invariant representation.

**Featuremap Attention.** Attention mechanism has achieved great success in natural language processing [35, 42], which captures the long-range information using a weighted sum of all the features in a sequence. Non-local neural network [45] brings the self-attention to the field of video classification and object detection in computer vision. A key difference between Co-occurrent Feature Model with

| Method | BaseNet | ACF | GP | Enc | pixAcc% | mIoU% |
|---|---|---|---|---|---|---|
| FCN | Res50 | | | | 76.3 | 46.3 |
| FCN | Res50 | ✓ | | | 79.0 | 49.8 |
| CFNet | Res50 | ✓ | ✓ | | 79.3 | 51.6 |
| CFNet | Res50 | ✓ | ✓ | ✓ | 79.8 | 52.4 |
| CFNet | Res101 | ✓ | ✓ | ✓ | 81.1 | 54.9 |

Table 2: Ablation study of CFNet on Pascal Context dataset. *ACF* indicates using Aggregated Co-occurrent Feature Module, *GP* means including global pooling feature branch, *Enc* represents Context Encoding Module [51]. Adding Co-occurrent Feature significantly improves the segmentation results, and including global pooling feature and Context Encoding can further boost the performance.

| pixAcc/mIoU | K=1 | K=2 | K=4 |
|---|---|---|---|
| H=1 | 77.1/49.4 | 79.3/51.6 | 79.5/51.8 |
| H=2 | 77.9/49.4 | 79.4/51.8 | 79.3/51.6 |
| H=4 | 79.2/51.2 | **79.6/52.1** | –[4] |

Table 3: Ablation Study of Contextual Prior and Multi-heads. We vary the number of mixtures K and number of multi-heads H and find H=4, K=2 gives the best performance.

self-attention or non-local network is that the proposed Co-occurrent Feature Model learns a prior distribution conditioned on the semantic context. In semantic segmentation, EncNet [51] calculates the pair-wise similarity between input and learnable codewords and predicts a set of channel-wise attention factors, which can be considered as co-occurrent features in channel dimension. PSANet [54] learns a long range context aggregation using a location sensitive non-local neural network strategy for semantic segmentation.

# 4. Experimental Results

In this section, we first explain the technical details for the implementation of the proposed CFNet and baseline FCN. Then we conduct a comprehensive ablation study of the proposed ACF Module and CFNet on Pascal Context dataset [33]. Then we report the performance of CFNet on Pascal VOC 2012 [12], ADE20K [56] and Cityscapes [9] datasets.

## 4.1. Implementation Details

For baseline FCN and proposed CFNet, we use ResNet [18] as the base network and apply dilation strategy for the pre-trained networks, resulting in stride-8 models. Following the prior work [51, 53], we use bilinear interpolation to upsample the network output logits for calculating the loss. We use standard SGD optimizer and set the momentum to 0.9 and weight decay to 0.0001. A "poly" like learning rate scheduling [5] is used $lr = base\_lr * (1 - \frac{iter}{total\_iter})^{power}$. We set the base learning rate as 0.004 for ADE20K and Cityscapes datasets and the power is set to 0.9. We use base learning rate of 0.004 for COCO pretraining and reduce it to 0.001 when fine-tuning on Pascal VOC. We use the "sync-once" implementation of Cross-GPU Batch Normalization provided by Zhang *et al.* [51]. As ACF Module is compatible with existing FCN based approaches, we also study the performance when adding Context Encoding Module and Semantic Encoding Loss with default settings in EncNet [51]. Following the prior work [53], an auxiliary loss is added after Res-4 by adding an additional FCN head to Res-4, which is applied to all the experiments.

| Method | BaseNet | mIoU% |
|---|---|---|
| FCN-8s [31] | | 37.8 |
| CRF-RNN [55] | | 39.3 |
| ParseNet [29] | | 40.4 |
| HO_CRF [1] | | 41.3 |
| Piecewise [27] | | 43.3 |
| VeryDeep [46] | | 44.5 |
| DeepLab-v2 [5] | Res101 + COCO | 45.7 |
| RefineNet [26] | Res152 | 47.3 |
| MSCI [25] | Res152 | 50.3 |
| EncNet [51] | Res101 | 51.7 |
| CFNet (ours) | Res50 | 51.5 |
| CFNet (ours) | Res101 | **54.0** |

Table 4: Segmentation results on PASCAL-Context dataset. (Note: mIoU on 60 classes w/ background.)

The networks are trained for 120 epochs for ADE20K dataset, 180 epochs on Cityscapes dataset, 30 epochs for COCO pretraining, 50 epochs on Pascal VOC and 80 epochs on Pascal Context dataset. The images and ground truth masks are randomly flipped and rescaled to the ratio of 0.5 to 2.0 and randomly cropped into the training sizes using zero padding if needed. We use the mini-batch size of 16 during the training. The samples are randomly shuffled, and the last batch is discarded if mini-batch size is less than 16.

**Evaluation and Metrics.** During the evaluation, we follow the best practice [51] to average the network predictions in multiple scales. We first resize the original image into different scales {0.5, 0.75, 1.0, 1.25, 1.5, 1.75}, then crop the scaled images into training image size and feed the images into the network with flipping. Finally, the predicted logits are averaged across different crops and scales. Since the multi-size evaluation improves the performance of all the methods, we adopts this strategy for all the experiments. We use the standard metrics of pixel accuracy (pixAcc) and mean intersection of union (mIoU) in this experiments. For the scene parsing results on Pascal Context and ADE20K validation sets, we ignore the background pixels in calculating the evaluation metrics, following the standard benchmark [56]. For the semantic segmentation results on Pascal VOC and Cityscapes datasets, we use the public server for the evaluation.

**ImageNet Pretraining.** Similar as in the prior work [51, 53], we modify the standard ResNet [17] by replacing the first 7×7 convolution with 3 consequent 3×3 convolution. We follow the best practise of ImageNet training [19] to train our base networks. The Top-1 and Top-5 accuracy on ImageNet validation set using center crop with the crop size of 224×224 and 320×320 are shown in Table 1. The pre-trained models will be included in the public code system.

### 4.2. Abalation Study on Pascal Context

Pascal Context dataset [33] is a scene parsing dataset, containing the semantic labels for the entire image, with 4,998 training and 5,105 validation images. Following the practice in prior work [5, 26, 33, 51], we use the 59 most frequent categories for this benchmark and consider all the other classes as background.

**Ablation Study of CFNet.** We first break down the improvements of CFNet over FCN, by conducting a set of experiments by adding individual components step-by-step to the baseline FCN. We use 4 mixtures and 2 multiheads in ACF Module with atrous-rate of 4 for the transformation $\Phi$ in this study. The baseline FCN achieves 76.3% pixAcc and 46.3% mIoU. Adding the ACF Module improves the pixAcc and mIoU by 2.7% and 3.6%. Including global pooling feature yields 0.9% boost in mIoU. Further improvements are from adding Context Encoding Module [51] and using deeper base network (See results in Table 2).

**Ablation Study of ACF Module.** To explore the best performance of Aggregated Co-occurrent Feature Module, we conduct the experiments with different hyper-parameters and settings. We first study different instantiations of the transformation $\Phi$ using different feed-forward network architectures and empirically find using the atrous rate of 4 gives best performance (detailed study in the supplementary material). We also explore the influence of contextual prior and multi-heads in the ACF Module in Table 3. To keep the comparison fair, we reduce the feature dimension, when increasing the number of mixtures or the number of multi-heads, so that the total computation of ACF Module remains roughly the same. Varying the number of mixtures K for contextual prior and the number of muli-heads H in ACF Module, we can see that using contextual prior significantly improves the expressiveness of the Softmax model, and empirically find K=2 and H=4 gives the best performance.

**State-of-the-art Comparisons.** We consider the background as one of the categories in order to compare with prior work (60 classes in total). The results are shown in Table 4. CFNet with ResNet-50 already outperforms most of the previous work even using much shallower base network. CFNet (ResNet-101) achieves 54.0% mIoU on validation set, which surpasses other approaches by a large margin even without using deeper base network or COCO pre-training.

### 4.3. Results on Pascal VOC 2012

Pascal VOC 2012 segmentation dataset [12] is one of the gold-standard benchmarks for object segmentation. Following the work [5, 53], we utilize the augmented set [14] with 10,582, 1,449 and 1,456 images in training, validation and test set. The CFNet is first trained on the train + val sets on the augmented set and then finetuned on the original Pascal VOC 2012 images as in previous work [51]. For fair comparision with prior work, we use ResNet-101 as the base network. CFNet-101 achieves 84.2% mIoU[5] on the test set, which outperforms all the previous work without COCO pre-training and achieves superior performance on most of the categories. State-of-the-art approaches typically pre-train the network using MS-COCO dataset [28]. We follow the prior work [6, 51] to generate semantic segmentation mask by merging the instance labels for the 20 categories shared with Pascal VOC 2012 dataset, and discard the labels for the other categories, which results in around 90K images with more than 1000 labeled pixels (from the training set of MS-COCO 2017). We first pre-train the CFNet on COCO dataset using learning rate of 0.004 and then finetune on the augmented and original training set. CFNet achieves the best result of 87.2%[6] on the test set. The per-class comparision is shown in Table 5, and CFNet achieves superior performance on many categories. (The entries using larger base model such as Xeption, or extra than COCO [28] & ImageNet [10] data for pre-training are not included in this benchmark [7, 25, 44].)

### 4.4. Results on ADE20K

ADE20K dataset [56] is a large scale scene parsing benchmark with 150 object and stuff categories, containing 20K training, 2K validation and 3K test images. We first train the baseline FCN and CFNet on the training set and evaluate the models on the validation set (results are shown in Table 6). Our baseline FCN using ResNet-50 achieves 39.28% mIoU using good pre-trained base network and multi-size evaluation. CFNet outperforms FCN by more

[5] http://host.robots.ox.ac.uk:8080/anonymous/ZFDFXP.html

[6] http://host.robots.ox.ac.uk:8080/anonymous/SOWG4O.html

| Method | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FCN [31] | 76.8 | 34.2 | 68.9 | 49.4 | 60.3 | 75.3 | 74.7 | 77.6 | 21.4 | 62.5 | 46.8 | 71.8 | 63.9 | 76.5 | 73.9 | 45.2 | 72.4 | 37.4 | 70.9 | 55.1 | 62.2 |
| DeepLabv2 [4] | 84.4 | 54.5 | 81.5 | 63.6 | 65.9 | 85.1 | 79.1 | 83.4 | 30.7 | 74.1 | 59.8 | 79.0 | 76.1 | 83.2 | 80.8 | 59.7 | 82.2 | 50.4 | 73.1 | 63.7 | 71.6 |
| CRF-RNN [55] | 87.5 | 39.0 | 79.7 | 64.2 | 68.3 | 87.6 | 80.8 | 84.4 | 30.4 | 78.2 | 60.4 | 80.5 | 77.8 | 83.1 | 80.6 | 59.5 | 82.8 | 47.8 | 78.3 | 67.1 | 72.0 |
| DeconvNet [34] | 89.9 | 39.3 | 79.7 | 63.9 | 68.2 | 87.4 | 81.2 | 86.1 | 28.5 | 77.0 | 62.0 | 79.0 | 80.3 | 83.6 | 80.2 | 58.8 | 83.4 | 54.3 | 80.7 | 65.0 | 72.5 |
| GCRF [43] | 85.2 | 43.9 | 83.3 | 65.2 | 68.3 | 89.0 | 82.7 | 85.3 | 31.1 | 79.5 | 63.3 | 80.5 | 79.3 | 85.5 | 81.0 | 60.5 | 85.5 | 52.0 | 77.3 | 65.1 | 73.2 |
| DPN [30] | 87.7 | 59.4 | 78.4 | 64.9 | 70.3 | 89.3 | 83.5 | 86.1 | 31.7 | 79.9 | 62.6 | 81.9 | 80.0 | 83.5 | 82.3 | 60.5 | 83.2 | 53.4 | 77.9 | 65.0 | 74.1 |
| Piecewise [27] | 90.6 | 37.6 | 80.0 | 67.8 | 74.4 | 92.0 | 85.2 | 86.2 | 39.1 | 81.2 | 58.9 | 83.8 | 83.9 | 84.3 | 84.8 | 62.1 | 83.2 | 58.2 | 80.8 | 72.3 | 75.3 |
| ResNet38 [47] | 94.4 | **72.9** | 94.9 | 68.8 | 78.4 | 90.6 | 90.0 | 92.1 | **40.1** | 90.4 | 71.7 | 89.9 | 93.7 | **91.0** | 89.1 | 71.3 | 90.7 | 61.3 | 87.7 | 78.1 | 82.5 |
| PSPNet [53] | 91.8 | 71.9 | 94.7 | 71.2 | 75.8 | 95.2 | 89.9 | **95.9** | 39.3 | 90.7 | 71.7 | 90.5 | 94.5 | 88.8 | **89.6** | 72.8 | 89.6 | **64.0** | 85.1 | 76.3 | 82.6 |
| EncNet [51] | 94.1 | 69.2 | 96.3 | **76.7** | **86.2** | **96.3** | **90.7** | 94.2 | 38.8 | 90.7 | **73.3** | 90.0 | 92.5 | 88.8 | 87.9 | 68.7 | 92.6 | 59.0 | 86.4 | 73.4 | 82.9 |
| CFNet (ours)[5] | **95.7** | 71.9 | **95.0** | 76.3 | 82.8 | 94.8 | 90.0 | **95.9** | 37.1 | **92.6** | 73.0 | **93.4** | **94.6** | 89.6 | 88.4 | **74.9** | **95.2** | 63.2 | **89.7** | **78.2** | **84.2** |
| **With MS-COCO Pre-training** | | | | | | | | | | | | | | | | | | | | | |
| CRF-RNN [55] | 90.4 | 55.3 | 88.7 | 68.4 | 69.8 | 88.3 | 82.4 | 85.1 | 32.6 | 78.5 | 64.4 | 79.6 | 81.9 | 86.4 | 81.8 | 58.6 | 82.4 | 53.5 | 77.4 | 70.1 | 74.7 |
| Dilation8 [50] | 91.7 | 39.6 | 87.8 | 63.1 | 71.8 | 89.7 | 82.9 | 89.8 | 37.2 | 84.0 | 63.0 | 83.3 | 89.0 | 83.8 | 85.1 | 56.8 | 87.6 | 56.0 | 80.2 | 64.7 | 75.3 |
| DPN [30] | 89.0 | 61.6 | 87.7 | 66.8 | 74.7 | 91.2 | 84.3 | 87.6 | 36.5 | 86.3 | 66.1 | 84.4 | 87.8 | 85.6 | 85.4 | 63.6 | 87.3 | 61.3 | 79.4 | 66.4 | 77.5 |
| Piecewise [27] | 94.1 | 40.7 | 84.1 | 67.8 | 75.9 | 93.4 | 84.3 | 88.4 | 42.5 | 86.4 | 64.7 | 85.4 | 89.0 | 85.8 | 86.0 | 67.5 | 90.2 | 63.8 | 80.9 | 73.0 | 78.0 |
| DeepLabv2 [5] | 92.6 | 60.4 | 91.6 | 63.4 | 76.3 | 95.0 | 88.4 | 92.6 | 32.7 | 88.5 | 67.6 | 89.6 | 92.1 | 87.0 | 87.4 | 63.3 | 88.3 | 60.0 | 86.8 | 74.5 | 79.7 |
| RefineNet [26] | 95.0 | 73.2 | 93.5 | 78.1 | 84.8 | 95.6 | 89.8 | 94.1 | 43.7 | 92.0 | 77.2 | 90.8 | 93.4 | 88.6 | 88.1 | 70.1 | 92.9 | 64.3 | 87.7 | 78.8 | 84.2 |
| ResNet38 [47] | 96.2 | 75.2 | **95.4** | 74.4 | 81.7 | 93.7 | 89.9 | 92.5 | 48.2 | 92.0 | 79.9 | 90.1 | 95.5 | 91.8 | 91.2 | 73.0 | 90.5 | 65.4 | 88.7 | 80.6 | 84.9 |
| PSPNet [53] | 95.8 | 72.7 | 95.0 | 78.9 | 84.4 | 94.7 | **92.0** | 95.7 | 43.1 | 91.0 | 80.3 | 91.3 | 96.3 | 92.3 | 90.1 | 71.5 | 94.4 | 66.9 | 88.8 | **82.0** | 85.4 |
| DeepLabv3 [6] | 96.4 | 76.6 | 92.7 | 77.8 | **87.6** | **96.7** | 90.2 | 95.4 | 47.5 | 93.4 | 76.3 | 91.4 | **97.2** | 91.0 | **92.1** | 71.3 | 90.9 | 68.9 | **90.8** | 79.3 | 85.7 |
| EncNet [51] | 95.3 | 76.9 | 94.2 | **80.2** | 85.2 | 96.5 | 90.8 | **96.3** | 47.9 | 93.9 | 80.0 | 92.4 | 96.6 | 90.5 | 91.5 | 70.8 | 93.6 | 66.5 | 87.7 | 80.8 | 85.9 |
| CFNet (ours)[6] | **96.9** | **79.7** | 94.3 | 78.4 | 83.0 | 96.7 | 91.6 | **96.7** | **50.1** | **95.2** | 79.6 | **93.6** | **97.2** | **94.2** | 91.7 | **78.4** | **95.4** | **69.6** | 90.0 | 81.4 | **87.2** |

Table 5: Per-class results on PASCAL VOC 2012 testing set. CFNet-101 outperforms existing approaches and achieves 84.2% and 87.2% mIoU w/o and w/ pre-training on COCO dataset. (The best two entries in each columns are marked in gray color. Note: the entries using larger base networks or extra data are not included [7, 25, 44].)

| **Method** | **BaseNet** | **mIoU%** |
|---|---|---|
| RefineNet [26] | Res152 | 40.7 |
| UperNet [48] | Res101 | 42.66 |
| PSPNet [53] | Res101 | 43.29 |
| DSSPN [24] | Res101 | 43.68 |
| SAC [52] | Res101 | 44.30 |
| EncNet [51] | Res101 | 44.65 |
| FCN (baseline) | Res50 | 39.28 |
| CFNet (ours) | Res50 | 42.87 |
| CFNet (ours) | Res101 | **44.89** |

Table 6: Results on ADE20K validation set. CFNet-101 outperforms all previous methods in mIoU using same base network.

than 3.5% mIoU using same base network. CFNet-101 achieves 44.89% mIoU and outperforms all previous methods in mIoU using the same base network. Visual comparison examples are shown in Figure 2. The proposed CFNet successfully captures and utilizes the semantic dependencies of the co-occurrent features across the entire image for making predictions, while the baseline FCN only utilizes the local feature representations.

### 4.5. Results on Cityscapes Dataset

Cityscapes dataset [9] is a high-resolution city street scene parsing dataset, including 5K high-quality labeled frames (fine data) and 20K weakly annotated ones (coarse data). We only use the fine data in this experiment with 2,975, 500 and 1,525 number of images for training, validation, and testing. 19 object/stuff categories are used in the evaluation. We use ResNet-101 as the base network, and train our CFNet on the training set using 768 crop size, then evaluate it on the validation set. CFNet achieves 79.56% mIoU on the validation set. For performance on test set, we retrain CFNet-101 on train and validation set and submit the prediction on test set to the evaluation server. CFNet achieves 79.60% mIoU (IoU classes) on the test set only using fine-label data. We have not used online hard example mining (OHEM) strategy in this experiment, which can further improve the performance.

## 5. Conclusion

To capture and utilize the co-occurrent features, we introduce a Co-occurrent Feature Model, which predicts the probability distribution of co-occurrent features for the given target. To further utilize co-occurrent features in semantic segmentation, we introduce an Aggregated Co-occurrent Feature Module to aggregate the co-occurrent context of the co-occurrent features. The proposed approach outperforms existing contextual modules achieving superior performance on gold-standard semantic segmentation benchmarks. We expect the co-occurrent feature representation and our state-of-the-art implementations will be beneficial to the segmentation work in the community.

## References


[1] A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision*, pages 524–540. Springer, 2016. 6

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 5

[3] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 60–65. IEEE, 2005. 4

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 8

[5] L.-C. 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. *arXiv:1606.00915*, 2016. 5, 6, 7, 8

[6] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 2, 4, 5, 7, 8

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018. 7, 8

[8] M. Cimpoi, S. Maji, and A. Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3828–3836, 2015. 5

[9] 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*, pages 3213–3223, 2016. 6, 8

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 5, 7

[11] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016. 5

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 6, 7

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 5

[14] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. 2011. 7

[15] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014. 5

[16] Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, pages 346–361. Springer, 2014. 2, 5, 7

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 2, 5, 6

[19] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li. Bag of tricks to train convolutional neural networks for image classification. *arXiv preprint arXiv:1812.01187*, 2018. 7

[20] J.-J. Hwang, T.-W. Ke, J. Shi, and S. X. Yu. Adversarial structure matching loss for image segmentation. *arXiv preprint arXiv:1805.07457*, 2018. 5

[21] T.-W. Ke, J.-J. Hwang, Z. Liu, and S. X. Yu. Adaptive affinity fields for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 587–602, 2018. 5

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[24] X. Liang, H. Zhou, and E. Xing. Dynamic-structured semantic propagation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 752–761, 2018. 8

[25] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang. Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 603–619, 2018. 6, 7, 8

[26] G. Lin, A. Milan, C. Shen, and I. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, July 2017. 5, 6, 7, 8

[27] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016. 6, 8

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 7

[29] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 1, 4, 5, 6

[30] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1377–1385, 2015. 8

[31] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1, 2, 5, 6, 8

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 2

[33] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 2, 6, 7

[34] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015. 5, 8

[35] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016. 5

[36] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe. Fullresolution residual networks for semantic segmentation in street scenes. *arXiv preprint*, 2017. 5

[37] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Computer vision, 2007. ICCV 2007. IEEE 11th international conference on*, pages 1–8. IEEE, 2007. 5

[38] 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*, pages 234–241. Springer, 2015. 5

[39] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009. 5

[40] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014. 4

[41] A. Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003. 5

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 4, 5

[43] R. Vemulapalli, O. Tuzel, M.-Y. Liu, and R. Chellapa. Gaussian conditional random field network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3224–3233, 2016. 8

[44] G. Wang, P. Luo, L. Lin, and X. Wang. Learning object interactions and descriptions for semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5859–5867, 2017. 7, 8

[45] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[46] Z. Wu, C. Shen, and A. v. d. Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*, 2016. 6

[47] Z. Wu, C. Shen, and A. v. d. Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv preprint arXiv:1611.10080*, 2016. 8

[48] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. *arXiv preprint arXiv:1807.10221*, 2018. 8

[49] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen. Breaking the softmax bottleneck: A high-rank rnn language model. *arXiv preprint arXiv:1711.03953*, 2017. 3, 4

[50] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 5, 8

[51] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 6, 7, 8

[52] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan. Scale-adaptive convolutions for scene parsing. In *Proc. 26th Int. Conf. Comput. Vis.*, pages 2031–2039, 2017. 8

[53] 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. 1, 2, 4, 5, 6, 7, 8

[54] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia. PSANet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 6

[55] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015. 5, 6, 8

[56] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proc. CVPR*, 2017. 2, 3, 6, 7