Class Guided Channel Weighting Network for Fine-Grained Semantic Segmentation

Xiang Zhang, Wanqing Zhao*, Hangzai Luo, Jinye Peng, Jianping Fan

Northwest University, Xi'an, China
{ZhangXiang2015@stumail., zhaowq@, hzluo@, pjy@, jfan@}nwu.edu.cn

Abstract

Deep learning has achieved promising performance on semantic segmentation, but few works focus on semantic segmentation at the fine-grained level. Fine-grained semantic segmentation requires recognizing and distinguishing hundreds of sub-categories. Due to the high similarity of different sub-categories and large variations in poses, scales, rotations, and color of the same sub-category in the fine-grained image set, the performance of traditional semantic segmentation methods will decline sharply. To alleviate these dilemmas, a new approach, named Class Guided Channel Weighting Network (CGCWNet), is developed in this paper to enable fine-grained semantic segmentation. For the large intra-class variations, we propose a Class Guided Weighting (CGW) module, which learns the image-level fine-grained category probabilities by exploiting second-order feature statistics, and use them as global information to guide semantic segmentation. For the high similarity between different sub-categories, we specially build a Channel Relationship Attention (CRA) module to amplify the distinction of features. Furthermore, a Detail Enhanced Guided Filter (DEGF) module is proposed to refine the boundaries of object masks by using an edge contour cue extracted from the enhanced original image. Experimental results on PASCAL VOC 2012 and six fine-grained image sets show that our proposed CGCWNet has achieved state-of-the-art results.

Introduction

Deep learning has achieved great success in semantic segmentation (Chen et al. 2018b,a; Wang et al. 2020; Arani et al. 2021), while semantic segmentation at the fine-grained level (i.e., sub-category level) has received little attention. Fine-grained semantic segmentation is a fundamental and challenging problem, whose goal is to recognize and distinguish multiple subordinate categories (e.g., "Boeing 737-200" and "Boeing 737-500") of a super-category (e.g., "Aircraft"). The study of this task can be applied to many practical applications, such as new retail, automatic driving, robot sensing, and image editing.

In fine-grained semantic segmentation, due to large variations in poses, scales, rotations, and color of the same sub-category, some parts of the prediction will be incorrect (see the first and second row of Figure 1). We describe this property as "intra-class inconsistency". Meanwhile, objects of different sub-categories from the same super-category usually have quite similar visual appearances, which will cause the prediction to be confused (see the third and fourth row of Figure 1). We describe this property as "inter-class indistinction". Most existing traditional semantic segmentation methods (Yu et al. 2018b; Zhang et al. 2019a; Li et al. 2019; Hou et al. 2020; Wang et al. 2020; Arani et al. 2021) have obtained promising results. However, in our experiments, we found that the performance of these methods will decline sharply in fine-grained semantic segmentation. The main reasons come from the following two aspects: 1) These methods regard the semantic segmentation as an intensive recognition problem and usually generate predictions based on the local receptive field of CNN. Among different sub-categories, some parts of objects usually have the same appearance, which will bring difficulties to traditional segmentation methods; 2) Since the goal of fine-grained semantic segmentation needs to distinguish complex boundaries from hundreds or thousands of sub-categories, the previous feature representation lacks the ability to distinguish such subtle differences. As shown in Figure 1, traditional segmentation methods like SPNet (Hou et al. 2020) may produce false predictions when suffering from the two confusions (i.e., intra-class inconsistency and inter-class indistinction).

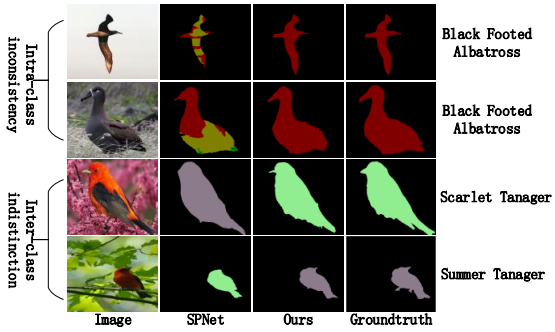

Figure 1: Main challenges of fine-grained semantic segmentation, i.e., intra-class inconsistency and inter-class indistinction. The second column is the SPNet based model. The third column is the output of our approach. The last column is the ground truth.

*Wanqing Zhao is the corresponding author.

To this end, we propose a Class Guided Channel Weighting Network (CGCWNet), which considers both the intra-class consistency and inter-class distinction to better enable fine-grained semantic segmentation.

To alleviate the problem of intra-class inconsistency, we proposed a Class Guided Weighting (CGW) module, which is designed based on the following two considerations: 1) Different appearances of the same sub-category may produce false predictions. Therefore, we use second-order feature statistics to learn robust feature representations for different appearances of the same sub-category; 2) The convolution operation is calculated from local receptive fields, which may lead to inconsistent predictions of semantic labels for a semantic object. The image-level category information can effectively provide global consistency guidance. To this end, we use the image-level fine-grained category probabilities as global information to guide fine-grained semantic segmentation. Besides, we build a Channel Relationship Attention (CRA) module, which enables the communication between channels to adaptively weight the features with high distinguishable ability to alleviate the problem of inter-class indistinction. Each channel of high-level features can be regarded as a class-specific response. Through the interaction between channels, the response channels of specific categories can be aggregated to highlight their feature representation, thereby enhancing the distinguishability of features for subtle differences between sub-categories. Specifically, we first obtain the channel relationship matrix by second-order feature statistics in the CGW module. Then, the channel relationship matrix is regarded as a channel crossing weight to enhance the distinguishable of features, so as to better complete the segmentation of similar sub-categories. Furthermore, a Detail Enhanced Guided Filter (DEGF) module is proposed to refine the boundaries of object masks by using an edge contour cue extracted from the enhanced original image. Three contributions of this paper are:

- To the best of our knowledge, this is the first study about fine-grained semantic segmentation which will be helpful for researchers in this field.
- Our CGW, CRA, and DEGF modules can support semantic segmentation at the fine-grained level by effectively alleviating the problems of intra-class inconsistency and inter-class indistinction. Moreover, those modules can easily be seamlessly integrated into most existing segmentation networks to improve their performance.
- We extend the fine-grained image classification datasets (i.e., FGVC Aircraft (Maji et al. 2013), CUB-200-2011 (Xiao et al. 2015), Stanford Cars (Krause et al. 2013), and "Orchid" Plant) to fine-grained segmentation datasets. In our experiments, the proposed CGCWNet has achieved state-of-the-art results on PASCAL VOC 2012 (Hariharan et al. 2011) and the expanded six fine-grained image sets.

Related Work

Encoder-Decoder. Encoder-Decoder architectures are widely used for semantic segmentation, where an encoder is used to reduce the feature maps and enlarge the receptive fields, and a decoder is used to recover the spatial information. Ronneberger et al. (Ronneberger, Fischer, and Brox 2015) introduce skip-connections to combine the low-level feature maps with the higher-level ones, which can enrich the details of segmentation results. SegNet (Badrinarayanan, Kendall, and Cipolla 2017) uses the pool indices to recover the reduced spatial information. Recently, DeepLabv3+ (Chen et al. 2018b) has achieved better performance by taking advantage of encoder-decoder architecture and atrous convolution. Some other works (Oliveira et al. 2020; Wang et al. 2020; Arani et al. 2021; Nirkin, Wolf, and Hassner 2021) also use the encoder-decoder architecture to improve the performance of semantic segmentation.

Global Context. Context can enlarge the receptive field to improve the performance of semantic segmentation. Yu et al. (Yu et al. 2018a,b) utilize global average pooling (GAP) to generate image-level context information. The atrous spatial pyramid pooling (ASPP) (Chen et al. 2017) is proposed to capture the spatial context based on different dilated rates. PSPNet (Zhao et al. 2017) uses the pyramid pooling module to partition the feature map into different scale regions. Several works (Liu, Rabinovich, and Berg 2015; Zhang et al. 2018; Liu et al. 2020; Chen et al. 2020) adopt global pooling to harvest the context. In contrast to the global context described above, in this paper, we propose a CGW module, which harvests the global context information from a categorical perspective. To be specific, we use category probabilities as global information to guide fine-grained semantic segmentation to unify semantic labels for all pixels of the same object, inherently considers intra-class consistency.

Attention Mechanism. Attention mechanism has shown its effectiveness in improving the performance of image recognition (Zhao, Jia, and Koltun 2020; Hou, Zhou, and Feng 2021; Vaswani et al. 2021), object detection (Hu et al. 2018; Zhang et al. 2020a), and semantic segmentation (Zhang et al. 2018; Fu et al. 2019; Huang et al. 2019; Zhong et al. 2020; Liu et al. 2021). For the task of semantic segmentation, Chen et al. (Chen et al. 2016) learn an attention mechanism to weight the multi-scale features softly. Zhong et al. (Zhong et al. 2020) design a squeeze-and-attention network architecture that leverages the squeeze-and-attention (SA) module to account for two distinctive characteristics (i.e., pixel-group attention and pixel-wise prediction) of segmentation. Numerous works (Zhang et al. 2018; Yu et al. 2018b,a) adopt channel attention to select the desired feature maps. Some researchers (Wang et al. 2018; Zhu et al. 2019) recently utilize self-attention to aggregate long-range contextual information. The previous works only explore first-order statistics, while ignoring the statistics higher than first-order, thus hindering the discriminative ability of the network. In this paper, we adopt second-order feature statistics to achieve the interaction between channels to enhance the distinction of features, thereby enlarging inter-class distinction.

Second-order Statistics. Second-order statistics have been studied in the context of texture recognition (Dai, Yue-Hei Ng, and Davis 2017) through so-called Region Covariance Descriptors (RCD), and further applied to image recognition (Gao et al. 2019; Koniusz and Zhang 2020) and image

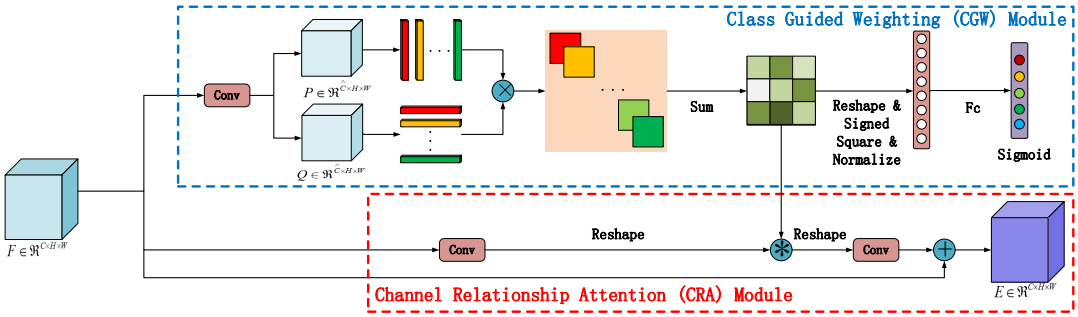

Figure 2: The flowchart for the CGW module and the CRA module, where $+$, $*$ and $\times$ represent the element-wise addition, matrix multiplication, and matrix outer products, respectively.

super-resolution (Dai et al. 2019). Some approaches (Lin, RoyChowdhury, and Maji 2015; Koniusz and Zhang 2020) perform second-order pooling for fine-grained image recognition. Li et al. (Li et al. 2017) propose a matrix normalized covariance (MPN-COV) for exploring the second-order statistics in large-scale classification. Dai et al. (Dai et al. 2019) design a second-order attention network to achieve more powerful feature representation and feature correlation learning, thereby achieving accurate image super-resolution. Our work is inspired by second-order statistics and we apply it to build CGW and CRA modules to alleviate the problems of intra-class inconsistency and inter-class indistinction in fine-grained semantic segmentation. Instead of computing second-order statistics of activations in the last convolutional layer as image representations in previous works, we use it to construct a channel relationship matrix to adaptive weight the original features to obtain the higher-level semantic features with high distinguishable ability.

Methodology

In this section, we first elaborate on the details of three proposed modules and how they can effectively handle the issues of intra-class inconsistency and inter-class indistinction. Then, we introduce our CGCWNet architecture.

Class Guided Weighting Module

Class Guided Weighting (CGW) module aims to alleviate the problem of intra-class inconsistency. In the design of the CGW module, we make the outer product of the feature maps as higher-level features for subsequent operations. The outer product of the feature maps can be considered as the second-order feature statistics, which can obtain higher-level semantic information to improve the identification ability and robustness of the network. Furthermore, we use the image-level fine-grained category probabilities as global information to improve the consistency of the same sub-category.

Specifically, given the input feature map (e.g., output of ResNet-101) $F \in \Re^{C\times H\times W}$ (C, H, W represent the number of channels, height, width), we apply one 1×1 convolution layer on F for channel reduction to obtain a low-dimensional feature map $A \in \Re^{\widehat{C}\times H\times W}$. Then A is copied to two same branches P, Q. Next, the channel relationship matrix $M_{(i,j)} \in \Re^{\widehat{C}\times\widehat{C}}$ at a location (i, j) in $H \times W$ can be calculated by an outer product as:

$$M_{(i,j)} = P_{(i,j)} \times Q_{(i,j)} \tag{1}$$

where $P_{(i,j)}, Q_{(i,j)} \in \Re^{1\times\widehat{C}}$ are the feature vectors P, Q at location (i, j). Then, the sum pooling aggregates the features across all the locations in the image to obtain an image descriptor $\Phi = \sum_{(i,j)} M_{(i,j)}, \Phi \in \Re^{\widehat{C}\times\widehat{C}}$. Φ is then passed through the square root step ($Z = sign(\Phi)\sqrt{|\Phi|}$), and followed by l_2 normalization ($X = Z/||Z||_2$) inspired by (Perronnin, Sánchez, and Mensink 2010). The Φ captures the pairwise correlations between feature channels and can model part-feature interactions, so that the response channels of a specific category can be aggregated to highlight its feature representation, thereby enhancing the discrimination of features and avoiding confusion between different subcategories.

Traditional semantic segmentation methods are limited by local perception, leading to inconsistent prediction results. We use the image-level fine-grained category probabilities as the weights to enhance the class-dependent feature maps, which encourages the prediction of our network to unify semantic labels for pixels of the same object and avoid confusion between different sub-categories. Specifically, we use a fully-convolutional layer on the top of the feature X with a sigmoid function, which outputs the category-based prediction probabilities $\gamma = \delta(\omega X)$, where ω denotes the layer weights, and δ is the sigmoid function. Then, the output of this module γ will be used as global information to weight semantic segmentation. The overall structure for the CGW module is shown in Figure 2.

Channel Relationship Attention Module

In fine-grained semantic segmentation, the prediction is always confused between different sub-categories with similar appearance. The inter-class indistinction problem is mainly due to the lack of distinguishing features. To this end, we build a CRA module, which enables the communication between channels to adaptively weight features with a high distinguishable ability to enlarge the inter-class distinction.

As illustrated in Figure 2, given the input feature map $F \in \Re^{C\times H\times W}$, we first feed it into a 1×1 convolution layer

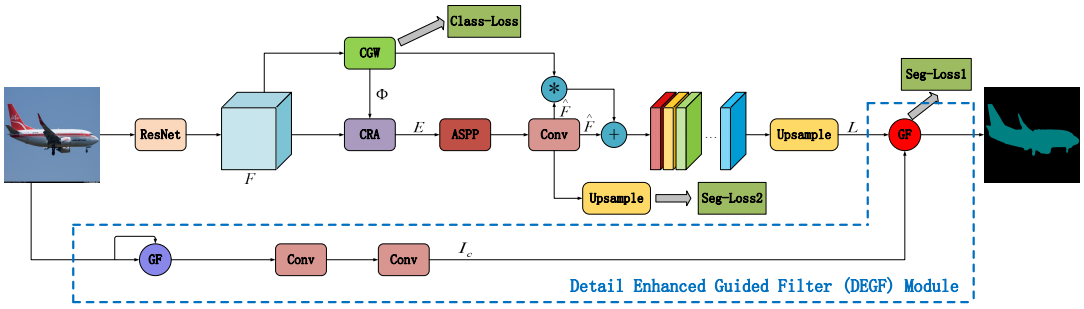

Figure 3: The flowchart of our proposed CGCWNet, where ResNet-101 is employed to extract dense features, $+$ and $*$ represent the element-wise addition and channel-wise multiplication, respectively.

to generate a new feature map $B \in \Re^{\widehat{C}\times H\times W}$. Then we reshape B into $\Re^{\widehat{C}\times N}$, where $N = H \times W$. After that, we perform a matrix multiplication between B and the channel relationship matrix Φ, and reshape the size to $\Re^{\widehat{C}\times H\times W}$. Φ comes from the CGW module, and it can aggregate response channels of the specific category to highlight its feature representation. In this way, the network will pay more attention to these channels (and their relationship) to enhance the distinction of features. Finally, in order to model the long-range semantic dependence between feature maps, we apply a 1×1 convolution operation to increase the channel dimension to C and perform an element-wise addition operation with the feature map F to obtain the final output $E \in \Re^{C\times H\times W}$.

Detail Enhanced Guided Filter Module

When max-pooling layers and sub-sampling operations are used in semantic segmentation networks (Chen et al. 2018b,a), their outputs are typically in low resolutions and result in coarse segmentation maps. The guided filter (He, Sun, and Tang 2013) is an edge-preserving operator, which can be applied to edge-aware smoothing, detail enhancement, image matting, etc. It can extract the edge contour information from the original image, and such edge contour information can be used as a cue to refine the contour of the object mask. However, due to the low contrast of the original image, the extracted edge contour is imprecise, making it difficult to refine the object mask. Based on these observations, we proposed the DEGF module. It first employs the guided filter to enhance the details of the original image, and then uses the enhanced image to extract edge contour information to refine the object mask. As illustrated in Figure 3, the original image is first processed by the guided filter to enhance the detail. Two 1×1 convolution layers are then applied to extract the low-level features I_c. Finally, the coarse segmentation map L and the low-level features I_c are fed to a guided filter to output the refined object mask.

Guided Filter for Image Detail Enhancement. We use a guided filter to extract the edge contour coefficients from the original image, then such coefficients are used to weight the input image for generating the enhanced one. The enhanced image e at a pixel i can be achieved as follows:

$$e_i = \sum_j R_{ij}(I) I_j \tag{2}$$

where i and j are pixel indexes. R_{ij} is the edge contour coefficient about the pixel i and pixel j. Following the similar derivations in (He, Sun, and Tang 2013), the expression of R_{ij} is:

$$R_{ij}(I) = \frac{1}{|w^2|} \sum_{k:(i,j)\in w_k} \left(1 + \frac{(I_i - \mu_k)(I_j - \mu_k)}{\sigma_k^2 + \varepsilon}\right) \tag{3}$$

where μ_k and σ_k^2 are the mean and variance of the image patch in the local square window w_k centered at the pixel k. $|w|$ is the number of pixels in w_k and ε is a regularization parameter. More details of the guided filter can be found in (He, Sun, and Tang 2013).

Guided Filter for Object Mask Refinement. After getting the enhanced image e, two 1×1 convolution layers are applied to extract the low-level features I_c. Here, we set the channel size of the first and second convolution layers to 64 and the number of categories, respectively. In the end, the coarse segmentation map L and the low-level features I_c are fed to a guided filter. The refined mask g at a pixel i is presented as follows:

$$g_i = \sum_j \widehat{R}_{ij}(I_c) L_j \tag{4}$$

where $\widehat{R}_{ij}$ is calculated similarly to Eq. (3).

Class Guided Channel Weighting Network

The overall network structure is depicted in Figure 3. The input image is first passed through a fully-convolutional network (e.g., ResNet-101) to produce a feature map F. After that, F is fed into the proposed CGW module, outputting a weight vector γ. At the same time, F is fed into the CRA and ASPP (Chen et al. 2017) modules to obtain the feature map $\widehat{F}$ with rich information. Then, we use global class weighting to obtain the coarse segmentation map $L = bilinear(\widehat{F} + \widehat{F} \cdot \gamma)$. $\widehat{F} \cdot \gamma$ is a channel-wise multiplication between the input feature map $\widehat{F}$ and the weight vector γ. $bilinear$ is used to upsample $\widehat{F} + \widehat{F} \cdot \gamma$ to the spatial size of the original image. Finally, the DEGF module is used to refine the coarse segmentation map.

We use three losses to jointly optimize our network. They are a class cross-entropy loss l_a behind CGW module and two class-balanced cross-entropy losses (Xie and Tu 2015)

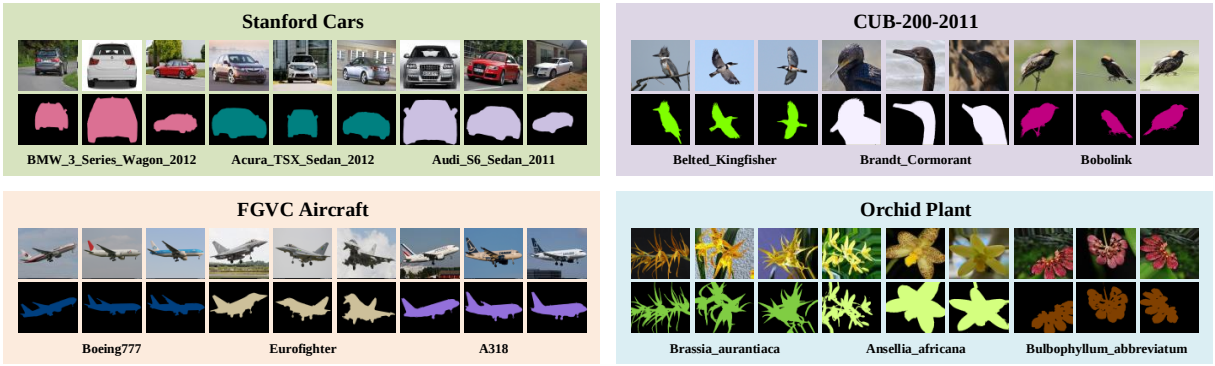

Figure 4: Some images and their corresponding pixel-level labellings in the Stanford Cars, CUB-200-2011, FGVC Aircraft, and our "Orchid" plant image sets.

l_c, l_f behind ASPP module and DEGF module for coarse segmentation and fine segmentation respectively. The overall loss can be formulated as shown in Eq. (5):

$$L = \lambda_a l_a + \lambda_c l_c + \lambda_f l_f \tag{5}$$

where λ_a, λ_c, and λ_f are balance parameters.

Experimental Results and Analysis

Base Network Architecture

We utilize three base networks to verify the generality and effectiveness of our proposed CGW, CRA, and DEGF modules. The first one is ResNet-101 (He et al. 2016) and the others are ResNet-101 with Atrous Spatial Pyramid Pooling (ASPP) (Chen et al. 2017) and DeepLabv3+ (Chen et al. 2018b). Note that the DeepLabv3+ uses ResNet-101 as the network backbone.

Datasets and Evaluation Metrics

PASCAL VOC 2012. The PASCAL VOC 2012 is a semantic segmentation benchmark with 20 foreground object classes and one background class. The dataset is augmented by the extra labellings provided by (Hariharan et al. 2011), which has 10,582, 1,449, and 1,456 images for network training, validation, and testing, respectively.

Fine-Grained Datasets. Existing datasets only provide pixel-level labellings at the super-category level rather than at the sub-category level. To achieve fine-grained semantic segmentation, we extend the fine-grained classification datasets to fine-grained segmentation datasets. Specifically, we use the method proposed by Zhang et al. (Zhang et al. 2020b) to automatically generate binary masks and manually fine-tune these masks. Finally, we use the image-level sub-category to label each pixel. Figure 4 shows some samples from the Stanford Cars, CUB-200-2011, FGVC Aircraft, and our "Orchid" plant image sets, with their corresponding masks at sub-category level. A detailed summary of these datasets is provided in Table 1.

Evaluation Metrics. We use Mean IoU (mean of class-wise intersection over union) as our main evaluation metric.

Datasets	#Class	#Training	#Testing
CUB-200-2011	200	5,994	5,794
FGVC Aircraft	100	6,667	3,333
Stanford Cars	196	8,144	8,041
"Orchid" Plant	252	103,179	4,300

Table 1: Statistics of fine-grained datasets used in this paper.

Implementation Details

For network training, we use min-batch stochastic gradient descent (SGD) optimizer with the batch size 6, initial learning rate $4e^{-3}$, weight decay 0.0002, and momentum 0.9 for Stanford Cars, CUB-200-2011, FGVC Aircraft, "Orchid" plant, and PASCAL VOC 2012 image sets. Following some previous works (Chen et al. 2018a; Yu et al. 2018b), we use the "poly" learning rate policy where the learning rate is multiplied by the factor $(1 - iter/max\_iter)^{0.9}$. In the DEGF module, the values of r and ε are first determined by grid search on the validation set, and then we use the same parameters to train the CGCWNet. In our network, C and $\widehat{C}$ are set to 2048 and 512 respectively. The loss weights λ_a, λ_c, and λ_f in Eq. (5) are set to 0.4, 0.6, and 1.0 respectively.

Ablation Studies

To verify the generality and effectiveness of our CGCWNet, we conduct some ablation experiments over PASCAL VOC 2012, CUB-200-2011, and Stanford Cars validation sets.

In Table 2, we evaluate the effectiveness of our proposed CGW, CRA, and DEGF modules by using different base networks. Specifically, we use ResNet-101, ResNet-101-ASPP, and DeepLabv3+ as the backbone architectures, respectively. By adding a CGW module, we achieve at least 1.11% improvement on PASCAL VOC 2012 validation set and obtain a better improvement (at least 2.04%) on fine-grained segmentation datasets (CUB-200-2011 and Stanford Cars). A reasonable explanation is that using image-level fine-grained probabilities to guide semantic segmentation can alleviate the intra-class inconsistency problem. By adding a CRA module, we achieve at least 1.09% improvement on PASCAL VOC 2012 validation set and at least

Methods	VOC	CUB	Cars
ResNet-101	66.46	43.87	48.46
ResNet-101 + DEGF	68.45	44.36	50.17
ResNet-101 + CRA	69.54	45.87	50.77
ResNet-101 + CGW	69.88	46.17	51.71
ResNet-101 + CRA + CGW + DEGF	70.31	46.58	52.29
ResNet-101 + ASPP	73.47	56.75	63.69
ResNet-101 + ASPP + DEGF	74.23	57.32	64.41
ResNet-101 + ASPP + CRA	74.65	58.64	66.20
ResNet-101 + ASPP + CGW	74.93	58.79	66.45
ResNet-101 + ASPP + CRA + CGW + DEGF	75.35	59.23	66.80
DeepLabv3plus	75.49	64.29	65.83
DeepLabv3plus + DEGF	76.49	65.19	67.62
DeepLabv3plus + CRA	76.58	65.88	68.31
DeepLabv3plus + CGW	76.60	66.51	68.66
DeepLabv3plus + CRA + CGW + DEGF	78.25	67.23	69.59

Table 2: Ablation studies for CGW, CRA, and DEGF modules on PASCAL VOC 2012, CUB-200-2011, and Stanford Cars validation sets. DeepLabv3plus represents DeepLabv3+.

Methods	VOC	CUB	Cars
DeepLabv3plus	75.49	64.29	65.83
DeepLabv3plus + GAP	76.04	64.37	66.55
DeepLabv3plus + SE	76.39	64.41	66.58
DeepLabv3plus + CGW	**76.60**	**66.51**	**68.66**
DeepLabv3plus + NL	75.57	64.64	66.92
DeepLabv3plus + APNB	75.63	64.87	67.23
DeepLabv3plus + Coordinate Attention	76.42	62.91	66.85
DeepLabv3plus + PSA	76.44	63.25	68.14
DeepLabv3plus + HaloAttention	76.53	62.17	67.12
DeepLabv3plus + CRA	**76.58**	**65.88**	**68.31**

Table 3: Comparison with different global context extraction and attention methods on PASCAL VOC 2012, CUB-200-2011, and Stanford Cars validation sets.

1.59% improvement on fine-grained segmentation datasets. It shows that using the channel relationship matrix to weight highly discriminative features is useful to cope with the inter-class indistinction problem in fine-grained datasets. By adding a DEGF module, we achieve at least 0.49% improvement on three datasets, proving that the DEGF module can improve segmentation results.

Experiments on PASCAL VOC 2012 and Fine-Grained Datasets

In this subsection, we present the comparison results with the different semantic segmentation modules and methods on PASCAL VOC 2012 and fine-grained datasets.

Methods	mIoU(%)	Methods	mIoU(%)
*DenseCRF	72.69	*DGF	73.58
AG-Net	60.90	**Our (ResNet-101+ASPP+DEGF)**	**74.23**

Table 4: Quantitative Comparison. * indicates implementation from DGF.

Comparison with other global context extraction modules Our proposed CGW module is related to some global context extraction modules. In this section, we compare our proposed CGW module with two different global context extraction modules (i.e., Global Average Pooling (GAP) (Yu et al. 2018a) and Squeeze-and-Excitation (SE) (Hu, Shen, and Sun 2018)) on PASCAL VOC 2012, CUB-200-2011 and Stanford Cars validation sets by using DeepLabv3+ as the base network. As shown in Table 3, adding GAP and SE can slightly improve the performance, which verifies the effectiveness of the global context. Meanwhile, our CGW module can achieve better performance than these modules. The reason is that SE and GAP focus on exploring different spatial strategies to capture richer global contextual information, but they cannot distinguish pixels from different classes explicitly when calculating the context. However, our CGW harvests the global context information from a categorical perspective, which can distinguish pixels from different classes and unify semantic labels for all pixels of the same object.

Our CGW module is also similar to CPNet (Yu et al. 2020). A major difference is that our CGW directly predicts image-level category probabilities as the class-wise encoding, while CPNet relies on pixel-level ground truth to encode the category relationship of pixels. This difference enables our method to directly perform a larger scale class-wise weighting in the attention map during the inference stage. Compared with DeepLabv3plus+GC (i.e, global context block in GC-Net), our DeepLabv3plus+CGW achieves better mIoU on CUB-200-2011 dataset (65.34% vs. 66.51%). Moreover, we compare DeepLabv3plus+CRA+CGW+DEGF with DeepLabv3plus+CRA+SE+DEGF. The experimental results show that our DeepLabv3plus+CRA+CGW+DEGF achieves better performance on the CUB-200-2011 (66.12% vs. 67.23%) and Stanford Cars (68.72% vs. 69.59%) datasets.

Comparison with other attention methods We further compare our proposed CRA module with several attention methods. The attention methods for comparison include: Non-local Block (NL) (Wang et al. 2018), Asymmetric Pyramid Non-local Block (APNB) (Zhu et al. 2019), Coordinate Attention (Hou, Zhou, and Feng 2021), PSA (Liu et al. 2021), and HaloAttention (Vaswani et al. 2021). Experimental results are presented in Table 3. One can easily observe that our CRA module can achieve better performance than all these attention methods. The main reason is that these modules only consider the feature interaction on the first-order feature statistics, while ignoring the statistics higher than first-order, thus hindering the discriminative ability of the network. However, we use second-order statistics to construct a channel relationship matrix to adaptive weight the original features to enhance the distinction of features, thereby enlarging inter-class distinction.

Comparison with other guided filter methods We further compare our DEGF with several different guided filter methods (i.e., DenseCRF (Krähenbühl and Koltun 2012), DGF (Wu et al. 2018), and AG-Net (Zhang et al. 2019b)) on

Methods	Aircraft	CUB	"Orchid" Plant	Cars	Multi-Object Aircraft
DeepLabv2 (Chen et al. 2018a)	54.90	54.21	71.50	58.34	53.17
DeepLabv3+ (Chen et al. 2018b)	60.15	64.29	70.53	65.83	58.44
Non-local (Wang et al. 2018)	59.58	63.64	70.14	64.92	56.93
DFN (Yu et al. 2018b)	60.37	63.48	70.73	66.15	59.21
ACFNet (Zhang et al. 2019a)	58.05	59.23	67.68	60.20	57.12
WASP (Artacho and Savakis 2019)	67.16	63.96	67.23	67.41	61.49
EMANet (Li et al. 2019)	64.50	63.21	68.75	66.23	60.28
SPNet (Hou et al. 2020)	55.94	65.67	65.53	68.44	56.57
HyperSeg-L (Nirkin, Wolf, and Hassner 2021)	60.47	64.50	71.16	66.10	58.33
DSRL (Wang et al. 2020)	68.33	66.72	72.01	68.65	61.81
Our-CTS	66.72	**76.42**	66.00	**74.68**	-
Our-CGCWNet	**68.62**	67.23	**75.97**	69.59	**62.93**

Table 5: Performance on five fine-grained image sets. "-" indicates that the result cannot be obtained.

Methods	mIoU(%)	Methods	mIoU(%)
DeepLabv3+	53.71	WASP	55.58
EMANet	53.57	SPNet	57.35
HyperSeg-L	52.86	DSRL	57.82
Our-CTS	56.10	**Our-CGCWNet**	**58.51**

Table 6: Performance on AIBD-Birds image set.

PASCAL VOC 2012 validation set. The results in Table 4 show that our method outperforms existing approaches. We attribute that our DEGF can enhance the details of the original image, thus making the extracted edge contour more precise, which is beneficial to refine the object mask.

Comparison with state-of-the-arts methods on fine-grained datasets We compare our proposed method with the state-of-the-art methods on five fine-grained datasets. Specifically, we train DeepLabv3+ with CRA, CGW, and DEGF (i.e., DeepLabv3plus+CRA+CGW+DEGF) as our method. The results are illustrated in Table 5. From these comparison results, one can easily observe that our method achieves better performance than other competitors. The reason may come from the following two aspects: 1) Unlike other methods that mainly solve traditional semantic segmentation task, our model is designed for the task of fine-grained semantic segmentation, which can effectively alleviate the problems of intra-class inconsistency and inter-class indistinction; 2) Our DEGF module can transfer the boundaries of the enhanced image to the segmentation output, so that the object mask can further be refined by using the enhanced image. It should be emphasized that the comparison methods in Table 5, except for ACFNet (Zhang et al. 2019a) and DFN (Yu et al. 2018b), are reproduced using the official codes and the parameters suggested in the original papers. Note that the Multi-Object Aircraft is our synthetic dataset, which has multiple objects with pixel-level labellings in each image. It contains 39,900 images for 100 fine-grained aircraft classes. We divide 39,900 images into 29,925 training images and 9,975 test images.

It is a very interesting experiment to compare CGCWNet with an approach (we called Classification To Segmentation (CTS)) that simply assigns fine-grained classification results to the foreground object of super-category segmentation. In our experiments, BCNN (Lin, RoyChowdhury, and Maji 2015) and Deeplabv3+, which have demonstrated outstanding performance in fine-grained image classification and semantic segmentation tasks, are used as classification and segmentation networks in CTS. As shown in Table 5, the proposed CGCWNet outperforms CTS on the FGVC Aircraft and "Orchid" Plant datasets, while performs worse than CTS on the CUB-200-2011 and Stanford Cars datasets. The reason is that the images in the CUB-200-2011 and Stanford Cars datasets have much simpler and clearer backgrounds than the "Orchid" Plant dataset. The highly recognizable appearances are easy for classifying and segmenting individually. However, once the image includes complex background or is difficult to classify by fine-grained classification methods, the result by CTS will be reduced sharply. The joint optimization of classifying and segmenting in our CGCWNet will benefit from each other to improve the performance of the fine-grained semantic segmentation, especially on hard images. Besides, CTS has the following potential disadvantages: 1) Once the classification result is wrong, the entire mask will become the wrong result; 2) When the variation between the sub-categories is large, the segmentation result of the super-category will be worse. 3) When there are multiple objects in an image, the CTS requires a detector to locate the position of the object. To verify our conclusion, we conduct experiments on a synthesized dataset (called AIBD-Birds) with complex backgrounds. The AIBD-Birds dataset contains 200 fine-grained bird species. It has 11,788 images, of which 5,994 images for training, 5,794 images for testing. The experiment results are shown in Table 6. The proposed CGCWNet can achieve the highest performance.

Conclusion

In this paper, a new approach, called Class Guided Channel Weighting Network (CGCWNet), is developed to achieve fine-grained semantic segmentation. Our CGCWNet can enhance feature representation by using CGW and CRA modules, which can alleviate both the intra-class inconsistency and inter-class indistinction. Besides, we embed a DEGF module in deep neural networks, which can further refine the predicted object masks. Ablation studies and performance comparison on several datasets have demonstrated the effectiveness of our proposed method.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. 61903300), Key Research and Development Program of Shaanxi (Program No. 2020ZDLGY04-07, 2021ZDLSF06-05), Innovation Capability Support Program of Shaanxi (Program No. 2021TD-05), and the fellowship of China Postdoctoral Science Foundation (Grant No. 2020M683695XB).

References

Arani, E.; Marzban, S.; Pata, A.; and Zonooz, B. 2021. Rgpnet: A real-time general purpose semantic segmentation. In *WACV*, 3009–3018.

Artacho, B.; and Savakis, A. 2019. Waterfall atrous spatial pooling architecture for efficient semantic segmentation. *Sensors*, 19(24): 5361.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE TPAMI*, 39(12): 2481–2495.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. 2018a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4): 834–848.

Chen, L.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv: Computer Vision and Pattern Recognition*.

Chen, L.; Yang, Y.; Wang, J.; Xu, W.; and Yuille, A. 2016. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 3640–3649.

Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 801–818.

Chen, W.; Zhu, X.; Sun, R.; He, J.; Li, R.; Shen, X.; and Yu, B. 2020. Tensor Low-Rank Reconstruction for Semantic Segmentation. In *ECCV*, 52–69. Springer.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *CVPR*, 11065–11074.

Dai, X.; Yue-Hei Ng, J.; and Davis, L. S. 2017. Fason: First and second order information fusion network for texture recognition. In *CVPR*, 7352–7360.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *CVPR*, 3146–3154.

Gao, Z.; Xie, J.; Wang, Q.; and Li, P. 2019. Global second-order pooling convolutional networks. In *CVPR*, 3024–3033.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *ICCV*, 991–998. IEEE.

He, K.; Sun, J.; and Tang, X. 2013. Guided image filtering. *IEEE TPAMI*, 35(6): 1397–1409.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hou, Q.; Zhang, L.; Cheng, M.; and Feng, J. 2020. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. In *CVPR*, 4003–4012.

Hou, Q.; Zhou, D.; and Feng, J. 2021. Coordinate attention for efficient mobile network design. *arXiv preprint arXiv:2103.02907*.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *CVPR*, 3588–3597.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *CVPR*, 7132–7141.

Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; and Liu, W. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 603–612.

Koniusz, P.; and Zhang, H. 2020. Power normalizations in fine-grained image, few-shot image and graph classification. *arXiv preprint arXiv:2012.13975*.

Krähenbühl, P.; and Koltun, V. 2012. Efficient inference in fully connected crfs with gaussian edge potentials. *arXiv preprint arXiv:1210.5644*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *ICCV*, 554–561.

Li, P.; Xie, J.; Wang, Q.; and Zuo, W. 2017. Is second-order information helpful for large-scale visual recognition? In *ICCV*, 2070–2078.

Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; and Liu, H. 2019. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, 9167–9176.

Lin, T.-Y.; RoyChowdhury, A.; and Maji, S. 2015. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 1449–1457.

Liu, H.; Liu, F.; Fan, X.; and Huang, D. 2021. Polarized Self-Attention: Towards High-quality Pixel-wise Regression. *arXiv preprint arXiv:2107.00782*.

Liu, J.; He, J.; Qiao, Y.; Ren, J. S.; and Li, H. 2020. Learning to Predict Context-adaptive Convolution for Semantic Segmentation. In *ECCV*, 769–786. Springer.

Liu, W.; Rabinovich, A.; and Berg, A. 2015. ParseNet: Looking Wider to See Better. *arXiv: Computer Vision and Pattern Recognition*.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Nirkin, Y.; Wolf, L.; and Hassner, T. 2021. Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. In *CVPR*, 4061–4070.

Oliveira, G. L.; Yogamani, S.; Burgard, W.; and Brox, T. 2020. Beyond Single Stage Encoder-Decoder Networks: Deep Decoders for Semantic Image Segmentation. *arXiv preprint arXiv:2007.09746*.

Perronnin, F.; Sánchez, J.; and Mensink, T. 2010. Improving the fisher kernel for large-scale image classification. In *ECCV*, 143–156. Springer.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Vaswani, A.; Ramachandran, P.; Srinivas, A.; Parmar, N.; Hechtman, B.; and Shlens, J. 2021. Scaling local self-attention for parameter efficient visual backbones. In *CVPR*, 12894–12904.

Wang, L.; Li, D.; Zhu, Y.; Tian, L.; and Shan, Y. 2020. Dual Super-Resolution Learning for Semantic Segmentation. In *CVPR*, 3774–3783.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *CVPR*, 7794–7803.

Wu, H.; Zheng, S.; Zhang, J.; and Huang, K. 2018. Fast End-to-End Trainable Guided Filter. In *CVPR*.

Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; and Zhang, Z. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 842–850.

Xie, S.; and Tu, Z. 2015. Holistically-nested edge detection. In *ICCV*, 1395–1403.

Yu, C.; Wang, J.; Gao, C.; Yu, G.; Shen, C.; and Sang, N. 2020. Context prior for scene segmentation. In *CVPR*, 12416–12425.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018a. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 325–341.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018b. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 1857–1866.

Zhang, F.; Chen, Y.; Li, Z.; Hong, Z.; Liu, J.; Ma, F.; Han, J.; and Ding, E. 2019a. ACFNet: Attentional Class Feature Network for Semantic Segmentation. In *ICCV*, 6798–6807.

Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; and Agrawal, A. 2018. Context encoding for semantic segmentation. In *CVPR*, 7151–7160.

Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. 2020a. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*.

Zhang, S.; Fu, H.; Yan, Y.; Zhang, Y.; Wu, Q.; Yang, M.; Tan, M.; and Xu, Y. 2019b. Attention guided network for retinal image segmentation. In *MICCAI*, 797–805. Springer.

Zhang, X.; Zhang, W.; Peng, J.; and Fan, J. 2020b. Automatic Image Labelling at Pixel Level. *arXiv preprint arXiv:2007.07415*.

Zhao, H.; Jia, J.; and Koltun, V. 2020. Exploring self-attention for image recognition. In *CVPR*, 10076–10085.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*, 2881–2890.

Zhong, Z.; Lin, Z. Q.; Bidart, R.; Hu, X.; Daya, I. B.; Li, Z.; Zheng, W.-S.; Li, J.; and Wong, A. 2020. Squeeze-and-attention networks for semantic segmentation. In *CVPR*, 13065–13074.

Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; and Bai, X. 2019. Asymmetric non-local neural networks for semantic segmentation. In *ICCV*, 593–602.