RGB-D Co-attention Network for Semantic Segmentation

Hao Zhou[1,3,*], Lu Qi[2,*], Zhaoliang Wan[1,3], Hai Huang[1,(✉)], and Xu Yang[3,(✉)]

[1] National Key Laboratory of Science and Technology of Underwater Vehicle, Harbin Engineering University, Harbin, China
zhouhao94@yahoo.com, wan.zhaoliang@icloud.com, haihus@163.com
[2] The Chinese University of Hong Kong, Hong Kong, China
luqi@cse.cuhk.edu.hk
[3] State Key Laboratory of Management and Control for Complex System, Institute of Automation, Chinese Academy of Sciences, Beijing, China
xu.yang@ia.ac.cn

Abstract. Incorporating the depth (D) information for RGB images has proven the effectiveness and robustness in semantic segmentation. However, the fusion between them is still a challenge due to their meaning discrepancy, in which RGB represents the color but D depth information. In this paper, we propose a co-attention Network (CANet) to capture the fine-grained interplay between RGB and D features. The key part in our CANet is co-attention fusion part. It includes three modules. At first, the position and channel co-attention fusion modules adaptively fuse color and depth features in spatial and channel dimension. Finally, a final fusion module integrates the outputs of the two co-attention fusion modules for forming a more representative feature. Our extensive experiments validate the effectiveness of CANet in fusing RGB and D features, achieving the state-of-the-art performance on two challenging RGB-D semantic segmentation datasets, i.e., NYUDv2, SUN-RGBD.

1 Introduction

Semantic segmentation aims to assign each pixel into different categories (e.g. desk, sofa, wall, floor). It is fundamental in computer vision and benefits a large number of applications, such as automatic driving, robotic sensing, visual SLAM and so on. Despite the communitys great achievement in semantic segmentation [1–8], most of the researches only used the RGB images. The RGB information provides models with robust color and texture but not geometric information. It makes hard to discriminate instances and context which shares the similar color and texture. As shown in Figure 1, pillows on a bed with similar color of the bed, cushion on a sofa with similar color of the sofa.

*Both Authors contributed equally to this work.

Fig. 1. Some hard samples of semantic segmentation only by RGB image. Left: pillows on a bed with similar color of the bed. Right: cushion on a sofa with similar color of the sofa.

To solve the problems described above, some researches begin to leverage depth information in assisting semantic segmentation [9–14]. The combination of RGB and depth images is vital significate in many aspects. On one hand, depth images provide necessary geometric information and can enrich the representation of RGB images. On the other hand, depth images are robust to environment disturbances, such as illumination, fog, etc. However, it is not trivial to fuse the color and depth images well duo to the data discrepancy between color and depth information, where depth image embedding geometric information and color image embedding texture information.

Remarkable efforts have been invested on this task RGB-D semantic segmentation. For example, [9–12] use depth image as an extra channel for the input. [13, 14] respectively extract features from RGB and depth images and then fuse them. [15–20] jointly learn the correlation between depth features and color features. Albeit efficient for these approaches, that mainly focus on the local feature fusing and do not take the long-range dependencies into consideration.

Instead of designing heuristic fusion module of local features, we prefer to design self-supervised fusion module for global information. Based on this idea, There are two requirements need to consider: the fused features should have strong representation ability and the fusion method can automatically learn the long-range dependencies between different modalities. According to the analysis, we propose a CANet that contains three parts, encoder, co-attention fusion part and decoder. The encoder has three parallel branches to extract depth, color and mixture features respectively. This parallel design avoids the influence between the extracted depth and color feature while brings a CNN learned mixture feature. The co-attention fusion part, inspired by self-attention [21], is proposed to sovle the data discrepany problem and effectively fuse RGB and depth features. The decoder is an up-sampled ResNet that decodes the fused feature for the final segmentation

The co-attention fusion part consists of three modules, position co-attention fusion module (PCFM), channel co-attention fusion module (CCFM) and final fusion module (FFM). PCFM and CCFM are proposed to fuse color and depth features in spatial and channel dimension. The FFM, is designed to effectively integrate the PCFM and CCFM and produces the final fused feature. For PCFM and CCFM, co-attention is first used to model long-range dependendencies between RGB and depth. Then, the learned long-range dependencies are used to

tranform the depth information into color feature space. Finally, the transformed depth feature is added with the orginal color feature. The key idea of co-attention fusion method can described as using a color feature query and a set of depth feature key-value pairs to transform the depth feature into color feature space and then fuse with corresponding local color feature.

The main contributions of our CANet can be summarized as below:

- We propose a novel Co-attention Network (CANet) for RGB-D semantic segmentation.
- The key part, co-attention fusion part, consisting of PCFM, CCFM and FFM. PCFM and CCFM are propposed to solve the data discrepany problem and effectively fusion color and depth features at position and channel dimensions respectively. FFM is used to integrate PCFM and CCFM.
- We perform extensive experiments on the NYUDv2 [22] and the SUN-RGBD [23] datasets. CANet significantly improves RGB-D semantic segmentation results, achieving state-of-the-art on the two popular RGB-D benchmarks.

2 Related works

2.1 Attention Modules

The attention mechanism [24–28] is widely used to model the global dependencies of features. There are many representations for attention mechanism. Among them, self-attention [29, 30] could capture the long-range dependencies in a sequence. The work [21] is the first one that proves simply using self-attention in machine translation models could achieve state-of-the-art results. Owing to the modeling capability of long-range dependencies, self-attention module benefits in many tasks [31–38].

Inspired by the great success in NLP, self-attention module also gets focuses in computer vision field [39–45]. SENet [40] proposes channel attention modules that adaptively recalibrate channel-wise feature responses. NLNet [39] proposes non-local operations for capturing long-range dependencies. GCNet [41] creates a simplified network of NLNet based on a query-independent formulation. SAGAN [42] uses position attention modules that models the long-range dependency in generative adversarial networks for image generation tasks. SCA-CNN [43], DANet [44] and ABD-Net [45] incorporate spatial and channel-wise attention on image captioning, sementic segmentation and person re-identification tasks respoecitvely. Different from previous works, we extend the attention mechanism to color-depth features fusion. We design two attention based fusion modules for long-range dependencies between features from different modalities.

We name our attention mechanism as co-attention. The concept of co-attention is widely used in Visual Question Answer (VQA) task. The work [46] presents a novel co-attention mechanism to inference for the question and the image consequently. The work [47] develops a co-attention mechanism to jointly learn both

the image and question attentions. The work [48] proves that co-attention mechanism enables dense, bi-directional interactions between image and text modalities. DANs [49] jointly leverages visual and textual attention mechanisms to create a fine-grained interplay between vision and language. The above-mentioned works learn the visual and textual attentions separately. Different from that, we use co-attention to acquire the global dependencies between color and depth modalities.

2.2 RGB-D Semantic Segmentation

Different from color image semantic segmentation, RGB-D semantic segmentation is provided with a piece of additional depth information by depth Image. In the early stage, works [50, 51, 22] design handcrafted features tailored for RGB with depth information. Recently, with the benefit of CNN in color image semantic segmentation, deep-learning-based RGB-D semantic segmentation methods [9–20] have been proposed. Some works [9–12] use depth information as an additional channel of RGB channels. However, simply using depth image as an extra channel of RGB image cannot take full advantage of the depth information.

To better exploit the depth context, multimodal feature fusion-based methods [13–19] are proposed for RGB-D semantic segmentation. FuseNet [13] introduces a fuse layer to fuse depth features into color features maps. RDFNet [15] uses multi-modal feature fusion blocks and multi-level feature refinement blocks to capture RGB-D features. LSD-GF [16] introduces a gated fusion layer to adjust the contributions of RGB and depth over each pixel. Depth-aware CNN [17] presents depth-aware convolution and depth-aware pooling to incorporate geometry information into color features. CFN [18] and SCN [19] use the available depth to split the image into layers with the common visual characteristics.

Nevertheless, the aforementioned existing works mainly focus on local feature fusion and do not take the long-range dependencies into consideratin. We propose a new idea, to use an attention mechanism to model the long-range dependencies between color and depth features. Then, the learned long-range dependencies are used to transform the depth feature into color feature space. Finally, the transformed depth feature is added with the original color feature.

3 Co-attention Network

In this section, we first present the overall architecture of CANet (Section 3.1), including a standard encoder-decoder structure and our proposed co-attention fusion part. Then we introduce the modules of co-attention fusion part in Section 3.2, 3.3 and 3.4 respectively. At last, we describe our multi-scale loss function in Section 3.5.

3.1 Network Architecture Overview

Inspired by Unet [53], our CANet adopts an encoder-decoder structure for RGB-D semantic segmentation. The encoder is used to extract latent features, and

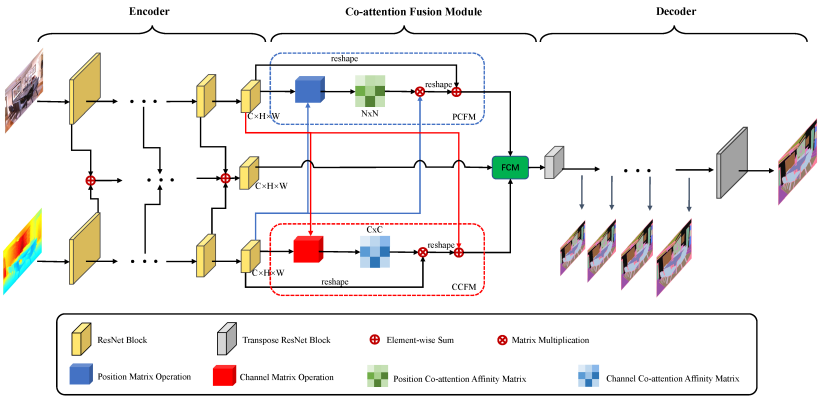


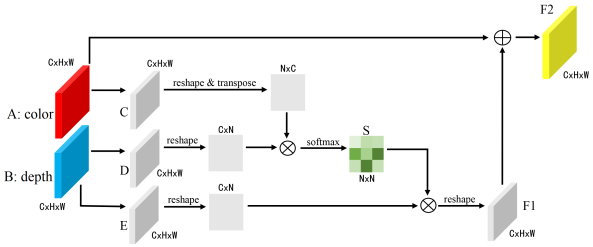
Fig. 2. Architecture of CANet. CANet mainly consists of three parts: 1) encoder (color encoder, depth encoder, mixture encoder). This paper adopts ResNet [52] as the backbone. 2) decoder, a upsample ResNet with standard residual building block. 3) co-attention fusion part, consists of PCFM, CCFM and FFM, are used to effectively fuse color features, depth features and mixture features.

then the decoder decodes them for final segmentation. For this structure, the robustness and effectiveness of latent features should directly influence the segmentation quality. As such, we proposed a co-attention fusion method to enhance our latent feature by fusing color and depth features. At first, we elaborately design the encoder to extract robust features from both color and depth information, which serves as the input for our co-attention method. Then, these features are adaptively fused by co-attention in different feature dimension.

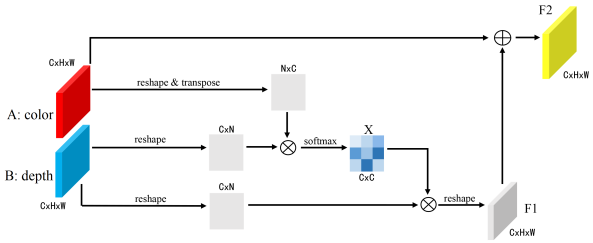
As shown in Figure 2, the encoder has three CNN branches. The first two, namely the RGB and D branch, are used to extract features of the color and depth image. Another CNN branch combines intermediate features from both RGB and D branches. For the decoder, it is an up-sampled ResNet with a series of the standard residual blocks. For fair comparison for other methods [14, 15], we extract the multiple up-sampled features to generate semantic maps for multi-scale supervision.

Apart from our multi-branch structure in encoder-decoder, we also propose a co-attention fusion method to fuse these encoded features. It has three modules, including position co-attention fusion module (PCFM), channel co-attention fusion module (CCFM) and final fusion module (FFM). The first two use co-attention to fuse color and depth features in spatial and channel dimension. And the last one wraps the features with high consistency.

The PCFM captures the spatial dependencies between any two positions from color and depth feature maps respectively. For the color feature at a certain position, it is aggregated by depth features at all spatial locations with a learnable weighted summation. It is similar to CCFM except for fusing features



(a) Position co-attention fusion module



(b) Channel co-attention fusion module

Fig. 3. The detailed structure of position co-attention fusion module and channel co-attention fusion module

among channels. Finally, FFM is designed to effectively integrate the outputs of these two co-attention modules.

3.2 Position Co-attention Fusion Module

Enriching the local features with context by attention has been widely used in RGB semantic segmentation. However, RGB and Depth features have different semantic information, meaning, making it hard to adopt a similar strategy for RGB-D images. Inspired by the independent embed features adopted in NLP [21], we introduce color, depth and mixture features by three branches. Each branch represents a unique embed feature. We fuse them step by step for better feature consistency.

We firstly introduce a position co-attention fusion module to adaptively fuse depth and color features. By this way, PCFM uses a spatial query of color feature and a set of spatial key-value pairs of depth feature to transform the global depth feature into color feature space and then fuse with corresponding local color feature.

The detail of PCFM is illustrated in Figure 3(a). The input color feature maps $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$ are fed into one convolution layer with batch normalization and ReLU activation to a feature map $\mathbf{C} \in \mathbb{R}^{C\times H\times W}$. C, H, W are channel, height and width of features respectively. A similar process is conducted in the input depth feature map $\mathbf{B} \in \mathbb{R}^{C\times H\times W}$ by two times, producing the two new feature maps $\mathbf{D}, \mathbf{E} \in \mathbb{R}^{C\times H\times W}$. Then we flatten the $\mathbf{C}, \mathbf{D}, \mathbf{E}$ feature maps in $C \times N$

format, where $N = H \times W$. All of $\mathbf{C}, \mathbf{D}, \mathbf{E}$ share the same feature embedding with the original feature but with different characteristics. As such, we could use them for forming our position co-attention affinity matrix.

For detail, the position co-attention affinity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ between $\mathbf{C}$ and $\mathbf{D}$ is calculated by the matrix multiplication and softmax layer:

$$s_{ji} = \frac{\exp\left(C_i^T \times D_j\right)}{\sum_{j=1}^{N} \exp\left(C_i^T \cdot D_j\right)}, i, j \in \{1, \cdots, N\} \tag{1}$$

Where s_{ji} represents the impact of i^{th} position of color feature maps on j^{th} position of depth feature maps. In other words, s_{ji} is the correlation for pixel-level features at i^{th} and j^{th} positions from different feature maps. Secondly, we obtain the co-attention feature maps $\mathbf{F}1$ by matrix production between $\mathbf{E}$ and $\mathbf{S}$. The $\mathbf{F}1$ adaptively aggregated depth feature of each position. At last, we perform the element-wise sum operation between the co-attention feature maps $\mathbf{F}1$ and original color feature maps $\mathbf{A}$:

$$F2_j = \alpha \sum_{i=1}^{N} (E_i s_{ji}) + A_j, j \in \{1, \cdots, N\} \tag{2}$$

We are noting that a learnable scale parameter α in this sum operation, dynamically balancing the contribution of these two features.

Equation (2) shows that each position of the fused feature maps $\mathbf{F}2$ is obtained by adding the local color feature with weighted sum of global depth features in spatial dimension. Hence, the fused feature maps have a global view of depth feature maps, and it selectively fuses spatial depth contexts according to the position co-attention affinity matrix.

3.3 Channel Co-attention Fusion Module

Each channel plays different role in RGB recognition tasks, which has been comprehensively explored in SENet [40]. Inspired by this, we propose a channel co-attention fusion module to fuse the channel features step by step. In this module, we adopt a similar method with our PCFM, except we operate and fuse features in channel dimension.

As illustrated in Figure 3(b), our channel co-attention fusion module is similar to our position fusion module in section 3.2. We flatten the input color and depth feature maps $\mathbf{A}, \mathbf{B}$ into the new feature maps with $C \times N$, where $N = H \times W$. They could be calculated to get the co-attention affinity matrix.

At first, the channel co-attention affinity matrix $\mathbf{X} \in \mathbb{R}^{C \times C}$ between $\mathbf{A}$ and $\mathbf{B}$ is calculated by matrix multiplication and softmax layer:

$$x_{ji} = \frac{\exp\left(B_i \times A_j^T\right)}{\sum_{j=1}^{N} \exp\left(B_i \cdot A_j^T\right)}, i, j \in \{1, \cdots, C\} \tag{3}$$

Where, x_{ji} represents the impact of the j^{th} channel of color feature maps on the i^{th} channel of depth feature maps. In other words, x_{ji} is the correlation for

the channel-level features at i^{th} and j^{th} channels from different feature maps. Secondly, we obtain the co-attention feature maps $\mathbf{F}1$ by matrix production between $\mathbf{X}$ and $\mathbf{B}$. The $\mathbf{F}1$ brings adaptively aggregated depth feature of each channel. Finally, we perform an element-wise sum operation between the co-attention feature maps $\mathbf{F}1$ and original color feature maps $\mathbf{A}$:

$$F2_j = \beta \sum_{i=1}^{N} (x_{ji} B_i) + A_j, j \in \{1, \cdots, C\} \tag{4}$$

Noting that a learnable scale parameter β is added in this sum operation to modify the contribution of these two features.

Equation (4) indicates that each channel of the fusion feature maps $\mathbf{F}2$ is obtained by adding the local color feature with weighted sum of global depth features in channel dimension. Hence, the fused feature maps have a global view of all the channel feature maps of depth, and it selectively aggregates channel feature map according to the channel co-attention affinity matrix.

3.4 Final Fusion Module

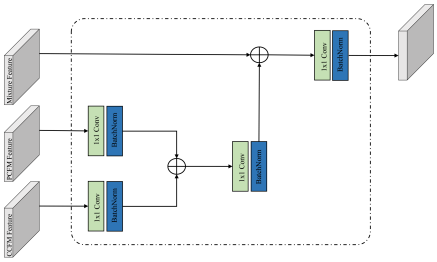


Fig. 4. Final Fusion Module (FFM)

FFM is used to integrate the output of PCFM, CCFM, and the mixture branch. The proposed FFM is implemented by four convolution layers followed by batch normalization and element-wise sum operation. The detailed structure of the FFM is shown in Figure 4. The features of PCFM and CCFM are first convolved followed by batch normalization (a conv unit). The output channel of these convolutions is 2048, expanding the channel dimension of original features. Then we fuse the expanded attentions features with element-wise addition. After that, we smooth the added features by an extra conv unit. By addition again between smoothed and mixture features and then convolution, we obtain the final features.

3.5 Loss Function

The Figure 2 illustrates our multi-scale loss function. At the training period, pyramid supervision introduces four intermediate side outputs from the features

of the four unsampled residual unit except of the final output. The side outputs have 1/2, 1/4, 1/8, and 1/16 the height and width of the final outputs, respectively. A cross-entropy loss function is used on the four side and final outputs as follows:

$$J\left(F_1, \ldots F_5\right) = \sum_{k=1}^{K} J_k\left(F_k\right) \tag{5}$$

where

$$J_k\left(F_k\right) = \sum_{(h,w)\in\Omega_k} L\left(y_k^*(h,w), y_k(h,w)\right) \tag{6}$$

J_k is the objective function for the side output or final output. The Ω_k denotes the set of pixels of side output or final output. The function L is cross entropy loss function. The whole network is trained by optimizing Equation (5) with back propagation.

4 Experiments

4.1 Comparison with the State-of-the-art

4.2 Datasets and Metrics

In this section, we evaluate our network through comprehensive experiments. We use two public datasets:

- NYUDv2 [22]: The NYUDv2 dataset contains 1449 RGB-D images. We follow the 40-class settings [51] and the standard split [22] by involves 795 images for training and 654 images for testing.
- SUN RGB-D [23]: The SUN RGB-D dataset consists of 10335 RGB-D image pairs with 37 categories. We use the standard training/testing split [23] with 5285 as training and 5050 as testing.

Three common metrics [1] are used for evaluation, including pixel accuracy (PixAcc.), mean accuracy (mAcc.) and mean intersection over union (mIoU).

4.3 Implementation Details

We implement our network using the PyTorch deep learning framework [58]. All the models are trained with Nvidia Tesla V100 GPU. We use the pre-trained ResNet-50/ResNet-101 [52] as our three backbone branches in the encoder. Except for the backbones, the weights of other layers in our network are initialized by a normal distribution with zero mean and 0.01 variance, while the biases are padded with zero. The SGD is used as our optimizer, with momentum 0.9 and weight decay 0.0005. The learning rate is 0.001 (NYUDv2) or 0.0005 (SUN RGB-D) for the backbone and 0.01 for the other parts at the early stage, and it decays by a factor of 0.8 in every 100 (NYUDv2) or 20 (SUN RGB-D) epochs.

In the training period, we resize the inputs including RGB images, depth images and ground truth labels to size 480×640. We are noting that, the ground

Table 1. Comparison with state-of-the-arts on the NYUDv2 dataset. Results are reported in terms of percentage (%) of pixel accuracy, mean accuracy, and mean IoU.

Method	Backbone	PixAcc.	mAcc.	mIoU
Gupta et al. [51]	-	60.3	35.1	28.6
Deng et al. [54]	-	63.8	-	31.5
FCN [1]	VGG-16	65.4	46.1	34.0
Eigen [55]	-	65.6	45.1	34.1
STD2P [56]	-	70.1	53.8	40.1
3DGNN [57]	ResNet-101	-	55.7	43.1
LSD-GF [16]	VGG-16	71.9	60.7	45.9
CFN [18]	ResNet-152	-	-	47.7
D-CNN [17]	ResNet-152	-	61.1	48.4
SCN [19]	ResNet-152	-	-	49.6
RDFNet [15]	ResNet-152	76.0	62.8	50.1
CANet	ResNet-50	75.7	62.6	49.6
CANet	ResNet-101	76.6	63.8	51.2

truth labels are further resized into four down-sampled maps from 240×320 to 30×40 for pyramid supervision of the side output. For fair comparison with other methods, we adopt the multi-scale and crop as our data augmentation strategy. Each image is also processed with random hue, brightness, and saturation adjustment. The mean and standard deviation of RGB and Depth images are calculated to normalize our input data.

Table 2. Comparison with state-of-the-arts on the NYUDv2 dataset. Results are reported in terms of percentage (%) IoU. The best performance for per class is marked in bold.

Method	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	blinds	desk	shelves	curtain	dresser	pillow	mirror	floormat
FCN [1]	69.9	79.4	50.3	66.0	47.5	53.2	32.8	22.1	39.0	36.1	50.5	54.2	45.8	11.9	8.6	32.5	31.0	37.5	22.4	13.6
Gupta et al. [51]	68.0	81.3	44.9	65.0	47.9	29.9	20.3	32.6	39.0	18.1	40.3	51.3	42.0	11.3	3.5	29.1	34.8	34.4	16.4	28.0
Deng et al. [54]	65.6	79.2	51.9	66.7	41.0	55.7	36.5	20.3	33.2	32.6	44.6	53.6	49.1	10.8	9.1	47.6	27.6	42.5	30.2	32.7
STD2P [56]	72.7	85.7	55.4	**73.6**	58.5	60.1	42.7	30.2	42.1	41.9	52.9	59.7	46.7	13.5	9.4	40.7	44.1	42.0	34.5	35.6
LSD-GF [16]	78.5	87.1	56.6	70.1	**65.2**	63.9	46.9	35.9	47.1	**48.9**	54.3	66.3	51.7	20.6	13.7	49.8	43.2	50.4	48.5	32.2
RDFNet [15]	79.7	87.0	60.9	73.4	64.6	**65.4**	**50.7**	39.9	**49.6**	44.9	61.2	67.1	**63.9**	**28.6**	14.2	**59.7**	49.0	49.9	**54.3**	**39.4**
CANet (ResNet-50)	79.6	87.5	61.1	70.7	63.7	64.7	46.8	44.6	46.5	46.9	61.2	68.9	58.0	22.4	14.1	56.1	47.0	48.6	49.1	32.0
CANet (ResNet-101)	**80.1**	**88.3**	**61.7**	72.8	63.9	**65.4**	48.0	**46.5**	48.3	44.4	**61.4**	**69.9**	59.5	27.2	**16.8**	59.3	**50.6**	**50.9**	51.3	38.6

Method	clothes	ceiling	books	fridge	tv	paper	towel	shower	box	board	person	nightstand	toilet	sink	lamp	bathtub	bag	ot. stuct.	ot. furn.	ot. props
FCN [1]	18.3	59.1	27.3	27.0	41.9	15.9	26.1	14.1	6.5	12.9	57.6	30.1	61.3	44.8	32.1	39.2	4.8	15.2	7.7	30.0
Gupta et al. [51]	4.7	60.5	6.4	14.5	31.0	14.3	16.3	4.2	2.1	14.2	0.2	27.2	55.1	37.5	34.8	38.2	0.2	7.1	6.1	23.1
Deng et al. [54]	12.6	56.7	8.9	21.6	19.2	28.0	28.6	22.9	1.6	1.0	9.6	30.6	48.4	41.8	28.1	27.6	0	9.8	7.6	24.5
STD2P [56]	22.2	55.9	29.8	41.7	52.5	21.1	34.4	15.5	7.8	29.2	60.7	42.2	62.7	47.4	38.6	28.5	7.3	18.8	15.1	31.4
LSD-GF [16]	24.7	62.0	34.2	45.3	53.4	27.7	42.6	23.9	11.2	58.8	53.2	**54.1**	80.4	59.2	45.5	52.6	15.9	12.7	16.4	29.3
RDFNet [15]	**26.9**	69.1	**35.0**	**58.9**	**63.8**	**34.1**	41.6	38.5	11.6	54.0	80.0	45.3	65.7	62.1	47.1	57.3	**19.1**	**30.7**	20.6	39.0
CANet (ResNet-50)	22.9	79.0	32.7	51.8	60.4	32.7	38.4	**41.3**	14.7	81.9	**81.0**	39.0	78.0	61.9	**49.5**	53.5	9.3	28.1	20.1	**39.3**
CANet (ResNet-101)	25.1	**79.5**	33.5	56.0	60.8	31.7	**47.7**	25.3	**14.8**	**83.7**	77.6	40.2	83.8	**67.3**	48.2	**66.2**	11.0	30.6	**21.2**	39.2

As shown in Tables 1, 2 and 3, we compare CANet with other state-of-the-art methods on the two RGB-D datasets. The performance is reported with different backbones ResNet-50 and ResNet-101.

NYUDv2 dataset. We evaluate the three aforementioned metrics on our network for 40 classes on the NYUDv2 dataset. As illustrated in Table 1, we achieve the new state-of-the-art results on all three metrics. We owe the better performance to the RGB-D co-attention fusion module. The two fusion modules could effectively fuse the two modality features by capturing the long-range dependencies between RGB and D information. On the most important metric mean IoU, we achieve 51.2% with a slightly 2.2% improvement over the recent state-of-the-art method RDFNet [15].

On the NYUDv2 dataset, the distribution of semantic labels is long tail, with the number of some labels are very few. To evaluate the performance of our model on the imbalanced distributed dataset, we also show the category-wise results on each category, as in table 2. Our method performs better than other methods over 18 classes (40 classes in total), especially in some hard categories (e.g., shelves, box, ot. furn.), which demonstrate the robustness of our method among different categories with imbalanced training data.

Table 3. Comparison with state-of-the-arts on the SUN RGB-D dataset. Results are reported in terms of percentage (%) of pixel accuracy, mean accuracy, and mean IoU

Method	Backbone	PixAcc.	mAcc.	mIoU
FCN [1]	VGG-16	-	-	35.1
FuseNet [13]	VGG-16	76.3	48.3	37.3
Jiang et al.[59]	-	76.6	50.6	39.3
D-CNN [17]	ResNet-152	-	53.5	42.0
3DGNN [57]	ResNet-101	-	57.0	45.9
LSD-GF [16]	VGG-16	-	58.0	-
RDFNet [15]	ResNet-152	81.5	60.1	47.7
CFN [18]	ResNet-152	-	-	48.1
CANet	ResNet-50	81.6	59.0	48.1
CANet	ResNet-101	82.5	60.5	49.3

SUN RGB-D dataset. Following the same test pattern on the NYUDv2 dataset. We also compare our method with state-of-the-art methods on the large-scale SUN RGB-D dataset. The test results are shown in Table 3. Our methods outperform existing RGB-D semantic segmentation methods and is the state-of-the-art with all three evaluation metrics. The comparison on this large-scale dataset again validates the effectiveness of our proposed method.

4.4 Ablation Study

To verify the performances of co-attention fusion modules, we conduct an ablation study on the NYUDv2 dataset. Each experiment is ablated with the same

Table 4. Ablation study of CANet on the NYUDv2 dataset. Results are reported in terms of percentage (%) of pixel accuracy, mean accuracy, and mean IoU.

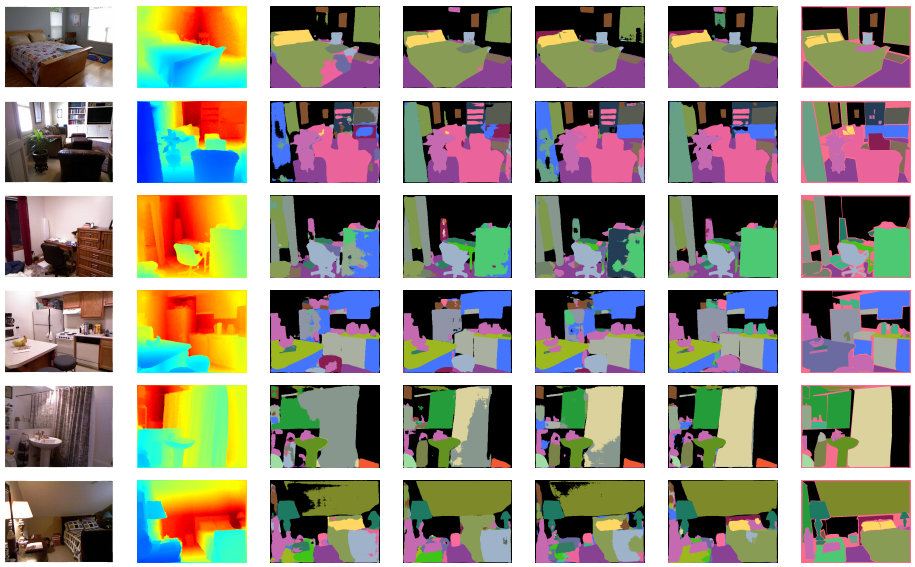
Method	Backbone	PixAcc.	mAcc.	mIoU
Basel.	ResNet-50	74.1	60.1	46.6
Basel. + PCFM	ResNet-50	75.4	61.6	48.9
Basel. + CCFM	ResNet-50	74.9	62.0	48.4
CANet	ResNet-50	75.7	62.6	49.6
Basel.	ResNet-101	75.2	62.2	48.5
Basel. + PCFM	ResNet-101	76.0	62.9	49.6
Basel. + CCFM	ResNet-101	75.5	63.0	50.2
CANet	ResNet-101	76.6	63.8	51.2

hypermeter setting at both training and testing period. For fair comparison, we regard the simple element-wise sum fusion as our baseline. The performance of each component is shown in Table 4. When using ResNet-50 as the backbone, the PCFM achieves 48.9 in mIoU with 5.0% improvement over the baseline method, and the CCFM improves 3.9% over the baseline method. When integrating the two modules, we gain further improvements 6.4% over the baseline, which demonstrating their complementary power over utilizing either alone. Furthermore, the usage of a deeper backbone network (ResNet-101) can still bring large improvements, which demonstrates our proposed modules are not limited to stronger backbones.

4.5 Visualizations

Semantic Segmentation Qualitative Visual Results: Figure 5 is the visualization for our sampled examples in RGB-D indoor semantic segmentation with Baseline, Baseline + PCFM, Baseline + CCFM, and Baseline + PCFM + CCFM (CANet) on the NYUDv2 dataset, which involves cluttered objects from various indoor scenes. Compared to Baseline we can see that both the PCFM and CCFM promotes the semantic segmentation results on details and misclassification problems. Moreover, the integration of PCFM and CCFM gets better segmentation results than the use of PCFM or CCFM individually.

Co-attention Affinity Matrix Visualization: The position co-attention affinity matrix has the shape of $HW \times HW$. For each point in the color features, a sub-co-attention affinity matrix ($1 \times HW$) is used to multiply with depth features maps. The aggregated depth feature then is added to the color feature of that point. In Figure 6, we choose two points (p1, p2) and visualize their sub-co-attention matrices. We could clearly see that sub-co-attention affinity matrix puts more attention on the areas with the same labels even when some of them are far away from that point. For example, in the first row, the co-attention affinity matrix of p1 focuses on most regions which is labeled as window, and p2 focuses the attention on the regions labeled as table. The visualization results



(a) RGB (b) Depth (c) Basel. (d) PCFM (e) CCFM (f) CANet (g) GT

Fig. 5. Semantic Segmentation Qualitative Visual Results on the NYUDv2 dataset. (a), (b) and (g) are the input color images, depth images and ground truth labels, respectively. (c) are the results of Baseline. (d) are the result of Baseline + PCFM. (e) are the results of Baseline + CCFM. (f) are the results of proposed CANet.

show the co-attention affinity matrix could capture long-range depth features with the same semantic label.

The channel co-attention affinity matrix is with the shape of $C \times C$. For the feature map of each channel, a sub-co-attention affinity matrix $(1 \times C)$ is a weight matrix of depth feature maps and is used to aggregate the depth features. The aggregated depth feature is added to the color feature of that channel. In Figure 7, we randomly select two channels (c1, c2) and visualize their aggregated depth feature maps. As Figure 7 shown, the attention of the fused channel features focuses more on areas with the same label. For example, in the first row, the attention of c1 are focused on the regions of wall, and c2 focus attentions on bed areas. The visualization results demonstrate that the channel co-attention affinity matrix could capture the class-aware long-range cross channel depth features.

5 Conclusion

In this paper, we propose a novel CANet method that learns more representative, robust and discriminative feature embeddings for RGB-D semantic segmentation. In CANet, co-attention is used to adaptively aggregate depth features with color features. Specifically, we introduce position co-attention fusion

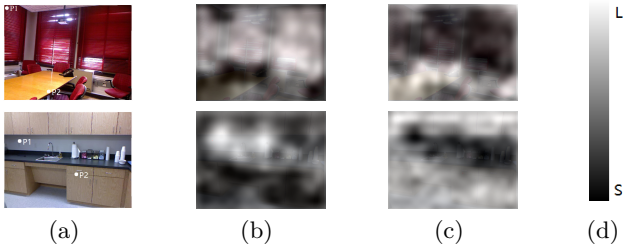


Fig. 6. Visualization of position co-attention affinity matrix. (a) Original image; (b) Sub affinity matrix of p1; (c) Sub affinity matrix of p2; (d) Color bar.

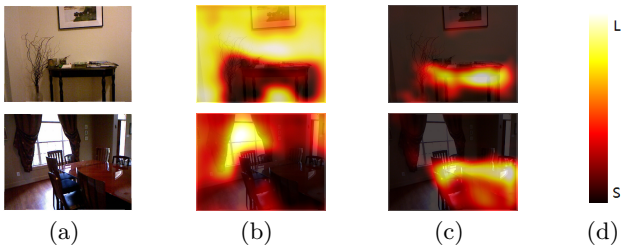


Fig. 7. Visualization of channel co-attention affinity matrix. (a) Original images; (b) Fused feature map of c1; (c) Fused feature map of c2; (d) Color bar.

module (PCFM) and channel co-attention fusion module (CCFM) to capture inter-modality long-range dependencies in spatial and channel dimensions respectively. Meantime, we design a final fusion module (FFM) to effectively integrate of position co-attention fusion module and channel co-attention fusion module. The ablation study and visualization results illustrate the importance of each component. The experiments on the NYUDv2 and the SUN RGB-D datasets demonstrate that the proposed CANet outperforms existing RGB-D semantic segmentation methods. The interdependency between different modality feature will be further explored in the future.

Acknowledge

This work is supported partly by the National Natural Science Foundation (NSFC) of China (grants 61973301, 61972020, 61633009, 51579053 and U1613213), partly by the National Key R&D Program of China (grants 2016YFC0300801 and 2017YFB1300202), partly by the Field Fund of the 13th Five-Year Plan for Equipment Pre-research Fund (No.61403120301), partly by Beijing Science and Technology Plan Project, partly by the Key Basic Research Project of Shanghai Science and Technology Innovation Plan (No.15JC1403300), and partly by Meituan Open R&D Fund.

References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3431–3440
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39** (2017) 2481–2495
3. Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J.: Amodal instance segmentation with kins dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3014–3023
4. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40** (2017) 834–848
6. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1925–1934
7. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 4353–4361
8. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
9. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: European conference on computer vision, Springer (2014) 345–360
10. He, Y., Chiu, W.C., Keuper, M., Fritz, M.: Rgbd semantic segmentation using spatio-temporal data-driven pooling. arXiv preprint arXiv:1604.02388 (2016)
11. Wang, J., Wang, Z., Tao, D., See, S., Wang, G.: Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. In: European Conference on Computer Vision, Springer (2016) 664–679
12. Husain, F., Schulz, H., Dellen, B., Torras, C., Behnke, S.: Combining semantic and geometric features for object class segmentation of indoor scenes. IEEE Robotics and Automation Letters **2** (2016) 49–55
13. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Asian conference on computer vision, Springer (2016) 213–228
14. Jiang, J., Zheng, L., Luo, F., Zhang, Z.: Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. arXiv preprint arXiv:1806.01054 (2018)
15. Park, S.J., Hong, K.S., Lee, S.: Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 4980–4989
16. Cheng, Y., Cai, R., Li, Z., Zhao, X., Huang, K.: Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 3029–3037
17. Wang, W., Neumann, U.: Depth-aware cnn for rgb-d segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 135–150

18. Lin, D., Chen, G., Cohen-Or, D., Heng, P.A., Huang, H.: Cascaded feature network for semantic segmentation of rgb-d images. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 1311–1319
19. Lin, D., Zhang, R., Ji, Y., Li, P., Huang, H.: Scn: Switchable context network for semantic segmentation of RGB-D images. IEEE transactions on cybernetics (2018)
20. Hu, X., Yang, K., Fei, L., Wang, K.: Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE (2019) 1440–1444
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. (2017) 5998–6008
22. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision, Springer (2012) 746–760
23. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 567–576
24. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
25. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. (2015) 2048–2057
26. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 21–29
27. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623 (2015)
28. Chen, X., Mishra, N., Rohaninejad, M., Abbeel, P.: Pixelsnail: An improved autoregressive generative model. arXiv preprint arXiv:1712.09763 (2017)
29. Cheng, J., Dong, L., Lapata, M.: Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733 (2016)
30. Parikh, A.P., Täckström, O., Das, D., Uszkoreit, J.: A decomposable attention model for natural language inference. arXiv preprint arXiv:1606.01933 (2016)
31. Tang, J., Jin, L., Li, Z., Gao, S.: Rgb-d object recognition via incorporating latent data structure and prior knowledge. IEEE Transactions on Multimedia **17** (2015) 1899–1908
32. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 3194–3203
33. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130 (2017)
34. Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., Zhang, C.: Disan: Directional self-attention network for rnn/cnn-free language understanding. In: Thirty-Second AAAI Conference on Artificial Intelligence. (2018)
35. Qi, L., Liu, S., Shi, J., Jia, J.: Sequential context encoding for duplicate removal. In: Advances in Neural Information Processing Systems. (2018) 2049–2058
36. Zhu, Y., Wang, J., Xie, L., Zheng, L.: Attention-based pyramid aggregation network for visual place recognition. In: Proceedings of the 26th ACM international conference on Multimedia. (2018) 99–107

37. Song, X., Zhang, S., Hua, Y., Jiang, S.: Aberrance-aware gradient-sensitive attentions for scene recognition with rgb-d videos. In: Proceedings of the 27th ACM International Conference on Multimedia. (2019) 1286–1294
38. Li, W., Tao, X., Guo, T., Qi, L., Lu, J., Jia, J.: Mucan: Multi-Correspondence Aggregation Network for Video Super-Resolution. arXiv preprint arXiv:2007.11803 (2020)
39. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7794–7803
40. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7132–7141
41. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2019) 0–0
42. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018)
43. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 5659–5667
44. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3146–3154
45. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 8351–8361
46. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances in neural information processing systems. (2016) 289–297
47. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision. (2017) 1821–1830
48. Nguyen, D.K., Okatani, T.: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6087–6096
49. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 299–307
50. Ren, X., Bo, L., Fox, D.: Rgb-(d) scene labeling: Features and algorithms. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2012) 2759–2766
51. Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from rgb-d images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 564–571
52. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
53. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241

54. Deng, Z., Todorovic, S., Jan Latecki, L.: Semantic segmentation of rgbd images with mutex constraints. In: Proceedings of the IEEE international conference on computer vision. (2015) 1733–1741
55. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. (2015) 2650–2658
56. He, Y., Chiu, W.C., Keuper, M., Fritz, M.: Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4837–4846
57. Qi, X., Liao, R., Jia, J., Fidler, S., Urtasun, R.: 3d graph neural networks for rgbd semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 5199–5208
58. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. (2019) 8024–8035
59. Jiang, J., Zhang, Z., Huang, Y., Zheng, L.: Incorporating depth into both cnn and crf for indoor semantic segmentation. In: 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), IEEE (2017) 525–530