Learning Multi-level Region Consistency with Dense Multi-label Networks for Semantic Segmentation

Tong Shen[1], **Guosheng Lin**[2*], **Chunhua Shen**[1], **Ian Reid**[1]
[1]School of Computer Science, The University of Adelaide, Australia
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
{tong.shen, chunhua.shen, ian.reid}@adelaide.edu.au
guosheng.lin@gmail.com

Abstract

Semantic image segmentation is a fundamental task in image understanding. Per-pixel semantic labelling of an image benefits greatly from the ability to consider region consistency both locally and globally. However, many Fully Convolutional Network based methods do not impose such consistency, which may give rise to noisy and implausible predictions. We address this issue by proposing a dense multi-label network module that is able to encourage the region consistency at different levels. This simple but effective module can be easily integrated into any semantic segmentation systems. With comprehensive experiments, we show that the dense multi-label can successfully remove the implausible labels and clear the confusion so as to boost the performance of semantic segmentation systems.

1 Introduction

Semantic segmentation is one of the fundamental problems in computer vision, whose task is to assign a semantic label to each pixel of an image so that different classes can be distinguished. This topic has been widely studied [Girshick *et al.*, 2014; Hariharan *et al.*, 2014; Yadollahpour *et al.*, 2013; Farabet *et al.*, 2013]. Among these models, Fully Convolutional Network (FCN) based models have become dominant [Dai *et al.*, 2015; Chen *et al.*, 2015; Lin *et al.*, 2015; Chen *et al.*, 2016]. These models are simple and effective because of the powerful capacity of Convolutional Neural Networks (CNNs) and being able to be trained end-to-end. However, most existing methods do not have the mechanism to enforce the region consistency, which plays an important role in semantic segmentation. Consider, for example, Figure 1, in which the lower left image is the output of a vanilla FCN, whose prediction contains some noisy labels that do not appear in the ground truth. With enforced region consistency, we can simply eliminate those implausible labels and clear the confusion. Our aim in this work is to introduce constraints to encourage this consistency.

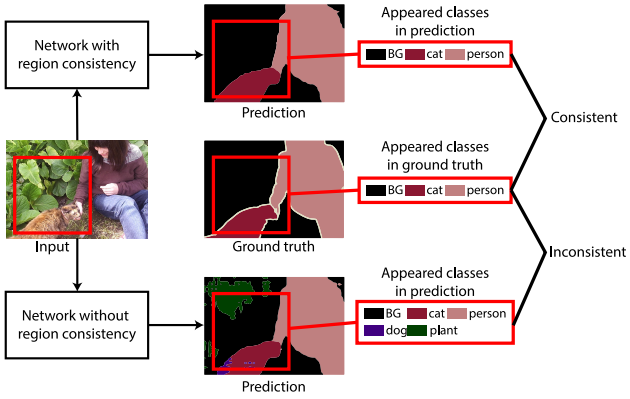

Figure 1: Illustration of region consistency. For a region in the input image, which is coloured in red, the corresponding part in the ground truth contains only three classes. In the network without region consistency, there are five classes that appear. If we explicitly encourage the consistency, those unlikely classes will be eliminated and the prediction will be better as shown on top.

Our proposal is both simple and effective: we argue that the region consistency in a certain region can be formulated as a multi-label classification problem. Multi-label classification has also been widely studied [Jiang, 2016; Wei *et al.*, 2016; Guo and Gu, 2011], whose task is to assign one or more labels to the image. By performing multi-label classification in a region, we can allow the data to suggest which labels are likely within the broad context of the region, and use this information to suppress implausible classes predicted without reference to the broader context, thereby improving scene consistency. While typical multi-label problems are formulated as whole-image inference, we adapt this approach to dense prediction problems such as semantic segmentation, by introducing dense multi-label prediction for image regions of various sizes.

Dense multi-label prediction is performed in a sliding window fashion: the classification for each spatial point is influenced by the network prediction and by the multi-label result for the surrounding window. By employing different window sizes, we are able to construct a multi-level structure for dense multi-label and enforce the region consistency at different levels both locally and globally. Figure 2 is an illus-

*The first two authors contributed equally and this work was done when Guosheng Lin was with The University of Adelaide.

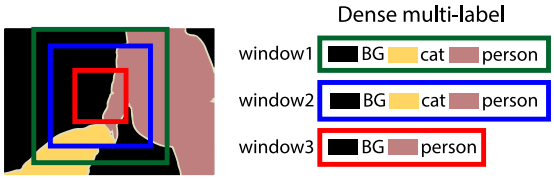

Figure 2: Illustration of dense multi-label with multi-level. Windows in different colours indicate different regions for dense multi-label classification.

tration of dense multi-label at multiple windows sizes.Here we use three windows of different sizes. The red window, the smallest, focuses more on the local region consistency, while the green window, the largest, is responsible for global region consistency. The other one, in blue, is for mid-level consistency. By sliding the windows to consider each spatial point, we perform multi-label densely at different level, encouraging the segmentation predictor to give predictions that are consistent with the dense multi-label prediction.

Our contributions are as follows:

- We address the problem of region consistency in semantic segmentation by proposing a dense multi-label module to achieve the goal of retaining region consistency, which is simple and effective. We also introduce a multi-level structure for dense multi-label to preserve region consistency both locally and globally.
- We evaluate our method on four popular semantic segmentation datasets including NYUDv2, SUN-RGBD, PASCAL-Context and ADE 20k, and achieve promising results. We also give analysis on how dense multi-label can remove the implausible labels, clear confusion and effectively boost the segmentation systems.

2 Related Work

Semantic segmentation has been widely studied [Girshick *et al.*, 2014; Hariharan *et al.*, 2014]. Early CNN based methods rely on region proposals or superpixels. They make segmentation prediction by classifying these local features. More recently, with Long *et al.*[Long *et al.*, 2015] introducing applying Fully Convolutional Networks(FCNs) to semantic segmentation, the FCN based segmentation models [Dai *et al.*, 2015; Chen *et al.*, 2015; Lin *et al.*, 2015; Chen *et al.*, 2016] have become popular.

Multi-label classification has also been widely studied. Traditional methods are based on graphical models [Xue *et al.*, 2011; Guo and Gu, 2011], while the recent studies benefit more from CNNs [Wei *et al.*, 2016; Jiang, 2016].

Here we propose a dense multi-label module to take advantage of multi-label classification and integrate it into semantic segmentation systems. Dense multi-label is performed in a sliding window fashion and treats all area in a window as multi-label classification. Experiments show that dense multi-label can help to keep the scene consistency, clear confusion and boost the performance of semantic segmentation.

3 Methods

3.1 Dense Multi-label

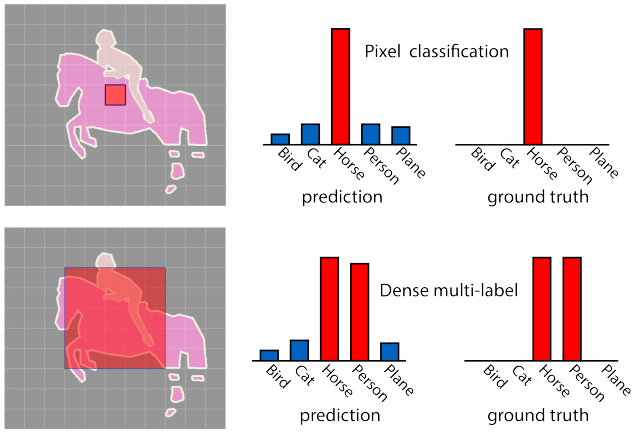

Figure 3: An illustration of differences between pixel classification and dense multi-label prediction. In pixel classification, we treat each spatial point as a single-label classification problem where only one class is supposed to get very high confidence; dense multi-label focuses on label concurrence where the labels that appear in the region will have equally high confidence.

Multi-label classification is a task where each image can have more than one label, unlike a multi-class classification problem [Simonyan and Zisserman, 2015; Szegedy *et al.*, ; He *et al.*, 2016] whose goal is to assign only one label to the image. This is more natural in reality because for majority of images, objects are not isolated, instead they are in context with other objects or the scene. Multi-label classification gives us more information of the image.

For a dense prediction task such as segmentation, it treats every spatial point as a multi-class classification problem, where the point is assigned with one of the categories. As shown in the upper part of Figure 3, the model predicts scores for each class and picks the highest one. The ground truth is an one-hot vector correspondingly. For a dense multi-label problem, each spatial point will be assigned with several labels to show what labels appear in the a certain window centered at this point. As shown in lower part of Figure 3, there are two classes being predicted with high confidence and the ground truth is given by a "multiple hot" vector.

Here we propose a method to learn a dense multi-label system and a segmentation system at the same time. We aim at using dense multi-label to suppress the implausible classes and encourage appropriate classes so as to retain the region consistency for the segmentation prediction both globally and locally. In the next section, more details of the whole framework will be provided.

3.2 Overview of Framework

An overview of the structure is shown in Figure 4, with the part in the dashed-line rectangle being the dense multi-label module. Without it, the network simply becomes a FCN. The input image is first fed into several low level feature layers

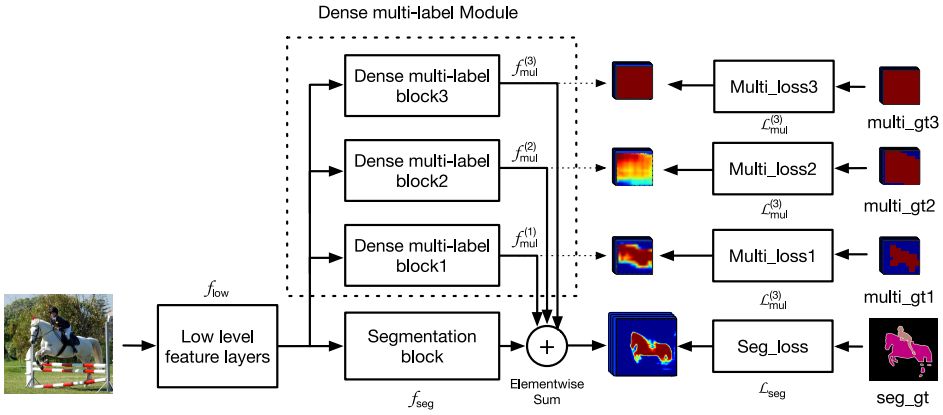

Figure 4: Illustration of the framework with dense multi-label module. The input image is first passed into low level feature layers, which are shared by the following blocks. Then the feature maps are fed into the segmentation block and three dense multi-label blocks. The element-wise sum will sum up the features from the blocks and make the final prediction. Apart from the segmentation loss, each dense multi-label block also has its own multi-label loss to guide the training.

that are shared by the following blocks. Then apart from going into the segmentation block, the features also enter three blocks for dense multi-label prediction. The outputs of theses blocks are merged element-wise for the final prediction.

In the training phase, the network is guided by four loss functions: the segmentation loss and three dense multi-label losses. We use softmax loss for the segmentation path, and use logistic loss for all the dense multi-label blocks.

The dense multi-label blocks have different window sizes for performing dense multi-label prediction within different contexts. With this multi-level structure, we are able to retain region consistency both locally and globally.

Let $\boldsymbol{x}$ denote the image. The process of the low level feature block can be described as:

$$\boldsymbol{o} = f_{low}(\boldsymbol{x}; \boldsymbol{\theta}_{low}), \tag{1}$$

where $\boldsymbol{o}$ is the output and $\boldsymbol{\theta}_{low}$ the layer parameters.

The dense multi-label blocks and the segmentation block are defined as:

$$\boldsymbol{m}^{(j)} = f^{(j)}_{mul}(\boldsymbol{o}; \boldsymbol{\theta}^{(j)}_{mul}), j \in \{1, 2, 3\} \tag{2}$$

$$\boldsymbol{s} = f_{seg}(\boldsymbol{o}; \boldsymbol{\theta}_{seg}), \tag{3}$$

where $\boldsymbol{m}^{(j)}$ and $\boldsymbol{s}$ denote the output of jth multi-label block and the output of segmentation respectively. $\boldsymbol{\theta}^{(j)}_{mul}$ and $\boldsymbol{\theta}_{seg}$ are layer parameters.

The final prediction is:

$$\boldsymbol{p} = \boldsymbol{s} + \boldsymbol{m}^{(1)} + \boldsymbol{m}^{(2)} + \boldsymbol{m}^{(3)}, \tag{4}$$

where $\boldsymbol{p}$ is the fused score for segmentation.

For the loss functions, we use logistic loss for the prediction of dense multi label blocks, $\boldsymbol{m}^{(1)}$,$\boldsymbol{m}^{(2)}$ and $\boldsymbol{m}^{(3)}$; softmax loss is used for final prediction $\boldsymbol{p}$. Let m_{ik} be the out of a dense multi-label block at ith position for kth class, and y^{mul}_{ik} be the ground truth for the corresponding position and class. The loss function for dense multi-label is defined as:

$$l_{mul}(\boldsymbol{y}^{mul}, \boldsymbol{m}) = \frac{1}{IK} \sum_{i}^{I} \sum_{k}^{K} y^{mul}_{ik} \log(\frac{1}{1+e^{-m_{ik}}}) + (1 - y^{mul}_{ik}) \log(\frac{e^{-m_{ik}}}{1+e^{-m_{ik}}}), \tag{5}$$

where $y^{mul}_{ik} \in \{0, 1\}$; I and K represent the number of spatial points and classes, respectively.

Similarly, let p_{ik} be the fused output at ith position for kth class, and y^{seg}_i be the ground truth for segmentation prediction at ith position. The loss function for segmentation is defined as:

$$l_{seg}(\boldsymbol{y}^{seg}, \boldsymbol{p}) = \frac{1}{I} \sum_{i}^{I} \sum_{k}^{K} \mathbb{I}(y^{seg}_i = k) \log(\frac{e^{p_{ik}}}{\sum_j e^{p_{ij}}}), \tag{6}$$

where $y^{seg}_i \in \{1 \dots K\}$.

Our goal is to minimize the objective function:

$$\min l_{seg} + \lambda(l^{(1)}_{mul} + l^{(2)}_{mul} + l^{(3)}_{mul}), \tag{7}$$

where λ controls the balance between the segmentation block and the dense multi-label blocks. I observe this parameter is not very sensitive. We set $\lambda = 1$ to treat each part equally.

3.3 Dense Multi-label Block

The details of the dense multi-label block are shown in Figure 5, where the input is feature maps at 1/8 resolution, due to the downsampling in the low level feature layers. After some convolutional layers with further downsampling, the dense multi-label is performed at 1/32 resolution with the sliding window and following adaptive layers. The reason for this setting is because dense multi-label requires a large sliding window, which will become a computational burden if we work at a high resolution. Downsampling can greatly reduce the size of feature maps and more importantly, the size of sliding window will shrink accordingly, thus making the computation more efficient. On the other hand, dense multi-label

requires more high level information. Therefore, working at a coarse level can capture the high level features better. The output of the dense multi-label is upsampled to be compatible with the segmentation block's output.

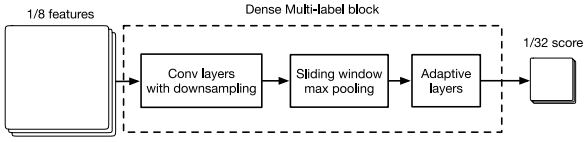

Figure 5: Details of a single dense multi-label block. The input features are fed into several convolutional layers and further downsampled. Then we perform sliding window with max pooling operation. After some adaptive layers, we have scores for dense multi-label at 1/32 resolution.

3.4 Ground Truth Generation

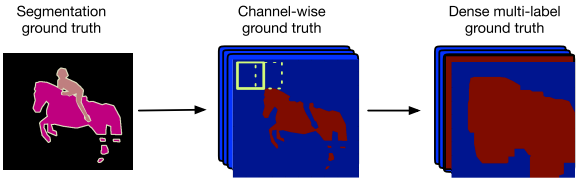

Figure 6: The segmentation ground truth is firstly converted to channel-wise labels, with 0 or 1 in each channel. The ground truth for dense multi-label can be obtained by performing max pooling on the channel-wise labels.

The ground truth for dense multi-label can be generated from the segmentation ground truth. The process is described in Figure 6. Firstly, the segmentation ground truth is converted to channel-wise labels, which means each channel only contains 1 or 0 to indicate whether the corresponding class appears or not. To generate a ground-truth mask for each class, for a given window size, we slide the window across each binary channel and perform a max-pool operation (this is equivalent to a binary dilation using a structuring element of the same size and shape as the window). We repeat this process for each window size. As noted in section 3.3, the dense multi-label classification is performed at 1/32 resolution while the segmentation is at 1/8. Therefore, we generate multi-label ground-truth data at 1/8 resolution with stride 4.

3.5 Network Configuration

The dense multi-label module is suitable for any segmentation system and it can be easily integrated. In this study, we use Residual 50-layer network [He *et al.*, 2016] with dilated kernels [Chen *et al.*, 2015]. In order to work at a relatively high resolution while keeping the efficiency, we use 8-stride setting, which means that the final output is at 1/8 resolution. As we mentioned in the last section, we perform dense multi-label at 1/32 resolution to make it more efficient and effective. The window sizes are then defined at 1/32 resolution. For example,let w be the window size. A window with $w = 17$ at 1/32 resolution means $4w = 68$ at 1/8 resolution. The corresponding window for the original image is $32w = 544$. We use $w_1 = 35$, $w_2 = 17$ and $w_3 = 7$ for all the experiments.

Block name	Initial layers	Stride
Low level feature block	conv1 to res3d	8
Segmentation block	res4a to res5c	1
Dense multi-label block	res4a to res5c	4

Table 1: Configuration for Res50 network. The low level feature block is initialized by layers "conv1" to "res3d" and has 8 stride. The segmentation block and dense multi-label blocks are initialized by layers "res4a" to "res5c" but do not share the weights with each other. The segmentation block does not have any downsampling, but the dense multi-label blocks have further 4 stride downsampling.

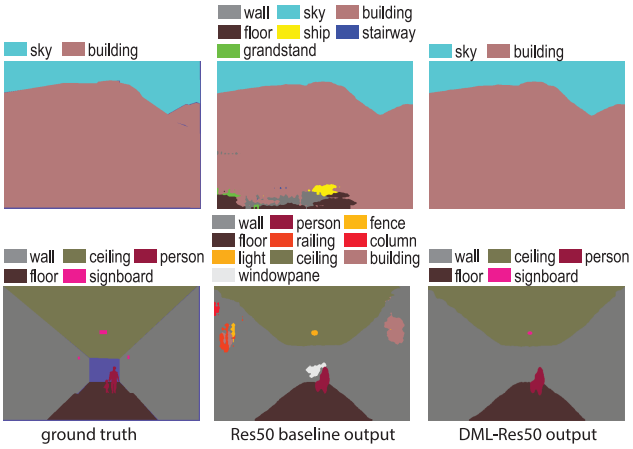

Figure 7: Example outputs of Res50 baseline and DML-Res50 on ADE-20k dataset.

Table 1 shows the configuration with 50-layer Residual net (Res50) as the base network. The low level feature block contains layers from "conv1" to "res3d". The segmentation block and dense multi-label blocks have layers from "res4a" to "res5c" as well as some adaptive layers. It is worth noting that it does not mean these blocks will share the weights even though they initialize the weights from the same layers. After initialization, they will learn their own features separately.

4 Experiments and Analysis

We evaluate our model on 4 commonly used semantic segmentation datasets: ADE-20k, NYUDv2, SUN-RGBD and PASCAL-Context. Our comprehensive experiments show that dense multi-label can successfully suppress many unlikely labels, retain region consistency and thus improve the performance of semantic segmentation.

The results are evaluated using the Intersection-over-Union (IoU) score [Everingham *et al.*, 2010]. Moreover, since our original motivation is to suppress noisy and unreasonable labels to keep labels consistent with the region, we also introduce new measurements to evaluate the number of classes that are not in ground truth, and further, the number of pixels that are predicted to be these wrong classes for each image.

We only use Res50 as base network to compare and analyse the performance. For all the experiments, we use batch size

Model	IOU	#Wrong class	#Wrong label
Res50 baseline	34.5	5.6	21836
DML Res50	**36.5**	**3.6**	**18294**

Table 2: Results on ADE-20k. The dense multi-label boosts the IOU by 2% and helps reduce the number of wrong class and label by 35% and 16% respectively.

Model	IOU
DilatedNet [Zhou *et al.*, 2016]	32.3
Cascade-DilatedNet [Zhou *et al.*, 2016]	34.9
DML-Res50(ours)	**36.5**

Table 3: Comparsion with other models on ADE-20k dataset. Our model achieves the best performance.

of 8, momentum of 0.9 and weight decay of 0.0005.

4.1 Results on ADE-20k

We first evaluate our result on ADE-20k dataset[Zhou *et al.*, 2016], which contains 150 semantic categories including objects such as person, car *etc*., and "stuff" such as sky, road *etc*.. There are 20210 images in the training set and 2000 images in the validation set.

As shown in Table 2, the model with dense multi-label (DML-Res50) yields a 2% improvement. To analyse the effectiveness of label suppression, we also use two criteria to evaluate this performance, which are shown as "Wrong class" and "Wrong labels". Wrong class means the number classes that are not supposed to appear but are mistakenly predicted by the model. Wrong labels describe how many pixels are assigned with those wrong classes. We observe that using Dense multi-label effectively reduces the wrong classes and labels, by 35% and 16% respectively. Some examples are shown in Figure 7. To make fair comparison, all the images are raw outputs directly from the network. The last column shows the outputs from the network with dense multi-label where we can observe great scene consistency compared with the output of the baseline network shown in the middle.

We achieve better results than the models reported in [Zhou *et al.*, 2016], as shown in Table 3.

4.2 Results on PASCAL-Context

PASCAL-Context dataset [Mottaghi *et al.*, 2010] is a set of additional annotations for PASCAL VOC 2010, which provides annotations for the whole scene with 60 classes (59 classes and a background class). It contains 4998 images in training set and 5105 images in validation set.

Figure 8 shows some typical examples on this dataset. We can also see clear scene consistency with dense multi-label involved. The outputs in the middle contain many noisy classes, especially the lower middle image contains "bird" and "sky", which are very unlikely in this scene. From Table 4, we can also see the great boost with dense multi-label. The wrong classes and labels are greatly reduced by 37% and 15%.

To compare with other models, we list several results on this dataset. Since different models have various settings such as multi-scale training, extra data, *etc*.we also explain it in Table 5. Considering all the factors involved, our method is

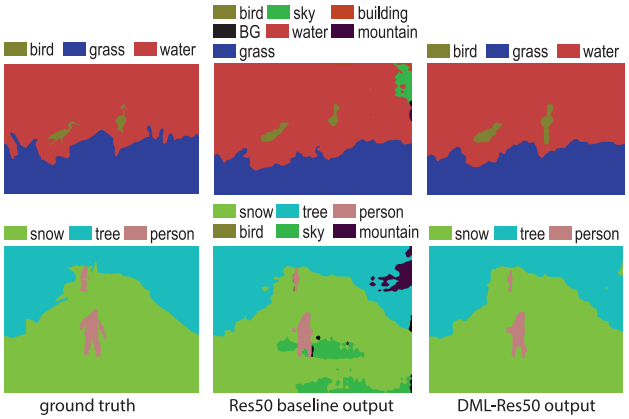

Figure 8: Example outputs of Res50 baseline and DML-Res50 on PASCAL-Context dataset.

Model	IOU	#Wrong class	#Wrong label
Res50 baseline	41.4	4.5	26308
DML-Res50	**44.4**	**2.8**	**22367**

Table 4: Results on PASCAL-Context dataset. The dense multi-label model increases the IOU by 3% and reduces the wrong classes and labels by 37% and 15%.

comparable since we only use Res50 as the base network and do not use multi-scale training and extra MS-COCO data.

4.3 Results on NYUDv2

NYUDv2 [Silberman *et al.*, 2012] is comprised of 1449 images from a variety of indoor scenes. We use the standard split of 795 training images and 654 testing images.

Table 6 shows the results on this dataset. With dense multi-label, the performance is improved by more than 1%, and the number of wrong class and label decrease by about 40% and 16%. Some examples are shown in Figure 9. Scene consistency still plays an important role in removing those noisy labels. Compared with some other models, we achieve the best result, as shown in Table 7.

4.4 Results on SUN-RGBD

SUN-RGBD [Song *et al.*, 2015] is an extension of NYUDv2 [Silberman *et al.*, 2012], which contains 5285 training images and 5050 validation images, and provides pixel labelling masks for 37 classes.

Figure 10 shows some output comparison on this dataset, where we can easily observe the effect of dense multi-label. The results are shown in Table 8. The network with dense multi-label helps improve the IOU by more than 3%. The wrong classes and labels also get decreased by 36% and 18% respectively. Compared with other methods, the network with dense multi-label reaches the best result, as shown in Table 9.

4.5 Ablation Study on PASCAL-Context

Table 10 shows an ablation study on the PASCAL-Context. The Res50 baseline yields mean IOU of 41.4%. Treating this as a baseline, we introduce dense multi-level module. Firstly, in the one level setting, we use the largest window

Model	Base	MS	Ex data	IOU
FCN-8s [Long *et al.*, 2015]	VGG16	no	no	37.8
PaserNet [Liu *et al.*, 2015]	VGG16	no	no	40.4
HO_CRF [Arnab *et al.*, 2015]	VGG16	no	no	41.3
Context [Lin *et al.*, 2016]	VGG16	yes	no	43.3
VeryDeep [Wu *et al.*, 2016]	Res101	no	no	44.5
DeepLab [Chen *et al.*, 2016]	Res101	yes	COCO	**45.7**
DML-Res50 (ours)	Res50	no	no	44.4

Table 5: Results on PASCAL-Context dataset. MS means using multi-scale inputs and fusing the results in training. Ex data stands for using extra data such as MS-COCO [Lin *et al.*, 2014]. Compared with state of the art, since we only use Res50 instead of Res101 and do not use multi-scale training as well as extra data, our result is comparable.

Model	IOU	#Wrong class	#Wrong label
Res50 baseline	38.8	8.2	27577
DML-Res50	**40.2**	**4.9**	**23057**

Table 6: Results on NYUDv2 dataset. Dense multi-label network has 1.4% higher IOU and 40% and 16% lower wrong classes and labels respectively.

size, which is basically global multi-label classification. Accordding to the results, the first level gives the biggest boost. With 2 levels involved, the global and mid-level window, the performance is improved further. The final level, the smallest window, brings 0.6% more improvement. The dense multi-label module helps improve the performance by 2.2% in total. After using CRF as post-processing, we can achieve IOU of 44.4 without using extra MS COCO dataset.

Model	IOU
FCN-32s [Long *et al.*, 2015]	29.2
FCN-HHA [Long *et al.*, 2015]	34.0
Context [Lin *et al.*, 2016]	40.0
DML-Res50 (ours)	40.2

Table 7: Comparison with other models on NYUDv2 dataset. Our method achieves the best result.

5 Conclusion

In this study, we propose a dense multi-label module to address the problem of scene consistency. With comprehensive experiments, we have shown that dense multi-label can enforce the scene consistency in a simple and effective way. More importantly, the dense multi-label is a module and can be easily integrated into other semantic segmentation systems.

Acknowledgements

This research was supported by the Australian Research Council through the Australian Centre for Robotic Vision (CE140100016). C. Shen's participation was supported by an ARC Future Fellowship (FT120100969). I. Reid's participation was supported by an ARC Laureate Fellowship (FL130100102).

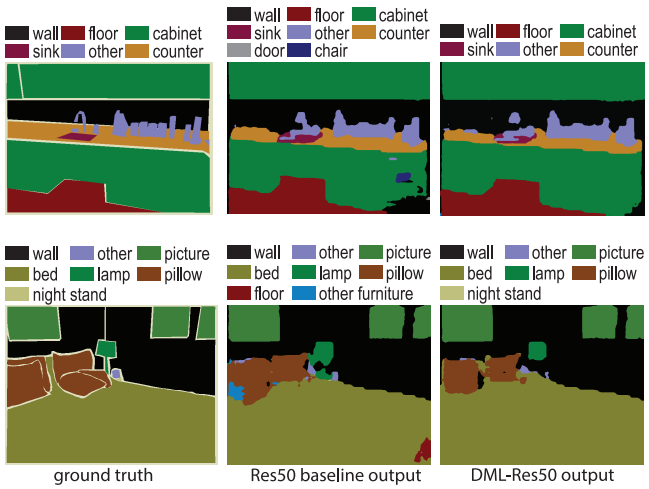

Figure 9: Example outputs of Res50 baseline and DML-Res50 on NYUDv2 dataset.

Model	IOU	#Wrong class	#Wrong label
Res50 baseline	39.3	5.3	24602
DML-Res50	**42.3**	**3.4**	**20104**

Table 8: Results on SUN-RGBD dataset. Dense multi-label helps increase the performance by more than 3% of IOU and decrease the wrong classes and labels by 36% and 18%.

Model	IOU
Kendall *et al.*[Kendall *et al.*, 2015]	30.7
Context [Lin *et al.*, 2016]	42.3
DML-Res50 (ours)	42.4

Table 9: Comparison with other models on SUN-RGBD dataset. We achieve the best result with dense multi-label network.

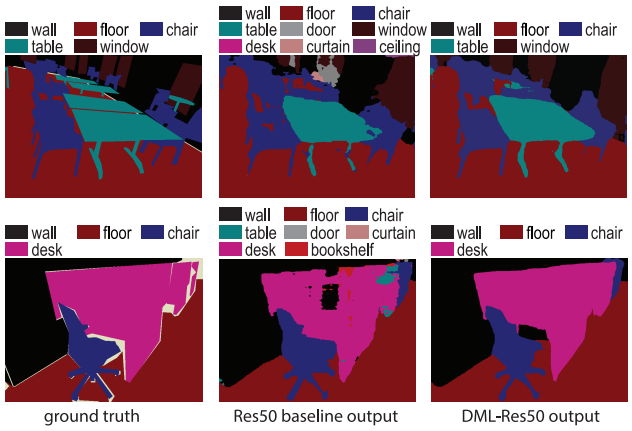

Figure 10: Example outputs of Res50 baseline and DML-Res50 on SUN-RGBD dataset.

Model	IOU
Res50 baseline	41.4
DML-Res50 1level	42.5
DML-Res50 2level	43.0
DML-Res50 3level	43.6
DML-Res50 3level + CRF	**44.4**

Table 10: Ablation study on PASCAL-Context.

References

[Arnab *et al.*, 2015] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip Torr. Higher Order Conditional Random Fields in Deep Neural Networks. *Arxiv*, page 10, 2015.

[Chen *et al.*, 2015] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *ICLR*, 2015.

[Chen *et al.*, 2016] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *CoRR*, abs/1606.0, 2016.

[Dai *et al.*, 2015] Jifeng Dai, Kaiming He, and Jian Sun. [M] BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. In *ICCV*, pages 1635–1643, 2015.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher K I Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[Farabet *et al.*, 2013] Clément Farabet, Camille Couprie, Laurent Najman, and Yann Lecun. Learning Hierarchical Features for Scene Labeling. *TPAMI*, 35(8):1915–1929, 2013.

[Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR*, pages 580–587, 2014.

[Guo and Gu, 2011] Yuhong Guo and Suicheng Gu. Multi-label classification using conditional dependency networks. In *IJCAI*, pages 1300–1305, 2011.

[Hariharan *et al.*, 2014] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous Detection and Segmentation. *ECCV*, pages 297–312, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, SHaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, 2016.

[Jiang, 2016] Wang Jiang. CNN-RNN : A Unified Framework for Multi-label Image Classification. *CVPR*, 2016.

[Kendall *et al.*, 2015] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv:1511.02680v1 [cs.CV]*, 2015.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir D Bourdev, Ross B Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft {COCO:} Common Objects in Context. {*arXiv*}*:1405.0312*, pages 740–755, 2014.

[Lin *et al.*, 2015] Guosheng Lin, Chunhua Shen, Anton van dan Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. pages 1–13, apr 2015.

[Lin *et al.*, 2016] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Exploring Context with Deep Structured models for Semantic Segmentation. *Arxiv 2016*, pages 1–14, 2016.

[Liu *et al.*, 2015] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. ParseNet: Looking Wider to See Better. *arXiv preprint: arXiv:1506.04579*, pages 1–11, 2015.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *CVPR*, pages 3431–3440, 2015.

[Mottaghi *et al.*, 2010] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-gyu Cho, Seong-whan Lee, Raquel Urtasun, and Alan Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. 2010.

[Silberman *et al.*, 2012] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7576 LNCS(PART 5):746–760, 2012.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. pages 1–14, 2015.

[Song *et al.*, 2015] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. *CVPR*, pages 567–576, 2015.

[Szegedy *et al.*,] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning.

[Wei *et al.*, 2016] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Shuicheng Yan, and Yao Zhao. HCP: A Flexible CNN Framework for Multi-Label Image Classification. *TPAMI*, 38(2):1901–1907, 2016.

[Wu *et al.*, 2016] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Bridging Category-level and Instance-level Semantic Image Segmentation. 2016.

[Xue *et al.*, 2011] Xiangyang Xue, Wei Zhang, Jie Zhang, Bin Wu, Jianping Fan, and Yao Lu. Correlative multi-label multi-instance image annotation. *ICCV*, pages 651–658, 2011.

[Yadollahpour *et al.*, 2013] Payman Yadollahpour, Dhruv Batra, and Gregory Shakhnarovich. Discriminative re-ranking of diverse segmentations. *CVPR*, pages 1923–1930, 2013.

[Zhou *et al.*, 2016] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic Understanding of Scenes through the ADE20K Dataset. *arXiv*, 2016.