Dual Super-Resolution Learning for Semantic Segmentation

Li Wang[1, *], Dong Li[1], Yousong Zhu[2], Lu Tian[1], Yi Shan[1]
[1] Xilinx Inc., Beijing, China.
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China.
{liwa, dongl, lutian, yishan}@xilinx.com, yousong.zhu@nlpr.ia.ac.cn

Abstract

Current state-of-the-art semantic segmentation methods often apply high-resolution input to attain high performance, which brings large computation budgets and limits their applications on resource-constrained devices. In this paper, we propose a simple and flexible two-stream framework named Dual Super-Resolution Learning (DSRL) to effectively improve the segmentation accuracy without introducing extra computation costs. Specifically, the proposed method consists of three parts: Semantic Segmentation Super-Resolution (SSSR), Single Image Super-Resolution (SISR) and Feature Affinity (FA) module, which can keep high-resolution representations with low-resolution input while simultaneously reducing the model computation complexity. Moreover, it can be easily generalized to other tasks, e.g., human pose estimation. This simple yet effective method leads to strong representations and is evidenced by promising performance on both semantic segmentation and human pose estimation. Specifically, for semantic segmentation on CityScapes, we can achieve $\geq$2% higher mIoU with similar FLOPs, and keep the performance with 70% FLOPs. For human pose estimation, we can gain $\geq$2% mAP with the same FLOPs and maintain mAP with 30% *fewer FLOPs. Code and models are available at https://github.com/wanglixilinx/DSRL.*

1. Introduction

Semantic segmentation is a fundamental task for scene understanding, which aims to assign dense labels for all pixels in the image. It has several potential applications in the fields of autonomous driving, robot sensing and so on. For most such applications, it is a challenge to simultaneously keep efficient inference speed and impressive performance, especially on mobile devices with limited resources.

Owing to the development of deep learning, semantic segmentation has also achieved significant improvements, in which high-resolution deep feature representation plays an essential role in achieving promising performance. Currently, there are two main lines to keep high-resolution representation. One is to explicitly maintain the high-resolution representations by using atrous convolution to replace strided convolution, such as DeepLabs [2, 3, 4]. The other one is to hallucinate higher-resolution feature maps by combining a top-down pathway and lateral connections, such as encoder-decoder frameworks like UNet [27]. However, these approaches often involve expensive computation cost. Besides, they usually take the original high-resolution image as input, which further increases the amount of calculation. Recently, compact segmentation networks have also attracted much attention due to their application advantages in resource-constrained devices. Nevertheless, their performance is far inferior to the state-of-the-art methods. In order to narrow the accuracy gap, these methods are often combined with a high-resolution input (e.g., 1024×2048 or 512×1024), which also brings notable computation costs. Once restricting the input size, regardless of large networks or compact networks, their performance degrades consider-

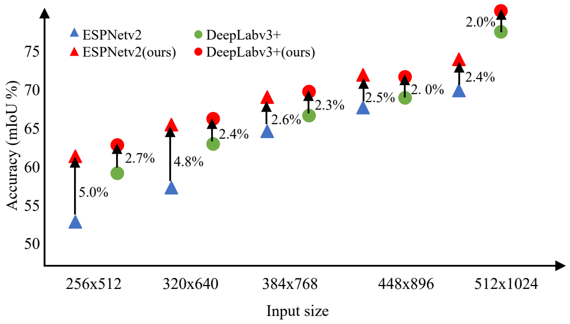


Figure 1. Accuracy vs. Input size for different networks on CityScapes validation set. Green points denote results for DeepLabv3+ with different input size: 256×512, 320×640, 384×768, 448×896, 512×1024 and 1024×2048, and blue triangles mark the results for ESPNetv2. Red ones represent the results with our method based on DeepLabv3+ and ESPNetv2, respectively.

[*] Corresponding author

ably. Figure 1 shows the performance of two representative segmentation networks: ESPNetv2 [24] and DeepLabv3+ [4] with various input resolutions. We can observe that when the input resolution decreases from 512×1024 to 256×512, the accuracy of both networks is reduced by more than 10%.

Therefore, in this paper, we design a clear and simple framework to alleviate this dilemma. Specifically, motivated by the image super-resolution, which aims to reconstruct a high-resolution image with a low-resolution input, we propose a novel Dual Super-Resolution Learning (DSRL) paradigm to keep high-resolution representation. Such a learning method is unified in a two-stream framework, which consists of Semantic Segmentation Super-Resolution (SSSR), Single Image Super-Resolution (SISR) and a Feature Affinity (FA) module. More specifically, we integrate the idea of super-resolution into existing semantic segmentation pipelines, thus formulating a Semantic Segmentation Super-Resolution (SSSR) stream. Then, the high-resolution features of the SSSR stream are further enhanced by the fine-grained structural representation from the SISR stream with Feature Affinity (FA) module. Moreover, these two streams share the same feature extractor, and the SISR branch is optimized with reconstruction supervision during training, and it will be freely removed from the network in the inference stage, thus causing cost-free overhead. We note that the proposed method can easily achieve a higher mIoU with similar FLOPs, and keep the performance with less FLOPs. As shown in Figure 1, our proposed DSRL can significantly improve the accuracy under different resolutions, especially for low-resolutions, thus can significantly reduce the computation cost with comparable performance. Compared to ESPNetv2 with an input size of 320×640, our method using a lower-resolution input image of 256×512 can gain 2.4% of mIoU and reduce 36% FLOPs at the same time. Last but not least, our framework can be easily extended to other tasks requiring high-resolution representation, like human pose estimation. Extensive experiments demonstrate the effectiveness and efficiency of the proposed method on both two challenging datasets, e.g., CityScapes [5] for semantic segmentation and MS COCO [19] for human pose estimation.

In summary, our main contributions include:

(1) We propose a dual super-resolution learning framework to keep high-resolution representation, which can improve the performance while keeping the inference speed;

(2) We validate the generality of the DSRL framework, which can be readily extended to other tasks requiring high-resolution representation, like human pose estimation.

(3) We demonstrate the effectiveness of our method both on semantic segmentation and human pose estimation. With a similar computation budget, we can improve $\geq 2\%$ accuracy, while reducing FLOPs with comparable performance.

2. Related Work

Semantic Segmentation. Semantic segmentation is a dense image prediction task, which plays a key role in high-level scene understanding. Driven by the rapid development of convolutional neural networks (CNNs), various works, FCN [21], DeepLabs [2, 3, 4], PSPNet [38] always adopt sophisticated feature extraction networks (e.g., ResNets [12] and DenseNets [13]) to learn discriminative feature representations for dense prediction. Besides, existing methods also develop critical strategies to further improve the performance, including atrous convolution [2, 3, 4], pyramid pooling module [38], attention mechanism [14], context encoding [37] and so on. However, these methods always involve expensive computation which limits their applications on resource-constrained devices.

Meanwhile, designing lightweight semantic segmentation models attracts much attention from the community. Most works focus on lightweight networks design by accelerating the convolution operations with factorization techniques. ESPNets [23, 24] exploit split-merge or reduce-expand principles to accelerate the convolution computation. Some others adopt efficient classification networks (e.g., MobileNet [28] and ShuffleNet [22]) or some compress techniques (e.g., pruning [9] and vector quantization [34]) to accelerate segmentation. In addition, [20] exploits knowledge distillation to help the training of compact networks by making use of large networks. However, their performance is far inferior to the state-of-the-art models.

Different from previous methods, we exploit the high-resolution features of single image super-resolution to guide the correlation learning of spatial dimension, thus benefiting the task of semantic segmentation. Our method can improve the performance significantly while keeping similar FLOPs.

Single Image Super-Resolution. SISR refers to the process of recovering high-resolution images from low-resolution images. Deep learning-based SISR methods have been widely proposed and achieve state-of-the-art performance on various benchmarks. Recently, there are four main kinds of supervised image super-resolution methods. (1) Pre-upsampling SR [7, 6] applies a traditional upsampling operation (e.g. bilinear or bicubic) to obtain a high-resolution image, and then refine it using a deep learning convolution network. This framework needs higher computation cost since most operations are performed in the high-dimensional space. (2) Post-upsampling SR [29, 16, 32] replaces the predefined upsampling operations with end-to-end learnable upsampling layers integrated at the end of the models, which can greatly reduce the computational complexity. (3) Progressive-upsampling SR [16, 15, 33] is introduced based on post-upsampling SR, it aims to reduce the learning difficulty by gradually reconstructing high-resolution images, and can cope with the need for multi-scale SISR. (4) Iterative up-and-down SR [13, 10, 31] ex-

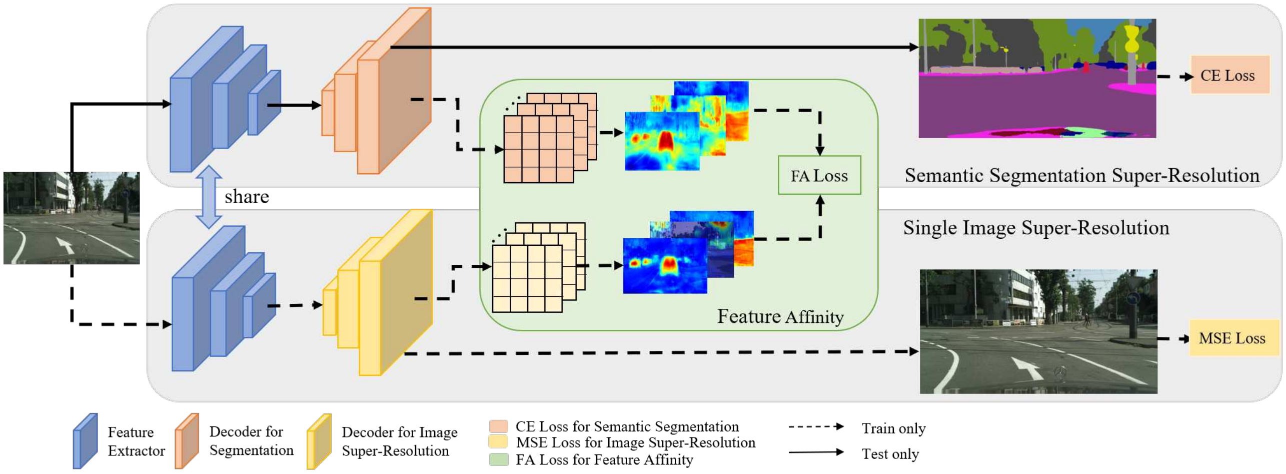

share
FA Loss
Feature Affinity
CE Loss
Semantic Segmentation Super-Resolution
Single Image Super-Resolution
MSE Loss
Feature Extractor
Decoder for Segmentation
Decoder for Image Super-Resolution
CE Loss for Semantic Segmentation
MSE Loss for Image Super-Resolution
FA Loss for Feature Affinity
Train only
Test only

Figure 2. The overview of the proposed DSRL framework, which includes three parts: Semantic Segmentation Super-Resolution (SSSR) branch, Single Image Super-Resolution (SISR) branch, and Feature Affinity (FA) module. The encoder is shared between the SSSR branch and the SISR branch. The architecture will be optimized with three terms: MSE loss for SISR branch, FA loss and a task-specific loss, e.g., the Cross-Entropy loss for semantic segmentation.

ploits iterative upsampling and downsampling layers to generate intermediate images and then combine them to reconstruct final high-resolution images. In consideration of high-quality results and low computation cost, we follow the spirit of the post-upsampling SR methods for our SISR branch in this work.

Multi-task Learning. Multi-task learning is generally used with CNNs to model related tasks jointly, e.g., pose estimation and action recognition [8], object detection and instance segmentation [11]. These methods usually treat multiple tasks equally both during training and testing phases. However, different from those methods with cross-task module design, we treat segmentation as the main task and SISR as an auxiliary task, where the image super-resolution branch will be removed during the inference stage and no extra computation cost is introduced.

3. Proposed Approach

In this section, we first review the most popular Encoder-Decoder architecture for semantic segmentation. We then present the proposed DSRL framework in detail and finally introduce the optimization function briefly.

3.1. Review of Encoder-Decoder Framework

We begin by briefly reviewing the traditional encoder-decoder architecture for semantic segmentation. As we know, the Encoder employs deep convolutional neural networks to extract hierarchical features with a scaling step of 2. Here, we denote output stride (OS) as the ratio of input image spatial resolution to the Encoder output resolution. In order to ensure high performance, OS usually equals 16 (or 8) by replacing the last one (or two) strided convolution block(s) with atrous convolutions correspondingly. Based on the down-sampled feature maps, the Decoder either directly exploits a bilinear upsampling layer with a scaling factor of OS (e.g., PSPNet [38]) or a simple designed sub-network (e.g., two consecutive upsampling layers in DeepLabv3+ [4]) to refine the segmentation results. However, most of existing methods can only upsample the feature maps to the same size as the input image for prediction, which might be smaller than the original image, e.g., sub-sampling the original image of 1024×2048 to 512×1024 as the network input in CityScapes, thus the ground truth needs to be down-sampled for supervision. On the one hand, this may result in the loss of effective label information. On the other hand, it is difficult to recover the original details only relying on the Decoder, which restricts the performance improvement.

3.2. Dual Super-Resolution Learning

To alleviate the above dilemmas, we propose a simple and effective framework, named as Dual Super-Resolution Learning (DSRL), to effectively improve the performance without computation and memory overload, especially with a low-resolution input. As shown in Figure 2, our architecture consists of three parts: (a) Semantic Segmentation Super-Resolution (SSSR); (b) Single Image Super-Resolution (SISR), and (c) Feature Affinity (FA) module.

Semantic Segmentation Super-Resolution. For the semantic segmentation, we simply append an extra upsampling module to produce the final prediction mask, the whole process named as Semantic Segmentation Super-Resolution (SSSR), e.g., as shown in Figure 4 (a), taking an input of 512×1024, we will generate an output of 1024×2048, which is 2× than the input image. Compared with most recent systems, which predict a 512×1024 mask for training and testing (then re-scaling to 1024×2048 in the post-processing stage), our method can make full use of the ground-truth and avoid the effective label information loss caused by pre-processing. Our extra semantic segmentation upsampling module consists of a stack of deconvolution layers, followed by BatchNorm and ReLU layers and it requires only a fewer parameters.

Single Image Super-Resolution. As discussed above, only relying on the decoder module is not enough to recover analogous high-resolution semantic feature representation obtained by using the original image as the input. Because the decoder is either a bilinear upsampling layer or a simple sub-network, it will not bring any additional information since the input is in a low-resolution (e.g., 512×1024). SISR aims to build a high-resolution image from the low-resolution input. This means that SISR can effectively reconstruct fine-grained structure information of the image under the low-resolution input, which is always helpful for semantic segmentation. To show a better understanding, we visualize the features of SSSR and SISR in Figure 3. By comparing (b) and (c) in Figure 3, we can easily find that the SISR contains more complete structures of objects. Although these structures do not explicitly imply the categories, they can be effectively grouped by the relationships between pixel and pixel or region and region. As we know, the relationships can implicitly deliver semantic information, thus benefiting the task of semantic segmentation. Therefore, we apply the high-resolution features recovered from SISR to guide the learning of high-resolution representation of SSSR, and these details can be modeled by the correlation or relationships between internal pixels. The relationship learning can make up for the simple design of the decoder. For the SISR branch, it shares the feature extractor with SSSR, as shown in Figure 4 (b), and we follow the design of [29] to reduce the computation and generate high-quality results. The whole branch is trained with the supervision of the original image and will be freely removed during the inference phase.

Feature Affinity Learning. Since SISR contains more complete structure information than SSSR, we introduce feature affinity learning to guide SSSR to learn high-resolution representation. As shown in Equation 1, FA aims to learn the distance of similarity matrix between SISR and SSSR branch, where the similarity matrix, as shown in Equation 2, mainly describes the pairwise relationship between pixels. Specifically, for a $W' \times H' \times C'$ feature map F, where $W' \times H'$ means the spatial dimension, we formulate the relationship between every two pixels, so the relation graph contains $W'H' \times W'H'$ connections, and S_{ij} denotes the relationship between the ith and the jth pixels

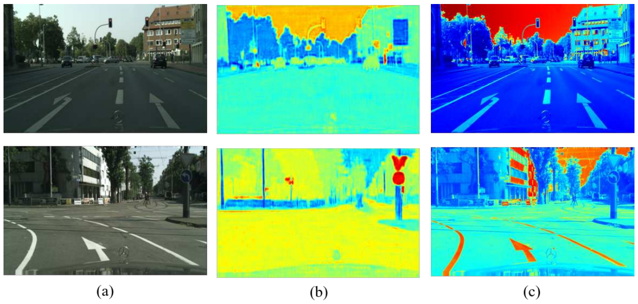
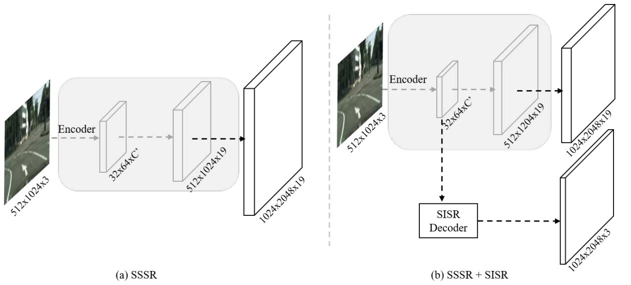


Figure 3. Feature-level visualization for SSSR and SISR under the same input (0.5×). (a) Input image, (b) SSSR feature visualization, and (c) SISR feature visualization.



(a) SSSR

(b) SSSR + SISR

Figure 4. (a) Semantic Segmentation Super-Resolution (SSSR) branch; (b) extends (a) with a Single Image Super-Resolution (SISR) branch. 'Encoder' denotes the shared feature extractor.

on feature map F. Theoretically, it's better to compute the affinity of every pair of pixels, but due to the high memory overheads, we subsample the the pairs of pixels to its $1/8$ in practice. Besides, in order to reduce the training instability caused by the discrepancy of feature distribution between SISR and SSSR branch, we append a feature transform module on the feature map of the SSSR branch before applying the FA loss, which consists of a 1×1 convolution layer followed by BatchNorm and ReLU layers.

$$L_{fa} = \frac{1}{W'^2 H'^2} \sum_{i=1}^{W'H'} \sum_{j=1}^{W'H'} ||S_{ij}^{seg} - S_{ij}^{sr}||_q \tag{1}$$

and

$$S_{ij} = (\frac{F_i}{||F_i||_p})^T \cdot (\frac{F_j}{||F_j||_p}) \tag{2}$$

S^{seg} and S^{sr} refer the semantic segmentation similarity matrix and SISR similarity matrix, respectively. p and q denote the norms used to normalize the features for stability. Here we set $p = 2$ and $q = 1$.

3.3. Optimization

The whole objective function, as shown in Equation 3, consists of a conventional multi-class Cross-Entropy loss L_{ce} in Equation 4 for semantic segmentation, a Mean Squared Error (MSE) loss L_{mse} in Equation 5 for SISR, and the structured relation term L_{fa} given in Equation 1.

$$L = L_{ce} + w_1 L_{mse} + w_2 L_{fa} \tag{3}$$

and

$$L_{ce} = \frac{1}{N} \sum_{i=1}^{N} -y_i log(p_i) \tag{4}$$

$$L_{mse} = \frac{1}{N} \sum_{i=1}^{N} ||SISR(X_i) - Y_i||^2 \tag{5}$$

where $SISR(\cdot)$ and Y refer the super-resolution output and its corresponding ground truth, p_i and y_i refer the segmentation predicted probability and the corresponding category for pixel i, N means the pixel number. w_1 and w_2 are set as 0.1 and 1.0, making these loss value ranges comparable. We minimize the whole objective function end-to-end.

4. Experiments for Semantic Segmentation

4.1. Datasets and Evaluation Metrics

The CityScapes dataset [5] focuses on urban visual scene understanding which consists of 2,975 training, 500 validation and 1525 test images with fine-grained annotations. It was captured across 50 cities in different seasons. The task is to segment an image into 19 classes. All images are in a resolution of 1024×2048. We perform detailed comparison experiments on its validation set and report the final results on the test set using the Online Server.

The CamVid dataset [1] is another automotive dataset. There are 11 valid different classes. The original frame resolution for this dataset is 960×720. We evaluate the performance on the validation and test set over 11 classes.

We use the common metric of mean Intersection over Union (mIoU) for semantic segmentation on both datasets. We also report the FLOPs of segmentation models to compare their computation cost.

4.2. Implementation Details

Networks. We conduct ablation studies on two representative segmentation architecture: ESPNetv2 [24] and DeepLabv3+ [4]. We also verify the effectiveness of our method on some other structures, such as PSPNet [38] and lightweight models: BiseNet [36] and DABNet [17].

Training setup. Our method is implemented in PyTorch. All the segmentation networks in this paper are trained by mini-batch stochastic gradient descent (SGD) with the momentum (0.9) and the weight decay (0.0005). The learning rate is initialized as 0.01, and we apply the poly learning rate strategy with power 0.9. Other than this, we follow the settings in the corresponding publications to reproduce the results of all the networks.

4.3. Ablation Study

Effect of algorithmic components. We first investigate the proposed method of our dual super-resolution learning system. The experiments are conducted on ESPNetv2 and DeepLabv3+ (with ResNet101 as the backbone) represent the compact and the large network, respectively, and we evaluate the mean Intersection over Union (mIoU) on the CityScapes validation set. Here, we resize the image to 256×512 as the input for accelerating the training of experiments. As shown in Table 1, taking ESPNetv2 as an example, we can see that the SSSR learning can improve the performance from 54.5% to 55.7% since it reduces the scaling times of ground truth. By adding the SISR branch, the mIoU can be effectively improved by 2.4%. While combining with the FA loss, the performance can be further improved to 59.5%(5.0% higher than the baseline), thus indicating that transferring the structure information between SISR and SSSR is necessary. The results on DeepLabv3+ can also draw the same conclusion, which consistently demonstrates the effectiveness of the proposed DSRL.

In order to better understand the DSRL, we also visualize the final segmentation features between the baseline ESPNetv2 and our DSRL. As shown in Figure 5 (c), our method can significantly enhance the sharpness of the boundary and improve the completeness of different categories, e.g., road, car and so on, thus undoubtedly strengthening the final discrimination ability of the model.

Method	Input size	Output size	Val. (%)
ESPNetv2 [24]	256×512	256×512	54.5
+ SSSR	256×512	512×1024	55.7
+ SSSR + SISR	256×512	512×1024	56.9
+ SSSR + SISR + FA	256×512	512×1024	**59.5**
DeepLabv3+ [4]	256×512	256×512	56.5
+ SSSR	256×512	512×1024	57.1
+ SSSR + SISR	256×512	512×1024	57.4
+ SSSR + SISR + FA	256×512	512×1024	**59.2**

Table 1. The effect of different components in the proposed method with an input in 256 × 512. + SSSR: adding the SSSR learning method. + SISR: adding the SISR learning method. + FA: adding the FA module between SSSR and SISR.

Effect of various input resolutions. We also compare the adaptability of the proposed method under different input resolutions, which ranges from 256×512 to 512×1024. As shown in Table 2, which proves the generality of our

Methods	256×512	320×640	384×768	448×896	512×1024
ESPNetv2 [24]	54.5	57.1	61.4	63.2	64.5
ESPNetv2(ours)	59.5	61.9	64.0	65.7	66.9
$\Delta mIoU(\%)$	**+5.0**	**+4.8**	**+2.6**	**+2.5**	**+2.4**
FLOPs(G)	1.35	2.11	3.04	4.13	5.40
DeepLabv3+ [4]	56.5	59.3	62.0	63.7	70.0
DeepLabV3+(ours)	59.2	61.7	64.3	65.7	72.0
$\Delta mIoU(\%)$	**+2.7**	**+2.4**	**+2.3**	**+2.0**	**+2.0**
FLOPs(G)	243.57	380.58	548.06	745.94	974.30

Table 2. Performance comparison with our DSRL on the CityScapes validation set with various input resolutions: 256×512, 320×640, 384×768, 448×896 and 512×1024, and we keep the same aspect ratio of height : width = 1 : 2. $\Delta mIoU$ refers the improved mIoU with the same input resolution.

method. With our method, two networks (ESPNetv2 and DeepLabv3+) outperform the accuracy of its corresponding baseline with the same input size. What's more, compared with a larger input resolution e.g., 448×896, our method with 384×768 can reduce 26% FLOPs and even improve the mIoU by 0.8% and 0.6% for ESPNetv2 and DeepLabv3+, respectively. Thus, our method can improve accuracy with the same computation budget, while reducing FLOPs with comparable performance, which consistently demonstrates the efficiency of the proposed DSRL.

4.4. Results on CityScapes

To further verify the effectiveness of our method, we apply our dual super-resolution learning method to some other architectures, e.g., PSPNet with ResNet101 as the backbone and several compact networks designed for real-time application, e.g., DABNet and BiseNet which is based on a ResNet18. Table 3 presents the segmentation accuracy and model complexity, respectively. We use FLOPs to evaluate the network complexity which is calculated at the resolution 512×1024. We can see that our approach can improve the results over different complexity networks: ESPNetv2, BiseNet, DABNet, PSPNet, and DeepLabv3+. For the network with a naive decoder such as PSPNet, the improvement is significant with 4.3% on the test set compared with the baseline of 69.3%. Compared to the distillation method [20], we provide another effective pipeline to achieve higher performance with the same FLOPs.

Figure 6 shows the IoU score for each class over DeepLabv3+. Our DSRL scheme improves the performance significantly, especially for the small categories with low IoU scores, e.g., 10.44% improvement for pole and 10.03% for rider. The qualitative segmentation results in Figure 8 demonstrate the effectiveness of our method, especially for structured objects, such as car, person and so on.

4.5. Results on CamVid

Table 4 shows the performance of ESPNetv2, BiseNet and DeepLabv3+ with and without the proposed DSRL schemes on CamVid. We train and evaluate the networks at the resolution 368×480 [1]. We can see that our method de-

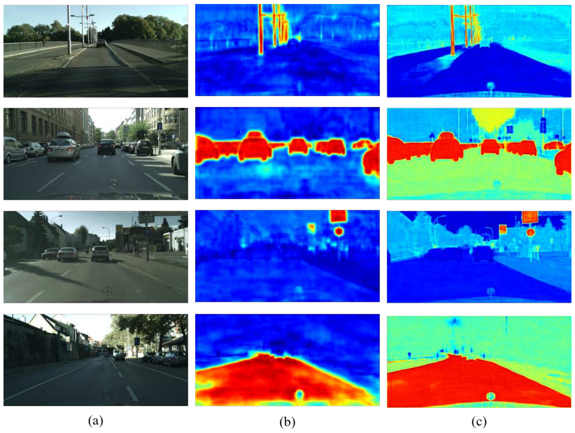


Figure 5. The visualization of segmentation features (better visualized in color). (a) input image. (b) the final segmentation features of baseline [24] method. (c) the final segmentation features of our DSRL.

Method	Val. (%)	Test (%)	GFLOPs
Current state-of-the-art results			
ENet [25]	-	58.3	7.24
ESPNet [23]	-	60.3	8.86
ERFNet [26]	-	68.0	25.60
PSPNet(ResNet18(0.5))[20]	-	54.1	133.40
PSPNet(ResNet18(0.5))[20]†	-	60.5	133.40
PSPNet(ResNet18(1.0))[20]	-	67.6	512.80
PSPNet(ResNet18(1.0))[20] †	-	71.4	512.80
FCN [21]	-	65.3	1335.60
RefineNet [18]	-	73.6	2102.80
Results w/o and w/ DSRL scheme			
ESPNetv2 [24]	64.5	65.1	5.40
ESPNetv2(ours)	**66.9**	**65.9**	5.40
DABNet [17]	62.6	65.0	20.44
DABNet(ours)	**65.4**	**66.2**	20.44
BiseNet [36]	62.6	61.8	49.20
BiseNet(ours)	**66.8**	**64.9**	49.20
DeepLabv3+ [4]	70.0	67.1	974.30
DeepLabv3+(ours)	72.0	69.3	974.30
DeepLabv3+(ours)‡	**73.4**	**71.8**	8464.23
PSPNet(ResNet101) [38]	71.5	69.1	287.48
PSPNet(ResNet101)(ours)	74.4	73.4	287.48
PSPNet(ResNet101)(ours)‡	**75.7**	**74.8**	3575.54

Table 3. The segmentation results comparison with other state-of-the-art methods on the CityScapes validation (Val.) and test (Test) set. We report GFLOPs at the same image resolution used for computing the accuracy. †: refers to with knowledge distillation method in [20]. ‡: refers the network is tested on multiple scales.

[1] Both ESPNetv2 and DeepLabv3+ require the input resolution to be a multiple of 16

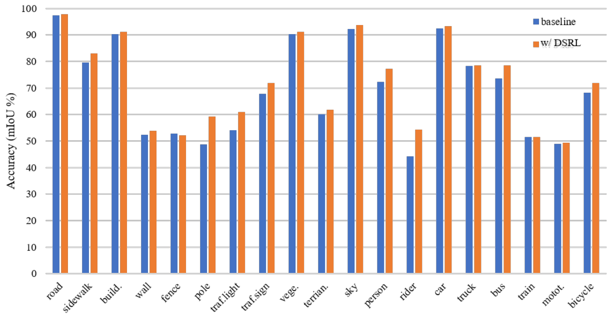
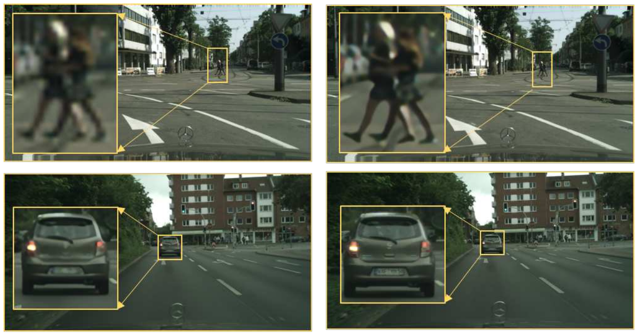


Figure 6. Illustrations of the effectiveness of DSRL method in terms of class IoU scores on the network DeepLabv3+ over the CityScapes validation set. The performance is improved significantly, especially for the hard classes with low IoU scores, such as pole and traffic light.



(a) (b)

Figure 7. Qualitative SISR results comparison on the CityScapes validation. (a) low-resolution images. (b) the SISR results.

livers a competitive performance. Under the same computational constraints, our method consistently outperforms the baseline methods by a large margin. Notably, our method based on ESPNetv2 provides 3.1% improvement compared with ENet [25], with only 50% FLOPs. The results demonstrate the effectiveness of the proposed DSRL method.

4.6. Results for Single Image Super-Resolution

We also present the results for image super-resolution with a scale factor of 2×. For SISR, Peak Single-to-Noise Ratio (PSNR) [35] is used to measure the image reconstruction quality, and Structural Similarity Index (SSIM) is proposed for measuring the structural similarity between output image and ground truth [35]. Our method achieves 0.35/0.78 for PSNR and SSIM on the CityScapes validation set, respectively. Figure 7 shows examples of our image super-resolution results.

Method	Test (%)	GFLOPs
ENet [25]	51.3	3.61
ESPNet [25]	57.8	3.09
ESPNet [20]	61.4	3.09
ESPNet-C [25]	56.7	3.09
ESPNet-C [20]	60.3	3.09
ESPNetv2 [24]	50.9	1.82
ESPNetv2 (ours)	**54.4**	1.82
BiseNet [36]	53.4	4.14
BiseNet (ours)	**57.0**	4.14
DeepLabv3+ [4]	60.4	326.13
DeepLabv3+ (ours)	**63.7**	326.13

Table 4. Segmentation results on the CamVid test set.

5. DSRL for Human Pose Estimation

Our framework can be readily extended to human pose estimation, which further exhibits the generality of our method. Human pose estimation is another challenging computation vision task, in which a high-resolution representation is also required for keypoint localization accuracy.

Dataset. The MS COCO dataset [19] contains over 200K images and 250K person instances labeled with 17 keypoints. We train our model on MS COCO train2017 dataset, including 57K images and 150K person instances, and do evaluation on the val2017 set containing 5K images.

Evaluation metric. The standard evaluation metric is based on Object Keypoint Similarity (OKS) [30]. We report standard average precision and recall scores: AP@0.5 (AP at OKS = 0.50), AP@0.75 (AP at OKS = 0.75), mAP (the mean of AP scores at 10 positions, OKS = 0.50, 0.55,..., 0.90, 0.95), AP(M) for medium objects, AP(L) for large objects, AR@0.5 (AR at OKS = 0.50), and AR@0.75 (AR at OKS = 0.75). More details can be referred to [30].

Implementation details. Models are trained on train2017 and evaluated on the val2017. We follow the same training settings as in [30]: we extend the human detection box in height or width to a fixed aspect ratio $height : width = 4 : 3$, and then crop the box from the image, which is resized to various fixed sizes, 256×192, 160×128, and 128×96.

Architecture. We adopt a two-stage top-down paradigm, HRNet-w32 [30] to verify our proposed method, which uses the offline person detection results to predict the human keypoint. HRNet is one of the state-of-the-art architectures for accurate human pose estimation. The network input is a person instance that is resized to a fixed size, e.g., 256×192, and the keypoint head produces a heatmap with the resolution of 64×48. Following the same design for semantic segmentation, we append an extra upsampling model at the end of the existing network to predict the keypoint, meanwhile, the SISR branch shares the same feature extractor. The total objective function is a weighted sum of three

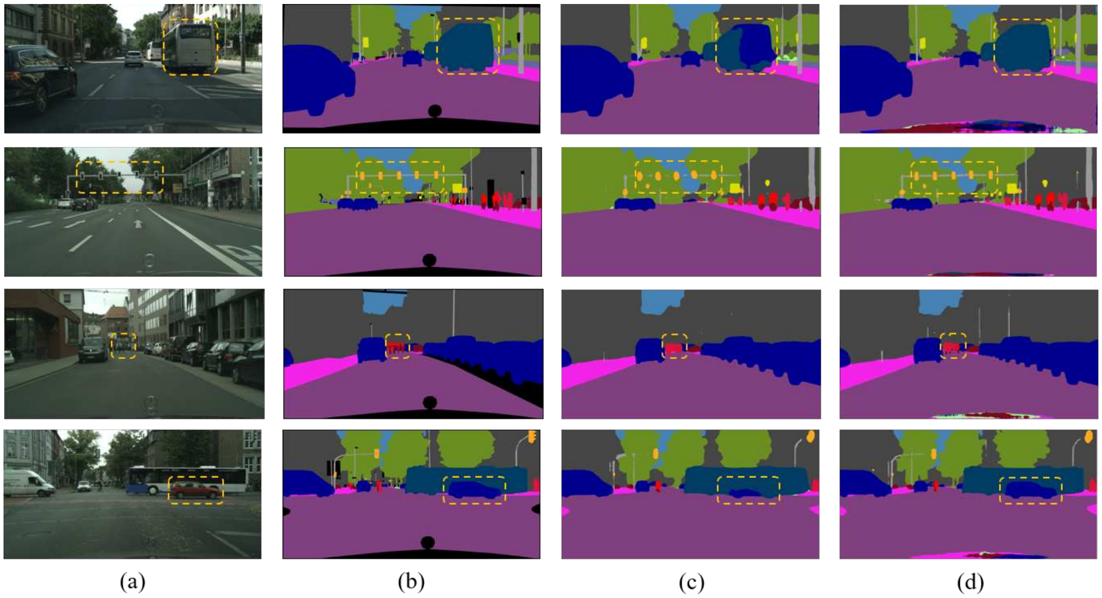


Figure 8. Comparison of segmentation results. (a) Input image. (b) Ground truth. (c) The baseline results of DeepLabv3+ [4] with 512×1024. (d) Results of DeepLabv3+ with our DSRL method.

Method	Input size	mAP(%)	AP@0.5	AP@0.75	AP(M)	AP(L)	AR	AR@0.5	AR@0.75	GFLOPs
HRNet-w32 [30]	256×192	74.4	90.5	81.9	70.8	81.0	79.8	94.2	86.5	7.12
HRNet-w32(ours)	256×192	**75.6**	**92.2**	**83.0**	**72.1**	**82.8**	**81.2**	**93.8**	**88.5**	7.12
HRNet-w32 [30]	160×128	69.2	89.3	78.1	66.7	75.2	75.7	93.6	83.7	2.97
HRNet-w32(ours)	160×128	**71.5**	**89.6**	**79.4**	**68.6**	**77.5**	**77.5**	**93.7**	**84.5**	2.97
HRNet-w32 [30]	128×96	64.6	87.8	73.9	62.7	69.8	71.7	92.8	80.2	1.78
HRNet-w32(ours)	128×96	**67.9**	**88.3**	**76.7**	**65.6**	**73.5**	**74.5**	**92.8**	**82.4**	1.78

Table 5. Human pose estimation results with HRNet-w32 on the MS COCO2017 validation set.

items: keypoints regression loss, MSE loss, and the FA loss.

Results on the validation set. Table 5 summarizes the performance comparations between the baseline HRNet method and the proposed DSRL method. With different resolutions of person instances as input, our method consistently surpasses HRNet by 1.2% to 3.3%. The experimental results demonstrate the effectiveness and generality of our method. Figure 9 shows some keypoint prediction results with our method on the COCO validation set.

6. Conclusion

In this work, we propose a dual super-resolution learning framework for semantic segmentation. The semantic segmentation super-resolution branch helps learn higher-resolution representations for dense label prediction, the single image super-resolution branch can recover detailed structure information, and the feature affinity module is introduced to enhance the high-resolution representations of semantic segmentation through the detailed structural information. We demonstrate the effectiveness of our approach with several recently-developed networks, and it can be readily extended to other tasks like human pose estimation, which further exhibits the generality of our method.

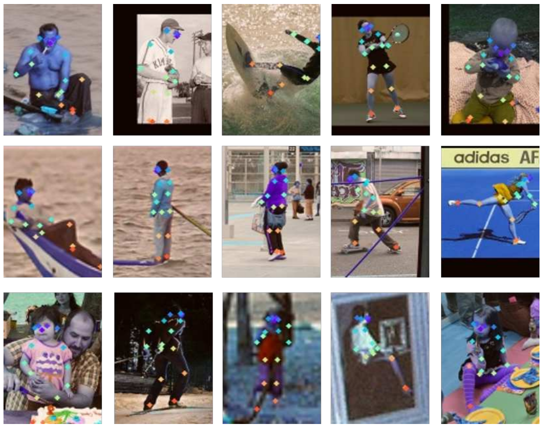


Figure 9. Qualitative results on the MS COCO 2017 validation set produced from HRNet-w32 with our proposed DSRL.

References

[1] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*. 5

[2] 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. 1, 2

[3] 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. *TPAMI*, 40(4):834–848, 2017. 1, 2

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1, 2, 3, 5, 6, 7, 8

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 5

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 38(2):295–307, 2015. 2

[7] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, 2016. 2

[8] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. R-cnns for pose estimation and action detection. In *CVPR*, 2014. 3

[9] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *ICLR*, 2016. 2

[10] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 2

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 2

[14] Haijie Tian Yong Li Yongjun Bao Zhiwei Fang and Hanqing Lu Jun Fu, Jing Liu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2

[15] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 2

[16] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *T-PAMI*, 2018. 2

[17] Gen Li, Inyoung Yun, Jonghyun Kim, and Joongkyu Kim. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. In *BMVC*, 2019. 5, 6

[18] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 6

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 7

[20] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *CVPR*, 2019. 2, 6, 7

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2, 6

[22] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018. 2

[23] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *ECCV*, 2018. 2, 6

[24] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In *CVPR*, 2019. 2, 5, 6, 7

[25] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. 6, 7

[26] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *TITS*, 19(1):263–272, 2017. 6

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 2

[29] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 2, 4

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. 2019. 7, 8

[31] Radu Timofte, Shuhang Gu, Jiqing Wu, and Luc Van Gool. Ntire 2018 challenge on single image super-resolution: Methods and results. In *CVPR*. 2

[32] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *ICCV*, 2017. 2

[33] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. A fully progressive approach to single-image super-resolution. In *CVPR*. 2

[34] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized convolutional neural networks for mobile devices. In *CVPR*, 2016. 2

[35] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *TMM*, 2019. 7

[36] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018. 5, 6, 7

[37] Yuan Yuhui and Wang Jingdong. Ocnet: Object context network for scene parsing. 2018. 2

[38] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2, 3, 5, 6