A General Two-Branch Decoder Architecture for Improving Encoder-Decoder Image Segmentation Models

Sijie Hu, Fabien Bonardi, Samia Bouchafa, Désiré Sidibé

A General Two-Branch Decoder Architecture for Improving Encoder-Decoder Image Segmentation Models

Sijie Hu[1], Fabien Bonardi[1], Samia bouchafa[1] and Désiré Sidibé[1]
[1]*University Paris-Saclay, Univ Evry, IBISC, 91020, Evry, France*
sijie.hu@universite-paris-saclay.fr

Keywords: multi-branch, encoder-decoder, complementary learning, supervised learning, semantic segmentation

Abstract: Recently, many methods with complex structures were proposed to address image parsing tasks such as image segmentation. These well-designed structures are hardly to be used flexibly and require a heavy footprint. This paper focuses on a popular semantic segmentation framework known as encoder-decoder, and points out a phenomenon that existing decoders do not fully integrate the information extracted by the encoder. To alleviate this issue, we propose a more general two-branch paradigm, composed of a main branch and an auxiliary branch, without increasing the number of parameters, and a boundary enhanced loss computation strategy to make two-branch decoders learn complementary information adaptively instead of explicitly indicating the specific learning element. In addition, one branch learns pixels that are difficult to resolve in another branch making a competition between them, which promotes the model to learn more efficiently. We evaluate our approach on two challenging image segmentation datasets and show its superior performance in different baseline models. We also perform an ablation study to tease apart the effects of different settings. Finally, we show our two-branch paradigm can achieve satisfactory results when remove the auxiliary branch in the inference stage, so that it can be applied to low-resource systems.

1 INTRODUCTION

Semantic segmentation can be formulated as the task of labeling all pixels in an image with semantic classes. Most state-of-the-art semantic segmentation models are based on the encoder-decoder architecture or its variants which can also be unified into two parts, namely encoder and decoder. Specifically, the encoder extract information from the original input, and the decoder integrate previously extracted information and recover semantic information from it. In recent years, researchers have been committed to exploring different technologies to learn a more general representation, such as using very large dataset (Sun et al., 2017; Deng et al., 2009) to train a model, or designing more complex model(Simonyan and Zisserman, 2014; He et al., 2016; Dosovitskiy et al., 2020). Then efficient feature extractors are selected as encoders and applied to the image segmentation task (Badrinarayanan et al., 2017; Zhao et al., 2017; Chen et al., 2018b; Wang et al., 2020). However, the more general representation extracted by the encoder, the more task dependency we need in the decoder to complete the efficient conversion between the feature representation and the specific task.

In order to improve the parsing ability of decoder to extracted features, DeeplabV3+ (Chen et al., 2018b) and PSPNet (Zhao et al., 2017) through pyramid pooling integrate the contextual information at multiple scales. U-Net (Ronneberger et al., 2015) and FCN (Long et al., 2015) use skip-connection to fuse feature maps of different layers. Some other models try to explore the interrelationships between features through attention mechanisms (Oktay et al., 2018; Li et al., 2019; Li et al., 2018). It is worth noting that some works have appeared recently to explore the two-branch structure in the decoder (Fu et al., 2019; Yuan et al., 2020). They capture meaningful information by carefully designing different branches. Unfortunately, existing two-branch structures were elaborately designed and are difficult to port to other types of decoders and the degradation of model performance caused by removing a branch is also unacceptable, or they were just designed for post-processing and are difficult to train end-to-end. On the other hand, with the continuous improvement of the encoder's representation ability, how to make full use of the information extracted by the encoder is still an open question. Therefore, we have reason to suspect that the existing encoder-decoder-based mod-

els do not fully integrate the information extracted by the encoder. And we verified this view through experiments.

To alleviate these problems, we propose a more general two-branch paradigm, composed of a main branch and an auxiliary branch for improving the structure of the decoder. At the same time, we design a simple yet efficient branch that can be flexibly integrated into existing encoder-decoder semantic segmentation systems to verify the effectiveness of the proposed two-branch structure. In order to enable two branches to learn complementary information, we customize a loss calculation method to supervise the learning process of each branch. With these ideas, different branches can learn complementary information adaptively instead of explicitly indicating the specific learning elements of different branches. In addition, learning complementary information can make the two branches compete with each other to a certain extent during the learning process which further improves the performance. To the best of our knowledge, the approach we proposed is the first to use the general two-branch paradigm to improve the analytical capabilities of the model. Moreover, compared with the counterpart of the original model, the ameliorated two-branch version reduces or maintains the number of parameters while improving performance.

Our main contributions can be summarized as follows:

- We propose a general two-branch paradigm to enhance the capability of decoder to parse the information extracted by the encoder without increasing the number of parameters.
- We propose the BECLoss that can supervise two-branch decoders to learn complementary information adaptively instead of explicitly indicating the specific learning elements to each branch.
- We design a simple yet efficient branch which can be flexibly integrated into existing encoder-decoder framework to form a two-branch structure.
- The ameliorated two-branch version outperform its original encoder-decoder counterpart by a large margin in Cityscapes dataset (Cordts et al., 2016) and Freiburg Forest dataset (Valada et al., 2016). Moreover, even if the auxiliary branches in the trained two-branch model are removed, our results are still far superior than the original encoder-decoder model.

2 RELATED WORK

Encoder-decoder and variants. As a general structural paradigm, encoder-decoder is widely used in the field of image segmentation. Such a structure usually extracts features from the input to a latent feature space by an encoder which utilized some popular classification networks as the backbone. Then, the spatial resolution is gradually restored while different tricks are employed to integrate the extracted features by a decoder. U-Net (Ronneberger et al., 2015) explored the potential relationship between the features of the encoding phase and their counterpart in the decoding phase through multiple skip-connections. SEMEDA (Chen et al., 2020) first learned to convert the label to an embedding space under the guidance of the boundary information, and then supervised the encoder-decoder structure under the learned subspace. For better learning the global context representation, multi-scale pyramid pooling and dilated convolution were adapted at different grid scales. PSPNet (Zhao et al., 2017) and Deeplab family (Chen et al., 2018b; Chen et al., 2017) introduced dilated convolution in encoder for increasing the receptive field while maintaining the resolution, then several parallel pyramid pooling were followed to integrate information at different scales. Inspired by (Chen et al., 2016; Hu et al., 2018; Woo et al., 2018), attention mechanism and its variants are adopted in encoders or decoders (Li et al., 2019; Chen et al., 2018a; Zhong et al., 2020) for improving performance. In (Li et al., 2018; Oktay et al., 2018), attention was deployed in the decoding stage for re-calibrating the feature maps with learnable weights. In addition, the application of self-attention or transformer (Vaswani et al., 2017) in encoder has gradually become popular due to its capability of encoding distant dependencies for better feature extraction. SETR (Zheng et al., 2020) adapted a pure transformer encoder to extract features from an image seen as a sequence of patches then followed a decoder to restore the semantic information.

Multi-branch. Learning different information through multiple parallel data streams has been proved to have more advantages for representation and generalization. Specifically, some existing works deployed multi-branch structure in the feature extraction stage (Wang et al., 2020; Tao et al., 2020; Takikawa et al., 2019), and other works deployed multi-branch to the prediction stage (Huang et al., 2017; Fu et al., 2019). HRNet (Wang et al., 2020) repeatedly exchanged the information across different resolutions by a series of parallel feature extraction streams in the encoding process to maintain high-resolution representations. Based on HRNet, (Tao

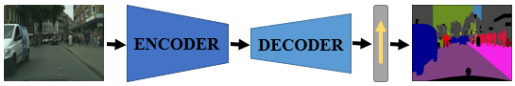

Figure 1: Encoder-Decoder paradigm.

et al., 2020) proposed a hierarchical multi-scale attention approach in which each data stream learned a certain image scale so that the model can consider the information of multiple input image scales when predicting. GSCNN (Takikawa et al., 2019) designed a two-stream structure, one for context information extraction, another one for boundary-related information extraction. Combined with attention, RAN (Huang et al., 2017) proposed a three-branch structure that performs the direct, and reverse-attention learning processes simultaneously. Similarly, DANet (Fu et al., 2019) used a two-branch encoder to learn the semantic relevance in spatial and channel feature spaces respectively. Unlike above works, SegFix (Yuan et al., 2020) proposed a post-processing scheme that predicted boundary and direction map by means of a two-branch decoder which were supervised by two boundary related losses.

Encouraged by multi-branch learning, we propose a more general and easy-to-deploy two-branch paradigm, in which a new branch can be easily inserted into the original decoder to form a two-branch decoder and as a result improve the discriminating ability. Different from previous works, we design a general paradigm and enable different branches to learn complementary information adaptively instead of explicitly indicating the specific learning elements of different branches.

3 METHODOLOGY

In this section, we first systematically describe the two-branch decoder paradigm, then we design a simple yet efficient branch that can be applied as a plug-in to existing encoder-decoder frameworks to turn them into our proposed two-branch architecture. Finally, we introduced a new loss calculation method that can be used to supervise branch learning complementary information.

3.1 Two-Branch Structure Prototype

In an image segmentation model, the encoder first converts the texture or color information into abstract high-dimensional embeddings through a series of non-linear transformations, then the decoder integrates different information and parses the high-level semantic information. (Badrinarayanan et al., 2017;

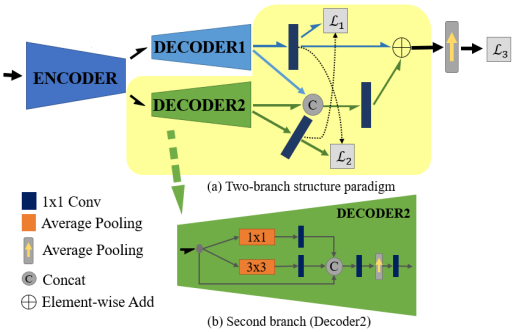

Figure 2: Overview of our proposed two-branch architecture. The output of the encoder is divided into two groups, which are represented by two 'half arrows'. Then each group is input to each branch separately and followed by a residual-liked module to fuse the outputs of two branches.

Chen et al., 2018b) are typical encoder-decoder networks. Structurally, existing encoder-decoder architectures can be simply represented by the diagram shown in Figure 1. Our proposed encoder-decoder based two-branch variant is depicted in Figure 2. As shown in Figure 2 (a) raw data is first input into the encoder for feature extraction, then encoded features are input to two branches separately, followed by a novel residual-liked (He et al., 2016) module to adaptively integrate information from different branches. For the fusion of two branch features, we use the output of the penultimate layer of each decoder instead of the last layer to retain more information. Specifically, in the residual path, we first concatenate the output features of two branches, next follow a 1×1 convolution to reduce the channels, and then features are combined with the output of the first branch by an element-wise addition operation. The final output is up-sampled to recover resolution if needed. We hope the ameliorated two-branch model can maintain the original encoder-decoder information stream through proposed residual-liked combination, while making additional branch more effective in providing supplementary information. Intuitively, this approach can easily keep the performance of the original encoder-decoder model.

3.2 Additional Branch Setting

In this part we design a simple branch which can be deployed into encoder-decoder framework to form a two-branch decoder architecture. As shown in Figure 2 (b), the branch takes the encoded features as input. Similarly to (Zhao et al., 2017), we utilize a parallel average pooling module, each path consisting of an average pooling operator and a 1×1 convolution operator. We concatenate the output of each path to get a multi-scale feature representation. Features are then processed by another 1×1 convolution. In the

end, we get the output of this branch through an upsampling operation and a 1×1 convolution operation. In addition, inspired by grouped convolution (Howard et al., 2017), we divide the encoded features into two groups along the channel axis and each grouped features is entered into a specific branch. We experimentally found that in this way we can greatly reduce the number of parameters in the decoder while maintaining the performance of the two-branch structure.

3.3 BECLoss

In supervised learning, loss function plays a crucial role in the optimization of the network. We further propose a novel loss computation strategy which can optimize this two-branch structure in an efficient way. Moreover, (Chen et al., 2020; Takikawa et al., 2019) have proved that introducing boundary information in the loss helps to improve the inherent sensitivity of the network to boundary pixels. Thus, we believe that introducing boundary information in the proposed loss can also help the model learns boundary features during the training stage. This is verified in ablation experiments. We name this well-designed loss BECLoss. Specifically, BECLoss takes three inputs, namely outputs of the first branch X^1 and the second branch X^2 and ground-truth map GT. And for simplicity, we assume batch size as 1, thus the shape of $X^k (k = 1,2)$ is $C \times H \times W$ and C, H and W indicate the number of predicted classes, high and width of input images respectively. First, we get the probability distribution $S^k \in \mathbb{R}^{H \cdot W \times C}$ which can be computed as:

$$S_i^k = \frac{exp(X_i^k)}{\sum_j^C exp(X_i^k[j])} \tag{1}$$

where $i = 0 \ldots H \times W - 1$ denotes the index of pixels, $j = 0 \ldots C - 1$ denotes the index of channels. Then, we compute the probability map of ground truth label $P^k \in \mathbb{R}^{H \cdot W \times 1}$ as:

$$p_i^k = S_i^k[gt_i] \tag{2}$$

where gt_i is the i^{th} pixel in GT. Following, we define a mask M^1 for indicating all the pixels whose probability in P^1 is less than a threshold τ. M^1 indicates the pixels that are difficult to predict in the first branch. With the computed M^1 and P^2, we filter out all pixels in X^2 whose probability is less than a threshold τ:

$$M_i^1 = \begin{cases} 1 & \text{if } P_i^1 < \tau \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $i = 0 \ldots H \times W - 1$ denotes the index of pixels.

In order to standardize the loss definition, we use $\mathcal{L}^1$ to indicate the boundary enhanced loss computed from X^1, and $\mathcal{L}^2$ to indicate a partial loss that we get from X^2. In $\mathcal{L}^1$ and $\mathcal{L}^2$ we only consider the pixels which are hard to predict in the first branch in order to utilize the additional branch to assist in the prediction of these pixels. In addition, we use a hyperparameter γ to control the influence of boundary information $B \in \mathbb{R}^{W \times H}$ (detailed in 3.4) to the loss of the first branch, we get $\mathcal{L}^1 \in \mathbb{R}^{H \cdot W \times 1}$:

$$\mathcal{L}_i^1 = -log(P_i^1) \times (1 + \gamma \cdot B_i) \times M_i^1 \tag{4}$$

where $i = 0 \ldots H \times W - 1$ denotes the index of pixels. Following, we compute the partial loss $\mathcal{L}^2 \in \mathbb{R}^{H \cdot W \times 1}$:

$$\mathcal{L}_i^2 = -\log(P_i^2) \times M_i^1 \tag{5}$$

Finally, the BECLoss can be written as a weighted average sum of $\mathcal{L}^1$ and $\mathcal{L}^2$:

$$\mathcal{L}_{BEC} = \frac{\sum_i (\mathcal{L}_i^1 + \eta \cdot \mathcal{L}_i^2)}{\sum M_i^1} \tag{6}$$

where η is a hyperparameter used to control the ratio of $\mathcal{L}^2$ in $\mathcal{L}_{BEC}$.

The two branches can automatically learn complementary information which helps the proposed model to further learn a more appropriate way to combine the outputs of the two branches.

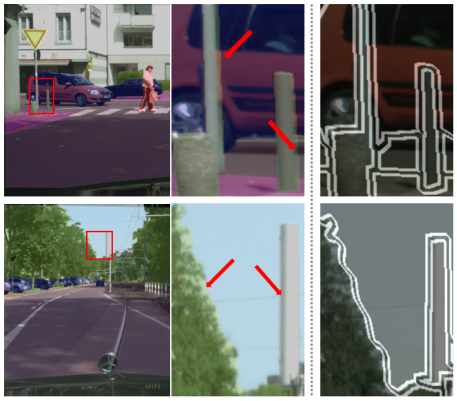

Figure 3: (a) Mis-labeled boundary pixels and (b) Extracted inner boundary.

3.4 Ground-Truth Boundary

In this part, we explain how we get ground-truth boundary map from ground-truth label map. Introducing approximate boundary information in the loss can improve the model's sensitivity to physical boundaries which results in an improvement of the prediction accuracy in boundary area. However, there are always labeled error pixels in the hand-labeled ground truth map, which are especially obvious at the boundary region as shown in Figure 3(a). In order to alleviate this problem, Figure 4 illustrates the inner

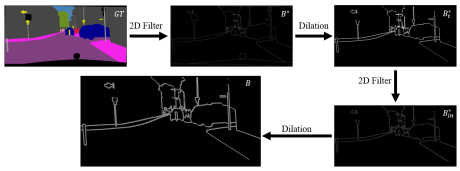

Figure 4: Ground-truth inner boundary extraction process.

boundary extraction process. Specifically, we first extract the boundary map B^* from the original ground-truth label map by a filter f that sets all pixels that do not have 8 identically-labeled neighbor pixels as 1, and other pixels as 0. Then we thicken the boundary by a 7×7 dilation operator and get boundary map B_t^*. Finally, we get the inner boundary B_{in}^* by applying the same filter f on B_t^* again and followed by another 3×3 dilation operator, as shown in Figure 3(b).

3.5 Joint Loss

The proposed BECLoss is designed for optimizing the network of two-branch paradigm. The purpose is to guide the two branches to learn complementary information. It can naturally be combined with other losses for training the whole network. Therefore, the network is trained to minimize a joint loss function:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \cdot \mathcal{L}_{BEC1} + \beta \cdot \mathcal{L}_{BEC2} \quad (7)$$

Specifically, $\mathcal{L}_{CE}$ is cross-entropy loss, $\mathcal{L}_{BEC1}$ and $\mathcal{L}_{BEC2}$ are proposed BECLoss for first and second branch, respectively. α and β are weights parameters of the two BECLoss.

4 EXPERIMENTAL RESULTS

In this section, we conduct experiments using Cityscapes dataset (Cordts et al., 2016) and Freiburg Forest dataset (Valada et al., 2016). In the following, we first modify some classic image semantic segmentation algorithms to build their two-branch decoder counterpart, then we compare the proposed two-branch architecture with the original network. Finally we carry out a series of ablation experiments on Freiburg Forest dataset. Our models are trained using Pytorch (Paszke et al., 2019) on one Nvidia Tesla P100 GPU with mixed precision settings.

4.1 Datasets

Cityscapes. The Cityscapes dataset is a large-scale database for urban street scene parsing. It contains 5000 finely annotated images captured from 50 cities with 19 semantic object categories, in which 2875 images are used for training, 500 images and 1525 images are used for validation and testing separately. All

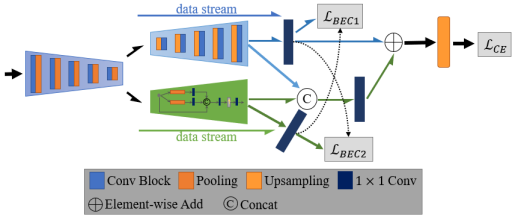

Figure 5: Architecture of modified SegNet with two decoders (SegNetT).

images are provided with the resolution of 2048×1024. We followed (Valada et al., 2019) and report results on the reduced 11 class label set.

Freiburg Forest. The Freiburg Forest dataset is an unstructured forested environments dataset. It contains 6 segmentation classes: sky, trail, grass, vegetation, obstacle and void. The dataset contains over 15000 images and 325 images with pixel level hand-annotated ground truth map. We follow (Valada et al., 2019) and use the same train and test splits provided by the dataset.

4.2 Implementation Details

In order to comprehensively test our proposed method, we deploy our proposed two-branch decoder on three classic baseline networks, namely, SegNet (Badrinarayanan et al., 2017), DeeplabV3+ (Chen et al., 2018b) and HRNet (Wang et al., 2020). Two-branch SegNet is shown in Figure 5. We divide the output of the encoder into two groups, one of which is input to the original data stream, in our two-branch implementation the original data stream means the upper branch of the decoder, and another is input to the additional data stream, the lower branch of the decoder. Next, we follow residual-liked module to fuse the two outputs, meanwhile deploy the BECLoss and cross-entropy loss during the training. More concretely, we supervise the learning process of the two branches through $\mathcal{L}_{BEC1}$ and $\mathcal{L}_{BEC2}$, and the combination of two outputs are guided by the auxiliary $\mathcal{L}_{CE}$. We follow the same way to implement the counterpart of DeeplabV3+ and HRNet. Note that we only take the backbone in the original model as an encoder, and the rest as the decoder. In practice, we use Res50 (He et al., 2016), Vgg16 (Simonyan and Zisserman, 2014) and HRNet-W18 (Wang et al., 2020) as backbones.

We initialize different encoders with the model pre-trained on ImageNet, this is totally the same as the original implementations (Badrinarayanan et al., 2017; Chen et al., 2018b; Wang et al., 2020). In order to speed up the convergence and reduce the interference of the initial learning rate setting, we employ a cyclical exponent learning rate policy (Smith, 2017) where the min_lr and max_lr are set to $1e-5$ and

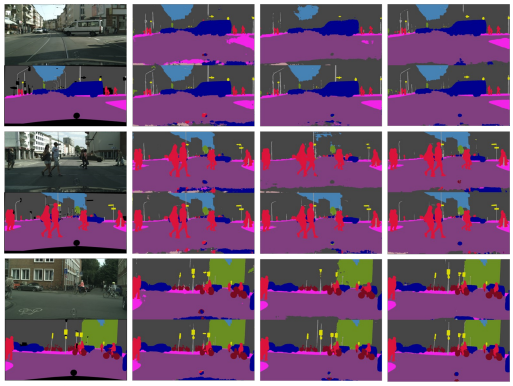

Figure 6: Qualitative results on the Cityscapes val set with 11 semantic class labels.

Table 1: Improvements with two-branch decoder on Cityscapes val set with 11 semantic class labels.

Methods	BaseNet	Mean IoU (%)	Parms. (M)
SegNet	Vgg16	75.82	29.4
SegNetT (ours)		80.64 (+**4.82**)	18.6
DeepLabv3+	Res50	80.31	26.6
DeepLabv3+T (ours)		82.45 (+**2.14**)	27.5
HRNet-W18	Hrnet-W18	82.34	9.6
HRNet-W18T (ours)		83.9 (+**1.56**)	9.6

$1e-2$, and cycle_length and step_size are set to 40 and 5 epochs respectively if not specified. Momentum and weight decay coefficients are set to 0.9 and 0.0005. Furthermore, we configure the hyperparameter γ and η in BECLoss as 10.0 and 0.3. The scale α and β in Equation 7 are simply set to 2.0. For Cityscapes dataset, we set input image size to 384×768, thus random cropping (cropsize 384×768) is applied during training, and during testing, we use the original image resolution of 1024×2048. For Freiburg Forest dataset, we resize the image to 384×768 during training and testing. All training images are augmented by random left-right flipping. We set 160 and 120 training epochs to Cityscapes datasets and Freiburg Forest dataset. And for both dataset we use the same mini batch size as 8. In addition, as we compare the original models with their two-branch encoder counterpart, so we perform the same settings for each comparison pair to ensure fairness.

4.3 Experimental Evaluation

In this section, we provide an extensive evaluation of each component of our framework on two challenging outdoor datasets, namely Cityscapes dataset and Freiburg Forest dataset. We use the widely used intersection over union (IoU) to evaluate the performance of our approach.

4.3.1 Results on Cityscapes Dataset

Table 1 summarizes the results of our two-branch decoder with different baselines. We can see that our approach significantly improves the mean IoU Specifically, our approach improves the mean IoU of original encoder-decoder frameworks namely SegNet, Deeplabv3+ and HRNet by 4.81, 2.14 and 1.56 respectively. Moreover, thanks to the reasonableness of proposed two-branch structure, our decoder is more efficient then the original decoder. In particular, our two-branch implementation of SegNet (SegNetT) greatly reduces the number of parameters while significantly improving the performance. DeepLabv3+T and HRNet only slightly increases the parameters (0.9M) or keep the number of parameters while improving the performance of the model. Our results also reflect that the original decoder does not make full use of the information extracted by the encoder. In addition, table 2 illustrates the category-wise comparison between various baselines and their two-branch variants. We surprisingly find that our method has a significant improvement in the prediction accuracy of small-scale targets, like "pole", "traffic sign" and "person". Several segmentation results are shown in Figure 6, we can see that our two-branch variants performs well on those small-size-object classes in the images compared to the baseline models. For "pole" and "traffic sign" in the images, the baseline models are more inclined to classify them as the surroundings and have difficult to distinguish targets with similar semantic labels. Moreover, dense targets are often taken as a whole and contain a lot of noise, and many details are overlooked, like "person" in the third row and "bike" in the fifth row. From the second, fourth and sixth rows, we can see that our method can correct these problems. Note that we may find the optimal hyperparameters to achieve better performances through grid search, but this is not the focus of this work.

4.3.2 Results On Freiburg Forest Dataset

We carry out experiments on the Freiburg Forest dataset to further evaluate the effectiveness of our method. Quantitative results of Freiburg Forest are shown in Table 3. The baselines (SegNet, DeepLabv3+, HRNet) yield mean IoU 69.99%, 77.48% and 78.29%. Our two-branch counterpart boost the performance to 81.79%, 82.73% and 83%. We can see that our methods outperforms their baselines with notable advantage, especially for the class of "obstacle", which is hardest to segment because of its serious class imbalance, and "obstacle" class consists of various types of objects which are diffi-

Table 2: Comparison in terms of IoU vs different baselines on the cityscapes val set with 11 semantic class labels.

Methods	sky	building	road	sidewalk	fence	vegetation	pole	vehicle	traffic sign	person	bicycle
SegNet	91.83	88.47	95.52	72.76	40.02	91.22	52.95	89.45	65.57	77.2	68.98
SegNetT (ours)	**93.35 (+1.52)**	**90.89 (+2.42)**	**96.65 (+1.13)**	**77.28 (+4.52)**	**49.72 (+9.7)**	**92.37 (+1.15)**	**61.54 (+8.59)**	**92.86 (+3.41)**	**75.64 (+10.07)**	**81.61 (+4.41)**	**75.06 (+6.08)**
DeepLabv3+	93.99	90.9	97.29	80.47	54.73	91.92	56.56	93.05	71.78	78.9	73.79
DeepLabv3+T (our	93.95	**91.99 (+1.09)**	**97.62 (+0.33)**	**82.31 (+1.84)**	**54.85 (+0.12)**	**92.55 (+0.63)**	**62.69 (+6.13)**	**94.17 (+1.12)**	**77.87 (+6.09)**	**82.4 (+3.5)**	**76.59 (+2.8)**
HRNet	94.31	92.06	97.67	82.31	54.94	92.59	63.31	94.34	76.37	82.27	75.4
HRNet-T (ours)	**94.83 (+0.52)**	**92.68 (+0.62)**	**97.98 (+0.31)**	**84.3 (+1.99)**	**56.28 (+1.34)**	**93.06 (+0.47)**	**67.17 (+3.86)**	**94.83 (+0.49)**	**79.98 (+3.61)**	**84.35 (+2.08)**	**77.09 (+1.69)**

Table 3: Improvements with two-branch decoder on Freiburg Forest val set.

Methods	BaseNet	Trail	Grass	Veg.	Sky	Obst.	Mean IoU (%)	Parms. (M)
SegNet	Vgg16	84.15	85.55	88.97	91.28	0	69.99	29.4
SegNetT (ours)		**88.55 (+4.4)**	**88.96 (+3.41)**	**0.91 (+1.94)**	**2.63 (+1.35)**	**47.93 (+47.93)**	**81.79 (+11.8)**	18.6
DeepLabv3+	Res50	83.03	86.11	89.96	92.16	36.1	77.48	26.6
DeepLabv3+T (ours)		**88.02 (+4.99)**	**88.93 (+2.82)**	**91.02 (1.06)**	**2.83 (+0.67)**	**52.87 (+16.77)**	**82.73 (+5.25)**	27.5
HRNet	Hrnet-W18	84.79	86.49	89.79	91.96	38.44	78.29	9.6
HRNet-T (ours)		**88.74 (+3.95)**	**89.35 (+2.86)**	**91.14 (+1.35)**	**92.6 (+0.64)**	**53.17 (+14.73)**	**83 (+4.71)**	9.6

cult to be unified into the same class. Furthermore, this dataset only contains large-scale targets such as tree, trail, etc. Therefore, it is difficult to optimize the segmentation accuracy of these classes through multi-scale learning. Benefiting from the efficient information integration capabilities of our proposed two-branch decoder paradigm, our method shows impressive advantages on the above issues. Several examples are shown in Figure 7.

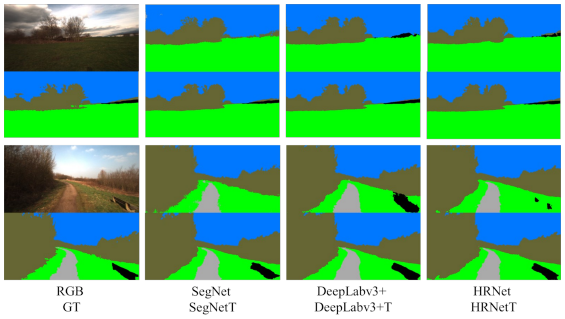

Figure 7: Qualitative results on the Freiburg Forest test set.

4.4 Ablation Study

4.4.1 BECLoss and Boundary

All two-branch variants are implemented by replacing the decoder of the original network with our proposed two-branch decoder, and through our well-designed BECLoss to explicitly supervise the learning process of the model, the two branches can learn complementary information. In addition, we introduce boundary information into BECLoss to improve the inherent sensitivity of our models to boundary pixels. To verify the validity of our method, we conduct a group of ablations to analyze the influence of various factors within our method. We report the results over the segmentation baseline SegNet on Cityscapes and Freiburg Forest dataset in Table 4.

As shown in Table 4, two-branch decoder improve the performance remarkably. Compared with the baseline SegNet, employing two-branch decoder yields a result of 78.54% mean IoU on Cityscapes dataset and 78.9% mean IoU on Freiburg Forest dataset, which brings 2.72% and 8.91% improvement. In addition, when we gradually replaced the cross-entropy loss CELoss of loss1 and loss2 with the BECLoss we designed, the performance further improves to 79.5% and 81.43%. Furthermore, we notice that when we use only one BECLoss, the result very slightly exceeds the result of using two BECLoss, as shown in the third row and the fifth row, the result from 79.54% goes to 79.5% on Cityscapes dataset. We further introduce boundary information to BECLoss, performance increased to 80.64%. Results show that our proposed two-branch decoder and boundary enhanced BECLoss bring great benefit to scene parsing.

Table 4: Ablation study on Cityscapes val set and Freiburg Forest test set. *Loss*1-*Loss*3 represent deployed loss in Figure 2, *B* indicates BECLoss enhanced by boundary information.

Methods	Loss1	Loss2	Loss3	B	Mean IoU (%)	
					Cityscapes	Freiburg
SegNet	\	\	CE	\	75.82	69.99
SegNetT	CE	CE	CE	\	78.54 (+2.72)	78.9 (+8.91)
SegNetT	BEC	CE	CE	N	79.54 (+3.72)	80.48 (+10.49)
SegNetT	CE	BEC	CE	N	79.07 (+3.25)	79.9 (+9.91)
SegNetT	BEC	BEC	CE	N	79.5 (+3.68)	81.43 (+11.44)
SegNetT	BEC	BEC	CE	Y	80.64 (**+4.82**)	81.79 (**+11.8**)

4.4.2 Single Branch

In many real scenarios, we have difficulties deploying complex models into practical applications due to computer resources and run-time limitations. Our approach can make our two-branch structure decoder adaptively learn complementary knowledge during

the training process. In addition, the two branches learn pixels that are difficult for each other to resolve during the training process, there is a certain competitive relationship between the two branches during the training process. This is what we expect, because on the one hand the auxiliary branch can learn complementary knowledge that can boost the main branch, on the other hand competition with each other can make each branch learn more efficiently. As a result, even if a branch is removed during the inference process, our results are far better than the original encoder-decoder structure and the number of parameters is less. As shown in Table 5, We use an extremely simple branch, as shown in Figure 2(b), retraining on the Cityscapes dataset, and we named the trained model '*ED*'. Moreover, we test the output results of each branch separately on the trained two-branch decoder model. Specifically, we take SegNet as an example. In the inference process, we only keep the upper branch of the model in Figure 5, and the output result obtained corresponds to 'O^*'. When we only keep the lower branch, the result corresponds to 'D^*'. '*O&D*' goes to the result of original two-branch model. The results of the main branch in our trained two-branch model are 80.49%, 82.35% and 83.87%, which is significantly exceed the counterparts of original encoder-decoder models (75.82%, 80.31% and 82.34%) while the number of parameters used dropped remarkably. Surprisingly, although the result of the auxiliary branch are sometimes not satisfactory, such as 65.34% with the encoder of Vgg16 or 76.67% with the encoder of Res50. However, we find that the residual-like module can effectively combine the outputs of the two branches to further improve the final result to 80.64%, 82.61% and 83.9%, as shown in '*O&D*' columns, which is mean that the final results are not adversely affected. This once again shows that our method can make each branch learns complementary information.

Table 5: Single branch test on Cityscapes val set with 11 semantic class labels. '*Enc.*' represent encoder, '*Dec.*' represent decoder. '*O*' indicates the decoder deployed in the original model. '*D*' the decoder in Figure 2(b), '*T*' indicates our two-branch decoder. 'O^*', 'D^*' and '*O&D*' mean the result from upper branch, lower branch and final branch separately.

Methods	Enc.	Dec.	Mean IoU (%)			Parms. (M)		
SegNet	Vgg16	*O*	75.82			29.4		
ED		*D*	65.34			15.3		
SegNetT (ours)		*T*	O^*	D^*	*O&D*	O^*	D^*	*O&D*
			80.49	67.34	80.64	18.4	14.9	18.6
DeepLabv3+	Res50	*O*	80.31			26.6		
ED		*D*	76.67			32.2		
DeepLabv3+ (ours		*T*	O^*	D^*	*O&D*	O^*	D^*	*O&D*
			82.35	77.3	82.61	25.3	25.7	27.5
HRNet	Res50	*O*	82.34			9.6		
ED		*D*	81.25			9.7		
HRNet-T (ours)		*T*	O^*	D^*	*O&D*	O^*	D^*	*O&D*
			83.87	76.83	83.9	9.6	9.6	9.6

5 CONCLUSION

In this paper, we have presented a general two-branch decoder paradigm, composed of a main branch and an auxiliary branch for scene segmentation. This decoder paradigm can be directly applied in an encoder-decoder framework to efficiently refine and integrate the information extracted by the encoder. With this two-branch decoder, we further propose a boundary enhanced complementary loss, named BECLoss. The insight is that the learning process of each branch can be supervised by our proposed BECLoss, so that different branches can adaptively learn complementary information without explicitly indicating the specific learning elements and compete with each other in the learning process. Moreover, in order to verify the effectiveness of our proposed methods, we design a simple yet efficient branch which is deployed as the auxiliary branch in our two-branch decoder. The comparative experiment shows that two-branch decoder paradigm and BECLoss can significantly improve the performance of the original encoder-decoder model consistently on challenge outdoor scenes, i.e. Cityscapes and Freiburg Forest datasets. It is worth mentioning that although we added an additional branch to decoder, it did not increase the number of parameters significantly. In addition, In the inference process, even if we delete a branch, we can still get performance far beyond the original counterpart. As a perspective for future work, we would like to design a more universal and general branch which can be applied in our two-branch decoder paradigm.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12):2481–2495.

Chen, H., Huang, Y., and Nakayama, H. (2018a). Semantic aware attention based deep object co-segmentation. In Asian Conference on Computer Vision, pages 435–450. Springer.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, L.-C., Yang, Y., Wang, J., Xu, W., and Yuille, A. L. (2016). Attention to scale: Scale-aware se-

mantic image segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3640–3649.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), pages 801–818.

Chen, Y., Dapogny, A., and Cord, M. (2020). Semeda: Enhancing segmentation precision with semantic edge aware loss. Pattern Recognition, 108:107557.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3213–3223.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., and Lu, H. (2019). Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3146–3154.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141.

Huang, Q., Xia, C., Wu, C., Li, S., Wang, Y., Song, Y., and Kuo, C.-C. J. (2017). Semantic segmentation with reverse attention. arXiv preprint arXiv:1707.06426.

Li, H., Xiong, P., An, J., and Wang, L. (2018). Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180.

Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., and Liu, H. (2019). Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9167–9176.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. In 2017 IEEE winter conference on applications of computer vision (WACV), pages 464–472. IEEE.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision, pages 843–852.

Takikawa, T., Acuna, D., Jampani, V., and Fidler, S. (2019). Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5229–5238.

Tao, A., Sapra, K., and Catanzaro, B. (2020). Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821.

Valada, A., Mohan, R., and Burgard, W. (2019). Self-supervised model adaptation for multimodal semantic segmentation. International Journal of Computer Vision, pages 1–47.

Valada, A., Oliveira, G., Brox, T., and Burgard, W. (2016). Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In International Symposium on Experimental Robotics (ISER).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pages 3–19.

Yuan, Y., Xie, J., Chen, X., and Wang, J. (2020). Segfix: Model-agnostic boundary refinement for segmentation. In European Conference on Computer Vision, pages 489–506. Springer.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2881–2890.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., et al. (2020). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. arXiv preprint arXiv:2012.15840.

Zhong, Z., Lin, Z. Q., Bidart, R., Hu, X., Daya, I. B., Li, Z., Zheng, W.-S., Li, J., and Wong, A. (2020). Squeeze-and-attention networks for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13065–13074.