# Stereo Superpixel Segmentation Via Decoupled Dynamic Spatial-Embedding Fusion Network

Hua Li*, Junyan Liang∗, Ruiqi Wu, Runmin Cong, Wenhui Wu, Sam Tak Wu Kwong, *Fellow, IEEE*

***Abstract*—Stereo superpixel segmentation aims at grouping the discretizing pixels into perceptual regions through left and right views more collaboratively and efficiently. Existing superpixel segmentation algorithms mostly utilize color and spatial features as input, which may impose strong constraints on spatial information while utilizing the disparity information in terms of stereo image pairs. To alleviate this issue, we propose a stereo superpixel segmentation method with a decoupling mechanism of spatial information in this work. To decouple stereo disparity information and spatial information, the spatial information is temporarily removed before fusing the features of stereo image pairs, and a decoupled stereo fusion module (DSFM) is proposed to handle the stereo features alignment as well as occlusion problems. Moreover, since the spatial information is vital to superpixel segmentation, we further design a dynamic spatiality embedding module (DSEM) to re-add spatial information, and the weights of spatial information will be adaptively adjusted through the dynamic fusion (DF) mechanism in DSEM for achieving a finer segmentation. Comprehensive experimental results demonstrate that our method can achieve the state-of-the-art performance on the KITTI2015 and Cityscapes datasets, and also verify the efficiency when applied in salient object detection on NJU2K dataset. The source code will be available publicly after paper is accepted.**

***Index Terms*—Stereo image, superpixel segmentation, stereo corresponding capturing, spatiality embedding.**

## I. Introduction

SUPERPIXEL segmentation aims at grouping the discretizing pixels into some high-level correlative units as input primitives in a variety of subsequent computer vision tasks, *e.g.*, salient object detection [2]–[5], image dehazing [6], image classification [7], object recognition [8], adversarial attack [9]. Nowadays, dual cameras have been widely used in extensive industrial applications, such as assistant driving and mobile phones. Compared with single images, stereo image pairs can obtain complementary information from the second viewpoint, which is beneficial to scene representation and object modeling [10]. However, how to effectively utilize complementary and correspondence information to generate superpixels for stereo images is still a challenging task.

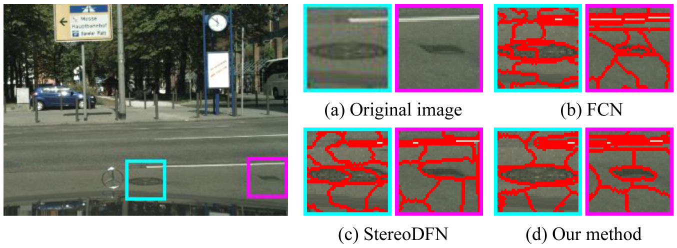

(a) Original image
(b) FCN
(c) StereoDFN
(d) Our method

Fig. 1: A simple illustration of the comparison between the state-of-the-art superpixel segmentation methods and our method.

For stereo superpixel segmentation, stereo image pairs are generally segmented separately as single views by traditional methods. By this way, the complementarity and correlation between the left and right views of stereo image pairs are ignored and cannot be explored sufficiently [11], [12]. Therefore, these methods cannot be regarded as an real implementation of stereo superpixel segmentation, since the intrinsic characteristics of stereo images are neglected. To take the collaborative relationship between left and right views into consideration, Li *et al.* [10] propose a collaborative optimization scheme to generate stereo superpixels with the parallax consistency, which is the first attempt to devise a specific superpixel segmentation method for stereo image pairs. The method first match the corresponding regions between the left and right view of a stereo image pair. Superpixels are initialized and matched in the corresponding regions. Then, the superpixels in the left and right views are refined simultaneously via a collaborative optimization strategy. Experimental results demonstrate it outperforms the methods that segment stereo image pairs separately. Nevertheless, this method extracts handcrafted feature instead of deep feature by an unsupervised way, which leads to the limitation of the performance.

Most recently, Wu *et al.* [1] propose an end-to-end dual attention fusion network (StereoDFN) for stereo superpixel segmentation, which extracts the deep features of stereo image pairs by convolutional neural networks instead of handcrafted features. Then it models the correspondence between the left and right views via the parallax attention module to integrate

*These authors contributed equally. Ruiqi Wu proposed the paper idea.

The preliminary work of this paper was published in ICME 2021 [1], which is accepted as oral presentation.

Hua Li is with the School of Computer Science and Technology, Hainan University, Hainan 570228, China (e-mail: lihua@hainanu.edu.cn).

Junyan Liang is with the School of Computer Science and Technology, Hainan University, Hainan 570228, China (e-mail: liangjunyan@hainanu.edu.cn).

Ruiqi Wu is with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China (e-mail: wuruiqi0722@gmail.com).

Runmin Cong is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: rmcong@bjtu.edu.cn).

Wenhui Wu is with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: wuwenhui@szu.edu.cn).

Sam Tak Wu Kwong is with the Department of Computer Science, City University of Hong Kong, Kowloon Hong Kong 999077, China, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, China (e-mail: cssamk@cityu.edu.hk).

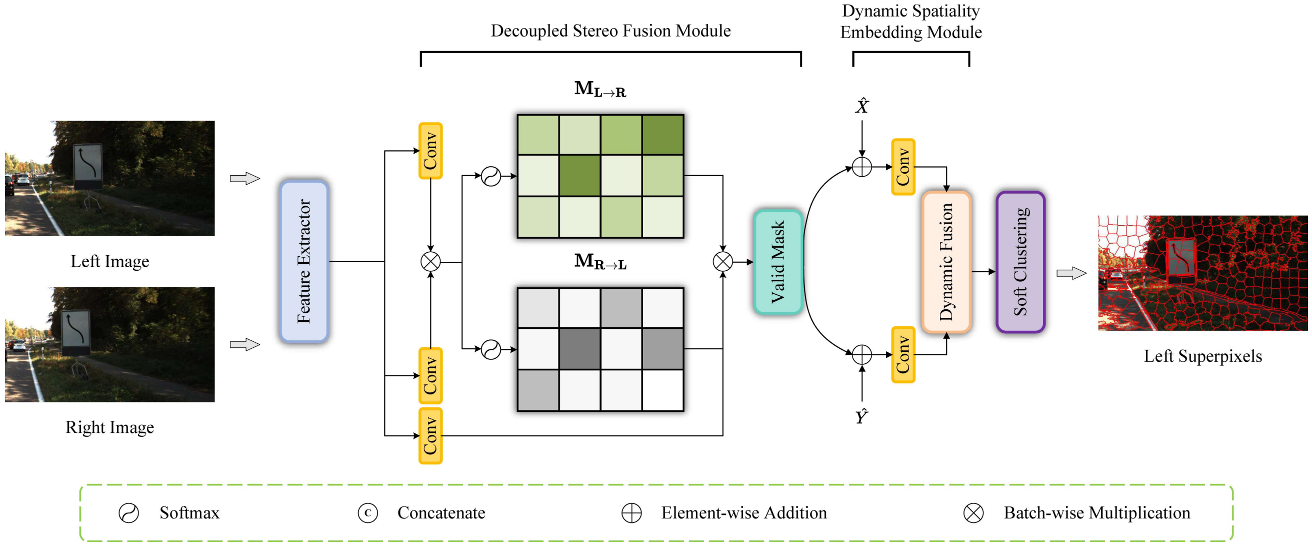


Fig. 2: Overall framework of the proposed method.

the complementary information of stereo image pairs and has achieved impressive performance. However, the existing superpixel segmentation methods like StereoDFN and superpixel sampling network (SSN) [13] utilize the five-dimensional features (including two-dimensional spatial features XY, and three-dimensional color features Lab) as input, which aims to extract high-dimensional deep features. For stereo superpixel segmentation task, the disparity information is often adopted to emphasize the correspondence between left and right views. If the spatial information is input before the feature fusion, this may result in strong constraints on spatial information, since the stereo features are coupled with spatial features, which will lead to degradation of image boundaries.

The presented work significantly extends StereoDFN with a decoupling mechanism of spatial information to only use three-dimensional color features (Lab) instead of five-dimensional features (XYLab) in modeling the correspondence between both views, thereby decoupling the stereo features (color features) and spatial features (XY) to relax the constrain of spatial information. Considering the importance of spatial information in superpixel segmentation, we further design a Dynamic Spatiality Embedding Module (DSEM) to re-add spatial information, the weighting of spatial information will be adaptively adjusted through the Dynamic Fusion (DF) mechanism in DSEM to fit images of different sizes, thereby obtaining a more accurate representation of spatial information and achieving a better performance. As the simple comparison shown in Fig. 1, we can see that our proposed method adhere to the object boundaries better than FCN [14] and StereoDFN.

In this paper, we improve upon our previous work in ICME [1]. The main contributions of the proposed work can be summarized as follows:

1) We propose a stereo superpixel segmentation method with a decoupling mechanism of spatial information to generate superpixels for stereo image pairs, which can integrate the correspondence from both left and right viewpoints to take the stereo features alignment and occlusion problems into account.
2) Since the coupling of stereo features and spatial features may impose strong constraints on spatial information while modeling the correspondence between stereo image pairs, we develop a spatial decoupling mechanism to model the correspondence with relaxed spatial constraint by decoupling the stereo features and spatial features, and postponing the embedding of spatial information after stereo features have been fused.
3) We design a Dynamic Spatiality Embedding Module (DSEM) to re-add spatial information for achieving a finer segmentation. The weighting of spatial information can be adaptively adjusted via the Dynamic Fusion (DF) mechanism in DSEM to fit images of different sizes, thereby achieving a better performance.
4) Our method achieves the state-of-the-art performance compared with previous works both quantitatively and qualitatively. Extensive ablation studies validate the effectiveness of the proposed strategy. With application in salient object detection, we also demonstrate that our method can achieve superior performance in downstream task.

The article is organized as follows. We briefly introduce the related work about existing superpixel segmentation algorithms in Section II. Then, we propose our model and detail each key component in Section III. The qualitative and quantitative experimental results and analyses are presented in Section IV. In Section V, we present the application of the proposed method in salient object detection. Finally, Section VII concludes this article.

## II. Related Work

The concept of “Superpixel” is first introduced in [15], which is an over-segmentation of images and generated by grouping pixels similar in low-level properties. Existing superpixel segmentation algorithms can be simply divided into two categories: unsupervised superpixel segmentation methods and supervised superpixel segmentation methods.

### A. Unsupervised Superpixel Segmentation Methods

Simple linear iterative clustering (SLIC) [16] is one of the most widely used unsupervised methods, which employs k-means clustering approach to generate superpixels efficiently by grouping nearby pixels based on five-dimensional color and position features of the images. Due to SLIC has fast runtime and impressive performance, many superpixel-based applications commonly use SLIC for superpixel segmentation. Linear spectral clustering (LSC) [17] generates superpixels based on kernel function instead of using the traditional eigen, which is not only able to produce compact superpixel, but also with low computational costs. Considering the irregular structure of superpixels, Li *et al.* [18] propose approximately structural superpixels (ASS), they regard superpixel segmentation as a square-wise asymmetric partition problem and generate ASS by an asymmetrically square-wise superpixel segmentation way, which can preserve semantics better and largely reduces data amount.

### B. Supervised Superpixel Segmentation Methods

In recent years, inspired by the success of deep learning techniques in a wide variety of computer vision tasks, some works try to use deep learning techniques for superpixel segmentation. Jampani *et al.* [13] propose the first deep learning-based end-to-end trainable superpixel segmentation network (SSN), which is enlightened by the SLIC method. To simplify the generation of superpixels, Yang *et al.* [14] propose a lightweight fully convolutional networks (FCN) that based on encoder-decoder structure, which generates superpixels efficiently by predicting the probability map between pixels and superpixels. More recently, Wu *et al.* propose an dual attention fusion network (StereoDFN), they attempt to take the collaborative relationship between stereo image pairs into consideration by modeling the correspondence between them, which is based on parallax attention mechanism.

## III. Proposed Method

In this work, we propose a stereo superpixel segmentation method with a decoupling mechanism of spatial information, the framework is illustrated in Fig. 2. In general, the proposed method can be divided into the following steps: First, stereo image pairs with Lab color space are input into fully convolutional network to extract the deep features. Then, the deep features of left and right views are fed to the Decoupled Stereo Fusion Module (DSFM), which integrates the features from both views. Moreover, Dynamic Spatiality Embedding Module (DSEM) is proposed to adaptively combine the spatial information with deep features. Finally, a soft clustering algorithm [13] is adopted to generate the superpixels.

In what follows, we detail the main components of the proposed method, which are feature extractor, Decoupled Stereo Fusion Module (DSFM), Dynamic Spatiality Embedding Module (DSEM) and loss functions respectively.

### A. Feature Extractor

A pair of weight-shared Convolutional Neural Networks (CNNs) is adopted to extract the deep feature of stereo image. The basic block is a ‘Conv-BN-ReLU’ block, which is composed of a convolution layer with $3 \times 3$ kernel size and 64 output channels, a batch-normalization layer and a ReLU activation function. Each of the two modules will be followed by a max-pooling layer for downsampling. For features captured from Block2, Block4 and Block6, we upsample them into the same resolution as the input image and concatenate them together. Block7 will fuse them and generate the final output. Through this way, the networks can effectively learn more multi-level and multi-scale features, which is benefit for both superpixel segmentation and capturing the correspondence of stereo image pairs. The schematic illustration of feature extractor has been shown in Fig. 3.

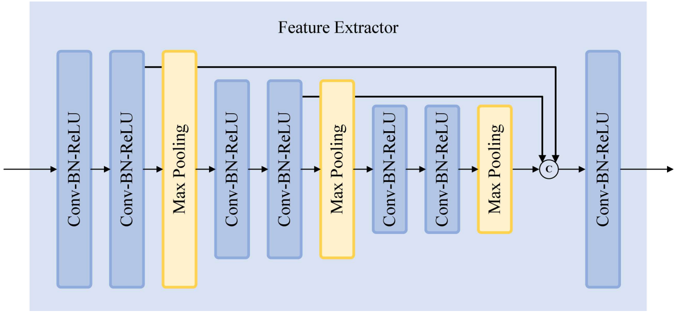
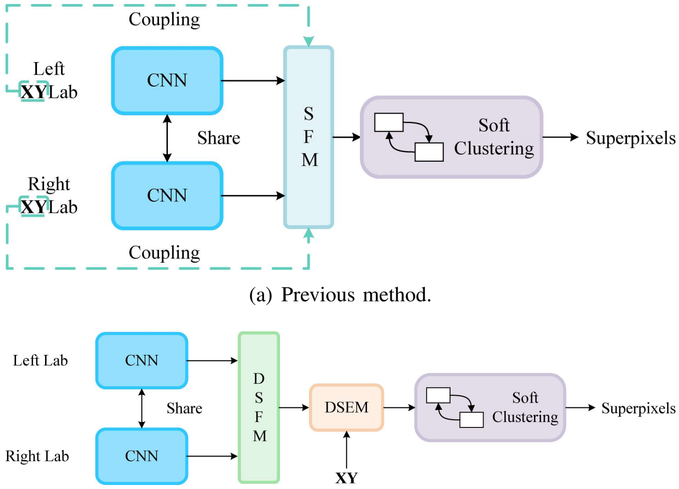


Fig. 3: The schematic illustration of Feature Extractor.



(a) Previous method.

(b) Proposed method.

Fig. 4: Differences between our method and StereoDFN. We only use Lab color information as input to decouple the stereo features and spatial features. After **D**ecoupled **S**tereo **F**usion **M**odule (DSFM), we design a **D**ynamic **S**patiality **E**mbedding **M**odule (DSEM) to re-add spatial information.

### B. Decoupled Stereo Fusion Module

Decoupled Stereo Fusion Module (DSFM) is the key component for fusing the stereo features. Considering the most significant problems in stereo features fusion, such as features alignment and occlusion problem, the proposed DSFM try to solve them via parallax attention mechanism and valid mask.

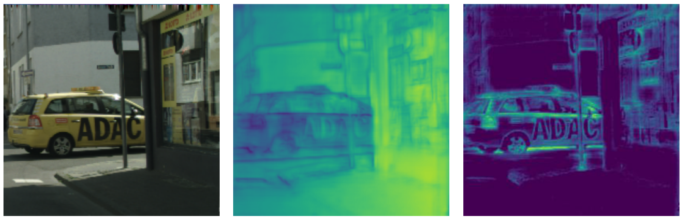

(a) Original image (b) Coupled (c) Decoupled

Fig. 5: The visualization of deep features after modeling the correspondence between stereo image pairs. Note that (b) is generated by StereoDFN, while (c) is generated by our method.

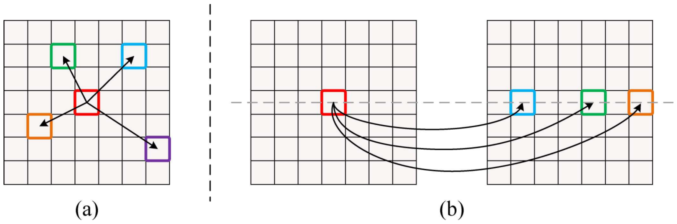


Fig. 6: The schematic illustration of the difference between self attention and parallax attention mechanism. (a) and (b) are self attention mechanism and our parallax attention mechanism respectively. We can find the computational complexity of parallax-attention mechanism is much smaller by comparing (a) and (b).

**Spatial Decoupling Mechanism.** Considering the coupling of stereo features and spatial features may impose strong constraints on spatial information while modeling the correspondence between stereo image pairs, thereby interfering with superpixels to adhere to the object boundaries. The spatial decoupling mechanism is proposed to model the correspondence between stereo image pairs with relaxed spatial constraint.More specifically, we remove the spatial information of input items for relaxing the constrain of spatial information. The input of StereoDFN is a five-dimensional features (XYLab), while the input of our proposed method is a three-dimensional features (Lab), this is the essential difference between our proposed method and StereoDFN, we also present the schematic illustration of the difference in Fig. 4. Furthermore, the effectiveness of our spatial decoupling mechanism has been shown in Fig. 5, benefiting from decoupling stereo features and spatial features, our method eliminates the interference of spatial information on modeling, and the boundary information is much more clearly.

**Stereo Features Alignment.** Since the corresponding pixels in stereo image pairs are located at different positions, it is extremely difficult to fuse the stereo features directly. Therefore, aligning the stereo features is necessary before fusion.

Inspired by [19], the parallax attention mechanism is utilized to model the correspondence between stereo image pairs. Since the left and right views of the images are only translated horizontally, while aligning in the vertical direction, the matching pixels in two views must be on the same horizontal line. For this reason, as the schematic illustration shown in Fig. 6(a) and Fig. 6(b), the parallax attention mechanism is only consider the correlation of pixels on the same horizontal line, instead of all the pixels in the image like traditional self attention mechanism. In this way, the computation complexity of attention can be largely reduced.

As the schematic illustration of aligning the features from the right view with the left view features shown in Fig. 2, for a pair of deep features $\mathcal{F}_L$ and $\mathcal{F}_R \in \mathcal{R}^{H\times W\times C}$, we can get $A$ and $B$ from a convolution layer with $1\times 1$ kernel size. Then, the parallax attention map $\mathcal{M} \in \mathcal{R}^{H\times W\times W}$ will be generated by:

$$\mathcal{M}_{R\to L} = softmax(A \otimes B^T), \tag{1}$$

$$\mathcal{M}_{L\to R} = softmax(B \otimes A^T), \tag{2}$$

where $\otimes$ denotes the batch-wise multiplication, and $T$ denotes the batch-wise transposition. $\mathcal{M}_{L\to R}(i,j,k)$ and $\mathcal{M}_{R\to L}(i,j,k)$ represent the contribution of the position $(i,j)$ in one view to position $(i,j)$ in another view. In this way, the supplementary information of one view can be obtained through another view, which can be formulated as follows:

$$\hat{\mathcal{F}}_L = \mathcal{M}_{R\to L} \otimes \mathcal{F}_{right}, \tag{3}$$

$$\hat{\mathcal{F}}_R = \mathcal{M}_{L\to R} \otimes \mathcal{F}_{left}, \tag{4}$$

where $\hat{\mathcal{F}}_L$ and $\hat{\mathcal{F}}_R$ denote the aligned features.

**Occlusion Problem.** Occlusion always exists in stereo images due to violent disparity variation, which will lead to an inaccurate stereo features fusion. Therefore, we further add the occlusion handling part in DSFM to take occlusion problem into consideration.

Taking the example of handling the occlusion in the right image. Assuming that a pixel located at $(i,j)$ is in the occlusion region, for any $k \in [1,W]$, $\mathcal{M}_{L\to R}(i,j,k)$ is an extremely low value since any pixel on the same horizontal line in the left view is not relevant to it. Thus, the valid mask $O_{L\to R}$ can be generated from parallax attention map $\mathcal{M}_{L\to R}$, which can be formulated as Eq. (5):

$$O_{L\to R}(i,j) = \begin{cases} 1, & \sum_{k\in[1,W]} \mathcal{M}_{L\to R}(i,k,j) > \tau \\ 0, & \sum_{k\in[1,W]} \mathcal{M}_{L\to R}(i,k,j) \leq \tau, \end{cases} \tag{5}$$

where $\tau$ is a threshold set to 0.1. Then, the occlusion handling part will fuse the stereo features. The left fused features $\tilde{\mathcal{F}}_L$ can be obtained as Eq. (6):

$$\tilde{\mathcal{F}}_L = Concat(\hat{\mathcal{F}}_L \circ O_{L\to R} + \mathcal{F}_L \circ (1 - O_{L\to R}), \mathcal{F}_L), \tag{6}$$

where $Concat(,)$ represents the concatenate operation on channel dimension, while $\circ$ represents the Hadamard product. Finally, $\tilde{\mathcal{F}}_L$ will be fed to a 'Conv-BN-ReLU' block for reducing the channel size to the size of origin features.

### C. Dynamic Spatiality Embedding Module

To prevent the spatial information influence the stereo correspondence modeling, the spatial information has been removed in the input items. However, spatial information is indispensable for superpixel segmentation method to adhere to

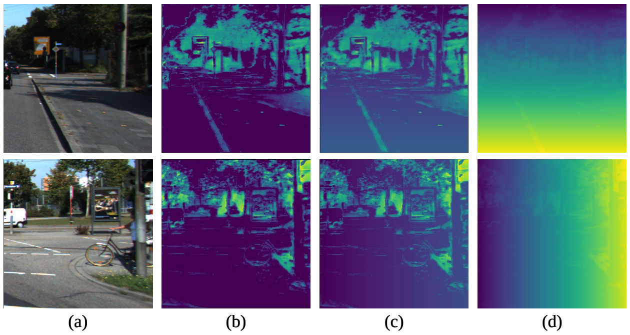


Fig. 7: (a) denotes the input images with $300 \times 300$ resolution, (b) the deep features, (c) the features after adding the spatial information with a value domain of $(0, 1)$, and (d) the features after adding the spatial information zoomed in the manner of [13], [16]. We can see our embedding strategy preserves more details than (d).

object boundaries more accurately, which is vital to achieve a better performance. Therefore, spatial information is re-added via the Dynamic Spatiality Embedding Module (DSEM) to take both conditions into consideration simultaneously, so that we can not only eliminate the disadvantage of spatial information in modeling the correspondence, but also utilize the advantage of spatial information. DSEM consists of two parts, which are Spatiality Embedding (SE) and Dynamic Fusion (DF). The architecture of DSEM can be seen in Fig. 2.

**Spatiality Embedding.** A reliable superpixel segmentation algorithms require the ability to handle images with different resolutions. However, the value of spatial information can be extremely large for a high-resolution image, which will pollute the image feature representation if the spatial information is embedded directly. In order to avoid such a disadvantage, we normalize the spatial information $X$ and $Y$ as Eq. (7):

$$\hat{X} = \frac{X}{max(X)}, \ \hat{Y} = \frac{Y}{max(Y)}, \tag{7}$$

where $X$ and $Y$ are spatial information on horizontal and vertical direction, respectively.

After normalizing, $\hat{X}$ and $\hat{Y}$ is added to fused features and get $\tilde{\mathcal{F}_X}, \tilde{\mathcal{F}_Y}$. Then, a convolution layer with $1 \times 1$ kernel size is followed to embed the spatial information. In this way, we can prevent an over-consideration of spatial information. Finally, $\tilde{\mathcal{F}_X}$, $\tilde{\mathcal{F}_Y}$ is concatenated with input features and send them to dynamic fusion part. Fig. 7 shows the effectiveness of our embedding strategy.

**Dynamic Fusion.** Although spatial information is indispensable in superpixel segmentation task, it does not always play the same important role of different regions in one image. For example, for regions with sparse textures, spatial information should be considered more to generate regular and compact superpixels. On the other hand, for regions with dense edges and complex contents, spatial information is relatively less important. Therefore, to achieve consistently excellent performance in different conditions, a dynamic fusion mechanism is designed to adaptively adjust the weighting of spatial information during fusion phase.

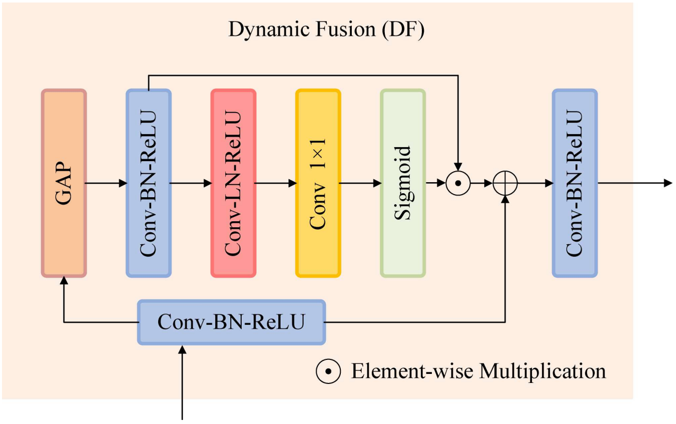


Fig. 8: The schematic illustration of Dynamic Fusion mechanism.

The dynamic fusion mechanism employs a channel-attention [20] way to adaptively aggregate and refine features. More specifically, we first use a 'Conv-BN-ReLU' block to fuse the features coarsely. Then, a global average pooling layer with another 'Conv-BN-ReLU' is followed to generate the global feature map. Finally, a series operations are utilized to produce the weighting map, which can be formulated as follows:

$$\mathcal{W} = g \cdot \sigma(C(ReLU(LN(C(g))))), \tag{8}$$

where $\mathcal{W}$ is the weighting map and $g$ is the global feature map. $\sigma, C, ReLU$ and $LN$ represents sigmoid function, a convolution layer with $1 \times 1$ kernel size, ReLU activation function and layer-normalization, respectively. In this way, a more effective representation of spatial information with the guidance of weighting map can be obtained. Finally, the weighting map is added to the input features and fed to the third 'Conv-BN-ReLU' block to generate the adjusted features. Fig. 8 presents the details of the Dynamic Fusion (DF) mechanism.

### D. Loss Functions

We design two loss functions for optimizing our model.

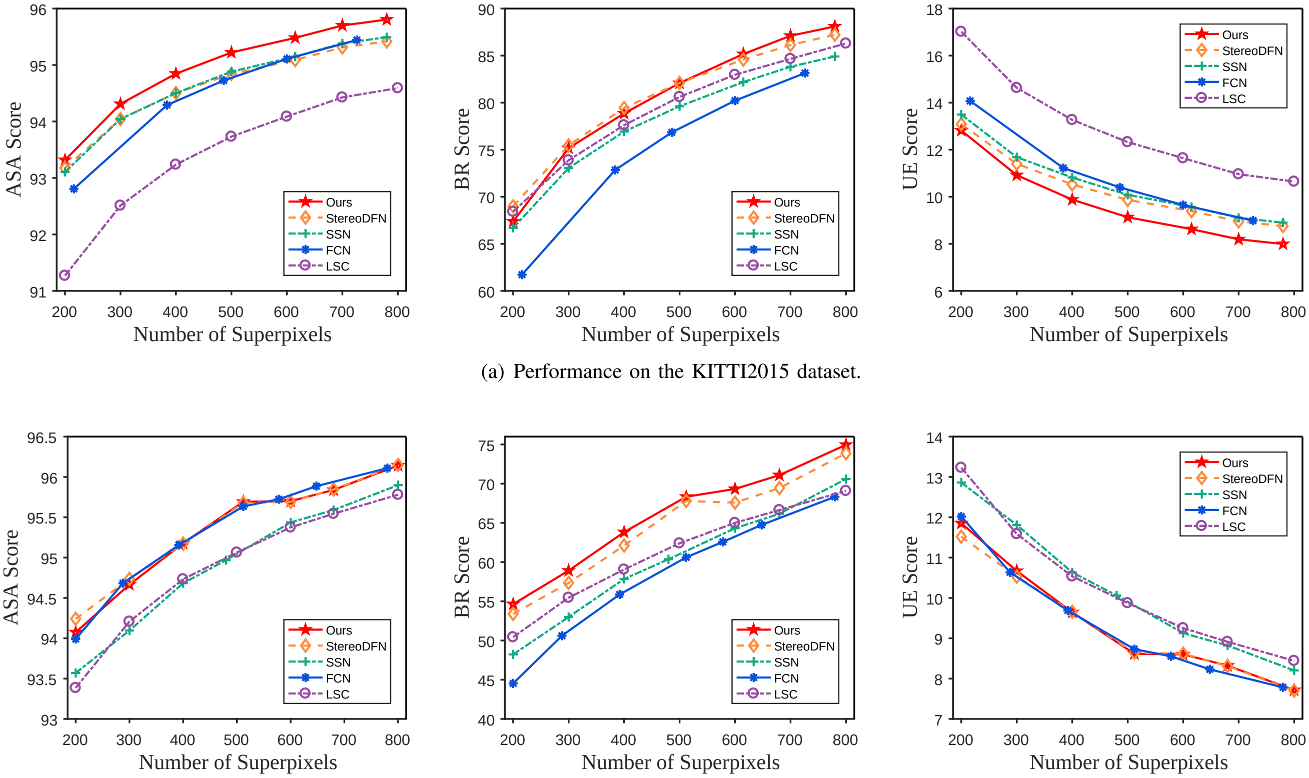
**Semantic Loss.** This function facilitates the superpixel adhere to semantic boundaries, which utilize the cross-entropy loss function $SE$ to measure the loss:

$$\mathcal{L}_{sem} = CE(S, S^*), \tag{9}$$

where $S$ denotes the one-hot semantic label of ground truth and $S^*$ is the reconstructed semantic label.

**Stereo Loss.** This loss function is designed to constrain the model to correctly estimate stereo correspondence. We also add valid mask to eliminate the problems caused by occlusion. Stereo loss is defined as:

$$\begin{aligned} \mathcal{L}_{stereo} = &\left\| O_{L \to R} \circ (I_L - \mathcal{M}_{R \to L} \otimes I_R) \right\|_1 + \\ &\left\| O_{R \to L} \circ (I_R - \mathcal{M}_{L \to R} \otimes I_L) \right\|_1, \end{aligned} \tag{10}$$



(a) Performance on the KITTI2015 dataset.

(b) Performance on the validation set of Cityscapes dataset.

Fig. 9: Quantitative comparison of the proposed method and other state-of-the-art methods.

where $I_L$, $I_R$ denotes the left and right image, respectively. $\circ$ denotes Hadamard product.

The total loss is the sum of these two functions:

$$\mathcal{L}_{total} = \mathcal{L}_{sem} + \lambda \mathcal{L}_{stereo}, \tag{11}$$

where $\lambda$ is empirically set to 1.0 for balancing the scales of different losses.

All above are the basic definitions of the metrics we used for evaluating, more details can be seen in [21].

## IV. Experiments and Results

### A. Experimental Setup

**Implementation Details.** We apply a batch-mode learning method with a batch size of 8 to train our model for 20K iterations. The Adam with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$) is utilized to optimize the network. In addition, the initial learning rate is $2 \times 10^{-4}$ and decreases by half every 2k iterations. After 8k iterations, the learning rate is fixed to $2 \times 10^{-5}$. During training phase, we randomly crop the images into size $200 \times 200$ for augmenting the training data. Following [13], [16], stereo image pairs with Lab color space is used as input, and the Lab color space information is zoomed by multiplying a coefficient $\beta = \eta \max(m_w/n_w, m_h/n_h)$, where $m$ and $n$ represent the number of superpixels and pixels, $\eta$ is equal to 2.5. All experiments are implemented by PyTorch framework on a PC with NVIDIA RTX A4000 GPU.

**Datasets.** Following the experiment settings in [1], we use KITTI2015 [22] and Cityscapes [23] datasets to train and test our model. KITTI2015 contains 200 stereo image pairs with semantic annotations of left images, we select 150 for training and 50 for testing. Moreover, to further indicate the superiority of the proposed method, we also use the Cityscapes dataset for evaluation. Cityscapes is a larger and more challenging dataset, which contains extensive stereo image pairs captured with diverse scenes, weathers and illumination conditions. Since the test set of Cityscapes is not public available, we use the validation set for comparing, which is consist of 500 stereo images. Furthermore, the image of Cityscapes has been scaled to quarter-resolution for convenience.

**Evaluation Metrics.** In our experiments, we use three widely used metrics to evaluate the performance of our model, which are achievable segmentation accuracy (ASA), undersegmentation error (UE), and boundary recall (BR). For superpixel map $S = \{S_i\}$ and ground truth of semantic label $G = \{G_j\}$, The detailed definitions of these metrics are as follows:

*Achievable segmentation accuracy (ASA):* ASA is a metric for evaluating the upper bound on the achievable segmentation accuracy, which can be formulated as:

$$ASA(S, G) = \frac{\sum_i max_j |S_i \cap G_j|}{\sum_j |G_j|}. \tag{12}$$

*Undersegmentaion error (UE):* UE essentially measures the error of superpixel segmentation with respect to the ground

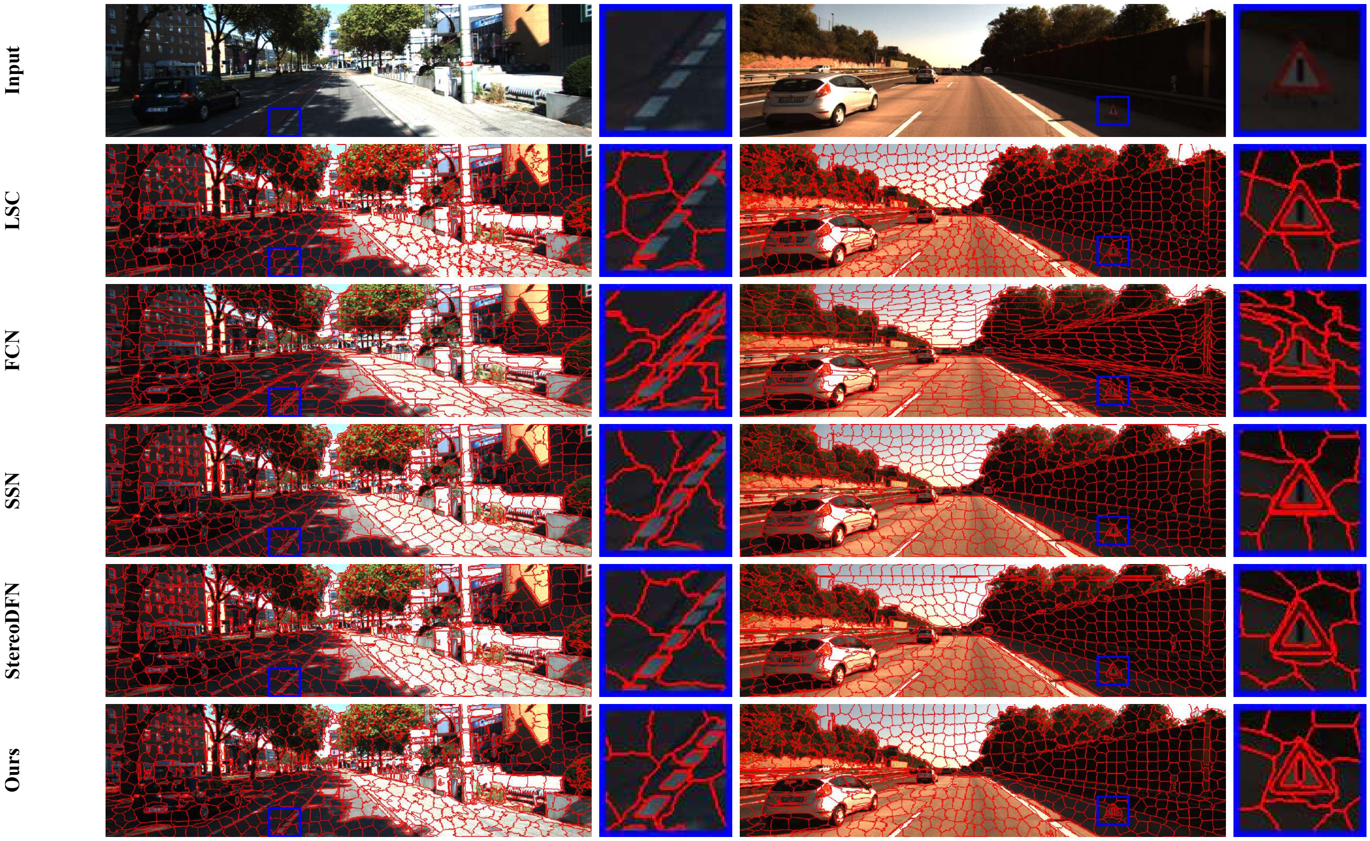


Fig. 10: Qualitative comparison of the proposed method and other state-of-the-art methods in KITTI2015.

truth. The UE is defined as:

$$UE(S,G) = \frac{1}{|G|}\sum_{G_j}\frac{(\sum_{S_i \cap G_j}|S_i|) - |G_j|}{|G_j|}. \tag{13}$$

*Boundary recall (BR):* This is a metric of how well the superpixel adhere to image boundaries. We use a coefficient $r$ to divide all pixels into two categories. $TP(S,G)$ is the boundary pixels in $G$ for which there is a boundary pixel in $S$ in range $r$, and $FN(S,G)$ is the opposite of it. Then BR can be formulated as:

$$BR(S,G) = \frac{TP(S,G)}{TP(S,G) + FN(S,G)}. \tag{14}$$

### *B. Comparison with State-of-the-Arts*

In this part, we compare our method with some state-of-the-art methods, including SSN [13], FCN [14], LSC [24], StereoDFN [1]. All of the compared methods are adopted the parameters setting of original works and implemented by the code released by [21] or the original authors.

**Quantitative Comparison.** Fig. 9(a) and Fig. 9(b) shows the quantitative comparison results of our proposed method and other state-of-the-art methods on KITTI2015 and Cityscapes, respectively. We can see that our method achieves the top score on KITTI2015 and comparaZble performance to FCN and StereoDFN on Cityscapes. Taking 700 superpixels for example, in terms of ASA, UE and BR, our method achieves the minimum percentage gain (computed with the highest score of other methods) on KITTI2015 is 0.4%, 9.3%, 1.1%, while the maximum percentage gain is 1.4%, 33.8%, 4.8%, respectively. On Cityscapes, our method achieves minimum and maximum percentage gain of BR is 2.4% and 9.8%, respectively, and also achieves a comparable performance to FCN and StereoDFN in terms of ASA and UE.

**Qualitative Comparison.** As the qualitative comparison results shown in Fig. 10 and Fig. 11, it is clear that our method achieves the best visual performance since it can adhere to object boundaries and preserve texture better. More specifically, on KITTI2015, we can see that our method can adhere to the boundary of various lane lines more accurately and capture the detail of the warning sign in low light while other methods cannot. For Cityscapes, only our method can capture the details on the warning sign and green light on the traffic light while adhering the image boundaries well.

In conclusion, through the quantitative comparison results based on standard evaluation criteria, we can see that our method outperforms other methods in most cases, and achieves the best visual performance in qualitative comparison. The impressive performance of our method also verifies the superiority of the proposed spatial decoupling mechanism.

### *C. Ablation Experiments*

In order to validate the effectiveness of each component in our proposed network, we perform extensive ablation experiments on KITTI2015. There are three types of ablation

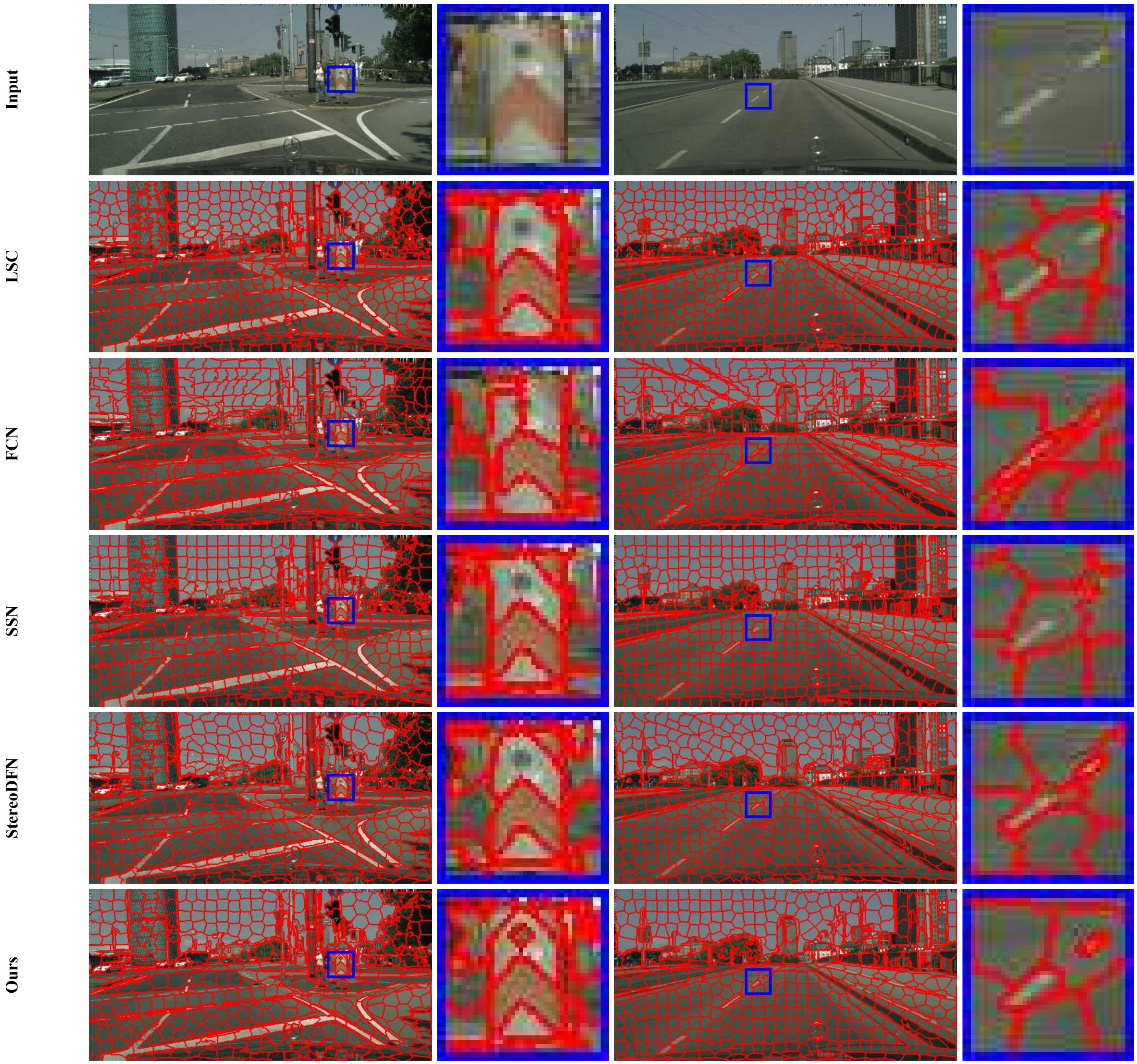


Fig. 11: Qualitative comparison of the proposed method and other state-of-the-art methods in Cityscapes.

models including T0, T1 and T2, T0 denotes the full model. For baselines in T1, they can indicate our DSEM can combine the spatial information better, while baselines in T2 show that our method makes a good use of information from another viewpoint and improve the performance of superpixel segmentation. In addition, T1 and T2 contain 3 ablation models respectively, More specifically,

- B1 denotes the ablation model without SE and DF modules.
- B2 stands for the ablation model with spatial information (XY) without SE and DF modules.
- B3 represents the ablation model without DF module.
- B4 means that the ablation model without stereo loss, and does not consider stereo features alignment and occlusion problem.
- B5 refers to the ablation model without considering occlusion problem.
- B6 is the ablation model without stereo loss

All of the ablation experiments are trained for 20K iterations. The specific structure of each ablation model also has been shown in TABLE I.

**Effectiveness of Each Component.** Fig. 13 reports the quantitative comparison results of ablation models on KITTI2015. We can see that adding spatial information directly can not make full use of it. However, the spatial information can be embedded into the network better through our SE and DF modules, resulting in higher performance gains. In addition, the DSFM module and Stereo Loss also play an

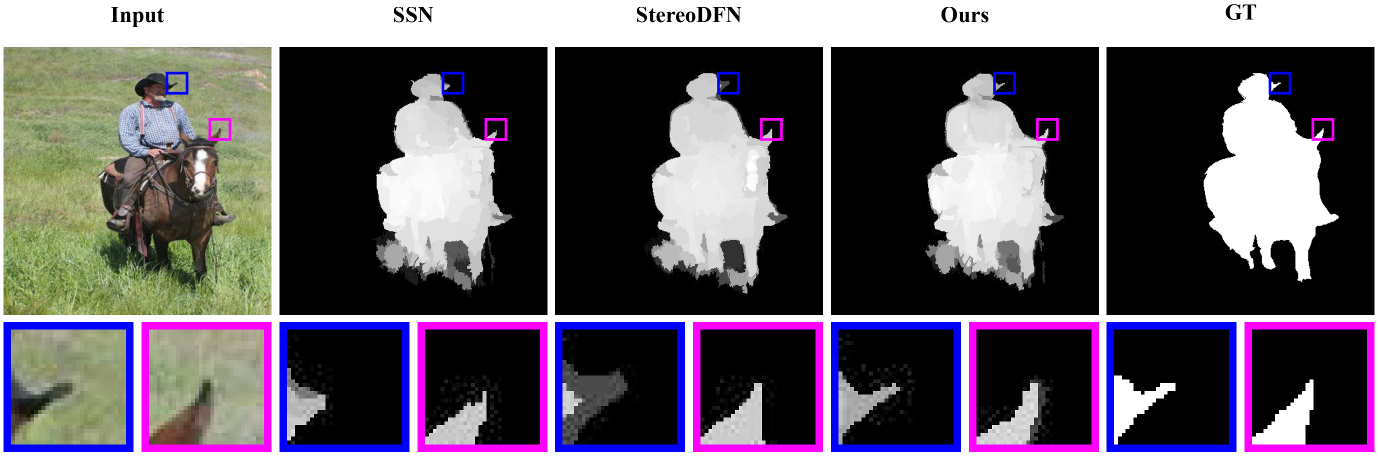


Fig. 12: Visual comparison of SOD results with different superpixel segmentation methods. Note that our method can preserve more details than others.

| Type | ID | Input | Component | | | | Stereo Loss |
|---|---|---|---|---|---|---|---|
| | | | SFA | OH | SE | DF | |
| T0 | B0 | Stereo | ✓ | ✓ | ✓ | ✓ | ✓ |
| T1 | B1 | Stereo | ✓ | ✓ | | | ✓ |
| | B2 | Stereo+XY | ✓ | ✓ | | | ✓ |
| | B3 | Stereo | ✓ | ✓ | ✓ | | ✓ |
| T2 | B4 | Single | | | ✓ | ✓ | |
| | B5 | Stereo | ✓ | | ✓ | ✓ | ✓ |
| | B6 | Stereo | ✓ | ✓ | ✓ | ✓ | |

TABLE I: Detailed setup for ablation experiments. B$n$ and T$n$ denote the $n$th baseline and the $n$th type respectively.

important role, which can solve the stereo features alignment and occlusion problem and constrain the model to correctly model stereo correspondence, respectively.

**Influence of Spatial Information.** From Fig. 13, we can observe that model with SE module tends to have a larger performance improvement than the model without SE module, which proves that the spatial information is helpful to generate regular and compact superpixels. Furthermore, adjusting the weighting of the spatial information adaptively through the DF module can make better use of it to further improve the performance.

## V. Application on Salient Object Detection

Salient object detection (SOD) has attracted increasing interest in recent years, since it plays a significant role in many popular computer vision tasks, including object recognition and detection [25], [26], image retargeting [27], [28], semantic segmentation [29], [30], etc. To improve the performance of salient object detection, Zhu *et al.* [31] propose an superpixel-based salient object detection method, they treat the saliency object detection problem as a saliency value optimization problem for all superpixels in an image. Moreover, they observe that background regions are more connected to image boundaries than salient object regions. Therefore, they propose a measure called boundary connectivity, which is utilized to perform salient object detection in their proposed method. The boundary connectivity is defined as follows:

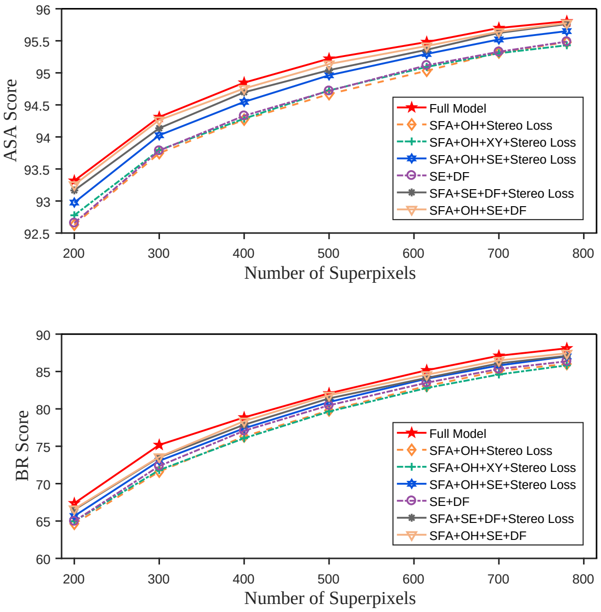


Fig. 13: **Ablation studies on KITTI2015.** The top figure shows the contributions of each component through ASA score, while the bottom figure through BR score.

$$BndCon(R) = \frac{|\{p \mid p \in R, p \in Bnd\}|}{\sqrt{|\{p \mid p \in R\}|}}, \tag{15}$$

where $p$ is a patch of an image and $Bnd$ is the set of all image boundary patches.

To indicate our method can perform better in downstream task, we use three state-of-the-art methods, which are our proposed method, StereoDFN [1] and SSN [13] to replace the default SLIC [16] as the superpixel segmentation method of [31]. In our experiments, we use NJU2K [32] for evaluation. NJU2K is a large dataset widely used for salient object detection of stereo images, which contains 2000 stereo image

| Method | SSN | StereoDFN | Ours |
|---|---|---|---|
| MAE ↓ | 0.1783 | 0.1794 | **0.1777** |
| E-measure ↑ | 0.6489 | 0.6427 | **0.6498** |

TABLE II: Results on NJU2K benchmark. ↑ denotes the higher is the better, and ↓ is contrary.

pairs, involving various objects and scenarios of different difficulty levels. Moreover, we select the first 400 images of NJU2K and resize all of them to size $400 \times 400$ for the ease of experimentation.

Following [31], we choose the mean absolute error (MAE) [33] to evaluate each method quantitatively, which is a metric to measure the average difference between the binary ground truth and saliency prediction map. However, MAE only focus on pixel-wise error. To consider structure cues, we also introduce Enhanced-alignment measure (E-measure) [34] as our evaluation metric.

The results of the quantitative evaluation are shown in Table II, we can see that our method achieves the best performance in terms of MAE and E-measure. In addition, the visual comparison results in Fig. 12 also illustrate the saliency map generated based on our method can focus on more details than other state-of-the-art methods, which validates that our method can perform well in downstream task both qualitatively and quantitatively.

## VI. Conclusion

Previously, stereo superpixel segmentation methods neglect the coupling of stereo features and spatial features, which may impose strong constraints on spatial information while modeling the correspondence between stereo image pairs. To solve this problem, we have presented an end-to-end stereo superpixel segmentation network with a decoupling mechanism of spatial information to eliminate such negative influence. In addition, spatial information is adjusted adaptively through our dynamic fusion mechanism in dynamic spatiality embedding module to generate regular and compact superpixels. Extensive experiments on several popular datasets have shown that our proposed method achieves the state-of-the-art performance and performs well in downstream task. The effectiveness of the components handling the spatial information and stereo features have also been verified in our ablation studies.

## References

[1] R. Wu, Y. Du, H. Li, and Y. Dai, "Stereo superpixel segmentation via dual-attention fusion networks," in *Proc. Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[2] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, "Going from rgb to rgbd saliency: A depth-guided transformation model," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3627–3639, 2020.

[3] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4819–4831, 2019.

[4] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 568–579, 2018.

[5] M. Xu, B. Liu, P. Fu, J. Li, and Y. H. Hu, "Video saliency detection via graph clustering with motion energy and spatiotemporal objectness," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2790–2805, 2019.

[6] M. Yang, J. Liu, and Z. Li, "Superpixel-based single nighttime image haze removal," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3008–3018, 2018.

[7] C. Shi and C.-M. Pun, "Multiscale superpixel-based hyperspectral image classification using recurrent neural networks with stacked autoencoders," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 487–501, 2019.

[8] C. Wang, Z. Liu, and S.-C. Chan, "Superpixel-based hand gesture recognition with kinect depth camera," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 29–39, 2014.

[9] X. Dong, J. Han, D. Chen, J. Liu, H. Bian, Z. Ma, H. Li, X. Wang, W. Zhang, and N. Yu, "Robust superpixel-guided attentional adversarial attack," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 12 892–12 901.

[10] H. Li, R. Cong, S. Kwong, C. Chen, Q. Xu, and C. Li, "Stereo superpixel: An iterative framework based on parallax consistency and collaborative optimization," *Inf. Sci.*, vol. 556, pp. 209–222, 2021.

[11] F. Cheng, H. Zhang, M. Sun, and D. Yuan, "Cross-trees, edge and superpixel priors-based cost aggregation for stereo matching," *Pattern Recognit.*, vol. 48, no. 7, pp. 2269–2278, 2015.

[12] L. Liu, W. Tao, and H. Liu, "Complementary saliency driven co-segmentation with region searching and hierarchical constraint," *Inf. Sci.*, vol. 372, pp. 72–83, 2016.

[13] V. Jampani, D. Sun, M. Liu, M. Yang, and J. Kautz, "Superpixel sampling networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 363–380.

[14] F. Yang, Q. Sun, H. Jin, and Z. Zhou, "Superpixel segmentation with fully convolutional networks," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 13 961–13 970.

[15] Ren and Malik, "Learning a classification model for segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2003, pp. 10–17 vol.1.

[16] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[17] J. Chen, Z. Li, and B. Huang, "Linear spectral clustering superpixel," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3317–3330, 2017.

[18] H. Li, S. Kwong, C. Chen, Y. Jia, and R. Cong, "Superpixel segmentation based on square-wise asymmetric partition and structural approximation," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2625–2637, 2019.

[19] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning parallax attention for stereo image super-resolution," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 12 242–12 251.

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020.

[21] M. Wang, X. Liu, Y. Gao, X. Ma, and S. Nouman, "Superpixel segmentation: A benchmark," *Signal Process.-Image Commun.*, vol. 56, pp. 28–39, 2017.

[22] M. Menze, C. Heipke, and A. Geiger, "Joint 3d estimation of vehicles and scene flow," in *Proc. ISPRS Workshop Image Sequence Anal.*, 2015.

[23] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3213–3223.

[24] J. Chen, Z. Li, and B. Huang, "Linear spectral clustering superpixel," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3317–3330, 2017.

[25] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 769–779, 2013.

[26] D. Zhang, D. Meng, L. Zhao, and J. Han, "Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016.

[27] Y. Ding, J. Xiao, and J. Yu, "Importance filtering for image retargeting," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 89–96.

[28] J. Sun and H. Ling, "Scale and object aware image retargeting for thumbnail browsing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1511–1518.

[29] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, 2017.

[30] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1354–1362.

[31] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 2814–2821.

[32] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Process.-Image Commun.*, vol. 38, pp. 115–126, 2015.

[33] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 733–740.

[34] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. Int. Joint Conf. Artif. Intell.*, 7 2018, pp. 698–704.