Stereo Superpixel Segmentation via Decoupled Dynamic Spatial-Embedding Fusion Network

Hua Li, Junyan Liang, Ruiqi Wu, Graduate Student Member, IEEE, Runmin Cong, Wenhui Wu, and 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 designed 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 [1], [2], [3], [4], image dehazing [5], image classification [6], object recognition [7], visual tracking [8], [9], adversarial attack [10]. 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 [11]. However, how to effectively utilize complementary and correspondence information to generate superpixels for stereo images is still a challenging task.

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 [12], [13]. 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. [11] 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. [14] 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 by the parallax attention module to integrate...
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 tasks.

II. RELATED WORK

The concept of “Superpixel” is first introduced in [17], 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. In addition, considering that some methods like StereoDFN and our method utilize the attention mechanism, some representative attention mechanisms are also introduced.

A. Superpixel Segmentation Methods

Unsupervised Methods: Simple linear iterative clustering (SLIC) [18] 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) [19] 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. [20] 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.

Supervised 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. [15] 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. [16] propose a lightweight fully convolutional networks (FCN) that
Fig. 2. Overall framework of the proposed method. The deep feature of stereo image pair is extracted by feature extractor, then the coupling of spatial and stereo feature is relaxed by DSFM. Finally, the spatial information is re-added by the DSEM and the superpixel segmentation result of the main view is generated by the soft clustering algorithm.

Based on encoder-decoder structure, which generates superpixels efficiently by predicting the probability map between pixels and superpixels. More recently, Wu et al. propose a 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.

### B. Attention Mechanisms

As is well known, attention mechanism acts a significant role in various computer vision tasks. Hu et al. [21] propose the channel attention mechanism, which utilizes the squeeze and excitation operations to perform channel-wise attention. Considering the lack of spatial-wise attention may produce suboptimal result, Woo et al. [22] propose convolutional block attention module (CBAM), the critical features have been emphasized along channel and spatial dimensions. In order to apply attention mechanism to stereo image, Wang et al. [23] propose parallax attention mechanism (PAM) to model the correspondence between left and right views.

### III. PRELIMINARIES

Aligning the stereo features is necessary before fusion. In our method, the features of stereo image pair are aligned based on the parallax attention module [23]. As the schematic illustration shown in Fig. 3, for a pair of deep features $F_L$ and $F_R \in \mathbb{R}^{H \times W \times C}$, we can get $A$ and $B$ from a convolution layer with $1 \times 1$ kernel size. Then, the parallax attention map $\mathcal{M} \in \mathbb{R}^{H \times W \times W}$ will be generated by:

$$\mathcal{M}_{R \rightarrow L} = \text{softmax}(A \odot B^T),$$  

$$\mathcal{M}_{L \rightarrow R} = \text{softmax}(B \odot A^T),$$

where $\odot$ denotes the batch-wise multiplication, and $T$ denotes the batch-wise transposition. The parallax attention map $\mathcal{M}_{L \rightarrow R}(i, k, j)$ represents the contribution of the position $(i, j)$ in left view to position $(i, k)$ in right view, while $\mathcal{M}_{R \rightarrow L}(i, j, k)$ represents the contribution of the position $(i, k)$ in right view to position $(i, j)$ in left view. Moreover, the parallax attention map also encode disparity information to help the stereo feature alignment, which can be formulated as follows:

$$\hat{F}_L = \mathcal{M}_{R \rightarrow L} \odot F_{\text{right}},$$  

$$\hat{F}_R = \mathcal{M}_{L \rightarrow R} \odot F_{\text{left}},$$

where $\hat{F}_L$ and $\hat{F}_R$ denote the aligned features, $F_{\text{left}}$ and $F_{\text{right}}$ are deep features of left and right view, respectively.

In addition, occlusion always exists in stereo image pairs due to violent disparity variation, which will lead to an inaccurate stereo features fusion. Therefore, the occlusion handling part in PAM is utilized to take occlusion problem into consideration.
Fig. 4. Schematic illustration of Feature Extractor.

More specifically, the valid mask $O_{L\rightarrow R}$ in PAM is used to handle the occluded regions in left view, which can be formulated as (5):

$$O_{L\rightarrow R}(i,j) = \begin{cases} 1, & \sum_{k\in[1,W]} M_{L\rightarrow R}(i,k,j) > \tau \\ 0, & \sum_{k\in[1,W]} M_{L\rightarrow R}(i,k,j) \leq \tau \end{cases}, \quad (5)$$

where $\tau$ is a threshold set to 0.1. More details regarding the PAM can be found in [23].

IV. 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 [15] 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. 4.

Fig. 5. Differences between our method and StereoDFN. We only use Lab color information as input to decouple the stereo features and spatial features. After Decoupled Stereo Fusion Module (DSFM), we design a Dynamic Spatiality Embedding Module (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 the parallax attention module [23], and also use the proposed spatial decoupling mechanism to optimize the process of stereo feature fusion.

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 and stereo channel attention module 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. 5. Furthermore, a simple case of visualization of deep feature has been shown in Fig. 6. We can see that benefiting from decoupling
stereo features and spatial features, our method eliminates the interference of spatial information on modeling, and the boundary information is much clearer and finer.

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 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 (6):

\[
\hat{X} = \frac{X}{\max(X)}, \quad \hat{Y} = \frac{Y}{\max(Y)},
\]

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 \( \hat{F}_X, \hat{F}_Y \). 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, \( \hat{F}_X, \hat{F}_Y \) is concatenated with input features and send them to dynamic fusion part. Fig. 7 simply 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.

The dynamic fusion mechanism employs a channel-attention [21] 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:

\[
W = g \cdot \sigma(C(ReLU(LN(C(g))))),
\]

where \( W \) 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. 2 presents the details of the Dynamic Fusion (DF) mechanism.

D. Loss Functions

The same with StereoDFN [14], two loss functions are used to optimize 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:

\[
\mathcal{L}_{sem} = CE(S, S^*),
\]

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:

\[
\mathcal{L}_{stereo} = ||O_{L \rightarrow R} \circ (I_L - \mathcal{M}_{R \rightarrow L} \otimes I_R)||_1 + ||O_{R \rightarrow L} \circ (I_R - \mathcal{M}_{L \rightarrow R} \otimes I_L)||_1,
\]

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:

\[
\mathcal{L}_{total} = \mathcal{L}_{sem} + \lambda \mathcal{L}_{stereo},
\]

where \( \lambda \) is set to 1.0 for balancing the scales of different losses, the influence of the value of \( \lambda \) can be found in Section V-C.
V. 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 20 K 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 2 k iterations. During training phase, we randomly crop the images into size 200 $\times$ 200 for augmenting the training data. Following [15], [18], 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 [14], we use KITTI2015 [24] and Cityscapes [25] 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:
  $$ASA(S, G) = \sum_i \max_j |S_i \cap G_j| / \sum_j |G_j|.$$  \hspace{1cm} (11)

- Undersegmentation error (UE): UE essentially measures the error of superpixel segmentation with respect to the ground truth. The UE is defined as:
  $$UE(S, G) = \frac{1}{|G|} \sum_{G_j} \frac{| \sum_i \mathbf{1}(S_i \cap G_j) | |S_i| - |G_j|}{|G_j|}.$$  \hspace{1cm} (12)

- 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 $r$, $FN(S, G)$ is the boundary pixels in $G$ for which there is no boundary pixel in $S$ in range $r$.
Fig. 9. Qualitative comparison of the proposed method and other state-of-the-art methods in KITTI2015. The first column is the superpixel segmentation result, and the second column is the display of segmentation details. Only our method can capture the detail of the warning sign.

and $FN(S, G)$ is the opposite of it. Then BR can be formulated as:

$$BR(S, G) = \frac{TP(S, G)}{TP(S, G) + FN(S, G)}$$  \hspace{1cm} (13)

All above are the basic definitions of the metrics we used for evaluation, more details can be seen in [26].

B. Comparison With State-of-The-Arts

In this part, we compare our method with some state-of-the-art methods, including SSN [15], FCN [16], LSC [19], SLIC [18], SNIC [27], StereoDFN [14]. All of the compared methods are adopted the parameters setting of original works and implemented by the code released by [26] or the original authors.

Quantitative Comparison: Fig. 8(a) and (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 comparable 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 2.98%, 65.8%, 26.1%, respectively. On Cityscapes, our method achieves minimum and maximum percentage gain of BR is 2.4% and 29%, 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 Figs. 9 and 10, 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
Fig. 10. Qualitative comparison of the proposed method and other state-of-the-art methods in Cityscapes. The first column is the superpixel segmentation result, and the second column is the display of segmentation details. Only our method can capture the detail of the warning sign.

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. Parameter Selection

To explore the influence of the weighting of \( \lambda \) on the performance of model, we conduct parameter selection experiments on KITTI2015. More specifically, the value of \( \lambda \) is set to 0.5, 0.75,
1.0, 1.25, 1.5, respectively. Then the model is trained based on the loss function of different values, the quantitative comparison results of each model has been shown in Fig. 11. Compared to the results with $\lambda$ set to 1.0, we can see that when the value of $\lambda$ is 0.5 or 1.5, the scores of ASA and BR are decreased, and when the value of $\lambda$ is 0.75 or 1.25, the scores of ASA and BR are decreased slightly, which shows that semantic loss and stereo loss are nearly the same important. Therefore, the value of $\lambda$ is finally set to 1.0.

**Table I**

| Type | ID | Input | Component | Loss | SCAM |
|------|----|-------|-----------|------|------|
| T0   | B0 | Stereo| SFA, OH, SE, DF | ✓    | ✓    |
| T1   | B1 | Stereo| SFA+OH+XY+Stereo Loss | ✓    | ✓    |
| T2   | B3 | Stereo| SFA+OH+SE+Stereo Loss | ✓    | ✓    |
| T3   | B7 | Stereo| SFA+OH+SE+OF | ✓    | ✓    |

**Fig. 11.** Parameter studies on KITTI2015. The top figure shows the influence of parameter $\lambda$ on ASA score, while the bottom figure on BR score.

**Fig. 12.** Ablation studies on KITTI2015. The top figure shows the contributions of each component through ASA score, while the bottom figure through BR score.

**D. 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 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. Besides, to verify whether removing SCAM in SFM has negative influence, we also add an additional type T4, which adds SCAM in the full model. 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.
- B7 is the full model with SCAM.

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

**Effectiveness of Each Component:** Fig. 12 reports the quantitative comparison results of ablation models on KITTI2015. We can see that adding spatial information directly is difficult to 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 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. 12, 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.
Influence of Stereo Channel Attention Module: StereoDFN use the SCAM to augment the stereo features by emphasizing informative representations. However, to relax the constraints on spatial information, the SCAM is removed in our full model. As the result of B7 shown in Fig. 12, we can see that when adding SCAM, the performance decreased.

VI. 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 [28], [29], image retargeting [30], [31], semantic segmentation [32], [33], etc. To improve the performance of salient object detection, Zhu et al. [34] 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:

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

(14)

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 [14] and SSN [15] to replace the default SLIC [18] as the superpixel segmentation method of [34]. In our experiments, we use NJU2K [35] for evaluation. NJU2K is a large dataset widely used for salient object detection of stereo images, which contains 2000 stereo image pairs, involving various objects and scenarios of different difficulty levels. In addition, we use the first 400 images of NJU2K and resize all of them to size $400 \times 400$ for the ease of experimentation.

Following [34], we choose the mean absolute error (MAE) [36] to evaluate each method quantitatively, which is a metric to measure the average difference between the binary ground truth and saliency prediction map. The MAE can be formulated as follows:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |SM(x, y) - GT(x, y)|,$$

(15)

where $SM$ and $GT$ are the saliency map and ground truth image, respectively. However, MAE only focus on pixel-wise error.

To consider structure cues, we also introduce Enhanced-alignment measure (E-measure) [37] as our evaluation metric, which can be formulated as follows:

$$EM = \frac{1}{W \times H} \sum_{x=1}^{w} \sum_{y=1}^{h} \phi_{FM}(x, y),$$

(16)

where $\phi_{FM}$ is the enhanced alignment matrix computed from saliency map and ground truth image. 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. Moreover, the visual comparison results in Fig. 13 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.

VII. 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. Cong et al., “Going from RGB to RGBD saliency: A depth-guided transformation model,” IEEE Trans. Cybern., vol. 50, no. 8, pp. 3627–3639, Aug. 2020.
[2] R. Cong et al., “Video saliency detection via sparsity-based reconstruction and propagation,” IEEE Trans. Image Process., vol. 28, no. 10, pp. 4819–4831, Oct. 2019.
[3] R. Cong et al., “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, Feb. 2018.
[4] 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, Nov. 2019.
[5] M. Yang, J. Liu, and Z. Li, “Superpixel-based single nighttime image haze removal,” IEEE Trans. Multimedia, vol. 20, no. 11, pp. 3008–3018, Nov. 2018.
[6] 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, Feb. 2020.
[7] 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, Jan. 2015.
[8] Z. Chen et al., “Siamban: Target-aware tracking with siamese box adaptive network,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 4, pp. 5158–5173, Apr. 2023.
[9] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, “Siamese box adaptive network for visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6667–6676.
[10] X. Dong et al., “Robust superpixel-guided attentional adversarial attack,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 12892–12901.
[11] H. Li et al., “Stereo superpixel: An iterative framework based on parallax consistency and collaborative optimization,” Inf. Sci., vol. 556, pp. 209–222, 2021.
[12] 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.
[13] 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.
[14] R. Wu, Y. Du, H. Li, and Y. Dai, “Stereo superpixel segmentation via dual-attention fusion networks,” in Proc. IEEE Int. Conf. Multimedia Expo, 2021, pp. 1–6.
[15] V. Jampuni, D. Sun, M. Liu, M. Yang, and J. Kautz, “Superpixel sampling networks,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 363–380.
[16] F. Yang, Q. Sun, H. Jin, and Z. Zhou, “Superpixel segmentation with fully convolutional networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 13961–13970.
[17] Ren and Malik. “Learning a classification model for segmentation,” in Proc. Int. Conf. Comput. Vis., 2003, vol. 1, pp. 10–17.
[18] R. Achanta et al., “SLIC superpixels compared to state-of-the-art superpixel methods,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[19] J. Chen, Z. Li, and B. Huang, “Linear spectral clustering superpixel,” IEEE Trans. Image Process., vol. 26, no. 7, pp. 3317–3330, Jul. 2017.
[20] 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, Oct. 2019.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Ruiqi Wu (Graduate Student Member, IEEE) received the B.S. degree from the School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China. His research interests including computer vision and machine learning, especially on image enhancement.

Runmin Cong (IEEE Senior Member) received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2019. He is currently a Professor with the School of Control Science and Engineering, Shandong University (SDU), Jinan, China. Before joining SDU, he was an Associate Professor at the Institute of Information Science, Beijing Jiaotong University, Beijing, China. He was a Visiting Student/Staff at Nanyang Technological University (NTU), Singapore, and City University of Hong Kong (CityU), Hong Kong. His research interests include computer vision, multimedia understanding, content enhancement, and machine learning. He has authored or coauthored more than 70 papers in prestigious international journals and conferences, which include 2 ESI hot papers (Top 0.1%), 11 ESI highly cited papers (Top 1%), Dr. Cong was the recipient of the Young Elite Scientist Sponsorship Program by the China Association for Science and Technology, Beijing Nova Program, Hong Kong Scholars Program, IEEE ICME Best Student Paper Runner-Up Award, ACM SIGWEB China Rising Star Award, First Prize for Scientific and Technological Progress Award of Tianjin Municipality, IEEE CVPR 2021 Outstanding Reviewer, 2021 IEEE Journal of Oceanic Engineering Outstanding Reviewer, Excellent Doctoral Dissertation Award from China Society of Image and Graphics (CSIG), and Excellent Scientific Paper Award for Beijing Youth. He is an Associate Editor for the Engineering Applications of Artificial Intelligence, Neurocomputing, IEEE Journal of Oceanic Engineering, Signal, Image and Video Processing, the Youth Editorial Board Member of the CAAI Transactions on Intelligence Technology, Guest Editor of IEEE Journal of Oceanic Engineering, Drones, Signal Processing: Image Communication, Multimedia Tools and Applications, and the Area Chair/SPC/PC Member of NeurIPS, CVPR, ICML, ICLR, ICCV, ECCV, SIGGRAPH, ACM MM, AAAI, IJCAI, ICME, ICPR, and ICMR.

Wenhui Wu received the B.S. and M.S. degrees from Xidian University, Xian, China, in 2012 and 2015, respectively, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, China, in 2019. She is currently an Assistant Professor with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. Her research interests include machine learning, image enhancement, and community detection.

Sam Tak Wu Kwong (Fellow, IEEE) received the B.S. degree in electrical engineering from the State University of New York, Buffalo, NY, USA, in 1983, and the M.S. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree from the University of Hagen, Hagen, Germany, in 1996. He is currently a Chair Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong. His research interests include video coding, pattern recognition, and evolutionary algorithms. He was elevated to an IEEE Fellow for his contributions on optimization techniques for cybernetics and video coding in 2014.