Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13487-7

1221: DEEP LEARNING FOR IMAGE/VIDEO COMPRESSION AND VISUAL QUALITY ASSESSMENT

Deep stereoscopic image saliency inspired stereoscopic image thumbnail generation

Yu Zhou[1] · Xiaotong Xiao[1] · Qiudan Zhang[2] · Xu Wang[1] · Jianmin Jiang[1]

Received: 30 June 2021 / Revised: 27 September 2021 / Accepted: 13 July 2022

Abstract
In this paper, we propose a stereoscopic image thumbnail generation method guided by the stereoscopic image saliency. Specifically, we utilize an uncertain-weighted fusion mechanism to combine the spatial saliency information with the saliency driven by depth cues, generating the dense stereoscopic saliency fixation map. Subsequently, the obtained dense fixation map is converted into a salient object map through a saliency optimization module, which provides the object-level saliency cues for the thumbnail generation task. Under the guidance of the obtained salient object map, a cropping window is employed to cut out the most salient region and generate the stereoscopic thumbnails, such that the disparity distribution of the original image can be well preserved, and avoid sharply deforming certain structured objects in the subsequent warping operation. Finally, the warping operation is utilized to adjust the aspect ratio of the stereoscopic thumbnail to the target size. Qualitative and quantitative results demonstrate that our proposed method achieves superior performance than the state-of-the-art benchmarks on the public datasets.

Keywords Stereoscopic saliency detection · Stereoscopic thumbnail generation · Energy minimization · Uncertain weighted fusion

1 Introduction

With the rapid growth of multimedia data, especially the 3D scene, there is an increasing demand for browsing the content quickly and briefly. Image display and quick browsing of VR, AR and other devices make stereoscopic thumbnail generation technology crucial. Stereoscopic thumbnail generation aims at producing thumbnails from a pair of stereoscopic images that represent the basic context and structural information in the original large

✉ Xu Wang
wangxu@szu.edu.cn

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, China

[2] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

stereoscopic images [33]. It plays an essential role in many applications including video browsing, image retrieval and image coding.

Thumbnail generation often needs to be guided by saliency maps that highlight the significant areas of the image. Therefore, an accuracy saliency map is crucial to this task. For example, in a simple image browsing and management system, thumbnail generation is usually processed by directly scaling the original image to achieve the target size. However, this can easily have a significantly negative impact on structured objects when the aspect ratio changes dramatically. Such rough operation usually causes important objects in the images to be excessively squeezed or stretched, which greatly affects the viewing experience of users. Moreover, there still exist several limitations in thumbnail generation technology, for example, over-compressed image and straight-line bending. In addition, some of the existing methods [8, 30, 33] cut too much background, leading poor visual experience with big salient object. Therefore, automatic generation of thumbnails with better visual perception is very urgent for the community.

Compared with 2D thumbnail generation, stereoscopic thumbnail generation is more challenging due to the additional requirements, including removing 3D visual discomfort and maintaining stereo perception. Most of the existing methods for stereoscopic thumbnail generation [29, 30, 33] did not carefully maintain the completeness of the objects in the scene perceived in front of the screen, leading to window conflict problem. Besides, the retinal rivalry also occurs due to without considering the monocular object constraint in these methods. A recent study [16] straightforwardly applies traditional 2D thumbnail method to the left and right views of the stereoscopic image, which may destroy the parallax distribution between left and right views of stereoscopic images. Moreover, the existing methods only utilize low-level hand-crafted features to generate the stereoscopic thumbnails, leading to the ignorance of high-level heuristic semantic information. In addition, the quality of depth map could also be another main concern, which may not be satisfactory in general. For image thumbnail generation, embedding accurate saliency cues that represent the significant regions in a scene is capable of improving the visual quality of thumbnails. However, the existing approaches on stereoscopic image saliency detection did not fully leverage the binocular characteristics of stereoscopic images. Therefore, it is urgent to develop an automatic stereoscopic image thumbnail generation method inspired by accurate stereoscopic saliency priors.

To tackle the above-mentioned issues, we propose a saliency inspired cropping-warping framework for stereoscopic images, enabling the stereoscopic thumbnail generation with arbitrary aspect ratio. Specifically, we first produce saliency distributions of stereoscopic images driven by the spatial and disparity correlation among left and right views. The salient object maps are subsequently generated under the optimization of the saliency distributions and a cropping operator is employed to cut off the non-salient regions at the edge of the stereoscopic image pairs. Finally, the stereoscopic thumbnail is produced by a warping operator that makes the thumbnail achieves the target aspect ratio by squeezing non-salient region.

In summary, the main contributions of this paper are listed as follows:

- We propose a novel saliency inspired cropping-warping framework for stereoscopic images, which leverages the guidance of the obtained stereoscopic saliency cues to generate stereoscopic thumbnails with arbitrary aspect ratio.
- We propose a deep learning based stereoscopic saliency detection method, which aims at producing the saliency maps by investigating the spatial and disparity correlations between the left and right views.

– We carry out several analyses to explore the effectiveness of saliency cues on stereoscopic image thumbnail generation. We believe these analyses are able to providing valuable insights to further facilitate the research of stereoscopic image thumbnail generation.

2 Related works

2.1 2D/3D thumbnail generation

Thumbnail generation can be divided into three categories, including continuous image retargeting, discrete image retargeting and thumbnail cropping. Specifically, the existing continuous retargeting works [30, 32] employed a special warping operator to resize images into arbitrary aspect ratios, meanwhile the visually prominent features are well preserved. However, it only shrinks the unimportant part of the background to achieve the target aspect ratio or resolution, and hardly modifies the important regions. For the discrete image retargeting, Avidan et al. [2] developed an image operator named seam carving that considers geometric constraints and image content at the same time when resizing the image. For thumbnail cropping, most of the existing studies [30, 33] focused on first detecting the salient regions, and subsequently finding a rectangle that contains the most important visual information only. In addition, there are several studies that generate thumbnails by combining various operators [39], such as seam carving combined with warping [35, 39], seam carving combined with cropping [22], and seam carving combined with scaling [11, 27].

The above mentioned continuous and discrete retargeting methods have their own advantages and limitations. There is no perfect retargeting operator that can generate high-quality thumbnails for input images with arbitrary aspect ratio. This is the reason for the development of multi-operator retargeting technologies. For example, for the 2D image retargeting, Rubinstein et al. [27] first proposed a multi-operator image retargeting method, which uses bidirectional warping (BDW) as a similarity measure to find the best operation sequence. Zhou et al. [41] recently regarded the multi-operator retargeting as a Markov decision process, and applied the reinforcement learning (RL) to achieve global optimization. However, there are few researches on stereoscopic image retargeting based on multi-operator. For instance, Chai et al. [5] used a warping operator to generate six-scale stereo pairs and finally produced the thumbnail pair by a cropping operator.

In recent year, the advanced deep learning tools accelerate the rapid development of thumbnail generation. However, the image retargeting task is an uncertain problem due to the uncertain way of generating the ideal retargeting results and the subjective evaluation method, which is difficult to be optimized by the supervised learning manner like other computer vision tasks such as target detection and image classification. Similarly, this also brings challenges to the objective evaluation method of the retargeting results of stereoscopic 3D images. Wang et al. [34] comprehensively considers the image quality, including the degree of geometric distortion and content loss of the retargeted image pair, and the depth perception measurement for SIR evaluation. In addition, limited-scale datasets are also one of the main factors that make it difficult for image retargeting task to be trained based on deep learning. There are few studies for image retargeting based on deep learning models. For instance, Cho et al. [9] proposed a deep learning framework learned by weak supervision manner, in which a shift map is learned during the model training stage to represent pixel-level map between origin quad and target quad. However, this work needs to fix the size of the input image, which greatly limits the practicability of this method.

Tan et al. [31] proposed a circular 2D image retargeting framework with the main idea that a ideal image retargeting model should have the ability to remap a retargeted image to be the original image. Inspired by above mentioned works, Fan et al. [13] proposed an unsupervised image retargeting network for stereoscopic image retargeting task. In that study, stereoscopic image retargeting model learns to generate retargeted results with target aspect ratio under the constraints of the inter-view correlation and disparity relationship of stereoscopic images.

The core of the existing stereoscopic image thumbnail generation technology is to incorporate the significant semantic information priors to guide the thumbnail generation. Therefore, how to leverage the characteristics of stereoscopic image and the powerful feature representative ability of deep neural networks to obtain an accurate saliency map is a main ingredient for stereoscopic image thumbnail generation.

2.2 Stereoscopic image saliency detection

Saliency detection aims to recognizing the most significant regions or objects in a scene, which plays an important role in many computer vision applications such as image automatic cropping, image segmentation, object recognition, etc. A pioneer study of saliency detection was proposed by Itti et al. [19], which produced the saliency maps by calculating the feature contrast between color, direction and intensity information. Subsequently, the saliency detection task is gradually divided into two branches, one dedicated to the location-based visual saliency detection, and the other is the object-level saliency detection. The former aims at recognizing the region-of-interest in human vision system, while the latter intends to detect the most salient objects from the scene. Actually, there is a certain connection between these two branches. As studied in [36], the salient objects of the scene can be inferred from the human eye-fixation distribution based saliency maps, providing useful guidance information for the thumbnail generation task.

The stereoscopic image includes a pair of left and right views. Limited works have been designed for predicting the location-based saliency of stereoscopic images. For instance, Fang et al. [14] utilized DCT coefficient to extract four features in terms of color, brightness, texture and depth to represent image energy, and constructed a stereoscopic saliency detection framework. Jiang et al. [20] proposed a stereoscopic 3D images saliency computation model guided by depth perception and visual comfort. Nguyen et al. [25] improved the traditional convolutional neural network to introduce different low-level information as the input of deep network, which obtained better performance for distorted stereoscopic image saliency detection. However, it is difficult to converge in the actual training process. Zhang et al. [37] designed a deep feature inspired stereoscopic image saliency detection method, which generated the saliency results by combining the saliency features in color and depth channels extracted from a pre-trained model. By considering the hierarchical contexture feature of both RGB and depth modalities, Mao et al. [24] proposed the CFPI-Net, in which the low-level and high-level integration features are fused further, solving the stereoscopic 3D images' saliency detection problem with spacial and depth information. Moreover, there are several works proposed for object-level saliency detection of stereoscopic images. For example, Cong et al. [10] proposed a depth explicit algorithm, which scored the depth map and reduced the influence of low-quality depth map on stereoscopic image saliency detection. Niu et al. [26] proposed a multi-cue-driven optimization (MCDO) algorithm to leverage depth, color and spatial cues to optimize the saliency maps obtained from the existing saliency detection methods.

However, the above methods do not fully investigate the disparity coherence between left and right views for stereoscopic image saliency detection. In this paper, our work specifically explores the disparity coherence among left and right views by a learning based manner, and then uses it to guide the stereoscopic saliency generation. Finally, the obtained stereoscopic saliency results are adopted to assist the generation of stereoscopic image thumbnails.

3 Proposed framework

The stereoscopic thumbnail generation task aims to create a stereoscopic image with user defined aspect ratio, by maintaining the original disparity distribution.

3.1 Overview

As shown in Fig. 1, our proposed saliency inspired cropping-warping framework consists of a stereoscopic image saliency detection (SISD) module and a stereoscopic thumbnail generation (STG) module. Given a stereo pair I_0 with size of $H_1 \times W_1$, desired weight ratio $r_w \in (0, 1]$ and desired height ratio $r_h \in (0, 1]$, the final generated stereoscopic thumbnail is with size of $(r_h H_1) \times (r_w W_1)$. Going through Stage a, the stereo saliency maps are generated by our proposed SISD network, which is described in detail in Sections 3.2.1–3.2.3. Subsequently, as showing in Stage b, we use the saliency optimization module, being introduced in Section 3.2.4, to obtain stereo salient object maps, which can guide the cropping operator to cut off the redundant edge area of original stereo pair. The cropping stage, generating both the cropped image and cropped salient object maps with size of $H_2 \times W_2$, where $H_2 \in [r_h H_1, H_1]$ and $W_2 \in [r_w W_1, W_1]$, is described in detail in Section 3.3.1. After that, a warping operator, described in Section 3.3.2, is applied on the cropped images with

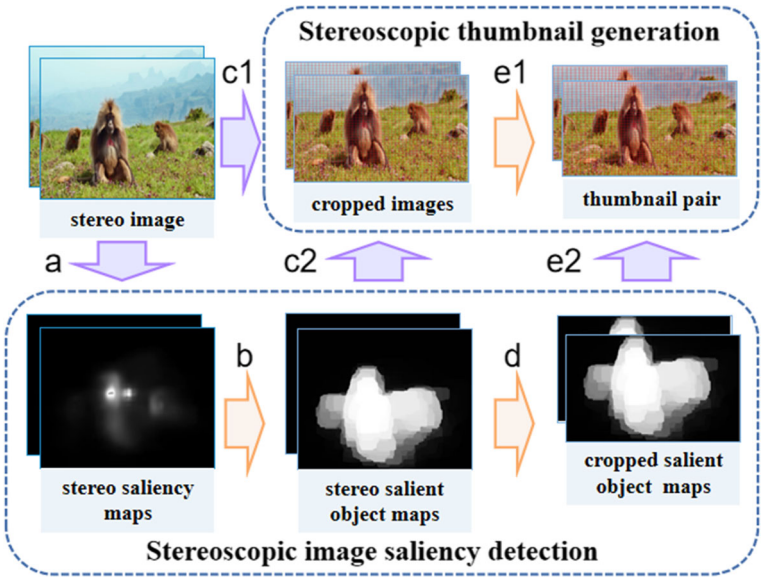

Fig. 1 The overall architecture of saliency inspired cropping-warping framework for stereoscopic images

the cropped salient object maps to generate the final thumbnail pair that meet the target size. The output of each stage in our method for four sample pictures is presented in Fig. 2. The details of each stage will be discussed in following section.

3.2 Stereoscopic image saliency detection

The stereoscopic thumbnail generation task can benefit from accurate saliency priors. As shown in Fig. 3, our proposed stereoscopic image saliency detection network consists of three parts: a spatial saliency module, a depth saliency module and an uncertain-weighted fusion module. In addition, we also implement an object salience module to obtain the refined salient object maps.

3.2.1 Spatial saliency module

The previous works [12, 18] have demonstrated that extracting image features in multi-scale pattern can improve the performance of saliency detection. Our spatial saliency module intends to learn the spatial saliency priors. We adopt the pre-trained model provided in [12] to extract spatial saliency priors, which can explore the emotional content of an image that attracts human attention. As the upper box shown in Fig. 3, the two former encoding branches extract the coarse and fine scale features f_c and f_f for the left image. The multi-scale spatial saliency feature f_{cf} is subsequently obtained by cascading these two features f_c and f_f. After the pooling and flatten operations as well as the computation of fully-connected layer, the feature f_{cf} is converted into a feature vector with 1024 dimensions, which learns the feature weights according to the spatial position and semantic information of the feature f_{cf}. The produced feature weights are used to make a weighting operation on each channel of the obtained multi-scale feature f_{cf}, highlighting the areas that cause the observer to experience emotional fluctuations. Finally, the spatial prior map P_L (or P_R) is produced for the left (or right) image by rescaling the resulting map of weighting to original size directly.

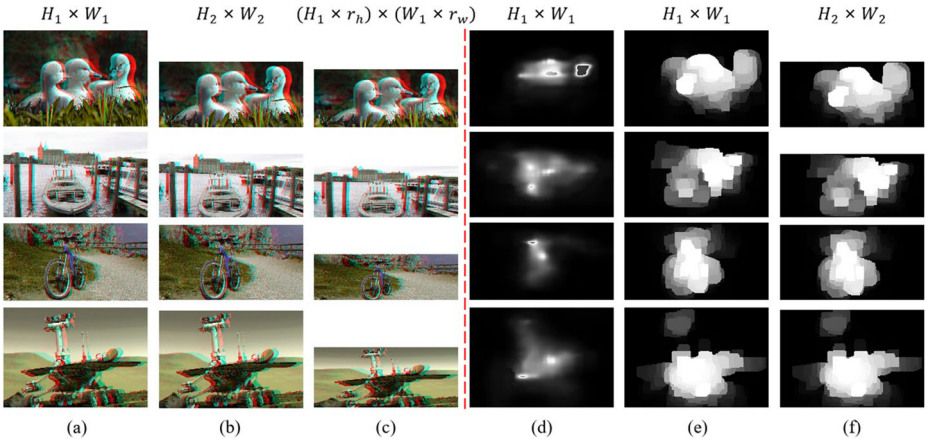

Fig. 2 Visualization of the phased output of stereoscopic thumbnail generation framework when the desired height ratio $r_h = 0.6$, desired weight ratio $r_w = 1$: The 3D anaglyph of (a) original image (b) cropped images after stage c (c) final thumbnail pair after stage e; The left (d) saliency map after stage a (e) salient object map after stage b (f) cropped salient object map after stage d

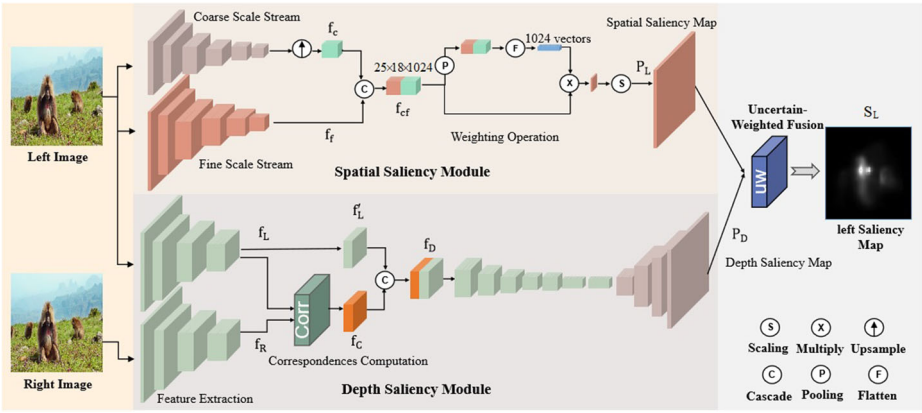

Fig. 3 The architecture of our proposed stereoscopic image saliency detection network

3.2.2 Depth saliency module

The depth prior related to saliency is also important for eye fixation prediction of stereoscopic images, for example, the closer the object is to the human eye, the more likely it is to attract human attention. Inspired by the study in [40], we investigate the relationship between left and right views, and utilize the obtained potential depth information to infer the depth prior map. The network structure is shown in the bottom box in Fig. 3. Specifically, we adopt two feature extraction flows to extract features f_L and f_R for the left and right views. Subsequently, the obtained features f_L and f_R are fed into a specially designed correlation layer termed as Corr, which enables to explore the horizontal distance correlation between the left and right views [40]. The calculation formula is as follows:

$$C(p_L, p_R) = \sum_{x \in [-k,k] \times [-k,k]} < f_L(p_L + x), f_R(p_R + x) >, \tag{1}$$

where p_L, p_R are the center of patches in f_L, f_R respectively, the patches' size are $2k + 1$. The symbol $< i, j >$ means the correlation analysis calculation between i and j. In addition, to produce the saliency feature f_D with implicit depth information, we combine the correlation feature f_C with the spatial feature f'_L obtained by using a convolution layer to reduce the output channel. Finally, the obtained feature f_D is fed into an encoder-decoder network to generate the depth prior map P_D guided by the depth information. The loss function of depth saliency module is defined as:

$$\mathcal{L}_{depth} = \frac{1}{N} \sum_{p=1}^{N} |P_D(p) - G_L(p)|, \tag{2}$$

where N is the number of pixels, and p is the index of pixels. P_D is the predicted depth saliency map, G_L is the ground-truth. It is worth noting that human fixation map of left image is regarded as the ground-truth during the network training.

3.2.3 Uncertain-weighted fusion module

Herein, we adopt an uncertainty weighting fusion method [15] to produce the final saliency map by combining the spatial prior map P_i ($i \in \{L, R\}$) with depth prior map P_D. For each

pixel $\mathbf{x}$ in prior map P_k ($k \in \{L, R, D\}$), the likelihood $f_k(\mathbf{x})$ of the pixel $\mathbf{x}$ being salient is defined as $f_k(\mathbf{x}) = 1 - \mathbf{exp}^{-(\frac{c_k(\mathbf{x})}{\beta})\gamma}$, where the connectedness $c_k(\mathbf{x})$, parameters β and γ are set default as suggested in [15]. The uncertainty $U_k(\mathbf{x})$ of current pixel $\mathbf{x}$ is measured as the entropy of the likelihood. For the spatial prior map P_i and depth prior map P_D, we can compute the final saliency map S_i via the following fusion rule:

$$S_i(\mathbf{x}) = \frac{U_i(\mathbf{x}) \cdot P_D(\mathbf{x}) + U_D(\mathbf{x}) \cdot P_i(\mathbf{x})}{2 \cdot (U_i(\mathbf{x}) + U_D(\mathbf{x}))}, i \in \{L, R\} \tag{3}$$

3.2.4 Saliency optimization module

Some previous works [36] have studied the difference and connection between fixation saliency and object saliency, and confirmed that the two can be transformed into each other. In previous section, we have obtained the precise fixation saliency. In order to obtain a reliable important map that used to guide the thumbnail generation operators, we transform the fixation saliency map to be object saliency map, which has complete outline of the object. The resulted object saliency map guides the cropping and wrapping operators keeping important object intact. Moreover, since the saliency maps of two views within a stereo image pair are estimated independently, thus salient and non-salient regions in these two views may not very consistent due to the slight difference of two views. Similar as in [33], we adopt the saliency optimization module to infer object-level saliency priors, by converting the pixel-wise saliency maps $\{S_L, S_R\}$ to the consistent salient object maps $\{\hat{S}_L, \hat{S}_R\}$.

For this stage's computational efficiency, the SLIC [1] algorithm is used to convert the stereo image pair to the set of super-pixels. Assuming that the original picture has N pixels, the SLIC algorithm converts the picture into an image with K superpixels according to the five-dimensional information of the color and XY distance in the image, and one super pixel has a size of $N/K (N >> K)$, so the image processing unit is changed from pixels to super pixels, which improves the subsequent calculation efficiency.

As for the generation of object-level salient, according to the saliency value of predicted fixation map, the foreground, background and uncertainty regions are defined. The principle concept of saliency optimization module within the stereoscopic saliency maps is to assign similar saliency values to the regions inside a homogeneous objects, and assign larger saliency values to the regions that far from the background region. The optimized salient object maps ensure the coherence of the left and right views. Considering that the stereoscopic salient object map is used for the downstream cropping task, we perform dilation on the salient object maps, so that the saliency object will have a certain distance from the cropping window.

3.3 Stereoscopic thumbnail generation

After obtaining the final salient object maps, we generate the stereoscopic image thumbnail via the following thumbnail cropping and energy minimization based warping operators. Specifically, the cropping operator is implemented to process the raw stereoscopic images with the guidance of salient object maps. If the salient regions are completely included in the cropped stereoscopic image with target aspect ratio, the thumbnail generation task is finished. Otherwise, the warping operator is implemented to achieve adaptive image warping with target aspect ratio. Details of each operator are discussed as follow.

3.3.1 Thumbnail cropping operator

This operator is designed to maintain the structure of salient objects and the disparity distribution between left and right views, while cut off the redundant regions according to the aspect ratio. The cropping process is completely guided by the salient object maps. Given the raw stereoscopic image with resolution $W \times H$ and desired ratio r_w, r_h, the new image is obtained by cropping the raw image along the horizontal (or vertical) direction when $r_w < 1 (r_h < 1)$. As shown in Fig. 4, take the vertical cropping as an example, the top boundary of the cropping window starts from the top side of raw image and slides to the bottom, while the bottom boundary starts from the bottom side of raw image and slides to top. The sliding operation will stop when the sum of saliency value of the j-th row $\mathrm{S}_{\mathrm{L}^j} > T$ or $\mathrm{S}_{\mathrm{R}^j} > T$. The threshold is defined as $T = \lambda S_{max}$, where S_{max} is the maximum row saliency value of the salient object map S_{L}. The parameter λ is set to 0.05 and 0.2 for vertical cropping and horizontal cropping, respectively. To ensure the height of cropped image pair larger than the target height and maintain the salient region center to be the center of cropped images, the sliding operation will go backward when the size of sliding window is smaller than the target size.

3.3.2 Energy minimization based warping operator

This operator intends to achieve adaptive image warping and outputs the re-targeted stereoscopic images $\{\hat{\mathrm{I}}_{\mathrm{L}}, \hat{\mathrm{I}}_{\mathrm{R}}\}$. Specifically, the cropped stereoscopic image pair $\{\mathrm{CS}_{\mathrm{L}}, \mathrm{CS}_{\mathrm{R}}\}$ is divided into regular quads for warping. The warping process rescales the images according to the importance of the quad, e.g., the less important quad will be squeezed more. We optimize the warping process via minimizing the following energy function:

$$\Phi = \Phi_q + \Phi_b + \Phi_a + \Phi_d. \tag{4}$$

The first term Φ_q in the energy function is the quad deformation energy term [32], which controls the deformation degree of the quad according to its saliency value, e.g., the quad with large saliency value will not deform as much as possible. Benefit from this loss item, the important area can be scaled uniformly, avoiding severe shape distortion. Specifically,

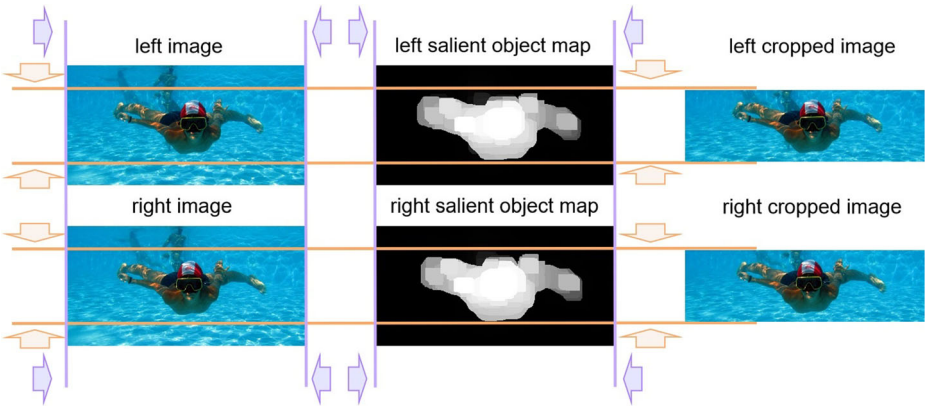

Fig. 4 Visualization of cropping stage when $r_h = 0.6, r_w = 1$

the quad deformation energy term [32] is defined as:

$$\Phi_q = \sum_{q \in Q} w_q \sum_{\{i,j\} \in E(q)}, ||(v_i' - v_j') - s_q(v_i - v_j)||^2 \tag{5}$$

where s_q is the scale factor of each quad. It can be initialized by taking $\frac{\partial \Phi_q}{\partial s_q}$ to zero. The w_q is the saliency value of quad. The $E(q)$ is the edge set of a quad, and the vertex within the quad is denoted as v, while the deformed version is v'.

The second term Φ_b reduces the bending degree of the grid line [32] which is defined as:

$$\Phi_b = \sum_{\{i,j\} \in E} ||(v_i' - v_j') - l_{ij}(v_i - v_j)||^2, \tag{6}$$

where l_{ij} is the scale factor of quad line. With this term, the wrapping result keeps grid lines being not bent as far as possible. The first two terms are all used to maintain important area in spatial dimension while changing aspect ratio.

The third term Φ_a is the feature consistent energy term [6], which keeps the corresponding pairs of feature points of left and right views align vertically:

$$\Phi_a = \frac{1}{n} \sum_{i=1}^{n} (\tilde{f_i^L}[y] - \tilde{f_i^R}[y])^2, \tag{7}$$

where n is the number of feature matching point pairs, $\tilde{f_i^L}[y]$, $\tilde{f_i^R}[y]$ are the vertical axis of deformed SIFT feature [23] of left and right views respectively. Under the feature consistent energy term, there will be no visual discomfort of excessive binocular asymmetry. The SIFT algorithm is used to extract feature matching pair between the stereoscopic image pair. Because of the feature matching points between left and right views have almost the same vertical axis in principle, we remove those points pairs with large vertical distance to ensure the accuracy of feature matching. And the importance value of each quad is computed as the average of all the saliency value of the pixel within the quad. Assuming a SIFT point p in quad that has vertex $\{v_i\}, i \in \{1, 2, 3, 4\}$, then axis of p can be expressed as the linear combination of v_i using barycentric coordinates, which are defined as follows,

$$f_p[y] = \sum_{k=1}^{4} \beta_k v_i[y] \tag{8}$$

and

$$f_p[x] = \sum_{k=1}^{4} \beta_k v_i[x]. \tag{9}$$

Besides, the disparity between the warped stereoscopic image will become smaller than the image pairs before warping. The smaller disparity, the weaker stereo perception which is not what we expect. Therefore, the disparity consistent energy term Φ_d [6], which keeps the disparity of warping stereoscopic image pair consistent to the raw one, is defined as:

$$\Phi_d = \frac{1}{n} \sum_{i=1}^{n} (\tilde{d_i} - d_i)^2, \tag{10}$$

where d_i is disparity, which is expressed as the difference between the abscissa of feature matching points, while $\tilde{d_i}$ represents the deformed version of d_i. Due to the deformation character of the warping operator, it is inevitable to cause severely distortion on the structural objects when the aspect ratio changes sharply. However, in our method, benefiting from

the previous cropping steps, the effect of the warping operator is to adjust the aspect ratio of images mildly. The resulting thumbnail will have more significant content and higher perceived quality.

4 Experimental results

In this section, we present the evaluation results of our proposed framework and compared of the proposed stereoscopic image saliency detection model and a stereoscopic thumbnail generation model with the state-of-the-art approaches. Besides, the optimized salient object maps are used to guide the stereoscopic thumbnail generation.

4.1 Performance on saliency detection

4.1.1 Dataset and implementation details

Our proposed method is evaluated on a newly built stereoscopic image saliency dataset (SIS) with 1086 stereoscopic scenes extracted from the movies and internet. All the stereoscopic images are resized into resolution 1920×1080 (3D side-by-side mode). The corresponding human eye fixation map is captured by a Tobii X3-120 eye tracker. For increasing the difficulty of saliency detection, the number of salient objects are varied. We divide the SIS dataset into a *train* set with 868 stereoscopic images and a *test* set with 218 stereoscopic images. We train the model on a server with NVIDIA TELSLA P100 GPU by using Tensorflow.

4.1.2 Quantitative and qualitative experiments

We compare the proposed saliency detection model with eight state-of-the-art methods in terms of six widely used metrics on the *test* set of SIS dataset. As shown in Table 1, the methods Itti et al.'s method [19] and GBVS [17] in the first group are two traditional 2D image saliency detection models that only use hand-crafted feature for saliency detection. The remaining methods (MSI-Net [3], GazeGAN [7] and CASNet [12]) in the first group are three learning based 2D image saliency detection approaches. From the Table 1, we can observe that the three learning based saliency detection methods achieve better performance than two traditional 2D saliency detection models. In addition, we also compare our proposed method with the traditional stereoscopic image saliency detection method (Fang et al.'s method) [14] and learning based stereoscopic image saliency detection methods including Depth-Sal [40] and LBVS-3D [4]. From the Table 1, it can be clearly seen that our proposed method is much better than Fang et al.'s method. Moreover, the two learning based methods (Depth-Sal [40] and LBVS-3D [4]) still inferior to our proposed method, especially the IG metric of Depth-Sal [40] method. This is possibly caused by the fact that the Depth-Sal [40] method pays more attention to exploring the relationship between the left and right views, while ignoring the spatial saliency clues. Different with the existing 3D methods, our proposed model combines the spatial and depth coherence originated from the left and right views to further facilitate the performance of saliency detection. We also provide the ROC curves of the above models in Fig. 5 to illustrate the superior performance of proposed model. From Fig. 5, we can clearly see that our proposed method obtains better performance than the existing benchmarks.

Table 1 Evaluation results on SIS dataset

Model	sAUC	AUC	CC	NSS	KLD	IG
Itti et al.'s method [19]	0.6617	0.7316	0.3111	0.9239	1.3218	2.5122
GBVS [17]	0.6786	0.8212	0.4809	1.3686	0.9884	2.9954
MSI-Net [3]	0.7363	0.8474	0.6293	1.8767	0.7683	3.3468
GazeGAN [7]	0.6968	0.7665	0.6366	1.9959	2.7685	1.1195
CASNet [12]	0.7448	0.8505	0.6794	2.0842	0.6436	3.5761
LBVS-3D [4]	0.6558	0.7445	0.4537	1.2951	1.3166	2.5407
Fang et al.'s method [14]	0.7054	0.8528	0.6160	1.7281	0.9170	3.0982
Depth-Sal [40]	0.7240	0.8266	0.7359	2.1402	1.9540	1.8788
Proposed	**0.7601**	**0.8614**	**0.8015**	**2.4065**	**0.4720**	**3.8401**

Moreover, we also provide the comparison of saliency maps obtained by the proposed saliency detection model with other state-of-the-art models in Fig. 6. From this figure, we can see that the traditional 2D and 3D saliency detection methods (Itti et al.'s method [19], GBVS [17] and Fang et al.'s method [14]) mistake the background as the salient regions, leading to poor performance in saliency detection. Compare the proposed model with other learning based 2D and 3D saliency detection models, we discover that the saliency maps obtained by our proposed model is closer to the ground-truth. This further demonstrates that our proposed saliency detection model achieves better performance than the existing counterparts.

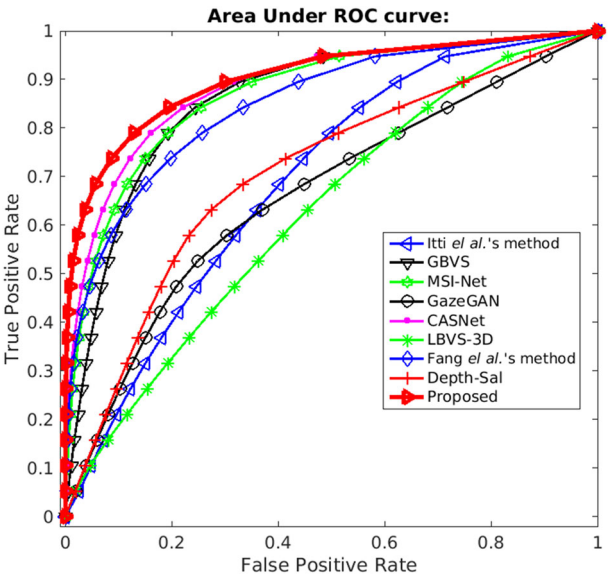

Fig. 5 ROC comparison of different models, including ground truth, Proposed, Depth-sal [40], Fang et al.'s method [14], LBVS-3D [4], CASNet [12], GazeGAN [7], MSI-Net [3], GBVS [17] and Itti et al.'s method [19] respectively

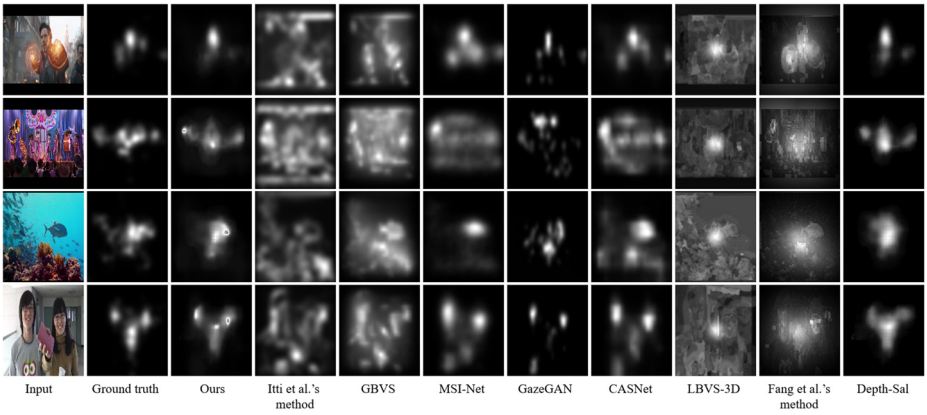

Fig. 6 From left to right are the left saliency map from input, ground truth, Ours, Itti et al.'s method [19], GBVS [17], MSI-Net [3], GazeGAN [7], CASNet [12], LBVS-3D [4], Fang et al.'s method [14] and Depth-Sal [40] respectively

4.2 Performance on stereoscopic thumbnail generation

4.2.1 Dataset

The NBU-SIRQA dataset [16] contains 45 source stereoscopic image pairs with different resolution. There are 720 re-targeted stereo image pairs generated by eight typical stereoscopic image retargeting methods when $r_w = 0.5$ or $r_w = 0.75$ (r_h both equal to 1.)

4.2.2 Quantitative and qualitative experiments

In Table 2, we provide the comparison results of our proposed stereoscopic thumbnail generation model with five state-of-the-art methods on NBU-SRIQA dataset in terms of the aspect ratio similarity (ARS) metric [38]. The involved five state-of-the-art methods are SCL method, SNS [32], SLWAP method [6], QoE method [28] and CPC method [33], where the SCL method directly scales the image to target aspect ratio. From this table, we can see that our proposed model achieves the best performance when $r_w = 0.5$, and comparable performance with QoE method when $r_w = 0.75$.

Furthermore, we also provide the comparison of the thumbnails generated by the proposed stereoscopic thumbnail generation model and the other re-targeting benchmarks in Fig. 7. From this figure, we can observe that the salient objects in the thumbnails generated by the SCL method have been squeezed heavy. The SNS method [32] is a 2D warping based approach that is utilized to process the left and right views of a stereoscopic image

Table 2 Performance comparsions on NBU-SIRQA Dataset in terms of ARS scores

r_w	2D method		3D method					
	SCL	SNS	SLWAP	QoE	CPC	only-warp	CPC-sal	Ours
0.5	63.85	68.84	73.04	72.89	73.45	69.48	72.67	**73.83**
0.75	76.50	77.08	77.69	**77.85**	77.61	77.08	77.72	77.63

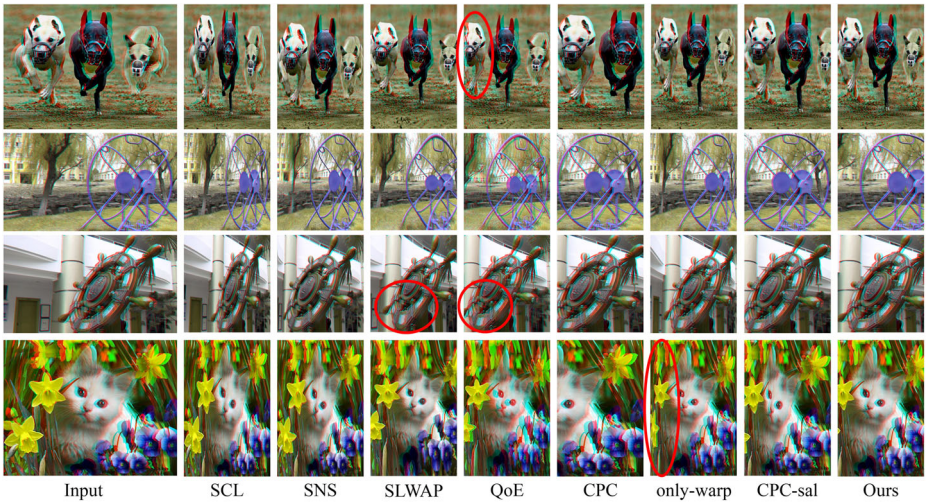

Fig. 7 The comparison results between our proposed method, four state-of-the-art methods and ablation module when $r_h = 1, r_w = 0.5$. We show the 3D anaglyph of each source stereo image pair as well as the 3D anaglyph of thumbnail image. From left to right are the input, thumbnail results from SCL(Uniform scaling), SNS [32], SLWAP [6], QOE [28], CPC [33] respectively. The only-warp is our model without cropping operator, the CPC-sal is our model with CPC's saliency map and Ours is our model with our saliency map

separately. Compared the methods SCL and SNS, we discover that the thumbnail results of SNS are squeezed less than SCL. However, the thumbnail results of SNS have low depth perception and vertical dis-alignment. That may be due to the lack of constraints between the left and right views. Different with methods SCL and SNS, SLWAP [6] is a 3D retargeting method. From the Fig. 7, we can see that the structured objects in the thumbnail results of SLWAP method are severely distorted, e.g., the circular objects in the second and third rows of Fig. 7. This is probably due to the inaccuracy of stereoscopic salient object map employed in SLWAP. As shown in Fig. 7, the provided thumbnail results of the 3D retargeting method QoE [28] tends to produce an obvious distortion at the boundary, e.g., the left dog in the first row of Fig. 7. This may be caused by the fact that the optimized disparity is beyond the original range of the image.

The CPC method [33] is a cropping based 3D thumbnail generation method and its cropping window can find the most salient region and cut it out efficiently, e.g., the complete circular object in the second row of Fig. 7. However, the CPC method is incapable of dealing with the scene with multiple salient objects, e.g., the right dog in the first row of Fig. 7. There are two reasons for these failure cases. First of all, this may be caused by the characteristics of the clipping operator. Secondly, the object integrity of the introduced salient object map is relatively ineffective with more context saliency interference, which leads to the truncation of salient objects. Compare with five counterparts, our proposed cropping-warping based thumbnail generation method reduces the risk of distortion at the boundary by benefiting from reliable stereoscopic image saliency priors. Therefore, our proposed model achieves better performance for stereoscopic thumbnail generation task against other methods. In addition, we also provide the thumbnail generation results in Figs. 8 and 9 in terms of the target width and height are 0.75 and 0.6 times to original width and height respectively. These provided thumbnail results further illustrate the advantage of our proposed model.

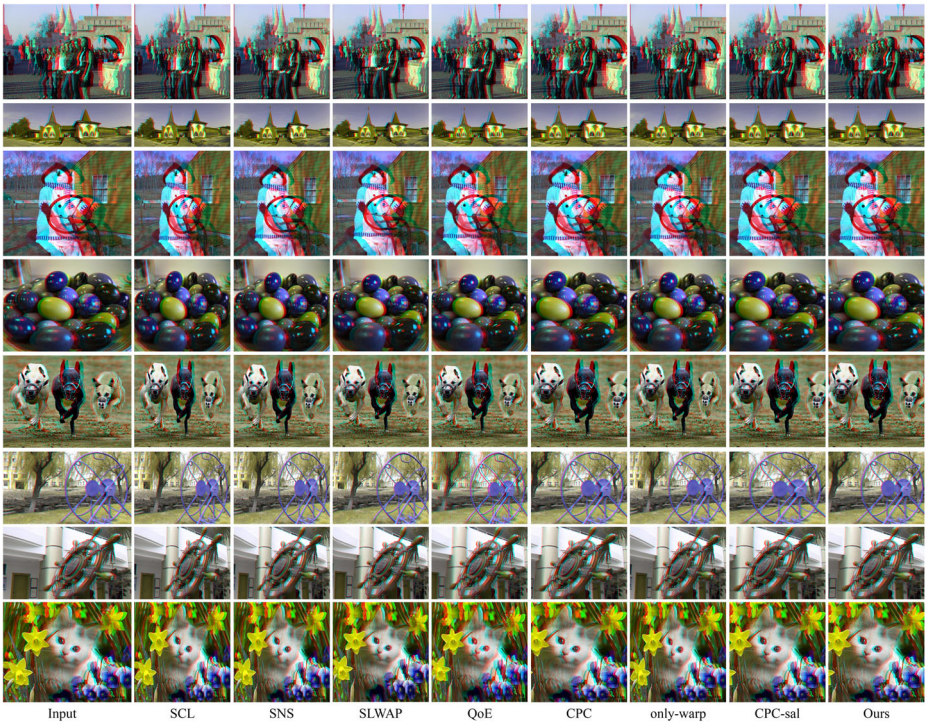

Fig. 8 The result's comparison when $r_h = 1, r_w = 0.75$ among four methods and ablation module. We show the 3D anaglyph of every source stereo image pair as well as the 3D anaglyph of thumbnail image. From left to right are the input, thumbnail results from SCL(Uniform scaling), SNS [32], SLWAP [6], QOE [28], CPC [33], only-warp, CPC-sal and Ours, respectively

4.2.3 Ablation studies

Ablation study on wrapping energy function As Fig. 10 shown, the ablation study on wrapping energy function (4) is conducted. Except for the quad deformation energy term, which is the basic constraint for wrapping operator, all other energy terms were subjected to ablation experiments. The w/o Bending, w/o Align and w/o Depth indicate the anaglyph of stereo thumbnail result that without grid line bending term Φ_b, feature consistent energy term Φ_a and disparity consistent energy term Φ_d, respectively. As the first picture in Fig. 10b shown, without Φ_b, what would have been straight in the image, especially straight objects or edges of the image, would have been bent. Comparing the second image of Fig. 10c with Fig. 10e, the results showing that Φ_a keep the image pair's features align vertically, avoiding binocular asymmetries [21]. The second image in Fig. 10d has a weaker 3D stereo perception than the original, while ours result keep original disparity with Φ_d.

Ablation study on saliency map Guided by the proposed salient object map, the only-warp method generates the stereoscopic thumbnail using warping operator directly. The CPC-sal method applies our thumbnail generation operator with the salient object map from CPC method [33]. As shown in the Table 2, compared to only-warp, our proposed method has

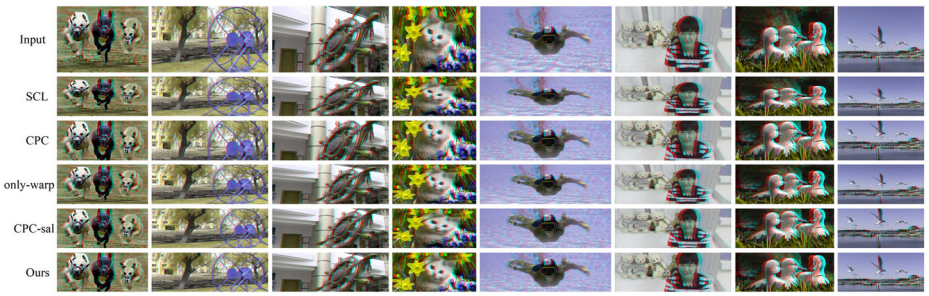

Fig. 9 The result's comparison when $r_h = 0.6$, $r_w = 1$ among one methods and ablation module. We show the 3D anaglyph of every source stereo image pair as well as the 3D anaglyph of thumbnail image. From top to bottom are the input, thumbnail results from SCL(Uniform scaling), CPC [33], only-warp, CPC-sal and Ours, respectively

better performance, highlighting the necessity of cropping operator. The results in Fig. 7 also demonstrate that without cropping operator, the thumbnail will be squeezed more seriously.

For the salient object map from CPC method, the disadvantage of ignoring high-level semantic information is not obvious when the changing degree of aspect ratio is small. Therefore, the ARS score of CPC-sal under 0.75-width requirement is higher than the proposed method mildly. However, due to the high accuracy of our salient object map, the advantage of our method is highlighted under more stringent requirements.

4.2.4 Computational complexity analysis

We measure the running time of the proposed thumbnail generation framework by matlab code on a PC with 2.6 GHz Intel xeon E5-2690 v4 CPU, NVIDIA Tesla P100 GPU and 256 GB RAM. The computational cost of our framework contains two parts. The first

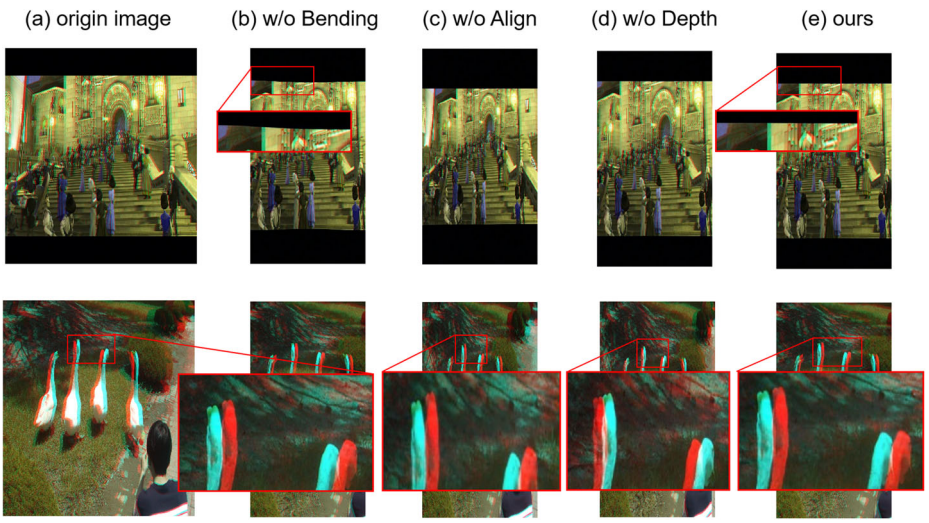

Fig. 10 Thumbnail obtained by w/o Bending, w/o Align, w/o Depth, and our proposed model with different wrapping energy constraint term

Table 3 Computational time of 1200 x 600 image

Method	Saliency prediction	Thumbnail generation	Total
OAC [33]	160.20 s	6.28 s	166.48 s
CPC [33]	160.20 s	1.73 s	161.93 s
Ours	5.65 s	3.45 s	9.1 s

part computes a stereo saliency object map, including spatial saliency map generation [12], depth saliency map generation [40], uncertainty weighting fusion [15] and saliency optimization [33]. The second part generates stereo thumbnails. Considering the stereo pair with a resolution of 1200×600, which is the sample image in Fig. 4, we implement our framework by an unoptimized python code of Tensorflow and a MATLAB code. In addition, we measure the running time of the work [33], which is one of the few stereoscopic thumbnail generation methods. The computational times are shown in Table 3. The results showing that ours method have significant advantages in the running time of saliency detection. The current implementation inputs the left and right images into the saliency model to obtain the left and right viewpoint saliency maps respectively. In order to reduce time consumption, we can use two saliency model to compute saliency map of left and right image in the meantime. In our thumbnail generation algorithm, the cropping and wrapping step takes 3.45s. So that our framework consumes a temperate time cost since the conversion from viewpoint map to saliency object map and the wrapping operator based energy minimization will occupy most of time.

5 Conclusion

In this paper, we propose a learning based saliency cues guided stereoscopic thumbnail generation framework. In which, a stereoscopic saliency detection network is designed to produce reliable saliency fixation map by combining the spatial and depth saliency information. Subsequently, the salient object map with complete contour of the object is obtained by the saliency optimization. Under the guidance of the obtained salient object map, we generate arbitrary aspect ratio stereoscopic thumbnails via the cropping-warping operator. Unlike the existing methods, our proposed framework avoids cutting or warping the image too much, leading to poor performance. Experimental results also demonstrate that our proposed framework outperforms the state-of-the-art methods.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant 61871270 and 62032015), in part by the Shenzhen Natural Science Foundation under Grants JCYJ20200109110410133 and 20200812110350001, in part by the National Engineering Laboratory for Big Data System Computing Technology of China.

References

1. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S (2012) Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 34(11):2274–2282
2. Avidan S, Shamir A (2007) Seam carving for content-aware image resizing. In: ACM SIGGRAPH 2007, pp 10–es

3. B AKA, B MSA, C KD, D RGAB (2020) Contextual encoder-decoder network for visual saliency prediction. Neural Netw 129:261–270
4. Banitalebi-Dehkordi A, Pourazad MT, Nasiopoulos P (2017) A learning-based visual saliency prediction model for stereoscopic 3D video (LBVS-3D). Multimed Tools Appl 76:23859–23890
5. Chai X, Shao F, Jiang Q, Ho YS (2019) MSTGAR: multioperator-based stereoscopic thumbnail generation with arbitrary resolution. IEEE Trans Multimed 22(5):1208–1219
6. Chang CH, Liang CK, Chuang YY (2011) Content-aware display adaptation and interactive editing for stereoscopic images. IEEE Trans Multimed 13(4):589–601
7. Che Z, Borji A, Zhai G, Min X, Guo G, Callet PL (2020) How is gaze influenced by image transformations? Dataset and model. IEEE Trans Image Process 29:2287–2300
8. Chen J, Bai G, Liang S, Li Z Automatic image cropping: a computational complexity study (2016)
9. Cho D, Park J, Oh TH, Tai YW, So Kweon I (2017) Weakly-and self-supervised learning for content-aware deep image retargeting. In: Proceedings of the IEEE international conference on computer vision, pp 4558–4567
10. Cong R, Lei J, Zhang C, Huang Q, Cao X, Hou C (2016) Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Process Lett 23(6):819–823
11. Dong WM, Bao GB, Zhang X, Paul JC (2012) Fast multi-operator image resizing and evaluation. J Comput Sci Technol 27(1):12–134
12. Fan S, Shen Z, Jiang M, Koenig BL, Xu J, Kankanhalli MS, Zhao Q (2018) Emotional attention: a study of image sentiment and visual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7521–7531
13. Fan X, Lei J, Liang J, Fang Y, Cao X, Ling N (2021) Unsupervised stereoscopic image retargeting via view synthesis and stereo cycle consistency losses. Neurocomputing 447:161–171
14. Fang Y, Wang J, Narwaria M, Le Callet P, Lin W (2014) Saliency detection for stereoscopic images. IEEE Trans Image Process 23(6):2625–2636
15. Fang Y, Wang Z, Lin W, Fang Z (2014) Video saliency incorporating spatiotemporal cues and uncertainty weighting. IEEE Trans Image Process 23(9):3910–3921
16. Fu Z, Shao F, Jiang Q, Chao M, Ho YS (2020) Subjective and objective quality assessment for stereoscopic 3d image retargeting. IEEE Transactions on Multimedia
17. Harel J, Koch C, Perona P (2006) Graph-based visual saliency. Adv Neural Inform Process Syst 19:545–552
18. Huang X, Shen C, Boix X, Zhao Q (2015) SALICON: reducing the semantic gap in saliency prediction by adapting deep neural networks pp 262–270 (2015). IEEE International Conference on Computer Vision, Santiago, pp 11–18. https://doi.org/10.1109/ICCV.2015.38
19. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259
20. Jiang Q, Shao F, Jiang G, Yu M, Peng Z, Yu C (2015) A depth perception and visual comfort guided computational model for stereoscopic 3d visual saliency. Signal Process: Image Commun 38:57–69
21. Jung YJ, Kim H, Ro YM (2016) Critical binocular asymmetry measure for the perceptual quality assessment of synthesized stereo 3d images in view synthesis. IEEE Trans Circuits Syst Video Technol 26(7):1201–1214
22. Kiess J, Guthier B, Kopf S, Effelsberg W (2012) Seamcrop for image retargeting. In: Multimedia on mobile devices 2012; and multimedia content access: algorithms and systems VI, vol 8304, p 83040K
23. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
24. Mao Y, Jiang Q, Cong R, Gao W, Shao F, Kwong S (2021) Cross-modality fusion and progressive integration network for saliency prediction on stereoscopic 3d images. IEEE Trans Multimedia, 1–1. https://doi.org/10.1109/TMM.2021.3081260
25. Nguyen AD, Kim J, Oh H, Kim H, Lin W, Lee S (2018) Deep visual saliency on stereoscopic images. IEEE Trans Image Process 28(4):1939–1953
26. Niu Y, Chen J, Ke X, Chen J (2019) Stereoscopic image saliency detection optimization: a multi-cue-driven approach. IEEE Access PP(99):1–1
27. Rubinstein M, Shamir A, Avidan S (2009) Multi-operator media retargeting. ACM Trans Graph (TOG) 28(3):1–11
28. Shao F, Lin W, Lin W, Jiang Q, Jiang G (2017) QoE-guided warping for stereoscopic image retargeting. IEEE Trans Image Process 26(10):4790–4805
29. Suh B, Ling H, Bederson BB, Jacobs DW (2003) Automatic thumbnail cropping and its effectiveness. In: Proceedings of the 16th annual ACM symposium on user interface software and technology, pp 95–104
30. Sun J, Ling H (2013) Scale and object aware image thumbnailing. Int J Comput Vis 104(2):135–153

31. Tan W, Yan B, Lin C, Niu X (2019) Cycle-ir: deep cyclic image retargeting. IEEE Trans Multimed 22(7):1730–1743
32. Wang YS, Tai CL, Sorkine O, Lee TY (2008) Optimized scale-and-stretch for image resizing. ACM Trans Graph 27(5):118
33. Wang W, Shen J, Yu Y, Ma KL (2016) Stereoscopic thumbnail creation via efficient stereo saliency detection. IEEE Trans Vis Comput Graph 23(8):2014–2027
34. Wang X, Shao F, Jiang Q, Fu Z, Chao M, Gu K, Ho YS (2021) Combining retargeting quality and depth perception measures for quality evaluation of retargeted stereopairs. IEEE Trans Multimedia, 1–1. https://doi.org/10.1109/TMM.2021.3081259
35. Wu L, Cao L, Xu M, Wang J (2014) A hybrid image retargeting approach via combining seam carving and grid warping. J Multimed 9(4):483
36. Wang W, Shen J, Dong X, Borji A, Yang R (2019) Inferring salient objects from human fixations. IEEE Trans Pattern Anal Mach Intell 42(8):1913–1927
37. Zhang Q, Wang X, Jiang J, Ma L (2016) Deep learning features inspired saliency detection of 3d images. In: Pacific rim conference on multimedia, pp 580–589. Springer
38. Zhang Y, Fang Y, Lin W, Zhang X, Li L (2016) Backward registration-based aspect ratio similarity for image retargeting quality assessment. IEEE Trans Image Process, 1–1
39. Zhang L, Li K, Ou Z, Wang F (2017) Seam warping: a new approach for image retargeting for small displays. Soft Comput 21(2):447–457
40. Zhang Q, Wang X, Wang S, Sun Z, Jiang J (2020) Learning to explore saliency for stereoscopic videos via component-based interaction. IEEE Trans Image Process 29(99):5722–5736
41. Zhou Y, Chen Z, Li W (2020) Weakly supervised reinforced multi-operator image retargeting. IEEE Trans Circuits Syst Video Technol 31(1):126–139