**Global and Local-contrast Guides Content-aware Fusion for RGB-D Saliency Prediction**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>SMCA-19-07-1346</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Regular Paper</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>30-Jul-2019</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>zhou, wujie; Zhejiang University of Science and Technology, the School of Information and Electronic Engineering; Zhejiang University; Lv, Ying; Zhejiang University of Science and Technology, the School of Information and Electronic Engineering; Lei, Jingsheng; Zhejiang University of Science and Technology, the School of Information and Electronic Engineering; Yu, Lu; Zhejiang University</td>
</tr>
<tr>
<td>Key Words:</td>
<td>Image analysis, Image processing, Multidimensional signal processing, Neural network architecture, Signal processing</td>
</tr>
</tbody>
</table>
Global and Local-contrast Guides Content-aware Fusion for RGB-D Saliency Prediction

Wujie Zhou, Ying Lv, Jingsheng Lei, Lu Yu, Member, IEEE

Abstract—Many RGB-D visual attention models (VAMs) have been proposed with diverse fusion methods; thus, the main challenge lies in the differences in the results between the different models. To address this challenge, we propose a local–global fusion method for fixation prediction on an RGB-D image; this method combines global and local information through a content-aware fusion module (CAFM) structure. First, it comprises a channel-based upsampling block for exploiting global contextual information and scaling up this information to the same resolution as the input. Second, our Deconv block contains a contrast feature module to utilize multilevel local features stage-by-stage for superior local feature representation. The experimental results show that the proposed model exhibits competitive performance on two databases.

Index Terms—RGB-D visual attention models, RGB-D image, local–global feature, contrast feature

I. INTRODUCTION

LIKE conventional RGB visual attention models (VAMs), RGB-D VAMs can be explicitly viewed as two branches: RGB-D eye fixation prediction and RGB-D salient object detection models [1]. The RGB-D eye fixation prediction model is used to obtain the pixel-level saliency intensity graph, which is a discrete grayscale graph, while the RGB-D saliency object detection model is used to detect and segment the most prominent target in a scene, which is usually a binary salient graph. In this work, we focus on the first type and aim to obtain a saliency intensity graph of an RGB-D image. This technology has a wide range of applications, including image classification [2], video/image segmentation [3–5], object recognition [6], visual tracking [7, 8], foreground map evaluation [9], person re-identification [10], and weakly supervised semantic segmentation [11]. Recently, convolutional neural network (CNN)-based methods [12–18] have become the main source for VAMs. Existing CNN-based VAM methods mainly deal with RGB images; this may produce unsatisfying results when the objects in images share a similar appearance with the background or when an image contains multiple interesting gaze targets.

The depth information from popular devices (e.g., Kinect and iPhone X) contains important complementary information for identifying salient regions, as shown in Fig. 1. Several RGB-D-based fusion methods [19–25] have been proposed in the last few years, and several studies have focused on the development of depth-induced saliency prediction models. However, in most such models, the fusion of features corresponding to the two above-mentioned modalities has been oversimplified. For example, existing RGB-D-based saliency prediction methods typically fuse RGB and depth input/features by simple concatenation, either via fusion at an early stage [19, 20], fusion at a late stage [21–23], or fusion at a middle stage [24, 25], as shown in Fig. 2. Therefore, certain complicated factors concerning said features and the strong correlations that exist between them are neglected. Such fusion strategies are inadequate to combine the complementary information obtained from two modalities. Therefore, it is necessary to effectively utilize depth information, particularly in the context of deep neural networks [26].

Instead of integrating RGB and depth information as done in existing fusion approaches [19–25], we show that the overall purpose of state-of-the-art CNN models (to use local-contrast and global features in the optimization) can be achieved using a simplified local–global feature model. The network of our model comprises convolution and Deconv blocks, where global features are generated by the final convolution through a channel-based upsampling block. A global feature can be used to accurately locate the position of saliency areas; however, it usually has a very low resolution. A local feature contains more details (such as texture, border, and outline) and has a higher resolution. To achieve this goal, we design a method for feature aggregation without limiting feature map resolution by using the channel-based upsampling block; this method can accurately restore pixel level prediction from relatively rough CNN deep outputs, and thus reduces the demand of the convolution decoder for an accurate response. Local contrast processing blocks are included in the Deconv block to promote local detail in features extracted by the model. The resulting global and local features are combined into a...
content-aware fusion module (CAFМ) block that gives the final saliency map output at the input resolution. To the best of our knowledge, this is the first work that explores global–local cross-modal complementarity in both top-down and bottom-up processes. Moreover, it is the first that introduces a content-aware weight distribution mechanism to reduce both cross-level and cross-modal fusion ambiguity in RGB-D data. In summary, our contributions of this study are as follows:

- We propose a novel global and local-contrast guides neural network, which can successfully identify useful contextual information in global and local contexts, and can effectively integrate the relevant information between different patterns for RGB-D saliency detection.
- The CAFМ is utilized to adaptively strengthen channel-wise salient map and emphasize key information inside the feature map while preserving the spatial structure of the feature map.
- Channel-based upsampling blocks are employed to effectively extract global information from low-resolution features and restore their resolution to the same size as the input data.
- The proposed model is validated on two publicly available datasets: NUS3D-Saliency [27] and NCTU-3DFixation [28]. The experimental results demonstrate that our model consistently outperforms state-of-the-art models on the two datasets.

II. RELATED WORK

A. Two-dimensional (2D) deep saliency models

The SALICON challenge [12] promoted the development of deep saliency models by providing the first large-scale saliency dataset. Most deep learning models use pre-trained networks as feature extractors to map from deep feature space to saliency space. They then fine-tune the network parameters using small datasets for specific tasks. For example, DeepNet [13] was an early use of the deep learning model, consisting of two versions, shallow and deep. Deep contains eight convolution layers and the first three layers are initialized from a pre-trained VGG network. SALICON [14] used the advanced semantic expression ability of the pre-trained deep neural network for target recognition, and then fine-tuned the neural network based on the objective function of the saliency evaluation index, and integrated information at different image scales. Deepfix [15] contained deep architectures of VGG [29], inception block of GoogleNet [30], and dilated convolutions [31] in location based convolution layers with a central bias. SalGAN [16] proposed a generative adversarial network (GAN) [32] with an encoding and decoding structure to generate pixel level saliency maps, and the binary cross entropy loss function was used to evaluate the difference between it and the groundtruth. DVA [17] used a series of deconvolution layers to form salient prediction of pixel-wise multi-layer depth features, and fuses them in the later stage. SAM [18] adopted an attention module with an ConvLSTM [33] block, which through a careful recursive mechanism enhances the salient features in turn. Kroner et al. [34] used a module with multiple convolutional layers at different dilation rates, called the ASPP module, to enhance context information containing advanced visual features at multiple spatial scales for simultaneously obtaining multiple scale features. Che et al. [35] introduced a GAN based on U-Net as a GazeGAN generator that combines the classic ‘skip connection’ with the ‘center surround connection (CSC)’ to take advantage of multilevel features.

B. Three-dimensional (3D) data driven neural networks

Fang et al. [36] used the contrast of weighted features between image blocks of the spatial distance of a Gaussian model to calculate five groups of significant features, and fused multiple groups of features. Liu et al. [37] extended a multimodal Bayesian fusion method for 3D stereoscopic videos. This Bayesian integration method successfully integrates saliency cues in a nonlinear manner. Li et al. [38] used a graph neural network to fuse four manually extracted saliency feature clues based on stereoscopic video input for 3D fixation prediction. Wang et al. [39] proposed a novel visual attention-driven model, where a module mimicking human attention behavior in a dynamic setting is used as a supervised
neural attention module to guide the subsequent module for fine-grained video object segmentation. Zhang et al. [40] proposed a pre-trained CNN to extract RGB and depth salient maps and a linear fusion method to combine color and depth feature maps in order to generate a final RGB-D saliency map. They added a center bias mechanism to enhance significance mapping.

For other 3D-induced tasks conducted with CNNs, Jiang et al. [41] introduced an attention mechanism to allocate weights on multi-level RGB and depth features, obtaining the final fusion feature graph by linear combination.

As shown in Fig. 2, existing CNN-based RGB-D approaches can be classified into five categories. The first, shown in Fig. 2(a), fuses the input in the earliest stage and regards the depth map as one channel of input directly [19, 20]. The second, Fig. 2(b), employs a ‘late fusion’ strategy. More specifically, individual predictions from both RGB and depth information are produced, and the results are integrated into a separate post-processing step such as pixel-wise summation and multiplication. For example, Fan et al. [21] used depth contrast and depth weighted color contrast to measure the saliency value of the regions, Cheng et al. [22] computed the saliency based on the laws of visually salient stimuli for both color and depth spaces. Desingh et al. [23] leveraged nonlinear support vector regression to fuse these predicted maps. The third scheme, shown in Fig. 2(c), combines the depth and RGB features extracted from different networks. For instance, Feng et al. [24] proposed a novel RGB-D saliency feature to capture the spread of angular directions. Similarly, Shigematsu et al. [25] proposed to capture background enclosure as well as low-level depth cues.

The fourth scheme, shown in Fig. 2(d), fuses depth and RGB features in the middle stage with bottom-up/top-down and skip-connection. For example, Zhu et al. [42] employed a prior-model guided master network to process the RGB information of images. The fifth scheme, shown in Fig. 2(e), shows two separate networks with a skip-connection, which incorporates cross-modal saliency information at the end of the network. Wang et al. [43] designed a dual-stream encoder-decoder neural network, each of which extracts features from either RGB or the depth modal and predicts a saliency map. A switch map is then used to select which of the features represent key information that can accurately locate salient objects in the fused feature map.

III. PROPOSED METHOD

Here, we provide a deep convolutional network architecture; the purpose is to effectively use RGB and depth features and accurately detect significant areas in the image (Fig. 1). As mentioned above, good saliency maps must take into account both local important details of the image and precise location information provided by the global context, as well as detailed knowledge from various resolutions. Toward this end, we implemented a novel CNN architecture.

A. Overall architecture

Fig. 3 depicts the overall CNN architecture of the proposed model. We used the VGGNet [29] as the backbone of the two bottom-up streams. Specifically, we removed its fully connected layers and retained its five convolutional blocks. In addition, the channel-based upsampling block follows the final convolutional blocks, which contain a dilation convolutional layer with 3 × 3 kernels and a hole size of 2 for global contextual reasoning without the constraint imposed by the resolution of the feature maps. Local features are then generated from the top-down stream where Deconv blocks are used to aggregate these features from the convolutional blocks stage-by-stage. Here, rich features within the local features contain objects in the background that may be of interest to human vision. The use of contrast features makes salient
Specifically, scale is the ratio of the parameters of the convolutional layers, and \( F \) is the non-linearity activation function, and \( W \) channels. In the following, layers have a kernel size of 1, with a total of 128 feature compute the global feature. The two ordinary convolutional 3 kernels and a hole size of 2 after the Conv-5 block to convolution layers and a dilation convolutional layer with 3 × channel-based upsampling block, which includes two ordinary significance prediction. To account for this, we added a resolution global information is indispensable for enhanced information to provide an accurate target location, and high final saliency map output at the input resolution.

**B. Capturing global context**

The final salient region in an image needs global context information to provide an accurate target location, and high resolution global information is indispensable for enhanced significance prediction. To account for this, we added a channel-based upsampling block, which includes two ordinary convolution layers and a dilation convolutional layer with 3 × 3 kernels and a hole size of 2 after the Conv-5 block to compute the global feature. The two ordinary convolutional layers have a kernel size of 1, with a total of 128 feature channels. In the following, \( W^g \) represents the weights from the \( c \) filter in the dilation convolutional layer and \( b^i \) represents its bias. Given the feature maps \( F \) from the Conv-5 block, the global features \( F^g \) of the \( c \) channel can be written as:

\[
F^g_c(x, y) = \sigma \left( \beta \left( \sum_{i,j} (W^g_c(x+i, y+j) \cdot F^i(j)) + b^i \right) \right)
\]

where \( \cdot \) denotes the dot product, \( \sigma \) denotes the ReLU non-linearity activation function, and \( \beta \) denotes BatchNorm operation.

We define the global features as:

\[
F_g = \{ F^1_g, F^2_g, \ldots, F^c_g \}
\]

where \( F_g \in \mathbb{R}^{H \times W \times C} \).

To obtain the global feature graph with the same resolution as the input, we need to compress the global feature graph in dimensions. Two simple convolution operations are applied to the linear transformation of \( F_g \); a compressed global sality map \( X_g \) can be written as follows:

\[
X_g = D(W_2(W_1F_g))
\]

where \( D \) represents the decompression process, \( W_1 \) and \( W_2 \) are the parameters of the convolutional layers, and \( W_1 \in \mathbb{R}^{1 \times 1 \times C} \) and \( W_2 \in \mathbb{R}^{1 \times 1 \times N} \), in which C is set to 256, and \( N \) is class \( \times \) scale \( \times \) scale. Specifically, scale is the ratio of \( F_g \) to \( X_g \), which is usually 16 or 32.

However, here, we set it as 8. The compression process is completed above. Fig. 4 shows the decompression.

**C. Local features**

1) Multi-Scale local features: As shown in Fig. 5, The Deconv blocks are connected to the VGGNet Conv-1 and then to Conv-5 processing blocks. The objective of these convolutional layers is to learn multi-scale local feature maps \( \{X_1, X_2, \ldots, X_5\} \) step-by-step. Each convolution block is stacked by two convolutional layers with kernel sizes of 1 × 1 and 3 × 3; the block has 128 channels.

2) Contrast features: Saliency is a unique area with attractive characteristics inside the foreground object, which can arouse people’s emotions more than other areas. Therefore, the significance of the feature must be in the foreground of the internal part of the contrast with other areas. To obtain such local detail contrast information, we manually extracted the contrast features associated with each local feature \( X_i \). Each contrast feature \( X^c_i \) is computed by subtracting the local average from its local maximum. The kernel size of the max pooling and the average pooling are all 3 × 3:

\[
X^c_i = \text{MaxPool}(X_i) - \text{AvgPool}(X_i)
\]

3) Deconvolution features: To obtain a prediction map with the same resolution as the original input of 224 × 224, we used several convolution and up-sampling layers to increase the size of feature maps \( X^c_i \). We adopted stepwise upward sampling, as shown in Fig 5. At each Deconv processing block, the resulting upsampled feature map \( U_i \) is calculated by combining the information of its local contrast feature \( X^c_i \), local feature \( X_i \), and the previous block’s upsampled feature \( U_{i-1} \).

\[
U_i = \text{Deconv}\left( \text{concat}(X_i, X^c_i, U_{i-1}) \right)
\]

The Deconv operation is implemented using a convolution layer with a stride of 1, a 1 × 1 kernel, and a bilinear upsampling with a scale factor of 2. The concat denotes the concatenation of \( X_i \), \( X^c_i \), and \( U_{i-1} \). The number of feature channels of \( U_i \) is equal to the sum of \( X_i \) and \( U_{i-1} \). Fig. 5 shows the details.

4) Final local saliency map: We used a convolution layer with a kernel size of 1 × 1 to obtain the final local saliency map \( X_l \) (contains RGB and depth feature maps). The input to that layer is the concatenation of \( X_i \), \( X^c_i \), and \( U_2 \):
Similarly, a global fusion map $X_{\text{g}}^F$ is generated as $X_{\text{g}}^F$ using Eq. (7) and Eq. (8).

Finally, the local–global fusion map $X_{\text{g}}^L$ can be generated as follows:

$$X_{\text{g}}^L = \alpha \left( \beta \left( \text{Conv}_1 \left( Z_{\text{g}} \right) \right) \right) \odot X_{\text{g}}^r + \alpha \left( \beta \left( \text{Conv}_1 \left( Z_{\text{g}} \right) \right) \right) \odot X_{\text{g}}^d$$  \hspace{1cm} (9)

where $\odot$ is the pixel-wise multiplication, and $\text{Conv}_1$ and $\text{Conv}_{\text{g}}$ represent convolution layers with a kernel size of $1 \times 1$ and a channel size of 1.

The CAFM can adaptively strengthen the channel-wise salient features and emphasize key information inside the feature map while preserving the spatial structure of the features. Fig. 6 shows the detailed architecture of the CAFM.

After the above derivation, the final saliency map $M$ can be obtained by a convolution layer followed by BatchNorm and Sigmoid function operations, formulated as follows:

$$M = \delta \left( \beta \left( \text{Conv}_6 \left( X_{\text{g}}^L \right) \right) \right)$$  \hspace{1cm} (10)

where $\text{Conv}_6$ denotes a convolution layer with a kernel size of $1 \times 1$ and a channel size of 1, and $\delta$ denotes the Sigmoid function.

E. Implementation detail

1) Data processing: We randomly sampled 70% of the images from the NUS and NCTU datasets as training sets, and 10% as verification sets; the rest were used as test sets. These two datasets contain a large number of images, some from real scenes and some from 3D movie scenes; they have rich visual content, including semantic content and complex backgrounds.

Before data entry model, the size of the image was adjusted to $224 \times 224$. In order to accelerate the convergence speed of the network and improve the computing performance of the network, pre-calculated parameters were used before input into the model to normalize each image area with the mean value as the center along the RGB channel into unit variance.

2) Training strategies: We trained our model by loading the pre-trained VGG-16 model on the ImageNet dataset. A batch size of 1 image was used in each iteration. The learning rate was initialized as $1 \times 10^{-4}$. The architecture parameters were learned by back-propagating the loss function using root mean square prop (RMSprop). We employed an early termination algorithm to prevent overfitting of the network. The experiments were performed on the publicly available Pytorch 1.1.0 [44] framework using a workstation equipped with a TITAN V GPU (with 12 GB memory).

IV. Experimental Results and Analyses

A. Datasets

To verify the performance of the proposed model, we evaluated all the saliency detection models on two datasets. Currently, there are few publicly available eye-tracking datasets for RGB-D saliency research. In this study, we
evaluated the saliency detection models on two state-of-the-art datasets: NUS3D-Saliency [27] and NCTU-3DFixation [28]. The detailed descriptions of these datasets are provided below.

1) The NUS3D-Saliency dataset (denoted as NUS) has 600 RGB-D images collected from 80 participants, containing several RGB-D scenes. This dataset provides RGB stimuli, depth maps, smooth depth maps, and 3D fixation maps.

2) The NCTU-3DFixation dataset (denoted as NCTU) consists of 475 RGB images and their depth maps. This dataset contains various scenes, mainly originating from existing RGB-D movies or videos.

### B. Evaluation metrics

The saliency prediction results are evaluated using several metrics: linear correlation coefficient (CC), similarity, AUC shuffled, AUC Borji, AUC Judd, normalized scanpath saliency (NSS), and Kullback-Leibler divergence (KL-Div). Some of these metrics compare the predicted saliency map with the ground truth.
with the ground-truth saliency map generated from the fixation points, whereas others directly compare the predicted saliency map with the fixation points [16, 17].

C. Comparison against state-of-the-art methods

We compared our proposed model in the two benchmark datasets above using six other state-of-the-art models: Fang [36], DeepFix [15], ML-net [45], DVA [17], iSEEL [46], and SAM [18]. These models have been proposed or are widely used for comparison in recent years. For these models, we used their publicly available code to obtain saliency maps with the parameters recommended by the authors. Table I provides a summary of these models, including our model. Most of the models require off-line training or are based on a deep learning framework. Fig. 6 shows the sample images of the seven saliency models (our model and the six other saliency models).

Table I presents the quantitative results obtained on the NUS and NCTU datasets. Compared with other models, it has strong robustness and is rarely distracted by high-contrast edges and complex background, and thus it can generate more accurate saliency prediction results. The proposed method can detect single and complex bottom-up saliency patterns with different scales and contrasts [see rows 1, 2, and 9] and can effectively handle global and local contrast features [particularly row 9, an image contains many complex, densely arranged, tiny objects]. More importantly, the face is a top-down factor that our network can detect very well [see row 5], along with people in a complex background [see rows 3 and 7], and long distance and short distance objects [see rows 4, 5, and 8]. It can also deal with the integration of contrasts and top-down factors. In particular, in rows 5, 6, 7, and 8, although the images contain too many eye-catching objects or complex backgrounds, our model can still detect the regions that excite human eyes.

D. Hybrid loss

As shown in Fig. 3, our network can output several saliency maps, including the local–global fusion map $X_L^f$, local fusion feature $X_L^f$, global fusion map $X_N^f$, local saliency map $X_L$, and global saliency map $X_g$ (the last two contain RGB and depth saliency maps, $X_L^d$, $X_D^d$, $X_L^g$, and $X_D^g$ respectively). During the training process, we applied a deep supervision mechanism [47] to impose a supervision signal (ground truth) for each saliency output, and thus we could compute the loss between each predicted and the ground truth saliency maps (supervision).

Our training loss is defined as the summation over all the outputs:

$$Loss = \psi L_{fg}^f + \varphi L_{lg}^f + \varphi L_{lg}^d + \omega L_{r}^g + \omega L_{l}^g + \omega L_{h}^g$$

(11)

where $L_{fg}^f$, $L_{lg}^f$, and $L_{lg}^d$ denote the KL loss [18, 48]; $L_{r}^g$, $L_{l}^g$, and $L_{h}^g$ denote the modified CC metric; and $\psi$, $\varphi$, and $\omega$ are three scalars that balance the seven loss functions. By conducting a large number of experiments, we found that the optimal combination of the parameters is $\psi = 1$ and $\varphi = \omega = 2$.

In this work, we used MSE combined with the CC metric as the loss $L_{lg}^g$ between the final local–global fusion map and the ground truth. However, the CC metric is slightly modified into the representation of dissimilarity without requiring empirical coefficients. The modified loss mimics the behavior of cross-entropy, which is widely used in image classification, approaching zero when there are no mistakes.

The CC computes the linear correlation between two distributions. The range of CC is $[-1,1]$, where the positive value indicates that the two distributions are completely correlated, and vice versa. To apply RMSprop more efficiently, we simply modified the CC metric as follows:

$$CC'(P,Q) = 1 - \frac{\sigma(P,Q)}{\sigma(P) \times \sigma(Q)}$$

(12)

We used $CC'$ to represent the modified metric. The $CC'$ converts the similarity metric into dissimilarity in the range of $[0,2]$. We simply take the summation of MSE and $CC'$ to formulate our final local–global fusion loss function:

$$L_{lg}^g = \frac{1}{N} \sum_{i,j} |P - Q|^2 + 1 - \frac{\sigma(P,Q)}{\sigma(P) \times \sigma(Q)}$$

(13)

E. Limitations

Fig. 7 shows some typical failure cases. The first row shows that our model does not perform well on images with people; both our model and iSEEL are affected by the presence of people. The second row shows that when there are many objects in the field of vision, our model prefers objects with a high color contrast and ignores smaller objects at the center. Similarly, iSEEL also ignores smaller objects and was more inclined to the center of the field of vision; however, the human eye does not always focus at the center of the field of vision. The third line shows that the distance between objects and the human eye affects the accuracy of eye focus positioning in our model, making the model pay more attention to objects at close range. However, iSEEL only pays attention to distant positions and ignores nearby brightly colored objects. This reflects its ineffectiveness for this type of detection. Although the Deepfix model can somewhat perform detection, its detection results do not reflect the areas that the human eye pays attention to.
V. Conclusions

In this study, we designed a global and local-context guides content-aware fusion model, which combines a content-aware fusion module (CAFM) and channel-based upsampling blocks. The proposed CAFM was utilized to adaptively strengthen channel-wise saliency map and emphasize key information inside the feature map while preserving the spatial structure of the feature map. While maintaining the feature space structure, the algorithm adaptively enhances the significant features of the channel direction and emphasizes the key information in the feature graph. The channel-based up-sampling block was used to extract global information from low-resolution features and restore their resolution to the same size as the input data. Moreover, the global features and local contrast features were effectively integrated. The accuracy of significance detection was improved. Our method achieved encouraging performance on two challenging datasets.

References


