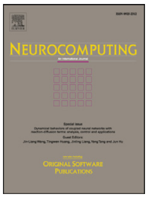
Neurocomputing 363 (2019) 46–57

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Salient object detection for RGB-D image by single stream recurrent convolution neural network

Zhengyi Liu[a,b,*], Song Shi[a,b], Quntao Duan[a,b], Wei Zhang[a,b], Peng Zhao[a,b]

[a] *Key Laboratory of Intelligent Computing Signal Processing of Ministry of Education, Anhui University, China*
[b] *School of Computer Science and Technology, Anhui University, China*

ARTICLE INFO

Article history:
Received 28 October 2018
Revised 21 May 2019
Accepted 4 July 2019
Available online 19 July 2019

Communicated by Prof. Junwei Han

Keywords:
Saliency object detection
RGB-D image
Recurrent convolutional neural network
Single stream network

ABSTRACT

Salient object detection for RGB-D images aims to utilize color and depth information to automatically localize objects of human interest in the scene and reduce the complexity of visual analysis. Different from existing saliency detection model with double-stream network, salient object detection by Single Stream Recurrent Convolution Neural Network(SSRCNN) is proposed. First RGBD four-channels input is fed into VGG-16 net to generate multiple level features which express the most original feature for RGB-D image. The coarse saliency map from the deepest features can detect and localize salient objects, but loss the boundaries and subtle structures. So Depth Recurrent Convolution Neural Network (DRCNN) is then applied to each level feature for rendering salient object outline from deep to shallow hierarchically and progressively. With the help of deeper level feature, original depth cue and coarse saliency map, each level feature can accurately predict the salient objects in different scales. At last all the saliency maps from each level are fused together to generate final results. Extensive quantitative and qualitative experimental evaluations on four dataset demonstrate that the proposed method outperforms most state-of-the-art methods.

1. Introduction

With the prevalence of RGB-D sensors such as Microsoft Kinect, Time-of-flight sensor and RealSense, it is convenient to capture RGB-D image composed of RGB color image and paired depth image. Salient object detection for RGB-D images can filter out irrelevant information and reduce the complexity of visual analysis, and it is served as an important pre-processing step in many problems such as object detection [1–3], ROI extraction [4], image retargeting [5], visual tracking [6], scene retrieval [7], stereoscopic image quality assessment [8], photography layout recommendation [9]. It aims to utilize both RGB view and depth view to automatically localize objects of human interest in the scene. Especially when the salient object and background share similar appearance, or salient objects share complex color itself, RGB data, which supplies appearance and texture information and is sensitive to light variations, is powerless to discriminate the salient objects from background. In this case, the paired depth data, which contains more affluent spatial structure, can contribute a lot of additional saliency cues.

For the RGB-D saliency detection task, how to get discriminative depth feature is the first key issue. Early works mainly focus on using original single dimensional depth value [3] or designing hand-crafted features such as depth contrast [10–13], ACSD [14], LBE [15], HOSO [16] which need strong understanding of domain specific knowledge. Indeed, the low-level features are unable to capture high-level semantic information about the objects and their surroundings and lack generalization ability when adopted for variant scenes especially when the salient objects have low depth contrast. Recently deep convolutional neural networks (CNN) are adopted in RGB-D saliency detection [17–27] for high-level representations of depth data. However, existing RGB-D saliency detection dataset only have thousands even hundreds of image pairs. To solve the discrepancy between the data hungry nature of CNNs and the insufficient data problem in RGB-D saliency detection domain, cross-modal transfer learning [18,20,25] which transfers the structure of the modality from RGB source view to be applicable for the depth target view is proposed. In order to use pre-trained models in RGB modality, depth values are encoded into three-channel HHA representations [28] (i.e., the horizontal disparity, height above ground, and the angle of the local surface normal

[*] Corresponding author at: School of Computer Science and Technology, Anhui University, China.
E-mail address: liuzywen@ahu.edu.cn (Z. Liu).

https://doi.org/10.1016/j.neucom.2019.07.012
0925-2312/© 2019 Elsevier B.V. All rights reserved.

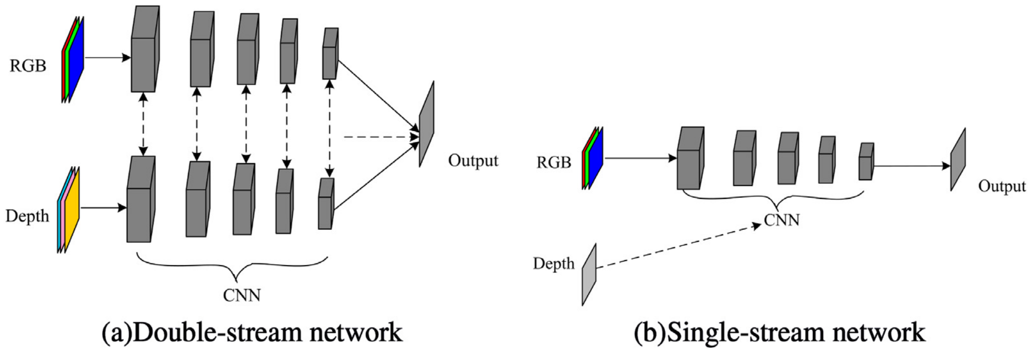


Fig. 1. Double-stream network and single-stream network.

with the inferred gravity direction) which has the same number of input channels as RGB representations. Different from this idea, depth data is directly used as the fourth channel of network input for data primitiveness and consistency in our paper.

For the RGB-D saliency detection task, how to fuse RGB and depth information sufficiently is another key issue. There are two main modes:

(1) RGB and Depth image parallel processing ("double-stream fusion")

Double-stream fusion generates RGB image saliency model and depth image saliency model respectively, and then stitches to optimize the eventual saliency detection results in the middle or end of two streams. Double-stream fusion has two distinct branches and includes RGB processing path and depth processing path, as shown in Fig. 1(a). Such characteristic is shown in traditional method [3,10,29–32]and deep learning method [17–20,24,25,27,33]. In other words, depth information is considered to be the same important as RGB information, and they are mutual influence and promotion.

(2) RGB and Depth image simultaneously processing ("single-stream fusion")

Single-stream fusion extends RGB image saliency model and fuses depth information from input channel part or feature part. Single-stream fusion has one master stream processing RGB image, and a branch processing depth image which can enhance the robustness of the master stream, as shown in Fig. 1(b). Such characteristic is shown in traditional method [12,34–39] and deep learning method [21,22,26,40]. In other words, depth information is considered to be supplement to RGB information, and they are different in importance.

Double-stream network needs two CNNs, which doubles the number of network parameters and computation cost. In our opinion single-stream network is effective when considering network parameters and computation cost.

Recently recurrent convolutional neural network(RCNN) has achieved state-of-the-art results in object recognition [41]. It has been applied to salient object detection for RGB image [42] and videos [43,44]. Inspired by its excellent performance, we propose Single Stream Recurrent Convolution Neural Network (SSRCNN) to detect salient object for RGB-D image. RGBD four-channels input is fed into VGG-16 net to generate multiple level features. The prediction by deeper features can coarsely detect and localize salient objects, but loss the boundaries and subtle structures. So Depth Recurrent Convolution Neural Network(DRCNN) is proposed to render salient object outline from deep to shallow hierarchically and progressively. With the help of deeper side output, original depth cue and coarse saliency map, each side outputs can refine the results from deeper side outputs by DRCNN and generate more accurate saliency maps. At last all the saliency map from each level are fused together to generate final results. Learning process are deep supervised from end to end.

In summary, the contributions of this work includes:

(1) Single-stream network with RGBD four-channels input is adopted to extract color and depth cue for saliency detection. It is different from double-stream network which considering cross-modal transfer learning from RGB modality to depth modality. It can utilize the original depth information to assist color information to detect salient objects with lower computational cost.

(2) Trunk net extracts multiple level feature from shallow to deep, and generates coarse saliency map. It locates salient object, but losses outline details. So DRCNN is presented to render salient object from deep to shallow. With the help of deeper layer side output, original depth cue and coarse saliency map, lower layer side outputs can generate the salient objects from multiple scales which retain more outline details.

(3) The proposed saliency detection method yields more accurate saliency maps and outperforms most state-of-the-art models on four benchmark dataset.

2. Related work

2.1. Salient object detection for RGB-D images

Recently RGB-D sensors provide excellent ability and flexibility to capture RGB-D images. Depth has been shown to be one of the practical cues for extracting saliency [45]. How to design discriminative depth feature and how to fuse RGB and depth information are two key issues in RGB-D salient object detection.

2.1.1. Design discriminative depth feature

Ouerhani et al. [3] propose depth, mean curvature and depth gradient to express depth feature. Ciptadi et al. [34] construct 3D layout and shape features from depth measurements. Desingh et al. [29], Cheng et al. [10], Ren et al. [12], Guo et al. [46], Tang et al. [32] measure depth contrast using depth difference between regions. Peng et al. [11] propose a multi-contextual contrast model including local contrast, global contrast and background contrast for detecting salient object from depth map. Ju et al. [14] define the depth saliency of a point as how much it outstands from surroundings which are measured using an anisotropic center-surround operator. Feng et al. [15] propose local background enclosure(LBE) feature which employs an angular density component and an angular gap component to measure the proportion of the object boundary in front of the background. Feng et al. [16] introduce the histogram of surface orientation (HOSO) feature to measure surface orientation distribution contrast for RGB-D saliency in the consideration that salient regions are often characterized by an unusual surface orientation profile with respect to the surroundings.

Although these methods demonstrate massive progress on RGB-D saliency detection, they do not go beyond the vein of resorting to hand-crafted low-level features and traditional bottom-up contrast-based saliency cues, which lack high-level representations and generalization ability.

Recently deep convolutional neural networks (CNN) are applied in various computer vision tasks [20–22] successfully. They are powerful in digging representative features. So CNNs are also adopted in RGB-D saliency detection for high-level representations of depth data.

However, existing RGB-D saliency detection dataset only have thousands even hundreds of image pairs. Current paradigms involve learning a generic feature representation on a large dataset with labeled images (e.g., RGB image dataset), and then specializing or fine-tuning the learned generic feature representation for the target dataset (e.g., depth image dataset). Due to difference between RGB and depth information in the number of the channels, depth information is often converted to HHA representation. Different from this viewpoint, single-stream network with RGBD four-channels input is proposed to extract depth and color cue synchronously.

2.1.2. Fuse RGB and depth cue

Ouerhani et al. [3], Desingh et al. [29] extract feature from RGB and depth data respectively, and generate conspicuity maps and then fuse them together. Cheng et al. [10] measure color contrast, depth contrast and spatial bias and fuse them by multiply operation at last. Xue et al. [30] propose a saliency object detection model integrating RGB and depth cues via mutual manifold ranking of RGB image and depth map, and take color and depth feature as the measurement of similarity, and define four corners as background queries. It ranks the saliency of the RGB image with composition feature including color and depth cues, and then makes use of the RGB saliency as the prior to rank the saliency of the depth map, and last fuses them. Guo et al. [31] propose a salient object detection method for RGB-D images using saliency evolution strategy. It first over-segments a RGB-D image into superpixels with the extended SLIC algorithm based on both color cue and depth cue. Then it estimates two saliency maps based on color cue and depth cue independently, and fuse them with refinement to obtain the initial saliency map with high precision. Finally, it iteratively propagates saliency over the whole image on a graph-based model and generates the final saliency map. Tang et al. [32] present a two-stage framework for detecting salient objects in challenging images and each stage combines color and depth features. In the object location stage, region contrast and depth prior measurement produce a noise-filtered salient patches, which indicate the location of the object. In the object boundary inference stage, boundary information is encoded in a graph using both depth and color information, and then heat diffusion is employed to infer more reliable boundaries. Chen et al. [17] first train the saliency detection networks for RGB and depth modalities separately, and then train a multi-modal network by fusing their deep representations in a late point. The original depth values are encoded into three-channel HHA representations to enable the use of pre-trained models in RGB modality. Chen et al. [18] pre-train RGB and depth modality independently and then stitch to initialize and optimize the eventual depth-induced saliency detection model. Chen et al. [19] train the saliency detection networks for RGB and depth modalities separately, and then train a multi-modal network in a late point. The fusion path is scattered to diversify the contributions of each modality from global and local perspectives. Han et al. [20] transfer the structure of the RGB-based deep neural network to be applicable for depth view using the task-relevant initialization and adding deep supervision in hidden layer, and fuse the deep representations of both views automatically using the task-relevant initialization and adding deep supervision in hidden layer. Wang et al. [25] estimate saliency from depth cues based on a convolutional neural network (CNN) trained by supervision transfer, and saliency detection model is based on the deep features of RGB images and depth images within a Bayesian framework.

Above works focus on double-stream network which RGB and depth information are parallel processed. But depth cue provides auxiliary but not particularly trustworthy information, single-stream network in which RGB and depth information are simultaneously processed is also an important fusion mode. Ren et al. [12], Ciptadi et al. [34], Zhang et al. [36], Jiang et al. [39] combine RGB and depth feature to measure low-level local contrast prior, global contrast prior, background prior, foreground prior, orientation prior for RGB-D saliency detection. Song et al. [37], Zhue et al. [38], Du et al. [47] combine color and depth feature to measure RGB-D saliency value by machine learning [21] to fuse different low level saliency cues including RGB and depth information and feed them to CNN for automatically detecting salient objects in RGBD images, rather than directly predicting saliency in an end-to-end manner with original images as inputs. This fusion strategy increases the difficulty to learn an efficient combination model, since the low-level features from each modality are noisy. Shigematsu et al. [22] fuse depth data by appending middle level and low level feature, such as background enclosure, depth contrast and histogram distance, into deep CNN architecture for RGB salient object detection. Zhu et al. [26] fuse depth sub-network, which is treated as an encoder convolution architecture, into RGB master network. Depth stream is supplementary to RGB stream.

Single-stream network need one stream, and it is effective when considering network parameters and computation cost.

2.2. Recurrent neural networks for salient object detection

Recurrent neural networks (RNN) are first designed to process sequential data [48,49], and later generalized to object recognition [41] and salient object detection tasks. Recurrent convolutional layer (RCL) which has the ability to learn more and more contextual features proposed by Liang and Hu [41] has a profound influence. RCL inherently incorporates multiple recurrent connections into each convolutional layer, and its receptive field expands when the iteration increases. Wang et al. [50] propose saliency detection with recurrent fully convolutional networks which take both the image and the saliency prior map as input to predict the saliency map. Network is recurrent in the sense that the output saliency map acting as a feedback signal is recursively passed through the network. It refines the prediction by iteratively correcting previous mistakes. Later they [51] further improve the network architecture by converting the fifth convolutional layer into a RCL enforcing long range of spatial-temporal consistency, thus more contextual information can be taken into account to get more robust feature maps. Tang et al. [52] incorporate the recurrent connections into each convolutional layer and add side-output layers to supervise the feature learning of the intermediate layers inspired by recurrent convolutional neural networks [41] and deeply-supervised nets [53]. They use RCL to replace the convolutional layers in the five blocks. Its aim is to extract contextual features. Le et al. [43] propose RCL3D which is an extension of RCL to handle video. The feature map at a deconvolution layer in the top-down pass and the feature map in its corresponding convolution layer in the bottom-up pass are concatenated across channel direction and refined by RCL3D unit to generate a stronger feature map in which the contextual information is enhanced and accumulated. Zhang et al. [54] propose multi-path recurrent connections inspired by Jin et al. [55]. Network is recurrent in the sense that the global semantic acting as a feedback signal is recursively concatenated to feature of each shallow layer for more effective features. Sun et al. [56], Xiao et al. [57] utilize RCL to enhance the local saliency information in some convolutional blocks, for example Conv2_2, Conv3_3 and Conv4_3, and then global feature

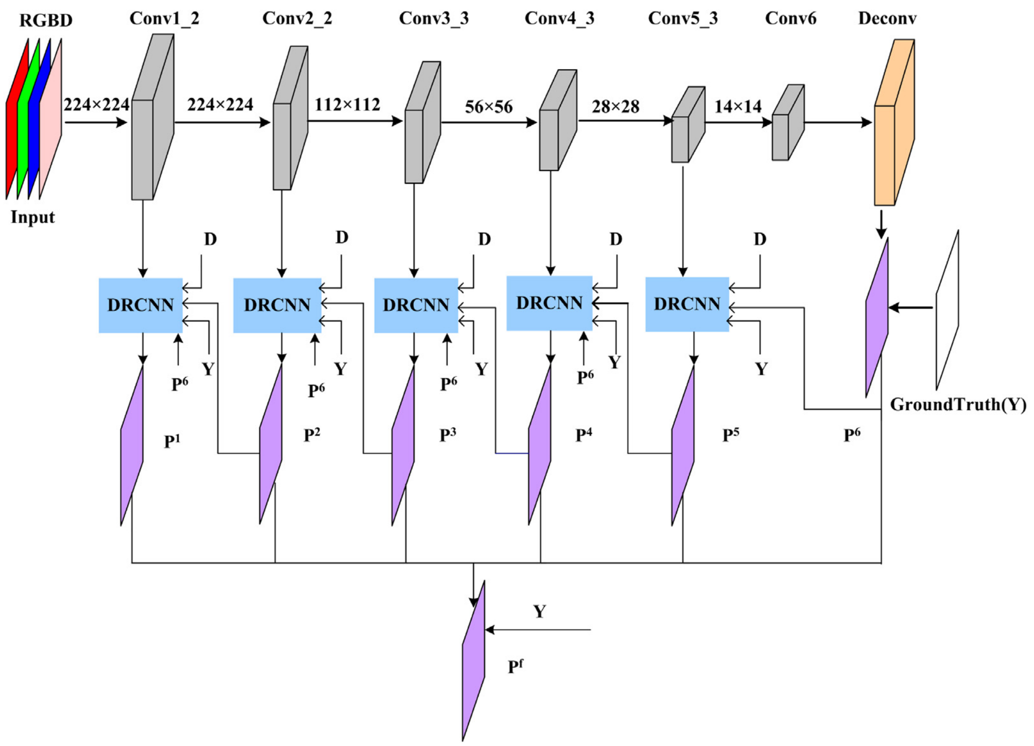


Fig. 2. Structure of Single Stream Recurrent Convolution Neural Network (SSRCNN).

is refined with the RCL enhanced side-outputs feature by self-attention mechanics or concatenation and upsampling operation.

Inspired by remarkable RCL, it is employed in our work for fusing original depth information, coarse saliency maps and deeper feature to improve current feature representation for sharp outlines.

3. Proposed method

Single Stream Recurrent Convolution Neural Network (SSRCNN) adopts VGG-16 Net [58] as trunk net which first coarsely detects salient objects in a global perspective, and then applies top-down side-output subnetwork Depth Recurrent Convolution Neural Network (DRCNN) which hierarchically and progressively refines the details of the salient objects from deep to shallow. SSRCNN with RGBD four channel inputs and DRCNN subnetwork is trained end-to-end by deep supervised learning.

3.1. Trunk net for coarse global prediction

As shown in Fig. 2, input RGB image and paired Depth image are concatenated into an input with four channels. They are wrapped to size 224 × 224 and fed into VGG-16 network which is removed all the fully connected layers and composed of conv1_2, conv2_2, conv3_3, conv4_3 and conv5_3 to extract deep features including color and depth feature. In order to meet requirements of the input, the convolution kernel in the Conv1_2 is modified from 64 × 3 × 3 × 3 to 64 × 3 × 3 × 4. Then coarse saliency map P^6 is produced by subsequent a convolutional layer, a de-convolutional layer, a convolutional layer and an activation function based on deep features from the output of conv5_3.

3.2. Depth recurrent convolution neural network for hierarchical refinement

Although coarse saliency map P^6 can detect and localize salient objects, but it losses the boundaries and subtle structures. The top-down side-output subnetwork Depth Recurrent Convolution Neural Network (DRCNN) is proposed to hierarchically and progressively render image details by combining the depth cues with the color features.

Each layer of trunk net connects a subnetwork DRCNN. For each DRCNN $DR^m (m = 1, \ldots, 5)$, as shown in Fig. 3, the coarse saliency map P^6, the prediction result P^{m+1} from deeper layer, the original depth information D and current convolutional layer feature F^m from the VGG-16 net compose the input. Convolution operation and sigmoid activation function are first adopted to guarantee the same size of P^6, P^{m+1}, and D, and then three convolutional layer and sigmoid activation function are applied to squash the features F^m step by step in order to prevent overwhelmed and retain the more details of the feature. Four inputs are concentrated together, and Recurrent Convolution Layer (RCL) [41] are applied to iteratively generate a finer saliency map P^m in which the contextual information is enhanced and accumulated.

The RCL with $t = 3$ can be unfolded to a feed-forward micro network of depth $t + 1 = 4$. Multiple recurrent connections make the micro network has multiple paths from the input layer to the output layer, which facilitates the learning. Besides, the effective receptive filed of an RCL unit expands when the time step increases, making the units to be able to accept larger and larger contexts. Thus RCL can help to incorporate local contexts, the depth cue and the high-level semantic information efficiently in DRCNN to get more accurate result.

RCL incorporates recurrent connections into each convolutional layer. The states of RCL units evolve over discrete time steps. Input $z_k^m(t)$ at time step t of the kth feature map in an DR^m is given by:

$$z_k^m(t) = \left(w_k^{f^m}\right)^T u^m + \left(w_k^{r^m}\right)^T x^m(t-1) + b_k^m \qquad (1)$$

where u^m and $x^m(t-1)$ are the feedforward input from the previous layer and the recurrent input from the current layer at time step $t-1$, $w_k^{f^m}$ and $w_k^{r^m}$ denote the vectorized feed-forward weights and recurrent weights respectively, b_k^m is the bias, and u^m

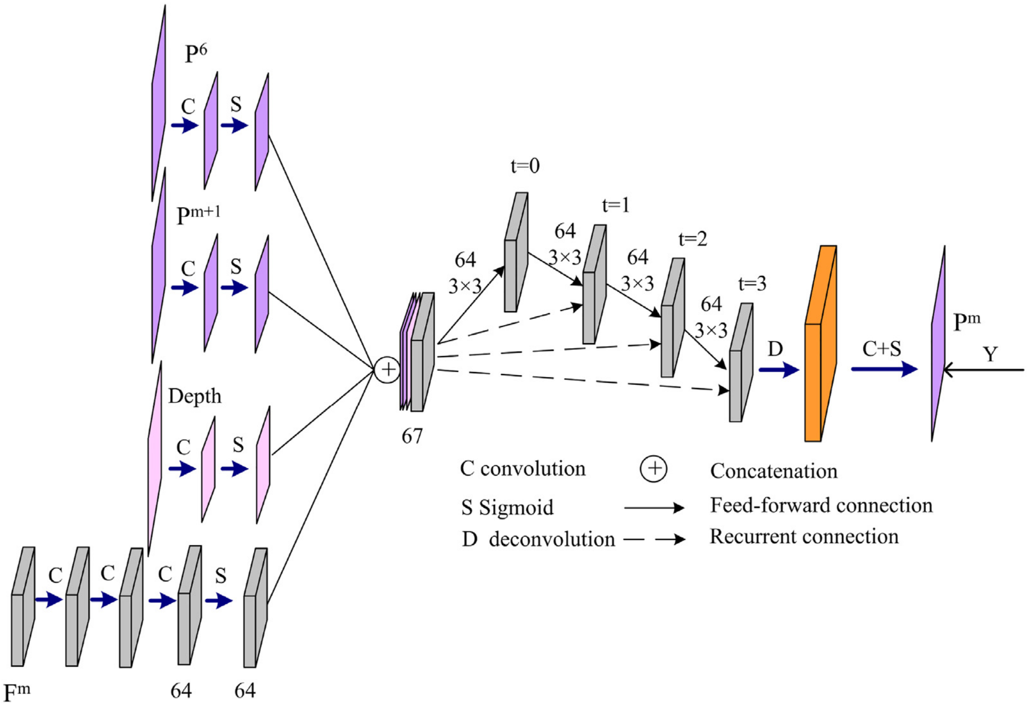


Fig. 3. Structure of Depth Recurrent Convolution Neural Network (DRCNN).

is defined as:

$$u^m = C(\varphi(F^m \oplus D \oplus P^6 \oplus P^{m+1})) \tag{2}$$

where F^m denotes the feature of the mth convolutional layer from the VGG-16 net, D denotes the original depth information, P^6 denotes coarse saliency map, and P^{m+1} denotes saliency map from the m+1th subnetwork DR^{m+1}, $\oplus$ denotes the concatenation operation, $\varphi(\cdot)$ denotes operation which removes duplicated input P^6 and P^{m+1} when $m = 5$, and C denotes the convolution operation with kernel size 3 × 3.

Net input $z_k^m(t)$ is processed by the rectified linear activation function (ReLU) [41] to increase nonlinearity and local response normalization(LRN) function [41] to prevent the states from exploding. Its output is defined as:

$$g(f_k^m(t)) = \frac{f(z_k^m(t))}{\left(1 + \frac{\alpha}{N}\sum_{k'=max(0,k-N/2)}^{min(K,k+N/2)} (f(z_{k'}^m(t)))^2\right)^\beta} \tag{3}$$

where $K = 64$ is the total number of feature maps in the current layer, $N = 7$ is the size of the local neighbor feature maps which are involved in the normalization, constant $\alpha = 0.001$ and $\beta = 0.75$ control the amplitude of normalization.

RCL in DRCNN can learn accurate features which contain more contextual information by four input P^6, D, F^m, P^{m+1}. In order to further recover image size to 224 × 224 size which is the same as the input image, a de-convolutional layer is adopted, and last a convolutional layer with 1 × 1 kernels and sigmoid activation function are applied to convert the feature maps to the saliency map.

3.3. Multi-scale and multi-level fusion

The deeper the trunk net, the smaller side-output-plane size, and the larger receptive field size.

All the output saliency maps $P^m (m = 1, \ldots, 6)$ with the same size 224 × 224 in SSRCNN are multi-scale and multi-level. The saliency map from shallow layers reflects the details of image, while the saliency map from deep layers shows the global semantic information. In order to highlight the detail of salient object, the feature map from the first and second layer are element-wise summed. Thus the difference between salient foreground and background is more obvious. Meanwhile in order to retain more global semantic, the feature maps from the third to sixth layer are concatenated. Then the summed part and concatenated part are concatenated and performed convolutional operation to generate the final saliency map, as shown in Fig. 4, and it is expressed by the following formula:

$$P^f = C(\sigma(P^1 + P^2) \oplus \sigma(P^3 \oplus P^4 \oplus P^5 \oplus P^6)) \tag{4}$$

where σ denotes sigmoid activation function, + denotes element-wise sum. The fusion process utilizes respective advantages of multi-scale and multi-level saliency maps and generates the better saliency map fusing local and global information of salient objects and robust to the size variation of salient objects.

3.4. Loss function

We enforce early supervision on all the hidden convolution layers to facilitate network parameter learning. To do so, we employ the deeply supervised learning scheme proposed in [53], which supervises both the hidden layers and the output layer to alleviate the problem of vanishing gradients during training.

For training of SSRCNN, the errors between all side-outputs and the ground truth should be computed and backward propagated. Therefore, we need to define a loss function to compute these errors.

As show in Fig. 2, the architecture of SSRCNN consists of five side-output subnetworks and a direct output subnetwork. For the sake of representation, we denote the direct output subnetwork as the sixth side-output subnetwork.

Give $T = \{(X^n, Y^n) | n = 1, \ldots, N\}$ denotes the input training dataset, where $X^n = \{x_i^n, i = 1, 2, \ldots, C\}$ denotes the nth input image, $Y^n = \{y_i^n | i = 1, 2, \ldots, C, y_i \in [0, 1]\}$ denotes the corresponding ground truth saliency maps, and N denotes the number of images in training dataset, C denotes the number of pixels in the

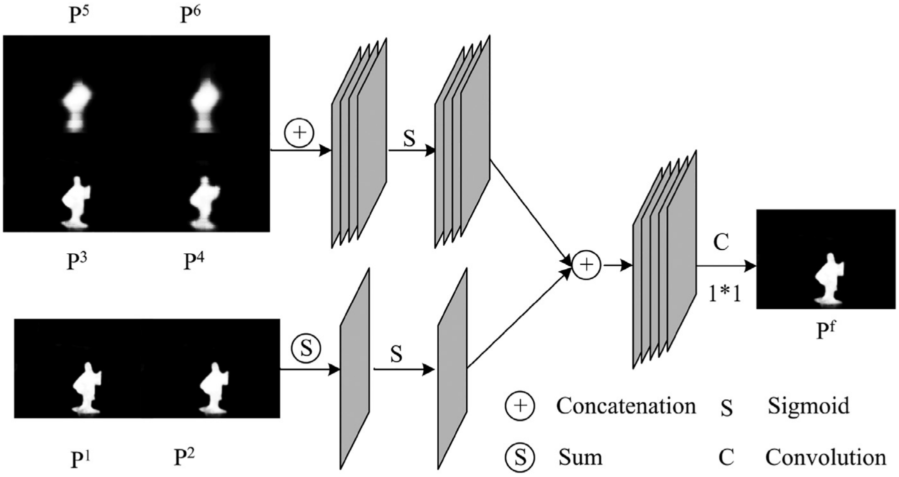


Fig. 4. Multi-scale and multi-level fusion.

input image. The parameters of all network layers in SSRCNN is set as W. Each side-output subnetwork DRCNN is associated with a classifier, and the corresponding weights can be represented as $w = \{w^m | m = 1, \ldots, M\}$, where $M = 6$ denotes the number of side outputs. We use a standard cross-entropy loss to compute the loss function over all pixels in the input image X^n and ground truth Y^n, the loss function can be represented by:

$$L^m_{side}(W, w^m) = -\sum_{i=1}^{C}(y^n_i logPr(y^n_i = 1|X^n; W, w^m) + (1 - y^n_i)logPr(y^n_i = 0|X^n; W, w^m)) \quad (5)$$

where $Pr(y^n_i = 1|X; W, w^m)$ represents the probability of the activation value at location i in the mth side output that measures how likely the pixel belong to the foreground, $Pr(y^n_i = 0|X; W, w^m)$ represents the probability of the activation value at location i in the mth side output that measures how likely the pixel belong to the background. In order to better capture the advantage of each side output, we add a weighted-fusion layer to connect each side output prediction, the loss function at the fusion layer can be represent by:

$$L_{fuse}(W, w, w^f) = -\sum_{i}^{C}(y^n_i logPr(y^n_i = 1|X^n; W, w, w^f) + (1 - y^n_i)logPr(y^n_i = 0|X^n; W, w, w^f)) \quad (6)$$

where w^f is the classifier parameter for the fused prediction. Thus, the joint loss function for all predictions can be written by

$$L(W, w, w^f) = \delta_f L_{fuse}(W, w, w^f) + \delta_m \sum_{m=1}^{M} L^m_{side}(W, w^m) \quad (7)$$

where δ_f and δ_m are loss weights to balance each loss term. For simplicity, we set $\delta_f = 1, \delta_m = 1$ as used in [2]. After computing all loses, we minimize the following objective loss function during training:

$$(W, w, w^f)^* = argmin(L(W, w, w^f)) \quad (8)$$

4. Experiments

4.1. Dataset

We conduct our experiments on four public benchmark dataset.

NLPR1000 [11]. NLPR1000 dataset contains 1000 images captured by Microsoft Kinect in various indoor and outdoor scenarios. It includes 11 types of indoor and outdoor scenes and more than 400 kinds of common objects under different illumination conditions.

NJU2000 [14]. NJU2000 dataset contains 2000 stereo images and photographs as well as the corresponding depth maps and manually labeled ground truths. The depth maps are generated using an optical flow method.

STEREO [59]. STEREO dataset has 797 stereoscopic images. These images are mainly collected from Internet and 3D movies. Depth images are generated by leveraging an optical method.

LFSD [60]. LFSD dataset contains 100 images with depth information and manually labeled ground truths. The depth information is captured using the Lytro light field camera.

Training dataset must be the same for fair comparison, but training dataset in MLFN [17], CTMF [20], MMCI [33], PCFN [24] and AF [27] are the same about 650 images from NLPR1000 and 1400 images from NJU2000, but they are different from PDNet [26] about 500 images from NLPR1000 and 1500 images from NJU2000. In our experiments, the training dataset which is the same as PDNet [26] is first adopted for all the ablation experiments and comparison experiments except for five models(MLFN [17], CTMF [20], MMCI [33], PCFN [24] and AF [27]). The testing dataset are the rest. Then our model is trained on the same training dataset as those five models for comparison with them. The testing dataset are the same as theirs, only including NLPR1000, NJU2000 and STEREO.

4.2. Evaluation metrics

Evaluation metrics are used to evaluate the performance of different saliency models.

PR Curve. PR curves are plotted by comparing with the ground truth by setting a group of thresholds on the saliency maps to achieve binary masks.

F-measure. F-measure is computed by:

$$F_\beta = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (9)$$

where we set $\beta^2 = 0.3$ the same as Martinet al. [61].

MAE. Mean absolute error(MAE) refers to the average pixel-wise error between the saliency map and ground truth. It is computed by:

$$MAE = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H}|S(x, y) - G(x, y)|}{W \times H} \quad (10)$$

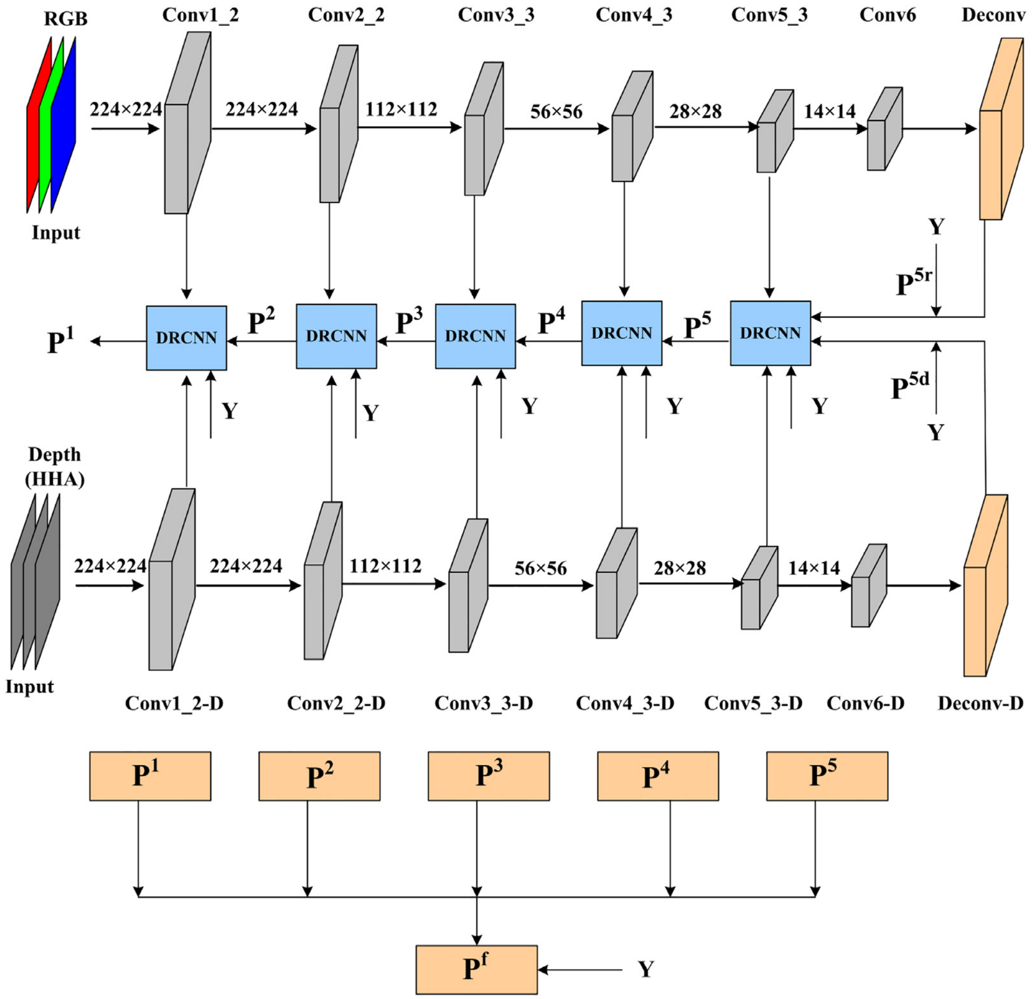


Fig. 5. Our double-stream framework with DRCNN module.

where H and W are height and width of the image, $S(x, y)$ and $G(x, y)$ denote the saliency map and ground-truth at the pixel (x, y), respectively.

S-measure [62]. Structural similarity measure simultaneously evaluates region-aware and object-aware structural similarity between saliency map and ground truth. S-measure is defined as:

$$S_\lambda = \lambda * S_o + (1 - \lambda) * S_r \tag{11}$$

where S_o and S_r are the object-aware and region-aware structural similarity respectively, λ is the balance parameter and set as 0.5 in our experiment.

E-measure [63]. Enhanced-alignment measure combines local pixel values with the image-level mean value in one term, jointly capturing image-level statistics and local pixel matching information.

In order to compare our method with other state-of-the-art methods fairly, all the evaluation metrics use the same codes[1] provided by Fan et al. [64].

4.3. Implementation details

The proposed model is implemented in python with Caffe toolbox [65]. It is evaluated on a machine equipped with one GTX Titan-x GPUs (with 12G memory). The momentum, learning rate, weight decay and mini-batch size are set as 0.99, 1e−10, 0.0005 and 1, respectively. Our SSRCNN is fine-tuned from an initialization with the pre-trained VGG-16 net [58]. In order to solve data insufficiency problem, data augmentation is adopted. Each training image is horizontally and vertically flipped and cropped from top, down, left and right 1/10 image parts. So the training dataset is increased by 16 times and meanwhile the global semantic information is kept almost unchanged. The training process costs 12 h for 10 epochs. During testing, the model runs about 12 fps with 224 × 224 resolution.

4.4. Ablation experiments

We first verify the effect of single-stream network with RGBD four-channels input strategy. The double-stream network with the same idea as our SSRCNN is designed and shown in Fig. 5. We can see that there are RGB and depth streams in the network equipped with our same DRCNN modules. The comparison results between double-stream network and single-stream network are shown in Table. 1. From comparison results we can see that single-stream network with RGBD four-channels input is superior to RGB and HHA double-stream network in detecting salient objects effectively. In addition single-stream network spends 12 h in training 10 epochs, while double-stream network spends 15 h in training the same epochs. So our single-stream network has some advantages in detection accuarcy and training time.

[1] https://mmcheng.net/zh/code-data.

Table 1
Single-stream network with RGBD four-channels input valuation.

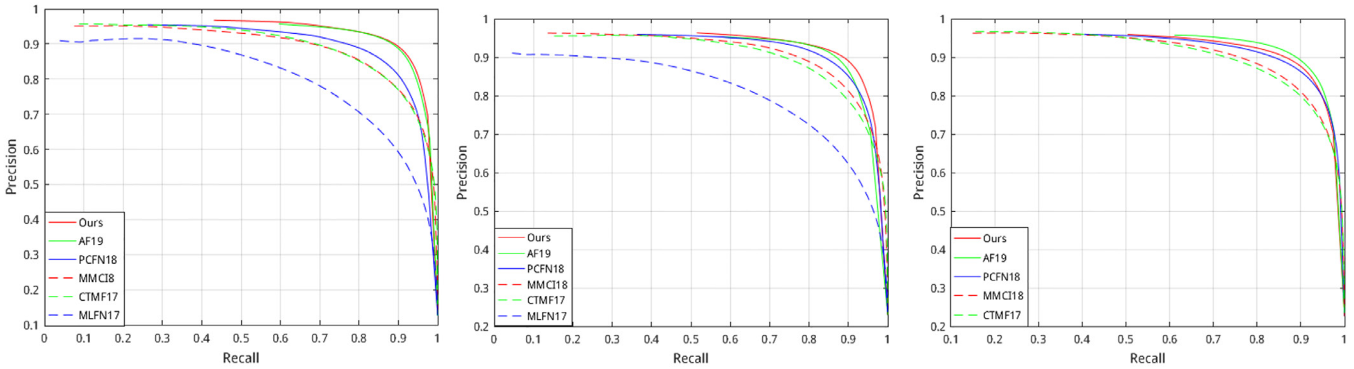
Model	NLPR1000				NJU2000				STEREO				LFSD			
	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑
double − stream	0.8310	0.0434	0.8660	0.9207	0.8615	0.0565	0.8676	0.9001	**0.8719**	0.0501	0.8804	**0.9161**	0.8183	0.0898	0.8207	0.8504
single − stream	**0.8526**	**0.0357**	**0.8932**	**0.9354**	**0.8746**	**0.0511**	**0.8827**	**0.9087**	0.8632	**0.0499**	0.8822	0.9146	**0.8480**	**0.0763**	**0.8587**	**0.8859**

Table 2
DRCNN module valuation.

Model	NLPR1000				NJU2000				STEREO				LFSD			
	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑
$DRCNN_N$	0.7717	0.0561	0.8561	0.8967	0.8077	0.0756	0.8502	0.8819	0.7911	0.0768	0.8454	0.8821	0.8203	0.1007	0.8286	0.8675
$DRCNN_Y$	**0.8526**	**0.0357**	**0.8932**	**0.9354**	**0.8746**	**0.0511**	**0.8827**	**0.9087**	**0.8632**	**0.0499**	**0.8822**	**0.9146**	**0.8480**	**0.0763**	**0.8587**	**0.8859**

Table 3
Comparison with DHSNet.

Model	NLPR1000				NJU2000				STEREO				LFSD			
	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑
$DHSNet_{4PD}$	0.8037	0.0429	0.8825	0.9111	0.8470	0.0604	0.8762	0.9007	0.8243	0.0644	0.8665	0.8920	0.8363	0.0869	0.8425	0.8714
$DRCNN_{4PD}$	**0.8526**	**0.0357**	**0.8932**	**0.9354**	**0.8746**	**0.0511**	**0.8827**	**0.9087**	**0.8632**	**0.0499**	**0.8822**	**0.9146**	**0.8480**	**0.0763**	**0.8587**	**0.8859**



(a)NLPR1000 dataset (b)NJU2000 dataset (c)STEREO dataset

Fig. 6. P–R curves comparison of different models based on PCFN training dataset.

Then we verify the effectiveness of DRCNN module. Model $DRCNN_N$ denotes single-stream network replaces DRCNN module with three convolution operations to generate side-output saliency maps. Model $DRCNN_Y$ denotes side-output is connected with DRCNN module for recursive refinement. Table 2 shows the valuation results about DRCNN module on four dataset. From the comparison results we can see that DRCNN module play an important role in recursively optimizing saliency maps.

4.5. Comparison with DHSNet

DHSNet [42] is very similar with ours which adopts the encoder–decoder architecture with skip-connection and recurrent convolutional layer(RCL) for salient object detection. But there are some differences in network structure and the manner of supervision. DHSNet adopts full connected layer with 784 nodes in the end of the encoder part and increases the parameters of network and the difficulty of convergence. DHSNet changes the size of ground truth to meet the supervision need of side-output saliency maps and increases the error of prediction result by downsampling ground truth. When RGBD four-channels input is fed into DHSNet, and original depth information and coarse saliency map are added into RCL like ours, its performance still inferior to ours, as shown in Table 3. Model $DHSNet_{4PD}$ denotes DHSNet with RGBD four-channels input equipped with coarse saliency map and original depth image in RCLs. $DRCNN_{4PD}$ denotes our network with RGBD four-channels input equipped with coarse saliency map and original depth image in RCLs. Comparison results demonstrate that our model achieves better performance than DHSNet on all the datasets.

4.6. Comparison with state-of-the-art RGB-D saliency models

We compare our method with state-of-the-art RGB-D saliency methods ACSD [14], SE [31], LBE [15], DF [21], PDNet [26], MLFN [17], CTMF [20], MMCI [33], PCFN [24] and AF [27]. Due to code or saliency maps absence of some models in some dataset, some curves or evaluation metrics values are missing. Training dataset of comparison models are different, so experiments are divided into two parts for fair comparison.

MLFN [17], CTMF [20], MMCI [33], PCFN [24] Method and AF [27] and our model trained in the same training dataset as PCFN [24] are grouped together. Their training dataset are the same and testing dataset are also the same, so comparison are conducted on their common testing dataset NLPR1000, NJU2000 and STEREO dataset and PR curves are demonstrated in the Fig. 6. The performance of our model beats the others on NLPR1000 and NJU2000 dataset. But it is inferior to AF [27] in STEREO

Table 4
F-measure, MAE, S-measure and E-measure comparisons of different models.

Model	NLPR1000				NJU2000				STEREO				LFSD			
	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑
ASCD15	0.5343	0.1787	0.6728	0.7417	0.6964	0.2021	0.6992	0.7863	0.6932	0.1956	0.7061	0.8048	0.7551	0.1881	0.7341	0.8011
SE16	0.6912	0.0913	0.7561	0.8385	0.7336	0.1687	0.6642	0.7722	0.7741	0.1452	0.7109	0.8308	0.7774	0.1670	0.6966	0.7834
LBE16	0.7355	0.0813	0.7619	0.8550	0.7401	0.1528	0.6953	0.7913	–	–	–	–	–	–	–	–
DF17	0.7348	0.0891	0.7909	0.8600	0.7703	0.1406	0.7596	0.8383	0.7650	0.1395	0.7664	0.8438	0.8133	0.1387	0.7898	0.8475
PDNet18	0.7968	0.0501	0.8864	0.8987	0.8228	0.0709	0.8770	0.8882	–	–	–	–	0.8283	0.1068	0.8471	0.8739
Ours (PDNet18)	**0.8526**	**0.0357**	**0.8932**	**0.9354**	**0.8746**	**0.0511**	**0.8827**	**0.9087**	**0.8632**	**0.0499**	**0.8822**	**0.9146**	**0.8480**	**0.0763**	**0.8587**	**0.8859**
MLFN17	0.6413	0.0889	0.7900	0.8208	0.7046	0.1374	0.7709	0.8087	–	–	–	–	–	–	–	–
CTMF17	0.7234	0.0561	0.8599	0.8690	0.7875	0.0847	0.8490	0.8638	0.7859	0.0867	0.8529	0.8699	–	–	–	–
MMCI18	0.7299	0.0591	0.8557	0.8717	0.8122	0.0790	0.8581	0.8775	0.8120	0.0796	0.8559	0.8896	–	–	–	–
PCFN18	0.7948	0.0437	0.8736	0.9163	0.8440	0.0591	0.8770	0.8966	0.8450	0.0606	0.8800	0.9054	–	–	–	–
AF19	0.8228	0.0329	0.9011	0.9316	0.8681	0.0534	0.8810	**0.9123**	**0.8725**	**0.0472**	**0.8921**	**0.9204**	–	–	–	–
Ours(PCFN18)	**0.8581**	**0.0302**	**0.9040**	**0.9438**	**0.8764**	**0.0474**	**0.8903**	0.9117	0.8611	0.0512	0.8817	0.9108	0.8432	0.0767	0.8501	0.8774

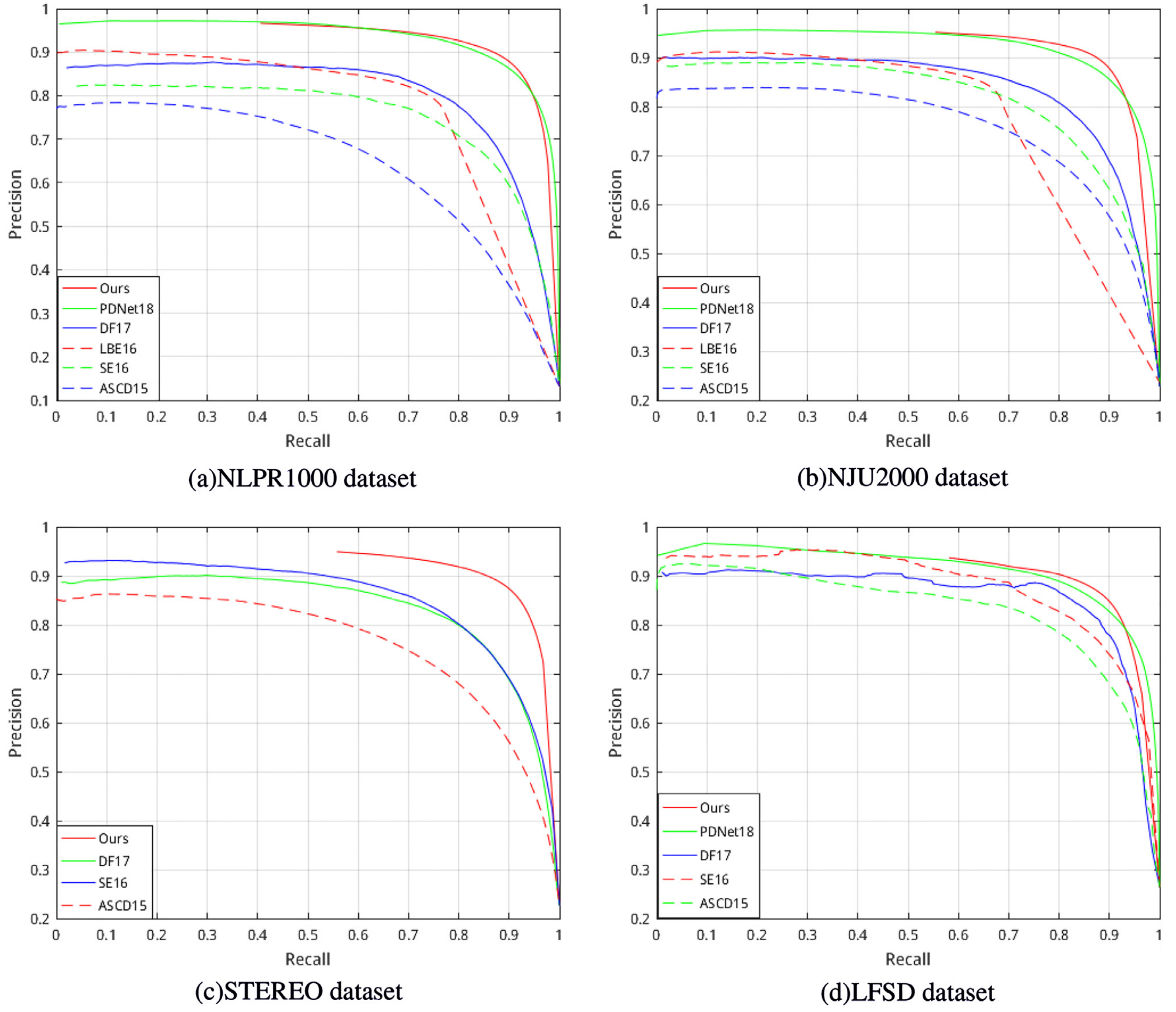


Fig. 7. P–R curves comparison of different models based on PDNet training dataset.

dataset. AF [27] can automatically choose the predictions from either RGB or depth modality and achieves outstanding performance in STEREO dataset which contains more depth data noise. So our method need to be further improved for removing the dirty depth information in the future. Their comparison results of evaluation metrics are shown in the bottom half part of Table 4.

The rest method and our model trained in the same training dataset as PDNet [26] are grouped together. PDNet and ours have the same training dataset, while the others don't belong to end-to-end deep learning method and testing dataset are tested directly. Fig. 7 demonstrates the comparison results. Our model outperforms other methods with a significant margin. Although there

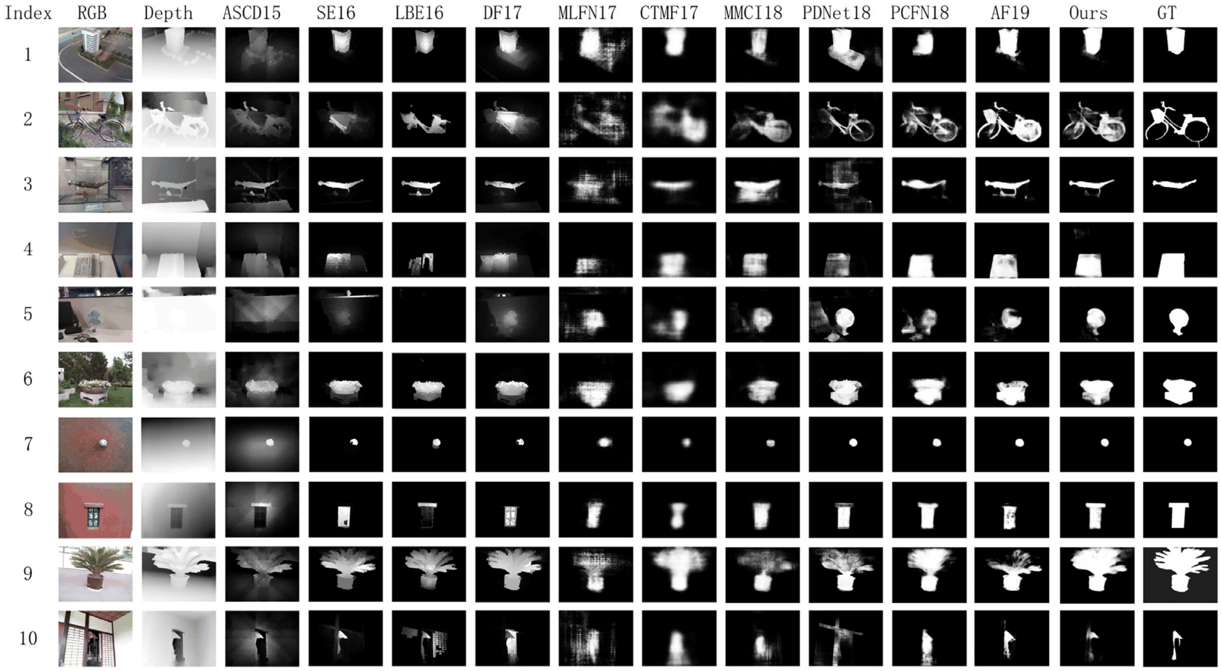


Fig. 8. Visual comparisons of different models.

is intersection between PDNet [26] and our model, our model wins it in more comprehensive evaluation metrics, as shown in the upper half part of Table. 4.

In addition Fig. 8 provides a visual comparison of our model with the other models. It can be observed that our model produces fine details and highlights the attention-grabbing salient regions.

5. Conclusion

In this paper, we propose the salient object detection method for RGB-D image by Single Stream Recurrent Convolution Neural Network with four-channels input and DRCNN subnetwork. RGBD four-channels input is presented to extract more original and accurate information to express RGB-D images. Ablation study shows that single-stream network with four-channels input is superior to double-stream network. DRCNN which is similar to RCL is proposed to take full advantage of original depth cue in RGB-D images and coarse saliency map from the deepest level feature for refining the outlines of salient objects from deep to shallow. Ablation study shows that DRCNN has the better effects than using convolutional operation to generate side-output saliency maps. By comparing with DHSNet, advantage of our model is further demonstrated. Even if DHSNet adopts the idea of DRCNN to progressively optimize saliency map, its performance is still inferior to ours. Meanwhile comparison experiments with the state-of-the-art methods demonstrate that our proposed method is excellent in extracting depth cue and fusing RGB and depth information. It achieves impressive performance and wins the others except for AF [27] in STEREO dataset. Inspired by its outstanding performance of AF [27], how to remove depth data noise will be further considered in the future. In addition, salient objects only correspond to partial regions of input image, so different spatial position of feature in CNNs have different importance. Different channels of feature in CNNs generate different response to foreground or background too. Our method treats all spatial position of feature in each channel and all channels without distinction. Future works will also focus more on adding different weights on image feature, which helps to generate more effective features for salient object detection.

Conflict of interest statement

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service or company that could be construed as the review of the manuscript.

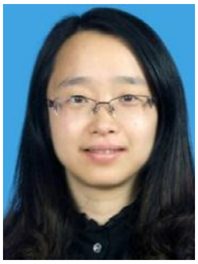
Acknowledgment

We thank Dr. Hao Chen from City University of Hong Kong for providing their result saliency maps. We also thank Prof. Mingming Cheng and Dr. Deng-ping Fan from Nankai University for providing the codes of all evaluation metrics. We further thank all anonymous reviewers for their valuable comments. This research is supported by National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2015BAK24B01), National Natural Science Foundation of China (61602004), Open issues on Co-Innovation Center for Information Supply & Assurance Technology, Anhui University (ADXXBZ201610), Natural Science Foundation of Anhui Province (1908085MF182).

References

[1] F. Zhang, B. Du, L. Zhang, Saliency-guided unsupervised feature learning for scene classification, IEEE Trans. Geosci. Remote Sens. 53 (4) (2015) 2175–2184.
[2] C. Craye, D. Filliat, J.-F. Goudou, Environment exploration for object-based visual saliency learning, in: Proceedings of the International Conference on Robotics and Automation (ICRA), IEEE, 2016, pp. 2303–2309.
[3] N. Ouerhani, H. Hugli, Computing visual attention from scene depth, Pattern Recognit. 1 (2000) 375–378. IEEE.
[4] W. Geng, R. Ju, X. Xu, T. Ren, G. Wu, Flat3d: browsing stereo images on a conventional screen, in: Proceedings of the International Conference on Multimedia Modeling, Springer, 2015, pp. 546–558.
[5] J. Wang, Y. Fang, M. Narwaria, W. Lin, P. Le Callet, Stereoscopic image retargeting based on 3d saliency detection, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 669–673.
[6] N. Imamoglu, Z. Wei, H. Shi, Y. Yoshida, M. Nergui, J. Gonzalez, D. Gu, W. Chen, K. Nonami, W. Yu, An improved saliency for RGB-d visual tracking and control strategies for a bio-monitoring mobile robot, in: Proceedings of the International Competition on Evaluating AAL Systems through Competitive Benchmarking, Springer, 2013, pp. 1–12.

[7] S. Naotoshi, T. Kanji, Y. Kentaro, Scene retrieval by unsupervised salient part discovery, in: Proceedings of the International Conference on Machine Vision Applications (MVA), IEEE, 2015, pp. 85–88.
[8] X. Wang, L. Ma, S. Kwong, Y. Zhou, Quaternion representation based visual saliency for stereoscopic image quality assessment, Signal Process. 145 (2018) 202–213.
[9] B. Zhang, R. Ju, T. Ren, G. Wu, Say cheese: personal photography layout recommendation using 3d aesthetics estimation, in: Proceedings of the Pacific Rim Conference on Multimedia, Springer, 2016, pp. 13–23.
[10] Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, in: Proceedings of the International Conference on Internet Multimedia Computing and Service, ACM, 2014, p. 23.
[11] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBd salient object detection: a benchmark and algorithms, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 92–109.
[12] J. Ren, X. Gong, L. Yu, W. Zhou, M. Ying Yang, Exploiting global priors for RGB-d saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 25–32.
[13] S.-T. Wang, Z. Zhou, H.-B. Qu, B. Li, Visual saliency detection for RGB-d images with generative model, in: Proceedings of the Asian Conference on Computer Vision, Springer, 2016, pp. 20–35.
[14] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, in: Proceedings of the International Conference on Image Processing (ICIP), IEEE, 2014, pp. 1115–1119.
[15] D. Feng, N. Barnes, S. You, C. McCarthy, Local background enclosure for RGB-d salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2343–2350.
[16] D. Feng, N. Barnes, S. You, Hoso: Histogram of surface orientation for RGB-d salient object detection, in: Proceedings of the IEEE Conference on Digital Image Computing: Techniques and Applications (DICTA), IEEE, 2017, pp. 1–8.
[17] H. Chen, Y. Li, D. Su, RGB-d saliency detection by multi-stream late fusion network, in: Proceedings of the IEEE International Conference on Computer Vision Systems, Springer, 2017, pp. 459–468.
[18] H. Chen, Y.F. Li, D. Su, RGB-d salient object detection based on discriminative cross-modal transfer learning, arXiv preprint arXiv:1703.00122 (2017).
[19] H. Chen, Y.-F. Li, D. Su, M 3 net: multi-scale multi-path multi-modal fusion network and example application to RGB-d salient object detection, in: Proceedings of the International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 4911–4916.
[20] J. Han, C. Hao, N. Liu, C. Yan, X. Li, CNNS-based RGB-d saliency detection via cross-view transfer and multiview fusion, IEEE Trans. Cybern. PP (99) (2017) 1–13.
[21] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, Q. Yang, RGBd salient object detection via deep fusion, IEEE Trans. Image Process. 26 (5) (2017) 2274–2285.
[22] R. Shigematsu, D. Feng, S. You, N. Barnes, Learning RGB-d salient object detection using background enclosure, depth contrast, and top-down features, arXiv preprint arXiv:1705.03607 (2017).
[23] X. Xu, Y. Li, G. Wu, J. Luo, Multi-modal deep feature learning for RGB-d object detection, Pattern Recognit. 72 (2017) 300–313.
[24] H. Chen, Y. Li, Progressively complementarity-aware fusion network for RGB-d salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3051–3060.
[25] S. Wang, Z. Zhou, W. Jin, H. Qu, Visual saliency detection for RGB-d images under a Bayesian framework, IPSJ Trans. Comput. Vis. Appl. 10 (1) (2018) 1.
[26] C. Zhu, X. Cai, K. Huang, T.H. Li, G. Li, Pdnet: prior-model guided depth-enhanced network for salient object detection, arXiv preprint arXiv:1803.08636 (2018).
[27] N. Wang, X. Gong, Adaptive fusion for RGB-d salient object detection, arXiv preprint arXiv:1901.01369 (2019).
[28] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from RGB-d images for object detection and segmentation, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 345–360.
[29] K. Desingh, K.K. Madhava, D. Rajan, C.V. Jawahar, Depth really matters: improving visual salient region detection with depth, in: Proceedings of the British Machine Vision Conference, 2013, pp. 98.1–98.11.
[30] H. Xue, Y. Gu, Y. Li, J. Yang, RGB-d saliency detection via mutual guided manifold ranking, in: Proceedings of the IEEE Conference on Image Processing (ICIP), IEEE, 2015, pp. 666–670.
[31] J. Guo, T. Ren, J. Bei, Salient object detection for RGB-d image via saliency evolution, in: Proceedings of the IEEE Conference on Multimedia and Expo (ICME), IEEE, 2016, pp. 1–6.
[32] Y. Tang, R. Tong, M. Tang, Y. Zhang, Depth incorporating with color improves salient object detection, Vis. Comput. 32 (1) (2016) 111–121.
[33] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-d salient object detection, Pattern Recognit. (2018).
[34] A. Ciptadi, T. Hermans, J. Rehg, An in depth view of saliency, in: Proceedings of the British Machine Vision Conference, 2013, pp. 112.1–112.11.
[35] H. Song, Z. Liu, H. Du, G. Sun, Depth-aware saliency detection using discriminative saliency fusion, in: Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 1626–1630.
[36] T. Zhang, Z. Yang, J. Song, RGB-d saliency detection with multi-feature-fused optimization, in: Proceedings of the International Conference on Image and Graphics, Springer, 2017, pp. 15–26.
[37] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, T. Ren, Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning, IEEE Trans. Image Process 26 (9) (2017) 4204–4216.
[38] L. Zhu, Z. Cao, Z. Fang, Y. Xiao, J. Wu, H. Deng, J. Liu, Selective features for RGB-d saliency, in: Proceedings of the IEEE Conference on Chinese Automation Congress (CAC), IEEE, 2015, pp. 512–517.
[39] L. Jiang, A. Koch, A. Zell, Salient regions detection for indoor robots using RGB-d data, in: Proceedings of the Conference on Robotics and Automation (ICRA), IEEE, 2015, pp. 1323–1328.
[40] J. Zhao, Y. Cao, D.-P. Fan, X.-Y. Li, L. Zhang, M.-M. Cheng, Contrast prior and fluid pyramid integration for RGBD salient object detection, in: Proceedings of the Conference on IEEE CVPR, 2019, pp. 1–10.
[41] M. Liang, X. Hu, Recurrent convolutional neural network for object recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2015, pp. 3367–3375.
[42] N. Liu, J. Han, Dhsnet: deep hierarchical saliency network for salient object detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2016, pp. 678–686.
[43] T.-N. Le, A. Sugimoto, Deeply supervised 3d recurrent FCN for salient object detection in videos, in: Proceedings of the British Machine Vision Conference, BMVC, 2017, pp. 1–13.
[44] D.-P. Fan, W. Wang, M.-M. Cheng, J. Shen, Shifting more attention to video salient object detection, in: Proceedings of the Conference on IEEE CVPR, 2019, pp. 1–11.
[45] C. Lang, T.V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, S. Yan, Depth matters: Influence of depth cues on visual saliency, in: Proceedings of the Conference on Computer Vision–ECCV, Springer, 2012, pp. 101–115.
[46] J. Guo, T. Ren, J. Bei, Y. Zhu, Salient object detection in RGB-d image based on saliency fusion and propagation, in: Proceedings of the Seventh International Conference on Internet Multimedia Computing and Service, ACM, 2015, p. 59.
[47] H. Du, Z. Liu, H. Song, L. Mei, Z. Xu, Improving RGBD saliency detection using progressive region classification and saliency fusion, IEEE Access 4 (2016) 8987–8994.
[48] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Proceedings of the 2013 IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 6645–6649.
[49] R. Socher, C.C. Lin, C. Manning, A.Y. Ng, Parsing natural scenes and natural language with recursive neural networks, in: Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 129–136.
[50] L. Wang, L. Wang, H. Lu, P. Zhang, X. Ruan, Saliency detection with recurrent fully convolutional networks, in: Proceedings of the Conference on European Conference on Computer Vision, Springer, 2016, pp. 825–841.
[51] L. Wang, L. Wang, H. Lu, P. Zhang, X. Ruan, Salient object detection with recurrent fully convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2018). 1–1.
[52] Y. Tang, X. Wu, W. Bu, Deeply-supervised recurrent convolutional neural network for saliency detection, in: Proceedings of the 2016 ACM Multimedia Conference, ACM, 2016, pp. 397–401.
[53] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: Proceedings of the Conference on Artificial Intelligence and Statistics, 2015, pp. 562–570.
[54] X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, Progressive attention guided recurrent network for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 714–722.
[55] X. Jin, Y. Chen, Z. Jie, J. Feng, S. Yan, Multi-path feedback recurrent neural networks for scene parsing, in: Proceedings of the Conference on AAAI, vol. 3, 2017, p. 8.
[56] F. Sun, W. Li, Y. Guan, Self-attention recurrent network for saliency detection, Multimed. Tools Appl. (2018) 1–15.
[57] F. Xiao, W. Deng, L. Peng, C. Cao, K. Hu, X. Gao, MSDNN: multi-scale deep neural network for salient object detection, arXiv preprint arXiv:1801.04187 (2018).
[58] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[59] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 454–461.
[60] N. Li, J. Ye, Y. Ji, H. Ling, J. Yu, Saliency detection on light field, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition, 2014, pp. 2806–2813.
[61] D.R. Martin, C.C. Fowlkes, J. Malik, Learning to detect natural image boundaries using local brightness, color, and texture cues, IEEE Trans. Pattern Anal. Mach. Intell. 26 (5) (2004) 530–549.
[62] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4558–4567.
[63] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, arXiv preprint arXiv:1805.10421 (2018).
[64] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, A. Borji, Salient objects in clutter: Bringing salient object detection to the foreground, in: Proceedings of the European Conference on Computer Vision, Springer, 2018, pp. 196–212.
[65] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the Twenty-second ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.

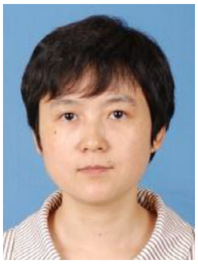


Zhengyi Liu is an associate professor in School of Computer Science and Technology, Anhui University, China. She received her B.S., M.S., and Ph.D. from Anhui University, China in 2001, 2004 and 2007, respectively. Her research interests include image and video processing, computer vision and deep learning.

Song Shi is a M.S. Candidate of Anhui University. He received his B.S. from Anhui University, China in 2016. His research interests include image and video processing and computer vision.

Quntao Duan is a M.S. Candidate of Anhui University. She received her B.S. from Jiangxi agricultural University, China in 2017. Her research interests include image and video processing and computer vision.

Wei Zhang is a M.S. Candidate of Anhui University. He received his B.S. from University of Science and Technology Liaoning, China in 2018. His research interests include image and video processing and computer vision.



Peng Zhao is an associate professor in School of Computer Science and Technology, Anhui University, China. She received her B.S. and M.S. from Anhui University, China in 1998 and 2003, respectively. She received Ph.D. from University of Science and Technology of China in 2006. Her research interests include image processing, and machine learning.