DOI: 10.1111/cgf.14441

COMPUTER GRAPHICS *forum*
Volume 41 (2022), number 1 pp. 377–388

GlassNet: Label Decoupling-based Three-stream Neural Network for Robust Image Glass Detection

Chengyu Zheng,[1] Ding Shi,[1] Xuefeng Yan,[1] Dong Liang,[1] Mingqiang Wei,[1] Xin Yang,[2] Yanwen Guo[3] and Haoran Xie[4]

[1]Nanjing University of Aeronautics and Astronautics, Nanjing, China
mqwei@nuaa.edu.cn
[2]Dalian University of Technology, Dalian, China
[3]Nanjing University, Nanjing, China
[4]Lingnan University, Hong Kong, China
hrxie@ln.edu.hk

Abstract

Most of the existing object detection methods generate poor glass detection results, due to the fact that the transparent glass shares the same appearance with arbitrary objects behind it in an image. Different from traditional deep learning-based wisdoms that simply use the object boundary as an auxiliary supervision, we exploit label decoupling to decompose the original labelled ground-truth (GT) map into an interior-diffusion map and a boundary-diffusion map. The GT map in collaboration with the two newly generated maps breaks the imbalanced distribution of the object boundary, leading to improved glass detection quality. We have three key contributions to solve the transparent glass detection problem: (1) We propose a three-stream neural network (call GlassNet for short) to fully absorb beneficial features in the three maps. (2) We design a multi-scale interactive dilation module to explore a wider range of contextual information. (3) We develop an attention-based boundary-aware feature Mosaic module to integrate multi-modal information. Extensive experiments on the benchmark dataset exhibit clear improvements of our method over SOTAs, in terms of both the overall glass detection accuracy and boundary clearness.

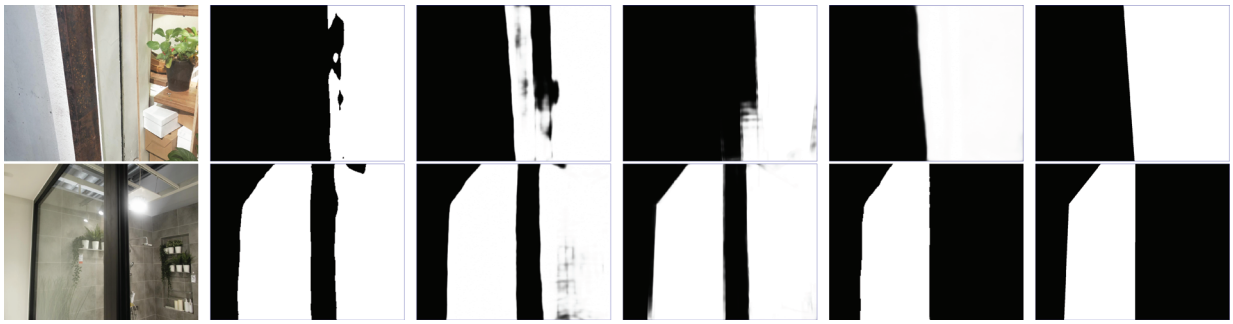
Keywords: image processing, image and video processing, image segmentation, image and video processing, computer vision–shape recognition, methods and applications

CCS Concepts: • Computing methodologies → Object detection

1. Introduction

Transparent glass is widely used in our daily life, such as glass windows/doors and many other glass products. However, it commonly hinders many vision-related tasks like depth prediction, instance segmentation, reflection removal, object detection and so forth. For example, when an intelligent robot or an unmanned plane operates automatically, they should avoid crashing into the glass. It is, therefore, essential to accurately detect the overall glass with its boundary clearly from single images. Unfortunately, most of the existing object detection methods generate inaccurate or even wrong regions of the glass with fuzzy boundaries, due to the fact that the glass is transparent. That means, a glass region nearly has no fixed patterns; the pattern is determined by the arbitrarily appeared object behind the glass. Therefore, unlike many other objects, which have relatively fixed patterns to detect more easily, the same appearance between the glass region and the objects behind it makes existing object detection methods work ineffectively. We list three representative detection approaches in Figure 1, i.e. DANet [FLT*19] for semantic segmentation, edge guidance network (EGNet) [ZLF*19] for edge-guided salient object detection (SOD), glass detection network (GDNet) [MYW*20] for glass detection as well as our proposed GlassNet. As shown, DANet wrongly considers the background as the glass; EGNet also yields wrong detection regions; although GDNet [MYW*20] pioneers to automatically detect glass from single images, it leads to inaccurate glass boundaries; while the proposed network operates smoothly on these two challenging images: The glass is exactly detected with its clearer boundaries by GlassNet.

Intuitively, like other vision tasks, a straightforward solution to enhance the glass detection ability, is to use boundaries of the glass as auxiliary supervision. However, in an image with glass in it, the glass-boundary pixels are much rarer than other pixels. Such a very unbalanced distribution of the glass-boundary pixels will introduce large prediction errors around the glass boundary. To this



(a) Input (b) DANet [FLT*19] (c) GDNet [MYW*20] (d) EGNet [ZLF*19] (e) GlassNet (ours) (f) GT

Figure 1: *GlassNet compares with its competitors on the public GDD dataset [MYW*20]. Current vision systems sense the presence of glass poorly, since a glass region has no fixed patterns (e.g. various objects will appear behind the glass, resulting in the same appearance of the glass and the objects behind the glass in an image). DANet commonly fails to detect the glass; glass detection network (GDNet) leads to inaccurate glass boundaries; edge guidance network (EGNet) yields wrong detection regions; while the proposed GlassNet operates smoothly on the two challenging images, where the glass is exactly detected and its boundaries are clearer. Please note that the white region corresponds to the detected glass, and the black region means the detected background.*

end, we arise an intriguing question that if the glass boundary diffuses itself into the glass's interior and the interior diffuses itself from its centre to boundary, a deep network can better focus on regions around the glass boundary and concentrate on centre areas of the glass object? To answer it, (1) we first use label decoupling (LD) [WWW*20] to explicitly decompose the original glass map into an interior-diffusion map and a boundary-diffusion map, where the first map is concentrated in the centre of glass objects and the second map focuses on regions around glass boundaries; (2) based on the three different types of label information, we propose a three-stream neural network for robust glass detection (GlassNet). For the interior-diffusion stream, we only use the highest two-level image features with rich semantic information to locate the glass region; for the boundary-diffusion stream, all levels of information are aggregated to make the detection result more accurate; for the original glass stream, we utilize the lowest two-level image features with more detailed information and highest-level image features to predict the final glass maps.

Meanwhile, we design a multi-scale interactive dilation (MID) module with a large receiving field to integrate the features from adjacent levels. And we propose an attention-based boundary fusion module to merge the boundary and glass features. We have tested all the approaches on the benchmark dataset GDD [MYW*20] and our GlassNet achieves a very competitive performance. In summary, our contributions are mainly four-fold:

1. We observe that in an image with glass in it, the glass-boundary pixels are much rarer than other pixels. Such a very imbalanced distribution of the glass-boundary pixels introduces large prediction errors around the glass boundary when performing object detection. To break such an imbalanced distribution between glass-boundary pixels and non-glass-boundary pixels, we utilize the LD procedure to decompose a glass label into an interior-diffusion map and a boundary-diffusion map to supervise the network training.
2. We propose a three-stream network, called GlassNet, which is enhanced by LD features to produce more precise glass maps.
3. We design a MID module to explore a wider range of contextual information and an attention-based boundary-aware feature Mosaic (BFM) module to integrate multi-modal information.
4. Extensive experiments on the benchmark dataset exhibit clear improvements of our method over SOTAs, in terms of both the overall glass detection accuracy and boundary clearness.

2. Related Work

In the past 2 years, glass detection had begun to attract much attention, but little work has been done on this topic. In this section, we briefly introduce the methods used in glass detection and the methods that can assist in solving this problem from relevant fields, including semantic segmentation, SOD and mirror detection.

Semantic segmentation. Semantic segmentation is a key problem in the computer vision community, which aims at assigning semantic class labels to each pixel in the given image. With the development of deep neural networks, an end-to-end training architecture method called fully convolutional networks (FCNs) [LSD15] has been proposed to solve this problem, which uses multi-scale context fusion to achieve high segmentation performance. However, the fixed geometric structures of convolution operations in those deep neural networks make the pixels capture local information and short-range contextual information inherently. Thus, Chen *et al*. [CPK*18] introduce an atrous spatial pyramid pooling module (ASPP) with multi-scale dilation convolutions for contextual information aggregation. Zhao *et al*. [ZSQ*17] further propose PSPNet to capture a wider range of contextual information by using a pooling operation and the pyramid structure. In addition, the encoder–decoder structures, like U-Net [RFB15], are widely used to fuse middle- and high-level semantic features.

However, the dilated convolution-based methods [DJS*18, CPSA17] fail to capture global contextual information and cause

sparse local information due to their structures. The pool-based methods [ZDS*18, ZTZ*17] aggregate context information in a non-adaptive way to make image pixels use the homogeneous contextual information. Therefore, Wang *et al*. [WGGH18] introduce non-local networks utilizing a self-attention mechanism [VSP*17, CDL16], which calculate the relationship between each pixel and all other pixels in an image, thus harvesting global contextual information. To solve the problem that self-attention-based methods have high computation complexity and occupy a huge amount of GPU memory, Huang *et al*. [HWH*19] and Fu *et al*. [FLT*19], respectively, propose CCNet and DANet to reduce the parameters. After that, Carion *et al*. [CMS*20] adopt transformer that is widely used in the NLP field, which replaces the convolution layer with the self-attention layer, for semantic segmentation.

SOD. SOD aims at identifying the most visually distinctive objects or regions in an image, which is widely applied as a pre-processing procedure for downstream tasks [XWL*18, 2021]. Early SOD methods are mainly based on hand-crafted features (e.g. colour, texture and contrast) to segment salient objects in the scene [YZL*13, ZLWS14, SB15, ZSL*15]. Recent convolutional neural networks (CNNs) [KSH12, SZ14, HZRS16] are extensively used and achieve very remarkable performance.

Ronneberger *et al*. [RFB15] propose U-Net, a representative network widely used in a variety of graphics processing tasks, which effectively generates more accurate detection results by using a skip connection operation and an encoder–decoder structure. Based on U-Net, many other methods adopt different decoders, combined with multi-level CNN features and have achieved remarkable performance. Zhang *et al*. [ZWL*17] introduce an AmuletNet for SOD that aggregates an another-level convolutional feature at each different level. Zhang *et al*. [ZWQ*18] add an attention module to the decoder, which can guide the network to selectively integrate multi-level features. Zhao *et al*. [ZW19] propose a pyramid feature attention network (PFAN) to enhance the high-level context features and the low-level spatial structural features. Pang *et al*. [PZZL20] propose aggregate interaction modules to integrate the features from adjacent levels by using a more complex decoder structure. Besides, more efforts utilize boundary information to improve the accuracy of saliency maps. Zhao *et al*. [ZLF*19] focus on the complementarity between salient edge information and salient object information and present an EGNet for SOD. Zhou *et al*. [ZXL*20] analyse the correlation between saliency and boundary and introduce an interactive two-stream decoder to explore multiple cues, including saliency, boundary and their correlation. Furthermore, Wei *et al*. [WWW*20] propose a label decoupling framework (LDF) that exploits more boundary information to enhance SOD performance.

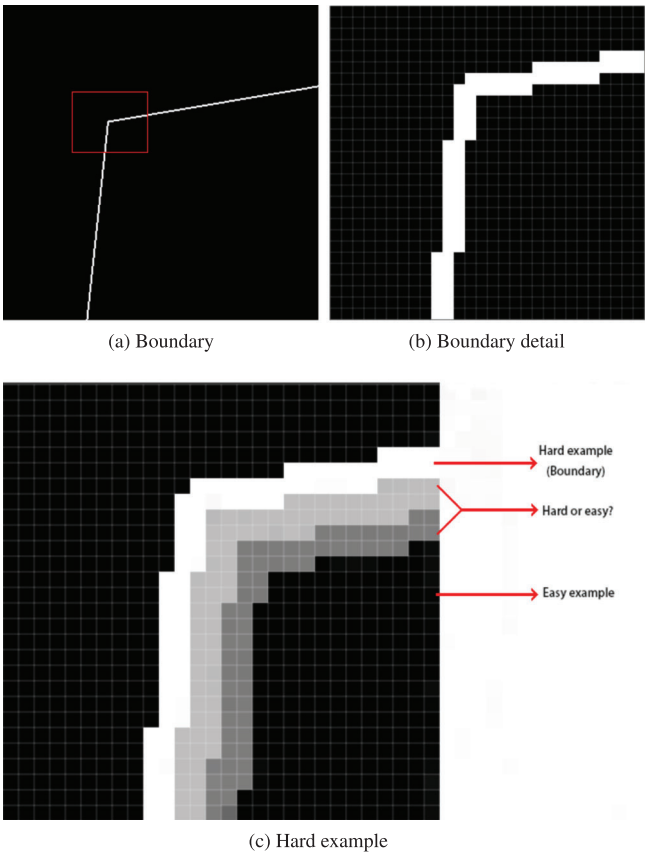
Mirror detection. Similar to other image detection tasks, mirror detection aims at segmenting mirror regions in single images. Yang *et al*. [YMX*19] make the first attempt to automatically detect mirrors and propose MirrorNet by utilizing inconsistencies between the inside and outside of the mirror region, called contextual contrasted features, to segment mirrors from the real scene. The reason is the performance difference between the mirror region and other non-mirror regions. The mirror region reflects the scene in front of the mirror, which makes the semantic and low-level discontinuities often occur at the boundary of the mirror. But not all mirrors have a great distinction between inside and outside. Some of them have little contextual contrasted information. Thus, Lin *et al*. [LWL20] introduce a model to progressively learn the content similarity between the inside and outside of the mirror while explicitly detecting the mirror boundaries. The scenes reflected by a mirror often exhibit similarities to scenes outside the mirror, which can aid to detect mirror regions by enlarging the receptive fields of the convolution operation. Glass detection is very similar to mirror detection that also has the problem of similar foreground and background.

Transparent object detection (TOD). Similar to glass detection, TOD aims to segment transparent object regions in single images. Xie *et al*. [XWW*20] propose a large-scale dataset for TOD named Trans10K and a novel boundary-aware segmentation method termed TransLab to address the TOD problem. However, there exists a difference between TOD and GD: TOD is a multi-label segmentation problem, while GD is a binary segmentation problem. This fact indicates that TOD does not operate smoothly on the GD task and vice versa. This is why we do not compare our method with those TOD methods.

Glass detection. Glass regions in an image do not have a fixed pattern since they depend on what appears behind the glass, and the content of the glass region is the content of the background region. This situation makes it difficult to distinguish between the glass and the background region, even using state-of-the-art segmentation methods. Meanwhile, other object detection methods are also not suitable for glass detection tasks on account of the difference between glass and other objects. Mei *et al*. [MYW*20] pioneer to propose a novel gGDNet) by exploring abundant contextual features from a large receptive field. They utilize multiple well-designed large-field contextual feature integration (LCFI) modules for the precise positioning of the glass region, but this method has poor performance in some cases where the glass boundary region or scene is very complex or the background inside and outside the glass is insufficient. Lin *et al*. [LHL21] observe that humans often rely on identifying reflections to sense the existence of glass and also rely on locating the boundary in order to determine the extent of the glass. They propose a rich context aggregation module (RCAM) to extract multi-scale boundary features and a reflection-based refinement module (RRM) to detect reflection. Then, they utilize two modules for glass surface detection to solve the problem of insufficient contexts in part of the scene.

3. Methodology

Motivation. Due to the unbalanced distribution between boundary and background pixels, only using boundary pixels for glass detection will lead to larger prediction errors of pixels close to the boundary than those far away from the glass. Therefore, the glass boundary should diffuse itself into the glass's interior to amplify its influence. Conversely, the glass's interior should diffuse itself from the centre to the boundary to loosen its influence. Based on this observation, we propose to decouple the glass label into the interior-diffusion component and the boundary-diffusion component, both of which are auxiliary supervisions to enhance the overall glass detection quality and boundary clearness. To make full use of the decoupled supervisions, we further present a three-stream network, which consists of the proposed MID modules to effectively integrate large-field contextual features for detecting glass of different sizes.



(a) Boundary (b) Boundary detail

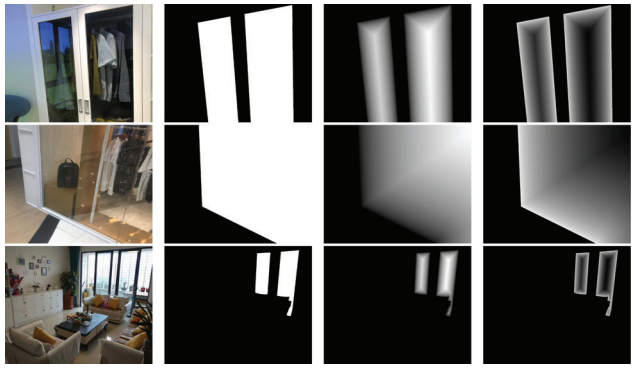
(c) Hard example

Figure 2: *Illustration of hard examples: (a) the boundary obtained by the boundary extraction algorithm in edge guidance network (EGNet) [ZLF*19], (b) the boundary detail, which only has two pixels wide and (c) it is challenging to determine the pixels near the boundary to be hard examples or not.*

Also, our proposed network learns to fuse multi-modal information to further enhance the performance.

3.1. Label decoupling

Many object detection methods pay attention to boundary information for the enhancement of detection accuracy, but the prediction difficulty of boundary pixels is closely related to their locations. Therefore, it is difficult to classify the pixels near the boundary correctly, which is called 'hard examples' [CZXL19]. In contrast, the consistency of interior regions makes the central region easier to detect. Therefore, a strategy of dealing with boundary and interior pixels differently will make the detection results more reasonable. However, it is difficult to claim which pixels are hard examples or not, as illustrated in Figure 2. We adopt the LD strategy proposed by Wang *et al*. [ZLF*19] to decouple the original glass map into an interior-diffusion map and a boundary-diffusion map, as shown in Figure 3. In more detail, LD uses the simple distance transformation (DT) to convert the ground-truth glass map into a new image, where the value of the foreground pixel is the minimum distance from the background obtained by the distance function. Please note



(a) Input (b) GT (c) Interior-diffusion label (d) Boundary-diffusion label

Figure 3: *Examples of label decoupling. In the interior-diffusion label (c) of GT (b), pixels close to the centre of the glass have larger values. In the boundary-diffusion label (d) of GT (b), pixels near the boundary of the glass have larger values. The sum of (c) and (d) is equal to (b).*

that the foreground herein refers to the glass region, and the background means the remaining non-glass region.

DT calculates the distance from the nearest zero points to itself for each non-zero point in an image. Its input is a binary graph such as the ground truth of the image detection task, which can be divided into two groups (i.e. the foreground I_{fg} and the background I_{bg}). The original metric function is defined as $f(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$ to calculate the distance between two pixels, and here we modify $f(p,q)$ to fit our approach. The new distance function is formulated as:

$$I'(p) = \begin{cases} \min_{q \in I_{bg}} f(p,q), & p \in I_{fg} \\ 0, & p \in I_{bg} \end{cases} \quad (1)$$

For the foreground pixel p, DT calculates the original distance function $f(p,q)$ by looking for its nearest pixel q in the background pixel, and directly sets the value to 0 for the background pixel. We use a linear normalization function $I' = \frac{I' - \min(I')}{\max(I') - \min(I')}$, which normalizes the image generated from the new distance function. Compared with the original image, the new image obtained by defining the distance function depends not only on its foreground or the background but also on its relative position. Therefore, the new image corresponds to the inner part of the original image, and the closer to the centre is, the larger the pixel value will be. The boundary images obtained by subtracting the new image from the original image can help deal with the hard examples. To remove the background interference, we process the new image and the original ground truth to generate the interior-diffusion label and boundary-diffusion label as:

$$Label \Rightarrow \begin{cases} BL = I * I' \\ DL = I * (1 - I') \end{cases} \quad (2)$$

where BL means the interior-diffusion label and DL represents the boundary-diffusion label. Thus, we decouple the original label into two different kinds of labels, to work in learning both interior and boundary features with different characteristics.

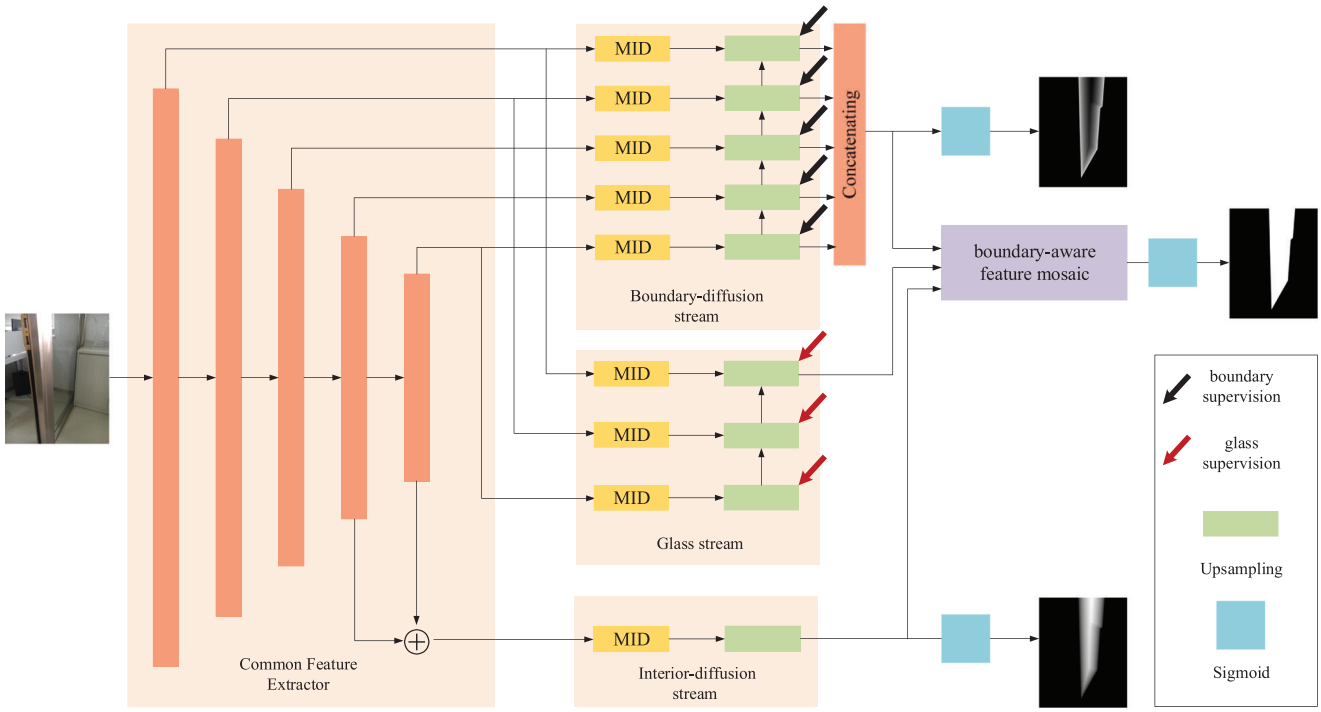


Figure 4: *Overview of the proposed GlassNet. The pre-trained ResNet-50 [HZRS16] is employed as the backbone network to extract multi-level image features. The extracted image features are fed into the three streams. In each of the three streams, we use multi-scale interactive dilation module (MID) to extract large-field contextual features, to obtain glass features, interior-diffusion features and boundary-diffusion features, respectively, through supervision. The three different features are fused through an attention-based boundary-aware feature Mosaic module (BFM) and fed themselves into the predict block to generate the final glass map.*

3.2. Network overview

The overview of the proposed model is illustrated in Figure 4, which consists of three parallel streams: an interior-diffusion stream, a boundary-diffusion stream and a glass stream. We first feed an image to the backbone network to extract multi-scale backbone features. Then, the features of each level are fed into the three streams supervised by the decoupled labels to generate different features. In each stream, we use the MID to extract large-field contextual features, then obtain the glass features, interior-diffusion features and boundary-diffusion features for each stream. Finally, we use the attention-based BFM module to integrate the boundary and interior-diffusion features into the glass prediction maps to generate the final glass map of the whole network. Details of the proposed approach are described as follows.

Feature encoder. We use ResNet-50 [HZRS16] as the backbone network to extract common multi-level image features for the three streams as suggested elsewhere [WBZ*17, WZW*18, LHY18]. In particular, as a backbone network, we remove the last global pooling and fully connected layers and only use the five residual blocks. For the sake of simplification, we represent these five blocks as $f_i(w_i)$, $i \in \{1, \ldots, 5\}$, where w_i is the weight parameters pre-trained on ImageNet [DDS*09] of the $f_i(\cdot)$ operation, and the output of the ith layer $f_i(\cdot)$ is the input of $f_{i+1}(\cdot)$, $\forall i \in \{1, \ldots, 4\}$. We feed an input image with the shape $H \times W$ into it to generate different-scale features denoted as $EF = \{EF_i | i = 1, 2, 3, 4, 5\}$ by utilizing $f_i(w_i)$, $i \in \{1, \ldots, 5\}$, i.e. $EF_{i+1} = f_i(EF_i)$. Then, we input the different levels of features into the three streams for processing. The features EF_5 and EF_4 are fed into the interior-diffusion stream decoder to roughly locate the glass region. In order to obtain a finer glass boundary, the features $\{EF_i | i = 1, 2, 3, 4, 5\}$ are fed into the boundary-diffusion stream decoder. In addition, we utilize the features $\{EF_i | i = 1, 2, 5\}$ for the glass map generation in the glass stream.

Three-stream decoder. As shown in Figure 4, we built a three-stream network to use the LD information. We utilize a LD procedure to decompose a glass label into an interior-diffusion map and a boundary-diffusion map to supervise the model separately. Through the supervision of these three different labels, better detection results can be obtained.

In each stream, we use a MID module to extract large-field contextual features. Figure 4 illustrates the detailed structure of the decoder. For the glass stream and the boundary-diffusion stream, we employ the short connections [HCH*17] to merge feature maps EF_i at different CNN layers, resulting in new feature maps (denoted as DF_i). Specifically, the merged feature map DF_i at the kth CNN layer ($i = 1, \ldots, 5$) is computed by

$$DF_i = Conv(Concat(MF_i, \ldots, MF_5)) \quad (3)$$

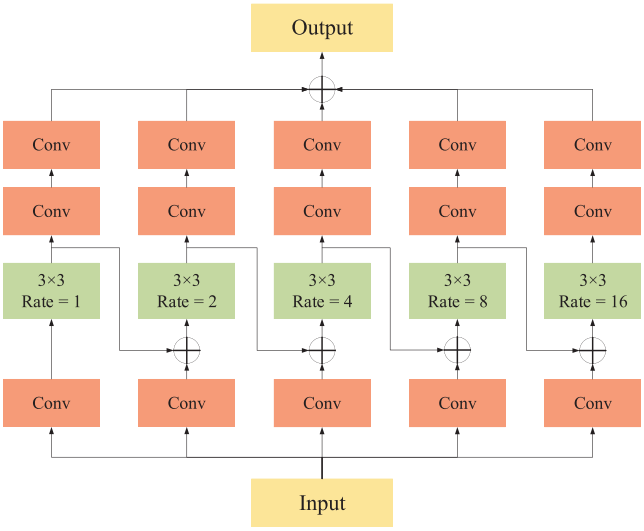


Figure 5: *The structure of the multi-scale onteractive dilation (MID) module. Conv represents a convolutional layer with a kernel size of* 3×3*, a normalized layer, and a ReLU layer.* 3×3 *means the convolutional kernel size, and rate represents the dilation rate in a dilated convolution.*

Then, we use MID to generate large-field contextual features MF_i, which can be formulated as:

$$MF_i = MID(DF_i) \tag{4}$$

where $MID(\cdot)$ is the MID module. Then, we integrate the different levels of MF through a decoder structure. In particular, we concatenate the output of each level in the boundary-diffusion stream to get the final boundary map. For the interior-diffusion stream, we add EF_4 and EF_5 by using the element-wise addition operation and entering them into MID to generate MF. In this way, different-level feature maps are jointly fused, which is beneficial for semantic segmentation.

3.3. MID module

For glass detection, the key to accurately locate glass regions is to aggregate a wide range of contextual features at different scales. GDNet [MYW*20] exploits the LCFI module to efficiently extract abundant contextual information from a large field. It utilizes convolutions with large kernels and dilated convolutions to enlarge the receiving field, and spatially separable convolution to reduce the parameters of convolution with large kernels. Inspired by them, we propose a multi-scale interactive dilation module (named as MID, as shown in Figure 5) to efficiently aggregate different-scale contextual information for enhancing the glass detection performance.

Specifically, we propose to utilize dilated convolution with different dilation rates to expand the receiving field of each pixel so that it can obtain a wider range of contextual information. After passing the feature map into this module, the maps first pass through a convolution layer for feature extraction at each of the branches. Then, we use dilated convolution to extract a wide range of context features. Different from LCFI [MYW*20], we use the dilation convolution with larger dilation rates, which are set to 2, 4, 8 and 16, respectively. And an additional branch is added, which uses a 3×3 convolution to obtain more dense local information. Besides, in order to reduce the parameters of the module in the whole network, we remove the large kernel convolution of LCEI. Meanwhile, we adopt Short Connections [HCH*17] to transfer the output of the smaller receiving field branch into the other larger branch for getting dense contextual information. Finally, we integrate the context feature maps of each branch through convolution layers and obtain the feature map MF by utilizing an element-wise addition operation.

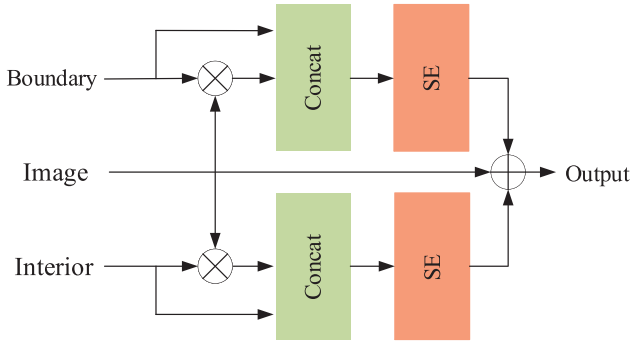


Figure 6: *Attention-based boundary-aware feature Mosaic module.*

3.4. Attention-based BFM module

After obtaining a high-quality boundary prediction map using the boundary-diffusion stream, we utilize a BFM module to integrate the boundary maps into the predicted glass maps generated by the glass stream, as shown in Figure 6. BFM first takes the predicted boundary, interior-diffusion and glass feature maps as inputs. Through using the predicted boundary and the interior-diffusion maps as attention maps, we integrate them into the feature maps of the glass stream by using an element-wise product operation $\otimes$. We concatenate the interior-diffusion and boundary-diffusion maps and input them into the SE module, respectively, to enhance the corresponding boundary- and interior-diffusion features. Here, the SE module is an architectural unit proposed by Hu *et al.* [HSS18], which is termed as the 'Squeeze-and-Excitation' (SE) block, that adaptively re-calibrates channel-wise feature responses by explicitly modelling inter-dependencies between channels. Finally, we add the enhanced features to the input glass feature map to generate the final output.

3.5. Loss function

We use two different loss functions, i.e. the binary cross-entropy (BCE) loss l_{bce}, and the intersection over union (IoU) loss l_{iou} [QZH*19], to supervise the network. The BCE loss is a widely used loss function in computer vision because of its robustness:

$$l_{bce} = -\sum_{(x,y)} [g(x,y)log(p(x,y)) + (1 - g(x,y))log(1 - p(x,y))] \tag{5}$$

IoU is an important metric to evaluate object detection quality by calculating the ratio of the intersection and union of the 'predicted box' and 'GT box'. Recently, it is widely used as the training loss:

$$l_{iou} = 1 - \frac{\sum_{x=1}^{H}\sum_{y=1}^{W} p(x,y)g(x,y)}{\sum_{x=1}^{H}\sum_{y=1}^{W}[p(x,y)+g(x,y)-p(x,y)g(x,y)]} \quad (6)$$

Therefore, the three different streams are supervised separately by combining different losses. In the interior-diffusion stream, the glass stream and the final output prediction maps, we adopt the BCE and IoU losses, which can be formulated as:

$$L_{glass} = \sum_{k=1}^{N_g}[l_{bce}(p_k, g_{glass}) + l_{iou}(p_k, g_{glass})] \quad (7)$$

$$L_{inner} = l_{bce}(p_k, g_{inner}) + l_{iou}(p_k, g_{inner}) \quad (8)$$

$$L_{final} = l_{bce}(p_k, g_{glass}) + l_{iou}(p_k, g_{glass}) \quad (9)$$

where L_{glass} is the sum of the different levels of losses in the glass stream, and L_{final} is the supervision loss of the final output of the whole network. p_k is the prediction map of the different branches in the three streams. g_{glass} is the ground-truth label, and g_{inner} is the interior-diffusion label decoupled by the original label. N_i and N_g are the numbers of branches in the interior-diffusion stream and the glass stream, respectively. Moreover, we only use the BCE loss in the boundary-diffusion stream:

$$L_{boundary} = \sum_{k=1}^{N_b} l_{bce}(p_k, g_{boundary}) \quad (10)$$

where $g_{boundary}$ is the boundary-diffusion label decoupled by the original label and N_b is the number of branches in the boundary-diffusion stream.

Therefore, the final loss function is formulated as:

$$Loss = L_{inner} + L_{boundary} + L_{glass} + L_{final} \quad (11)$$

4. Experiments

4.1. Datasets and evaluation metrics

Currently, there is only a dataset available, i.e. GDD [MYW*20], which is the first large-scale benchmark for glass detection and has 4018 mirror images with their corresponding masks. We use five metrics widely used by other computer vision tasks to evaluate the performance of our model and existing state-of-the-art methods. First, we use two popular metrics, i.e. the pixel accuracy (acc) and the intersection of union (IoU). Besides, we apply the F-measure [AHES09] and mean absolute error (MAE) metrics from the SOD field, which are widely adopted elsewhere [CTWH18, LYC*18, LHY18, HCH*17]. The F-measure is the weighted harmonic mean of precision and recall. We use the maximum F-measure (F_β) version as:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{beta^2 Precision + Recall} \quad (12)$$

where β^2 is set to 0.3 as suggested in Achanta *et al.*[AHES09]. MAE is the mean absolute error, i.e. the mean value of the absolute error between the prediction and the ground truth, which is defined as:

$$MAE = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}|p(i,j) - g(i,j)| \quad (13)$$

where $g(x, y) \in [0, 1]$ is the ground-truth label of the pixel (x, y) and $p(x, y) \in [0, 1]$ is the predicted probability of being glass. In addition, we select the balance error rate (BER) [VHS15] from the shadow detection field as our last metric, which can be obtained as:

$$BER = 100 \times \left(1 - \frac{1}{2}\left(\frac{TP}{N_p} + \frac{TN}{N_n}\right)\right) \quad (14)$$

where TP, TN, N_n and N_p are the numbers of true positives, true negatives, glass pixels and non-glass pixels, respectively.

4.2. Implementation details

We implement the proposed network GlassNet based on the PyTorch framework [PGM*19] and train it on the benchmark dataset GDD. The pre-trained ResNet-50 network [HZRS16] on ImageNet [DDS*09] is used to initialize the parameters of the backbone, and the other parameters are initialized randomly. We train the whole network by using the stochastic gradient descent (SGD) with a momentum of 0.9 and the weight decay of 5×10^{-4}. The initial learning rate is set to 0.0001 and is adjusted by poly decay strategies [YWP*18] with a power of 0.9. The network with a batch setting of 4 is trained on an NVIDIA GTX 1080 Ti graphics card. During testing, images are adjusted to the resolution of 512×512 for inference without any post-processing.

4.3. Comparison with the SOTAs

Compared methods. It only has one deep learning-based method for glass detection from single images. Thus, we compare to this method and other 14 state-of-the-art methods, which are PSPNet [ZSQ*17], DenseASPP [YYZ*18], PSANet [ZZL*18], DANet [FLT*19] and CCNet [HWH*19] chosen from the semantic segmentation field, R^3Net [DHZ*18], CPD [WSH19], BASNet [QZH*19], EGNet [ZLF*19] and LDF [WWW*20] chosen from the SOD field, DSC [HZF*18] and BDRAR [ZDH*18] chosen from the shadow detection field, MirrorNet [YMX*19] and PMD [LWL20] from the mirror segmentation field and GDNet [MYW*20] used for glass detection. For a fair comparison, we retrain all the other methods on the GDD dataset by using their publicly available codes.

Quantitative comparison. We compare the proposed network with state-of-the-art methods from the relevant fields mentioned above, which are shown in Table 1. The first, second and third best results are marked in bold, red and blue, respectively. Obviously, compared with other methods in related fields, our method is better than the SOTA methods.

Qualitative evaluation. Some prediction examples of the proposed method and state-of-the-art approaches have been shown in Figure 7. We observe that the proposed method not only highlights

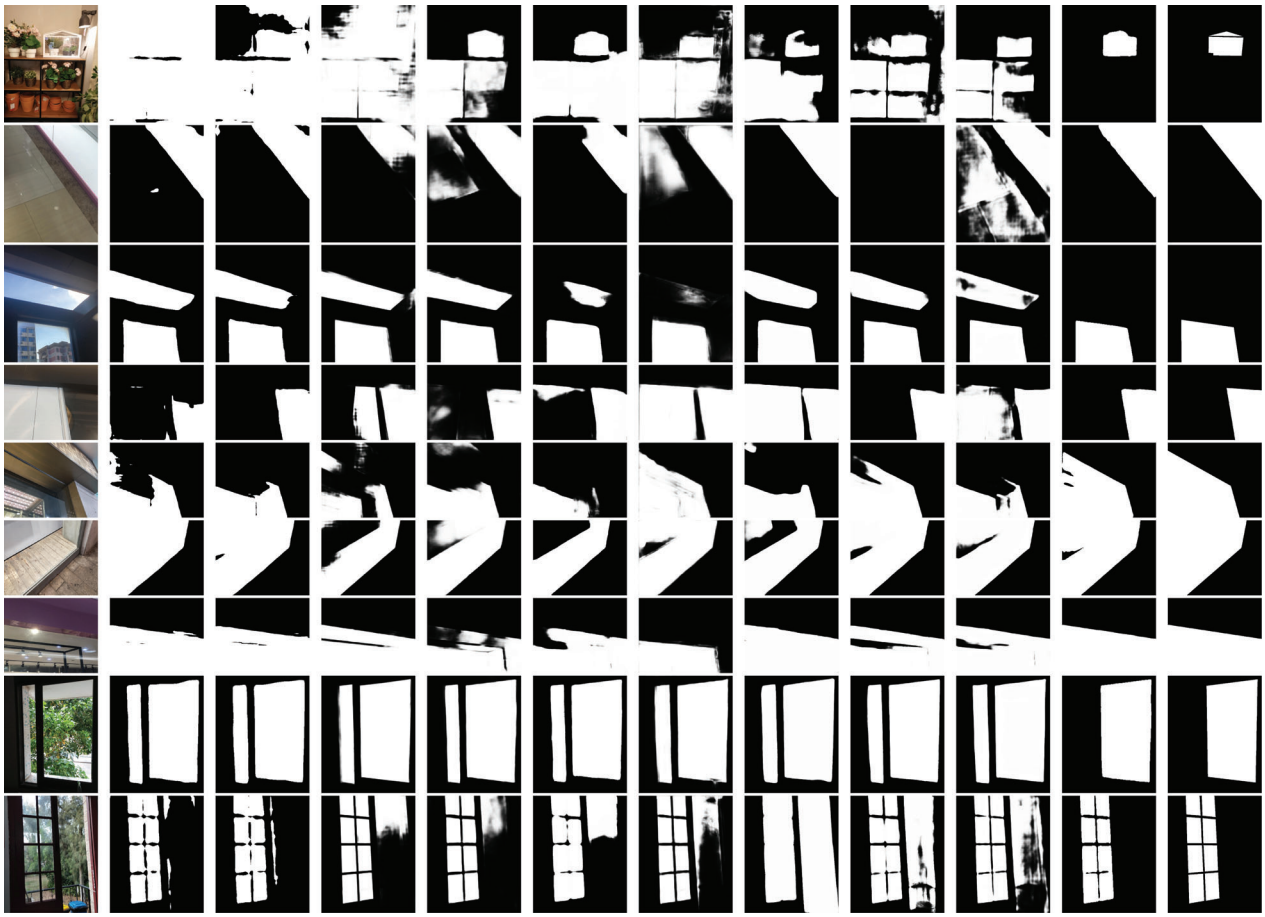


Figure 7: *Visual comparison of our GlassNet with SOTAs on the GDD testing set.*

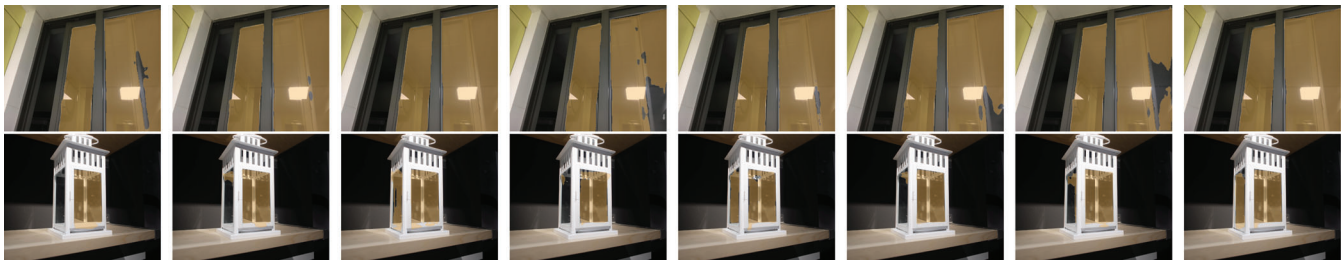
the glass regions clearly but also well suppresses the background noise. It can be seen that our method can accurately detect small glass (e.g. the first four rows), large glass (e.g. the fifth to seventh rows) and others (e.g. the eighth and ninth rows). Although GDNet can locate these regions well, it has low detection accuracy for the boundary regions and cannot even detect the boundary regions correctly. In contrast, our method has higher detection accuracy in the boundary region because we use boundary information to force the network to pay more attention to the boundary region.

4.4. Ablation studies

Table 2 demonstrates the effectiveness of each component in our model. From the first line to the third line, we can see that both boundary- and interior-diffusion branches can effectively improve the performance. Moreover, the effect of the network without the boundary-diffusion stream is worse than that without the interior-diffusion stream, which is consistent with our observation: Boundaries significantly improve the detection ability, which should be specially considered. In addition, the final detection accuracy can also be improved by the proposed multiple mixing losses, as shown in the 4th and 5th lines, where each row omits a loss, i.e. BCE and IOU, respectively. Finally, w/o MID and w/o BFM, respectively, indicate that we do not use any one of the two modules each time in our network, which shows their contributions on improving the glass detection quality. Figure 8 shows a visual example, proving that our method successfully addresses the glass detection problem with the help of boundaries.

4.5. Failure cases

GlassNet has two limitations, as shown in Figure 9: (1) In the case of very large-scale glass, e.g. the area of the glass occupies more than 95% or even 100% of the whole image, it may operate poorly on such extreme cases, due to the lack of sufficient contextual information; and (2) it is nearly impossible to detect the glass in the very weak light, since under the very weak-light condition, the boundary area of the glass and the background will share very similar properties, i.e. they are all black regions with pixel values approximating to (0,0,0).



(a) W/o interior and boundary (b) W/o boundary (c) W/o interior (d) W/ only BCE loss (e) W/o IoU loss (f) W/o MID (g) W/o BFM (h) Our full scheme

Figure 8: *Visual comparison of our GlassNet with its variants.*

Table 1: *Quantitative results on the GDD dataset. CRF indicates whether CRF [KK11] is used as a post-processing step. The first, second and third best results are marked in bold, red and blue, respectively.*

Method	CRF	acc↑	IoU↑	F_β ↑	MAE↓	BER↓
PSPNet[ZSQ*17]	×	0.916	0.841	0.906	0.084	8.79
DenseASPP[YYZ*18]	×	0.919	0.837	0.911	0.081	8.66
PSANet[ZZL*18]	×	0.918	0.835	0.909	0.082	9.09
CCNet[HWH*19]	×	0.915	0.843	0.904	0.085	8.63
DANet[FLT*19]	×	0.911	0.842	0.901	0.089	8.96
R^3Net[DHZ*18]	√	0.869	0.767	0.869	0.132	13.85
CPD[WSH19]	×	0.907	0.825	0.903	0.095	8.87
BASNet[QZH*19]	×	0.907	0.829	0.896	0.094	8.70
EGNet[ZLF*19]	×	0.885	0.788	0.858	0.115	10.87
LDF[WWW*20]	×	0.921	0.843	0.908	0.079	7.52
BDRAR[ZDH*18]	√	0.902	0.800	0.908	0.098	9.87
DSC[HZF*18]	×	0.914	0.836	0.911	0.090	7.97
MirrorNet[YMX*19]	√	0.918	0.851	0.903	0.083	7.67
PMD[LWL20]	√	0.921	0.836	0.894	0.078	8.34
GDNet[MYW*20]	×	0.939	0.876	0.920	0.061	5.62
GlassNet (ours)	×	**0.946**	**0.887**	**0.937**	**0.054**	**5.42**

Table 2: *Ablation study results. Best results are highlighted in bold.*

Strategy	acc↑	F_β ↑	BER↓
w/o interior and boundary	0.936	0.925	6.37
w/o boundary	0.938	0.927	6.25
w/o interior	0.941	0.932	5.98
w/only BCE loss	0.937	0.928	6.31
w/o IoU loss	0.940	0.933	5.92
w/o MID	0.943	0.932	5.53
w/o BFM	0.940	0.923	5.80
Our GlassNet	**0.946**	**0.937**	**5.42**

(a) Input (b) Poor detection

Figure 9: *Failure cases.*

5. Conclusion

In this paper, we propose a three-stream network for glass detection from single images, called GlassNet. GlassNet consists of a LD procedure that decouples the ground truth into an interior-diffusion label and a boundary-diffusion label, a MID module for extracting and capturing contextual features, and a three-stream network integrating multi-scale and multi-modal information to generate the final prediction map. Besides, GlassNet utilizes an attention-based BFM module to integrate multi-modal information for further improving the glass detection quality. Experiments on the benchmark datasets demonstrate that our GlassNet outperforms the state-of-the-art methods under different evaluation metrics.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Nos. 62032011, 62172218), in part by the Joint Fund of National Natural Science Foundation of China and Civil Aviation Administration of China (No. U2033202), in part by the Free Exploration of Basic Research Project, Local Science

and Technology Development Fund Guided by the Central Government of China (No. 2021Szvup060) and in part by the Key Program of Jiangsu Provincial Department of Culture and Tourism (No. 20ZD06).

References

[AHES09] Achanta R., Hemami S., Estrada F., Susstrunk S.: Frequency-tuned salient region detection. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), IEEE, pp. 1597–1604.

[CDL16] Cheng J., Dong L., Lapata M.: Long short-term memory-networks for machine reading. EMNLP (2016): 551–561.

[CMS*20] Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S.: End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision* (2020), Springer, pp. 213–229.

[CPK*18] Chen L. C., Papandreou G., Kokkinos I., Murphy K., Yuille A. L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40*, 4 (2018), 834–848.

[CPSA17] Chen L. C., Papandreou G., Schroff F., Adam H.: Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

[CTWH18] Chen S., Tan X., Wang B., Hu X.: Reverse attention for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 234–250.

[CZXL19] Chen Z., Zhou H., Xie X., Lai J.: Contour loss: Boundary-aware learning for salient object segmentation. *arXiv preprint arXiv:1908.01975* (2019).

[DDS*09] Deng J., Dong W., Socher R., Li L. J., Li K., Fei-Fei L.: Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), IEEE, pp. 248–255.

[DHZ*18] Deng Z., Hu X., Zhu L., Xu X., Qin J., Han G., Heng P. A.: R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (2018), AAAI Press, pp. 684–690.

[DJS*18] Ding H., Jiang X., Shuai B., Liu A. Q., Wang G.: Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 2393–2402.

[FLT*19] Fu J., Liu J., Tian H., Li Y., Bao Y., Fang Z., Lu H.: Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 3146–3154.

[HCH*17] Hou Q., Cheng M. M., Hu X., Borji A., Tu Z., Torr P. H.: Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 3203–3212.

[HSS18] Hu J., Shen L., Sun G.: Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7132–7141.

[HWH*19] Huang Z., Wang X., Huang L., Huang C., Wei Y., Liu W.: CCNet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 603–612.

[HZF*18] Hu X., Zhu L., Fu C. W., Qin J., Heng P. A.: Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7454–7462.

[HZRS16] He K., Zhang X., Ren S., Sun J.: Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.

[KK11] Krähenbühl P., Koltun V.: Efficient inference in fully connected CRFs with gaussian edge potentials. *Advances in Neural Information Processing Systems 24* (2011), 109–117.

[KSH12] Krizhevsky A., Sutskever I., Hinton G. E.: Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25* (2012), 1097–1105.

[LHL21] Lin J., He Z., Lau R. W.: Rich context aggregation with reflection prior for glass surface detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 13415–13424.

[LHY18] Liu N., Han J., Yang M. H.: PiCANet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3089–3098.

[LSD15] Long J., Shelhamer E., Darrell T.: Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3431–3440.

[LWL20] Lin J., Wang G., Lau R. W.: Progressive mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3697–3705.

[LYC*18] Li X., Yang F., Cheng H., Liu W., Shen D.: Contour knowledge transfer for salient object detection. In *Proceedings*

of the European Conference on Computer Vision (ECCV) (2018), pp. 355–370.

[MYW*20] Mei H., Yang X., Wang Y., Liu Y., He S., Zhang Q., Wei X., Lau R. W.: Don't hit me! glass detection in real-world scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3687–3696.

[PGM*19] Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., Desmaison A., Köpf A., Yang E. Z., DeVito Z., Raison M., Tejani A., Chilamkurthy S., Steiner B., Fang L., Bai J., Chintala S. Pytorch: An imperative style, high-performance deep learning library. NeurIPS (2019): 8024–8035.

[PZZL20] Pang Y., Zhao X., Zhang L., Lu H.: Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9413–9422.

[QZH*19] Qin X., Zhang Z., Huang C., Gao C., Dehghan M., Jagersand M.: BASNet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 7479–7489.

[RFB15] Ronneberger O., Fischer P., Brox T.: U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention* (2015), Springer, pp. 234–241.

[SB15] Srivatsa R. S., Babu R. V.: Salient object detection via objectness measure. In *Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP)* (2015), IEEE, pp. 4481–4485.

[SZ14] Simonyan K., Zisserman A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[VHS15] Vicente T. F. Y., Hoai M., Samaras D.: Leave-one-out kernel optimization for shadow detection. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3388–3396.

[VSP*17] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I.: Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[WBZ*17] Wang T., Borji A., Zhang L., Zhang P., Lu H.: A stagewise refinement model for detecting salient objects in images. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 4019–4028.

[WGGH18] Wang X., Girshick R., Gupta A., He K.: Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7794–7803.

[WLF*21] Wang W., Lai Q., Fu H., Shen J., Ling H., Yang R.: Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1.

[WSH19] Wu Z., Su L., Huang Q.: Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 3907–3916.

[WWW*20] Wei J., Wang S., Wu Z., Su C., Huang Q., Tian Q.: Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 13025–13034.

[WZW*18] Wang T., Zhang L., Wang S., Lu H., Yang G., Ruan X., Borji A.: Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3127–3135.

[XWL*18] Xin Y., Wang S., Li L., Zhang W., Huang Q.: Reverse densely connected feature pyramid network for object detection. In *Asian Conference on Computer Vision* (2018), Springer, pp. 530–545.

[XWW*20] Xie E., Wang W., Wang W., Ding M., Shen C., Luo P.: *Segmenting Transparent Objects in the Wild. LNCS* (vol. 12358). Springer Science and Business Media Deutschland GmbH, 2020, pp. 696–711.

[YMX*19] Yang X., Mei H., Xu K., Wei X., Yin B., Lau R. W.: Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 8809–8818.

[YWP*18] Yu C., Wang J., Peng C., Gao C., Yu G., Sang N.: BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 325–341.

[YYZ*18] Yang M., Yu K., Zhang C., Li Z., Yang K.: DenseASPP for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3684–3692.

[YZL*13] Yang C., Zhang L., Lu H., Ruan X., Yang M. H.: Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3166–3173.

[ZDH*18] Zhu L., Deng Z., Hu X., Fu C. W., Xu X., Qin J., Heng P. A.: Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 121–136.

[ZDS*18] Zhang H., Dana K., Shi J., Zhang Z., Wang X., Tyagi A., Agrawal A.: Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7151–7160.

[ZLF*19] Zhao J. X., Liu J. J., Fan D. P., Cao Y., Yang J., Cheng M. M.: EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 8779–8788.

[ZLWS14] Zhu W., Liang S., Wei Y., Sun J.: Saliency optimization from robust background detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 2814–2821.

[ZSL*15] Zhang J., Sclaroff S., Lin Z., Shen X., Price B., Mech R.: Minimum barrier salient object detection at 80 fps. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1404–1412.

[ZSQ*17] Zhao H., Shi J., Qi X., Wang X., Jia J.: Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 2881–2890.

[ZTZ*17] Zhang R., Tang S., Zhang Y., Li J., Yan S.: Scale-adaptive convolutions for scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2031–2039.

[ZW19] Zhao T., Wu X.: Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 3085–3094.

[ZWL*17] Zhang P., Wang D., Lu H., Wang H., Ruan X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 202–211.

[ZWQ*18] Zhang X., Wang T., Qi J., Lu H., Wang G.: 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.

[ZXL*20] Zhou H., Xie X., Lai J. H., Chen Z., Yang L.: Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9141–9150.

[ZZL*18] Zhao H., Zhang Y., Liu S., Shi J., Loy C. C., Lin D., Jia J.: Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 267–283.