Don't Hit Me! Glass Detection in Real-world Scenes

Haiyang Mei[1] Xin Yang[1,4,⋆] Yang Wang[1] Yuanyuan Liu[1] Shengfeng He[2]
Qiang Zhang[1] Xiaopeng Wei[1,⋆] Rynson W.H. Lau[3,†]
[1] Dalian University of Technology [2] South China University of Technology
[3] City University of Hong Kong [4] Advanced Institute of Information Technology Peking University
https://mhaiyang.github.io/CVPR2020_GDNet/index

Abstract

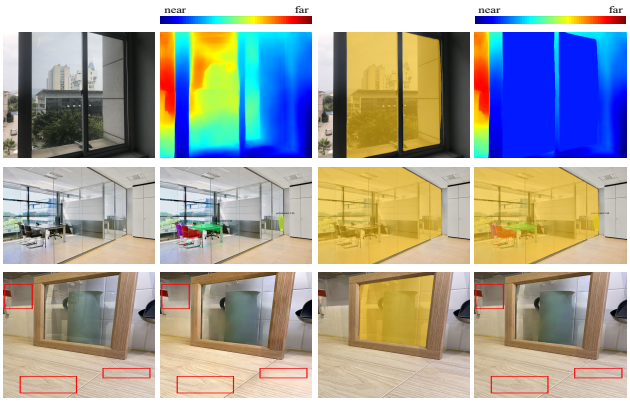
Glass is very common in our daily life. Existing computer vision systems neglect it and thus may have severe consequences, e.g.*, a robot may crash into a glass wall. However, sensing the presence of glass is not straightforward. The key challenge is that arbitrary objects/scenes can appear behind the glass, and the content within the glass region is typically similar to those behind it. In this paper, we propose an important problem of detecting glass from a single RGB image. To address this problem, we construct a large-scale glass detection dataset (GDD) and design a glass detection network, called GDNet, which explores abundant contextual cues for robust glass detection with a novel large-field contextual feature integration (LCFI) module. Extensive experiments demonstrate that the proposed method achieves more superior glass detection results on our GDD test set than state-of-the-art methods fine-tuned for glass detection.*

1. Introduction

Glass is a non-crystalline, often transparent, amorphous solid that has widespread practical and decorative usages, *e.g.*, window panes, glass doors, and glass walls. Such glass objects can have a critical impact to the existing vision systems (*e.g.*, depth prediction and instance segmentation) as demonstrated in Figure 1, and would further affect intelligent decisions in many applications such as robotic navigation and drone tracking, *i.e.*, the robot/drone might crash into the glass. Hence, it is essential for vision systems to be able to detect and segment glass from input images.

Some small glass-made objects such as cup and wine glass can be detected well by the existing methods as they have relatively fixed patterns. However, automatically detecting glass from images like the ones shown in Figure 1(a) is an extremely challenging task. Due to the fact that a glass region does not have a fixed pattern, *i.e.*, arbitrary objects/scene can appear behind the glass, and the content presented in the glass region is typically similar to that behind the glass. This makes the glass fundamentally different from other common objects that have been well-addressed by the state-of-the-art segmentation methods [9]. Meanwhile, directly applying existing salient object detection methods [19, 24] to detect glass is inappropriate, as not all glass regions are salient. Besides, a recent mirror segmentation method [38] may segment mirrors by detecting content discontinuity at the mirror boundary. However, the content behind the glass is part of the real scene that often exhibits weak content discontinuity with the scene outside the glass, making the glass detection problem more difficult.

To address the glass detection problem, a straightforward solution is to apply a reflection/boundary detector for



(a) glass images (b) w/o correction (c) GDNet (d) w/ correction

Figure 1. Problems with glass in existing vision tasks. In depth prediction, existing method [16] wrongly predicts the depth of the scene behind the glass, instead of the depth to the glass (1*st* row of (b)). For instance segmentation, Mask RCNN [9] only segments the instances behind the glass, not aware that they are actually behind the glass (2*nd* row of (b)). Besides, if we directly apply an existing singe-image reflection removal (SIRR) method [36] to an image that is only partially covered by glass, the non-glass region can be corrupted (3*rd* row of (b)). GDNet can detect the glass (c) and then correct these failure cases (d).

⋆ Xin Yang and Xiaopeng Wei are the corresponding authors.
† Rynson W.H. Lau leads this project.

glass detection. Unfortunately, this may fail if the glass has only weak/partial reflections or ambiguous boundary due in some complex scene, *e.g*., the second image in Figure 1(a). In general, humans can identify the existence of glass well. We observe that humans typically would combine different contextual information to infer whether and where glass exists. These contexts not only include low-level cues (*e.g*., the color difference between inside and outside of the glass, blur/bright spot/ghost caused by reflection), but also high-level contexts (*e.g*., relations between different objects). This inspires us to leverage abundant contextual features for glass detection.

In this paper, we address the glass detection problem from two aspects. First, we construct a large-scale glass detection dataset (GDD), which consists of 3,916 high-quality images with glass and corresponding glass masks, covering various daily-life scenes. Second, we propose a glass detection network (GDNet), in which multiple well-designed large-field contextual feature integration (LCFI) modules are embedded to harvest abundant low-level as well as high-level contexts from a large receptive field, for accurately detecting glass of different sizes in various scenes.

To sum up, our contributions are as follows:

- We contribute the first large-scale glass detection dataset (GDD) with glass images in diverse scenes and corresponding manually labeled glass masks.
- We propose a novel network with a well-designed large-field contextual feature integration module for glass detection, by exploring abundant contextual features from a large receptive field.
- We achieve superior glass detection results on the GDD test set, by comparing with state-of-the-art models fine-tuned for glass detection. We further demonstrate the capability of our network to extract abundant contexts in the mirror segmentation task.

2. Related Work

In this section, we briefly review state-of-the-art methods from relevant fields, including semantic/scene/instance segmentation, salient object detection, specific region detection/segmentation, and single image reflection removal.

Semantic/scene/instance segmentation. Semantic segmentation aims to segment and parse a given image into different regions associated with semantic categories of discrete objects. Scene segmentation further considers stuff when assigning a label for each pixel. Recently, great progress has been achieved benefited by the advances of deep neural networks. Based on fully convolutional networks (FCNs) [22], state-of-the-art model variants typically leverage multi-scale context aggregation or exploit more discriminative context to achieve high segmentation performance. For example, Chen *et al*. [1] introduce an atrous spatial pyramid pooling (ASPP) to capture multi-scale context information. Zhao *et al*. [46] employ a pyramid pooling module to aggregate local and global context. Ding *et al*. [5] explore contextual contrasted features to boost the segmentation performance of small objects. Zhang *et al*. [40] introduce a channel attention mechanism to capture global context. Fu *et al*. [7] leverage channel- and spatial-wise non-local attention modules to capture contextual features with long-range dependencies. Huang *et al*. [12] further propose a criss-cross attention module to efficiently capture information from long-range dependencies.

Instance segmentation aims to differentiate individual instances of the each category. A typical method is Mask-RCNN [9], which adds a branch of the object detection network Faster-RCNN [25] and achieves good results. PANet [20] further adds bottom-up paths to aggregates multi-level features for detection and segmentation.

However, applying the above segmentation approaches for glass detection (*i.e*., treating glass as one of the object categories) may not be appropriate as arbitrary objects/scenes can appear behind the glass, making glass fundamentally different from other objects. In this paper, we focus on the glass detection problem and formulate it as a binary classification problem (*i.e*., glass or non-glass).

Salient object detection (SOD). Early methods mainly based on low-level hand-crafted features, such as color and region contrast. Many state-of-the-art deep models are devoted to fully utilizing the integration of different levels of features to enhance network performances. Specifically, Liu *et al*. [18] progressively integrate local context information to predict saliency maps. Zhang *et al*. [42] propose a generic framework to integrate multi-level features at different resolutions. Zhang *et al*. [44] introduce an attention guided network to selectively integrate multi-level information in a progressive manner. Zhang *et al*. [41] design a bi-directional message passing module with a gated function to integrate multi-level features. Wang *et al*. [30] integrate high-level and low-level features by performing both top-down and bottom-up saliency inferences in an iterative and cooperative manner.

In general, the content presented in the glass region is from a real scene, instead of just one or multiple salient objects. Therefore, existing SOD methods may not be able to detect the whole glass region well.

Specific region detection/segmentation. Here, we briefly review three binary classification tasks: shadow detection, water hazard detection, and mirror segmentation.

Shadow detection aims to detect shadows for better scene understanding. Hu *et al*. [11] address the shadow detection problem by analyzing image context in a direction-aware manner. Zhu *et al*. [52] combine local and global

Figure 2. Example glass image/mask pairs in our glass detection dataset (GDD). It shows that GDD covers diverse types of glass in daily-life scenes.

contexts for shadow detection. Zheng *et al*. [50] consider shadow distractions. In general, there is an intensity difference between shadow region and non-shadow region, while glass typically does not have such obvious intensity difference between inside and outside of the glass, making the glass detection problem more difficult to address.

Water hazard detection is to detect water in puddles and flooded areas, on and off the road, to reduce the risk to autonomous cars. The reflection on the water surface typically is an inverted and disturbed transform of the sky or nearby objects above the water surface. Han *et al*. [8] present a reflection attention unit to match this pattern in the vertical direction. However, reflections on the glass can be generated from arbitrary directions and thus applying this method may not be suitable.

Mirror segmentation is a newly proposed research topic that aims to segment mirror regions from a single RGB image. Yang *et al*. [38] observe that there exists both high-level and low-level discontinuities between inside and outside of the mirror and leverage contextual contrasted features to segment mirrors. As the contents presented in the mirror is actually the scene in front of the mirror, both semantic and low-level discontinuities often occur at the boundary of the mirror. For the glass, the scene behind it is part of the real scene and thus there may have less content discontinuity between the glass region and its surrounding. Therefore, utilizing contextual contrasted features to detect glass may not obtain the desired results.

Single image reflection removal (SIRR). Reflection is a frequently-encountered source of image corruption when shooting through a glass surface. Such corruptions can be addressed via a single-image reflection removal (SIRR) task. Traditional SIRR methods employ different priors (*e.g*., sparsity [14], smoothness [29, 15], and ghost [26]) to exploit the special properties of the transmitted and reflection layers. In recent deep-learning-based methods, edge information [6, 28], perceptual loss [43] and adversarial loss [32] are used to improve the recovered transmitted layer.

SIRR can be seen as an image enhancement problem. It aims to recognize where the reflections are and then remove them to enhance the visibility of the background scene. The ultimate goal of our glass detection problem is not to recognize only the reflections but to detect the whole glass region, which may contain only partial or weak reflections.

3. A New Dataset for Glass Detection - GDD

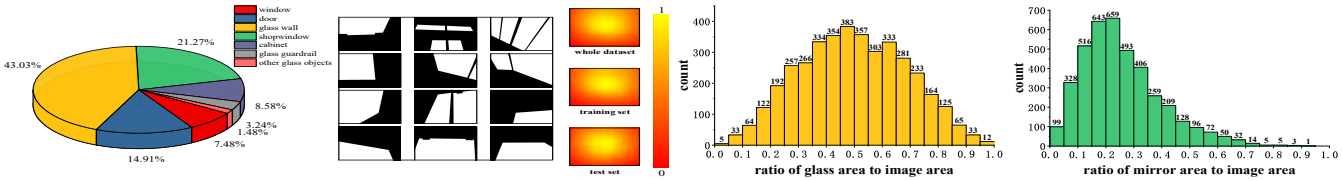
To facilitate the study of the glass detection problem, we contribute a large-scale glass detection dataset (GDD). It contains 3,916 pairs of glass and glass mask images. To the best of our knowledge, GDD is the first large-scale benchmark specifically for glass detection.

Dataset construction. The glass images are captured with some latest cameras and smartphones, and the pixel-level glass masks are labeled by professional annotators. Our constructed glass detection dataset GDD covers diverse daily-life scenes (*e.g*., bathroom, office, street, and mall), in which 2,827 images are taken from indoor scenes and 1,089 images are taken from outdoor scenes. Figure 2 shows some example glass and glass mask images in GDD. More examples can be found in the Supplemental. For dataset split, 2,980 images are randomly selected for training and the remaining 936 images are used for testing.

Dataset analysis. To validate the diversity of GDD and how challenging it is, we show its statistics as follows:

- **Glass type.** As shown in Figure 3(a), there are various types of common glass in GDD (*e.g*., shopwindow, glass wall, glass door, glass guardrail, and glass on window and cabinet). Other relatively small glass objects such as glass bulbs and glass clocks are also included. The reason that such glass objects occupy only a small ratio in GDD is that in this work, we aim to detect relatively large transparent glass that could contribute critical effect to scene understanding. The small glass objects are mainly to add diversity.

- **Glass location.** Our GDD has glass located at different positions of the image, as illustrated in Figure 3(b). We further compute probability maps that indicate how likely each pixel belonging a glass region, to show the location distributions of glass in GDD. The overall spatial distribution tends to be centered, as glass is typically large and covers the center. Besides, the glass spatial distributions for the training/test splits are consistent to those of the whole dataset.



(a) glass type distribution (b) glass location distribution (c) glass area distribution (d) mirror area distribution

Figure 3. Statistics of our dataset. We show that GDD has glass with reasonable property distributions in terms of type, location and area.

- **Glass area.** We define the size of the glass region as a proportion of pixels in the image. In Figure 3(c), we can see that the glass in our GDD varies in a wide range in terms of size and the majority of them fall in the range of [0.2, 0.8]. Glass falling in the range of (0, 0.2] represents small glass objects or glass corners. Such small glass regions can be easily cluttered with diverse background objects/scenes. Glass falling in the range of (0.8, 1.0) is typically located close to the camera. In this situation, the content of the image is dominated by the complicated scene behind the glass. Extreme cases, *i.e.*, glass area equals to 0 or 1, are not included in GDD. Compared with the mirrors in the mirror segmentation dataset MSD [38] (Figure 3(d)), glass in our GDD typically has a larger area, which means more objects/scenes would be presented inside the glass, making GDD more challenging.

4. Methodology

We observe that humans can identify the existence of glass well, by considering contextual information, in terms of low-level cues (*e.g.*, color difference between inside and outside of the glass, blur/bright spot/ghost caused by reflection) as well as high-level contexts (*e.g.*, relations between different objects). This inspires us to leverage abundant contextual features for glass detection.

To this end, first, we propose a novel Large-field Contextual Feature Integration (LCFI) block to extract abundant contextual features from a large field for context inference and glass localization. Second, based on the LCFI block, a novel LCFI module is designed to effectively integrate multi-scale large-field contextual features for detecting glass of different sizes. Third, we embed multiple LCFI modules to the glass detection network (GDNet) to obtain large-field contextual features of different levels for the robust glass detection under various scenes.

4.1. Network Overview

Figure 4 presents the proposed glass detection network (GDNet). It employs the LCFI module (Figure 5) to learn large-field contextual features. Given a single RGB image, we first feed it into the multi-level feature extractor (MFE) to harvest features of different levels, which are further fed into four proposed LCFI modules to learn large-field contextual features. The outputs of the last three LCFI modules are fused to generate high-level large-field contextual features, which will be used to guide the low-level large-field contextual features extracted by the first LCFI module to focus more on the glass regions. Finally, we fuse high-level and attentive low-level large-field contextual features to produce the final glass detection result.

4.2. Large-field Contextual Feature Integration

Figure 5 illustrates the structure of our LCFI module. Given the input features, the LCFI module aims to efficiently and effectively extract and integrate multi-scale large-field contextual features, for the purpose of detecting glass of different sizes.

LCFI block. The LCFI is designed to efficiently extract abundant contextual information from a large field for context inference and glass location. The common practice to obtain larger context information is to use convolutions with large kernels or dilated convolutions. However, large kernels would result in heavy computation and a large dilation rate would lead to sparse sampling. Non-local operations [31] could provide long-range dependencies but also suffer from huge computation. Here, we propose to use spatially separable convolutions to achieve the goal of efficiently extracting abundant contexts from a large field:

$$F_c = \aleph(conv_h(conv_v(F))), \tag{1}$$

where F denotes the input features. $conv_v$ and $conv_h$ refer to vertical convolution with kernel size $k \times 1$ and horizontal convolution with kernel size $1 \times k$, respectively. $\aleph$ represents the batch normalization (BN) and ReLU operations. F_c denotes the extracted large-field contextual features.

As the content inside a glass region is typically complicated, contextual features with different characteristics are needed to eliminate the ambiguity. Thus, we use another spatially separable convolution with reverse convolution order, *i.e.*, $\aleph(conv_v(conv_h(F)))$, to extract complementary large-field contextual features. Besides, we adopt dilated spatially separable filters to ensure that more contexts can be explored in a larger field. Finally, the large-field contextual features extracted from two parallel paths are fused by a 3×3 convolution followed by BN and ReLU. The tasks of

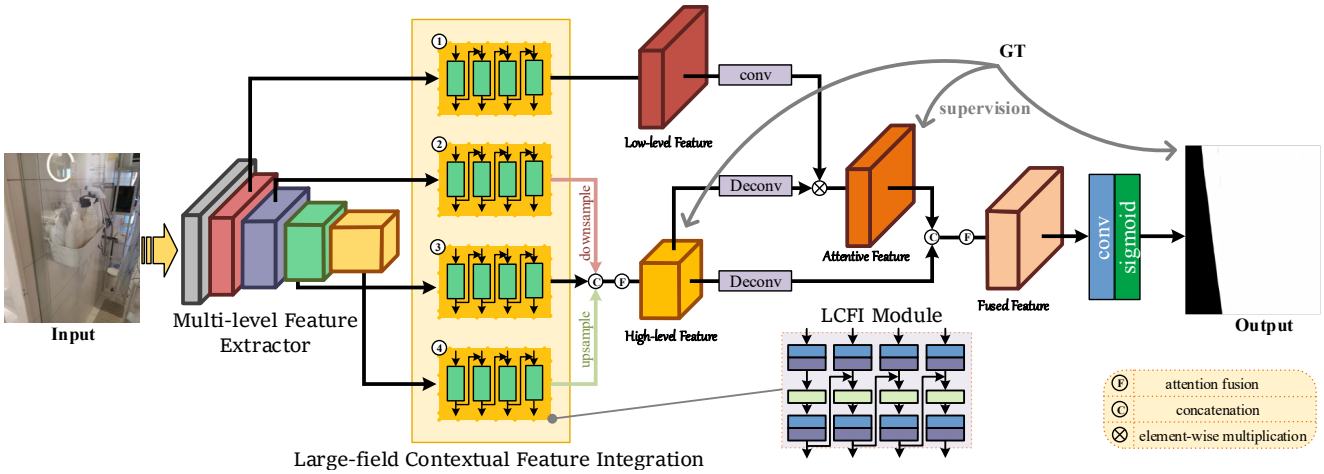


Figure 4. The pipeline of the proposed GDNet. First, we use the pre-trained ResNeXt101 [35] as a multi-level feature extractor (MFE) to obtain features of different levels. Second, we embed four LCFI modules to the last four layers of MFE, to learn large-field contextual features at different levels. Third, the outputs of last three LCFI modules are concatenated and fused via an attention module [33] to generate high-level large-field contextual features. An attention map is then learned from these high-level large-field contextual features and used to guide the low-level large-field contextual features, *i.e.*, the output of the first LCFI module, to focus more on glass regions. Finally, we combine high-level and attentive low-level large-field contextual features by concatenation and attention [33] operations to produce the final glass map.

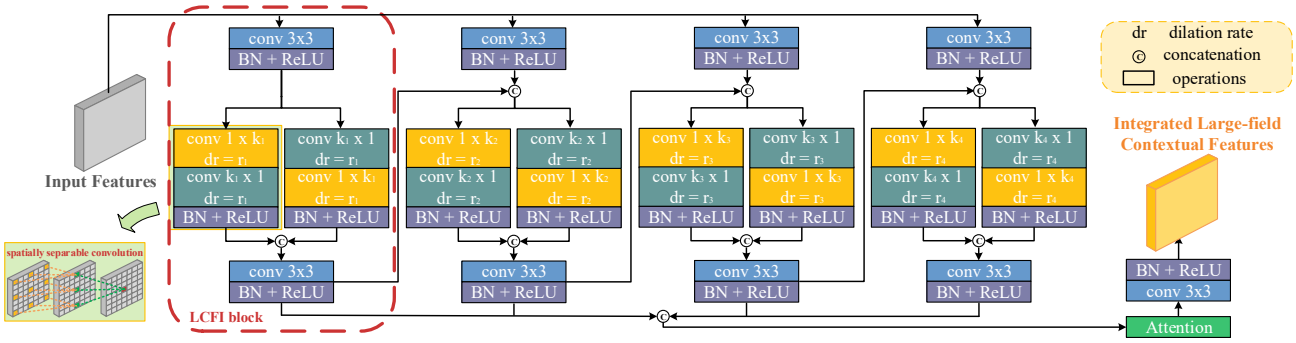


Figure 5. The structure of the LCFI module. The input features are passed through four parallel LCFI blocks, and the outputs of all LCFI blocks are fused to generate multi-scale large-field contextual features. In each LCFI block (red dashed box), input features are fed to two parallel spatially separable convolutions with opposite convolution orders to obtain large-field contextual features with different characteristics. The output of the current LCFI block is then fed to the next LCFI block to be further processed in a larger field.

the LCFI block can be formulated as:

$$\begin{aligned} F_{lcfi} = \aleph(conv_2(concat(\aleph(conv_v(conv_h(F_l))), \\ \aleph(conv_h(conv_v(F_l)))))), \\ F_l = \aleph(conv_1(F_{in})), \end{aligned} \quad (2)$$

where F_{in} denotes the input features of the LCFI block and F_{lcfi} denotes the integrated large-field contextual features. $conv_1$ and $conv_2$ denote the local convolutions with a 3×3 kernel.

LCFI module. Glass captured in an image can vary in size (Figure 3(a)). Given the kernel size k and dilation rate r, the LCFI block extracts contextual features from a large field of a fixed size. On the one hand, if this field is not large enough to cover the whole glass region, incomplete detection may occur. On the other hand, if this field is too large for a small glass region, too much noise would be introduced and cause a false positive detection. To address this problem, contexts of different scales should be considered. Hence, based on the LCFI block, we propose a LCFI module to harvest contextual features from large fields of different scales. Specifically, we feed the input features into four parallel LCFI blocks and fuse their outputs using an attention module [33]. To further explore more contextual features, we add information flow between adjacent LCFI blocks, *i.e.*, we feed the output of a current LCFI block to the next LCFI block. By doing so, local features F_l^i and large-field contextual features from the previous block F_{lcfi}^{i-1} are combined and further processed by the current LCFI

block. Practically, for the spatially separable convolutions in four LCFI blocks, the kernel size k is set to 3, 5, 7, 9, and the dilation rate dr is set to 1, 2, 3, 4, respectively.

Although we draw inspiration (*i.e.*, adding information flow between different paths/blocks) from the integrated successive dilation (ISD) module in [27] in our module design, the proposed LCFI module is inherently different from ISD in both motivation and implementation. The ISD module aims to extract invariant features for the salient object embedded in various contexts while our LCFI module is designed to locate glass of different sizes by exploring contextual information from a large field of different scales. Besides, ISD uses 3×3 convolutions with a large dilation rate (*e.g.*, r=16) to capture large-field contexts. We argue that the contexts extracted in this way are insufficient for complete glass detection (Figure 6). Instead, in each LCFI block, we leverage spatially separable convolutions to explore abundant contextual cues from the large field.

4.3. Loss Function

We adopt three types of losses, binary cross-entropy (BCE) loss l_{bce}, edge loss l_{edge} [49] and IoU loss l_{iou} [24], to optimize the network during the training process. Specifically, for high-level large-field contextual features, we combine BCE loss and IoU loss, *i.e.*, $L_h = l_{bce} + l_{iou}$, to force them to explore high-level cues for complete glass detection. For attentive low-level large-field contextual features, we want them to provide low-level cues for predicting glass maps with clear boundaries. Thus, we combine BCE loss and edge loss, *i.e.*, $L_l = l_{bce} + l_{edge}$. The edge loss would implicitly help the network find the boundaries belonging to the glass. For the final output, the complete detection with clear glass boundary is desired. So, we combine BCE loss, IoU loss and edge loss, *i.e.*, $L_f = l_{bce} + l_{iou} + l_{edge}$. Finally, the overall loss function is:

$$Loss = w_h L_h + w_l L_l + w_f L_f, \tag{3}$$

where w_h, w_l and w_f represent the balancing parameters for L_h, L_l and L_f, respectively.

5. Experiments

5.1. Experimental Settings

Implementation details. We have implemented GDNet on the PyTorch framework [23]. For training, input images are resized to a resolution of 416×416 and are augmented by horizontally random flipping. The parameters of the multi-level feature extractor are initialized by the pre-trained ResNeXt101 network [35] and the other parameters are initialized randomly. Stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 5×10^{-4} is used to optimize the whole network for 200 epochs. We adjust the learning rate by the poly strategy [21] with abasic learning rate of 0.001 and a power of 0.9. The batch size is set to 6 and the balancing parameters w_h, w_l and w_f are empirically set to 1. It takes about 22 hours for the network to converge on an NVIDIA GTX 1080Ti graphics card. For testing, images are also resized to the resolution of 416×416 for inference. There is no post-processing, such as the fully connected CRFs [13] needed for the final glass detection results.

Evaluation metrics. For a comprehensive evaluation, we adopt five metrics for quantitatively evaluating the glass detection performance. The first two metrics are the intersection of union (IoU) and pixel accuracy (PA), which are widely used in the semantic segmentation field. We also adopt the F-measure and mean absolute error (MAE) metrics from the salient object detection field. F-measure is a harmonic mean of average precision and average recall, formulated as: $F_\beta = \frac{(1+\beta^2) Precision \times Recall}{\beta^2 Precision + Recall}$. We set $\beta^2 = 0.3$ to emphasize more on precision over recall, as suggested in [3]. The last metric is the balance error rate (BER), which is a standard metric in the shadow detection field. It is defined as: $BER = (1 - \frac{1}{2}(\frac{TP}{N_p} + \frac{TN}{N_n})) \times 100$, where TP, TN, N_p and N_n represent the numbers of true positive pixels, true negative pixels, glass pixels, and non-glass pixels, respectively. Note that unlike the first three metrics, for MAE and BER, the lower their values, the better the detection results are.

5.2. Comparison with the State-of-the-arts

Compared methods. As a first attempt to detect glass from a single RGB image, we validate the effectiveness of our GDNet by comparing it with 18 state-of-the-art methods from other related fields. Specifically, we choose ICNet [45], PSPNet [46], DenseASPP [37], BiSeNet [39], PSANet [47], DANet [7] and CCNet [12] from the semantic segmentation field, DSS [10], PiCANet [19], RAS [2], R^3Net [4], CPD [34], PoolNet [17], BASNet [24] and EGNet [48] from the salient object detection field, DSC [11] and BDRAR [52] from the shadow detection field, and MirrorNet [38] from the mirror segmentation field. For a fair comparison, we use either their publicly available codes or the implementations with recommended parameter settings. All methods are retrained on the GDD training set.

Evaluation on the GDD test set. Table 1 reports the quantitative results of glass detection on the proposed GDD test set. We can see that our method outperforms all the other state-of-the-art methods on all five metrics. Figure 7 shows the qualitative comparison of our method with the others. It can be seen that our method is capable of accurately detecting both small glass (*e.g.*, the first three rows) and large glass (*e.g.*, 4-7*th* rows). This is mainly because multi-scale contextual features extracted by the LCFI module can help the network better locate and segment glass. While

Methods	IoU↑	PA↑	F_β↑	MAE↓	BER↓
Statistics	40.75	0.584	0.564	0.451	39.31
ICNet [45]	69.59	0.836	0.821	0.164	16.10
PSPNet [46]	84.06	0.916	0.906	0.084	8.79
DenseASPP [37]	83.68	**0.919**	**0.911**	**0.081**	8.66
BiSeNet [39]	80.00	0.894	0.883	0.106	11.04
PSANet [47]	83.52	0.918	0.909	**0.082**	9.09
DANet [7]	84.15	0.911	0.901	0.089	8.96
CCNet [12]	84.29	0.915	0.904	0.085	8.63
DSS [10]	80.24	0.898	0.890	0.123	9.73
PiCANet [19]	83.73	0.916	0.909	0.093	8.26
RAS [2]	80.96	0.902	0.895	0.106	9.48
R^3Net* [4]	76.71	0.869	0.869	0.132	13.85
CPD [34]	82.52	0.907	0.903	0.095	8.87
PoolNet [17]	81.92	0.907	0.900	0.100	8.96
BASNet [24]	82.88	0.907	0.896	0.094	8.70
EGNet [48]	**85.04**	**0.920**	**0.916**	0.083	**7.43**
DSC [11]	83.56	0.914	0.911	0.090	7.97
BDRAR* [52]	80.01	0.902	0.908	0.098	9.87
MirrorNet* [38]	**85.07**	0.918	0.903	0.083	**7.67**
GDNet (ours)	**87.63**	**0.939**	**0.937**	**0.063**	**5.62**

Table 1. Quantitative comparison to state-of-the-arts on the GDD test set. All methods are re-trained on the GDD training set. * denotes using CRFs [13] for post-processing. "Statistics" means thresholding glass location statistics from our training set as a glass mask for detection. The first, second and third best results are marked in red, green, and blue, respectively. Our method achieves the state-of-the-art under all five common evaluation metrics.

the state-of-the-arts are typically confused by the non-glass regions, which share similar boundaries/appearances with the glass regions, our method can successfully eliminate such ambiguities and detect only the real glass regions (*e.g.*, 1*st*, 7*th* and 8*th* rows). This is mainly contributed by the proposed large-field contextual feature learning, which provides abundant contextual information for context inference and glass localization.

5.3. More Glass Detection Results

Figure 8 further shows some glass detection results on images beyond the GDD test set, *i.e.*, images selected from the ADE20K dataset [51] (the first three columns) and images downloaded from the Internet (4-12*th* columns). We can see that GDNet performs well under these various cases, demonstrating the effectiveness of GDNet.

5.4. Component Analysis

Table 2 evaluates the effectiveness of the proposed LCFI module. We can see that multi-scale convolutions can improve the detection performance. In addition, using dilated convolution in the LCFI module (*i.e.*, LCFI w/ sparse) performs better than using local one (*i.e.*, LCFI w/ local), as contexts can be explored from a larger receptive field.

Networks	IoU↑	F_β↑	BER↓
base	84.89	0.923	7.40
base + LCFI w/ one scale	86.22	0.931	6.51
base + LCFI w/ two scales	86.78	0.932	6.34
base + LCFI w/ local	86.93	0.932	6.36
base + LCFI w/ sparse	87.13	0.933	5.88
base + LCFI w/ one path	87.31	0.935	5.81
GDNet	**87.63**	**0.937**	**5.62**

Table 2. Component analysis. "base" denotes our network with all LCFI modules removed. "one scale" and "two scales" denote that there are one and two LCFI blocks in the LCFI module. "local" denotes replacing spatially separable convolutions in LCFI with local convolutions and keeping the parameters approximately the same. Based on "local", "sparse" adopts dilated convolutions to achieve a similar receptive field as spatially separable convolutions. "one path" denotes that there is only one spatially separable convolution path in each LCFI block. Our LCFI module contains four LCFI blocks and each of them contains two parallel paths.

image base base + LCFI w/ local base + LCFI w/ sparse GDNet

Figure 6. Visual comparison of GDNet with variations.

method	IoU↑	PA↑	F_β↑	MAE↓
MirrorNet* [38]	78.95	0.935	0.857	0.065
GDNet (Ours)	**80.31**	**0.943**	**0.876**	**0.058**

Table 3. Comparison to MirrorNet [38] on MSD test set.

With approximately the same number of parameters, using spatially separable convolution (*i.e.*, LCFI w/ one path) can harvest more contexts from a large field to further boost performance. Finally, large-field contextual features with different characteristics can be obtained by two parallel spatially separable convolution paths and help GDNet achieve the best detection results. Figure 6 shows a visual example. We can see that our method successfully addresses the glass under-segmentation problem with the help of abundant contextual features extracted from the large field.

5.5. Mirror Segmentation

With the help of a well-designed large-field contextual feature integration module, our GDNet can explore abundant contextual information from a large field, and thus has the potential to handle other challenging vision tasks. Here, we take the mirror segmentation as an example. We re-train our GDNet on the mirror segmentation dataset MSD [38], and show the results in Table 3. These results demonstrate that large-field contextual information can effectively boost the performance of mirror segmentation.

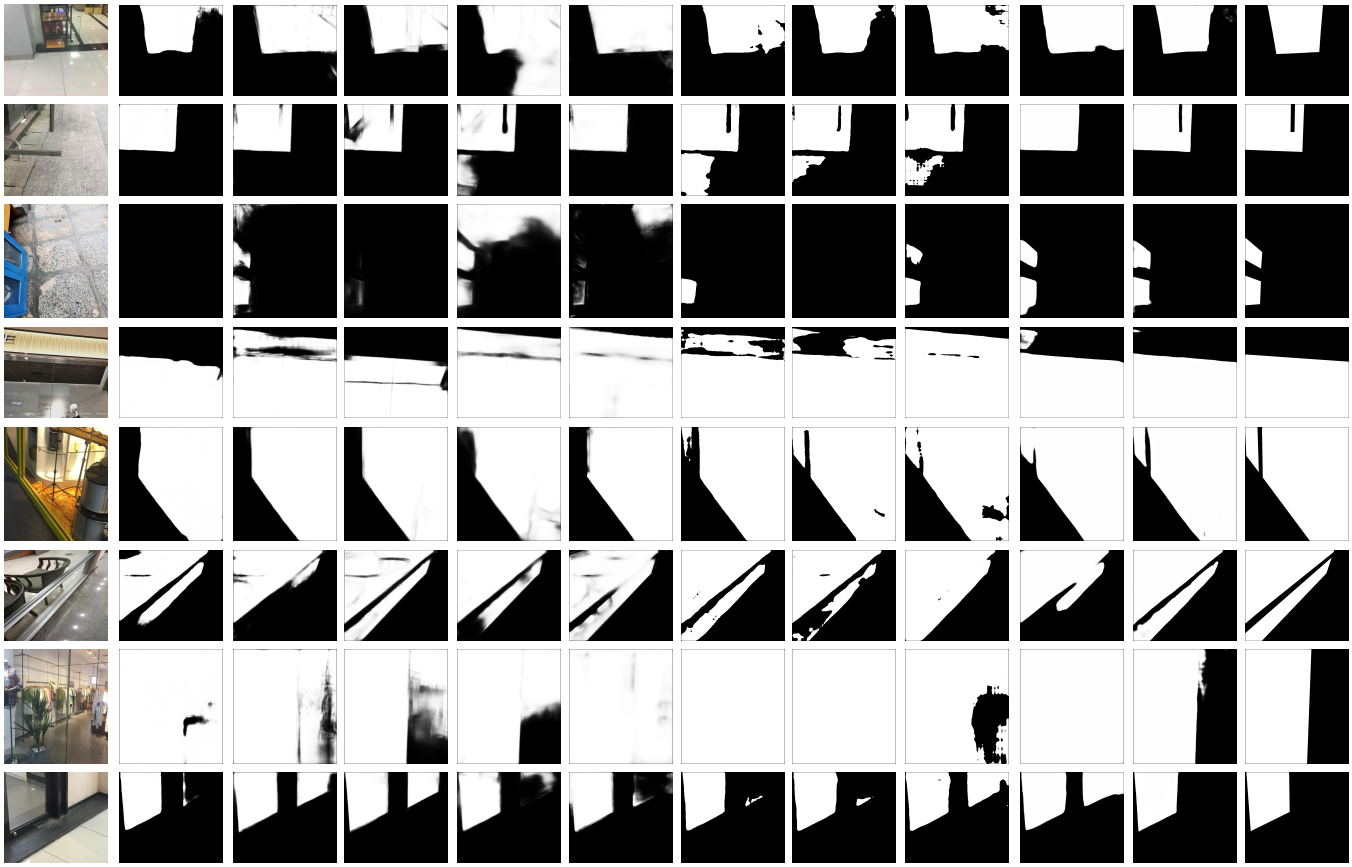
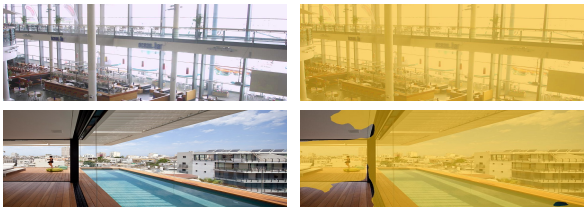


Image BDRAR[52] DSC[11] EGNet[48] PoolNet[17] RAS[2] CCNet[12] DANet[7] PSANet[47] MirrorNet[38] GDNet GT

Figure 7. Visual comparison of GDNet to the state-of-the-art methods on the proposed GDD test set.

Figure 8. More glass detection results on images beyond the GDD test set.



input images our results

Figure 9. Failure cases.

6. Conclusion

In this paper, we have proposed an important problem of detecting glass from a single RGB image and provided a large-scale glass detection dataset (GDD) covering diverse scenes in our daily life. A novel network is also proposed to address this challenging task. It leverages both high-level and low-level contexts extracted from a large field to detect glass of different sizes in various scenes. Extensive evaluations on the images in and beyond the GDD test set verify the effectiveness of our network. Our method would fail in some cases where the scene is very complex or provides insufficient contexts both inside and outside of the glass, as shown in Figure 9. As the first attempt to address the glass detection problem with a computational approach, we focus in this paper on detecting glass from a single RGB image. As a future work, we would like to explore how to address the above failure scenarios.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China, Grants 91748104, 61972067, 61632006, U1811463, U1908214, 61751203, and in part by the National Key Research and Development Program of China, Grants 2018AAA0102003 and 2018YFC0910506.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017.

[2] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, 2018.

[3] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 2014.

[4] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R3net: Recurrent residual refinement network for saliency detection. In *IJCAI*, 2018.

[5] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2018.

[6] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *ICCV*, 2017.

[7] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[8] Xiaofeng Han, Chuong Nguyen, Shaodi You, and Jianfeng Lu. Single image water hazard detection using fcn with reflection attention units. In *ECCV*, 2018.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[10] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017.

[11] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *CVPR*, 2018.

[12] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

[13] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[14] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE TPAMI*, 2007.

[15] Yu Li and Michael S. Brown. Single image layer separation using relative smoothness. In *CVPR*, 2014.

[16] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018.

[17] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019.

[18] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016.

[19] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018.

[20] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018.

[21] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv:1506.04579*, 2015.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[24] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 2017.

[26] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T. Freeman. Reflection removal using ghosting cues. In *CVPR*, 2015.

[27] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong Tian. Selectivity or invariance: Boundary-aware salient object detection. In *ICCV*, 2019.

[28] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. Crrn: Multi-scale guided concurrent reflection removal network. In *CVPR*, 2018.

[29] Renjie Wan, Boxin Shi, Tan Ah Hwee, and Alex C Kot. Depth of field guided reflection removal. In *ICIP*, 2016.

[30] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *CVPR*, 2019.

[31] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[32] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *CVPR*, 2019.

[33] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[34] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019.

[35] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[36] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *ECCV*, 2018.

[37] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018.

[38] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In *ICCV*, 2019.

[39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018.

[40] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018.

[41] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, 2018.

[42] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017.

[43] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, 2018.

[44] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 2018.

[45] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, 2018.

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[47] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018.

[48] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet:edge guidance network for salient object detection. In *ICCV*, 2019.

[49] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, 2019.

[50] Quanlong Zheng, Cao Ying Qiao, Xiaotian, and Rynson W.H. Lau. Distraction-aware shadow detection. In *CVPR*, 2019.

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

[52] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, 2018.