Learning Semantic Associations for Mirror Detection

Huankang Guan Jiaying Lin Rynson W.H. Lau*
City University of Hong Kong

Abstract

Mirrors generally lack a consistent visual appearance, making mirror detection very challenging. Although recent works that are based on exploiting contextual contrasts and corresponding relations have achieved good results, heavily relying on contextual contrasts and corresponding relations to discover mirrors tend to fail in complex real-world scenes, where a lot of objects, e.g., *doorways, may have similar features as mirrors. We observe that humans tend to place mirrors in relation to certain objects for specific functional purposes,* e.g., *a mirror above the sink. Inspired by this observation, we propose a model to exploit the semantic associations between the mirror and its surrounding objects for a reliable mirror localization. Our model first acquires class-specific knowledge of the surrounding objects via a semantic side-path. It then uses two novel modules to exploit semantic associations: 1) an Associations Exploration (AE) Module to extract the associations of the scene objects based on fully connected graph models, and 2) a Quadruple-Graph (QG) Module to facilitate the diffusion and aggregation of semantic association knowledge using graph convolutions. Extensive experiments show that our method outperforms the existing methods and sets the new state-of-the-art on both PMD dataset (f-measure: 0.844) and MSD dataset (f-measure: 0.889). Code is available at* https://github.com/guanhuankang/Learning-Semantic-Associations-for-Mirror-Detection.

1. Introduction

Mirrors appear everywhere in our daily life. In general, they lack a consistent appearance, as their appearances mainly depend on their surroundings. Due to this special property, the presence of mirrors can affect many vision tasks [7, 53]. They are also considered as a potential hazardous factor for computer vision tasks [49]. Hence, mirror detection is important.

There are a few works that address the mirror detection problem [24, 30, 47]. Yang *et al.* [47] propose a model to learn contextual contrasts/discontinuities between mirror and non-mirror regions, while Lin *et al.* [24] propose a model to learn object correspondences between mirror and non-mirror regions. On the other hand, Mei *et al.* [30] propose to learn depth discontinuities between mirror and non-mirror regions in RGBD images. All these methods are essentially based on learning a binary relationship between mirror and non-mirror regions. While they may be effective in simple scenes, relying on learning a binary relationship between mirror and non-mirror regions can easily fail in real-world scenes, in which there are many objects (or *distractors*) that may possess similar properties, such as doorways, windows, wardrobe doors and photo frames.

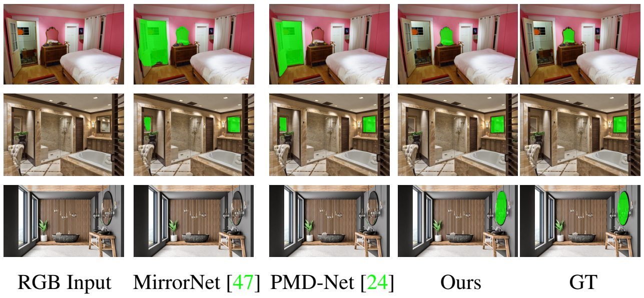

Figure 1. Existing mirror detection methods based on learning contextual contrasts [47] or corresponding relations [24] falsely identify some distractors (*e.g.*, the doorway in the 1^{st} row and the painting in the 2^{nd} row) as mirrors, and miss the mirror (3^{rd} row) when the mirror is captured at an oblique angle to the camera along with some occluding lights. In contrast, our method considers the semantic associations between mirrors and their surrounding objects (*e.g.*, the vanity table in the 1^{st} row and the sink in the 2^{nd} and 3^{rd} rows), yielding accurate results.

In this work, we observe that humans typically place mirrors in certain relationships with some specific objects for functional purposes. For example, we usually put a mirror above the vanity table in the bedroom to help with the morning makeup (the top image in Figure 1) or above a sink in the washroom to allow us to check our look after washing our faces or hands (the second and third images in Figure 1). This observation aligns with studies in cognitive neuroscience [2, 3, 18, 32] that the surrounding objects

[*] Rynson W.H. Lau is the corresponding author, and he leads this project.

can provide a complementary and effective source of contextual information, helping the visual system to locate the target object quickly and confidently. Based on this observation, we propose to learn the semantic knowledge of mirrors and their surrounding objects, and explore the semantic associations between them to help detect and locate mirrors more reliably. As shown in Figure 1, since existing methods, *e.g.*, [24, 47], rely on learning the contextual contrasts or correspondence relations, they can easily be confused by objects that are similar to mirrors, such as the doorway in the top image and the painting in the second image. In contrast, our approach based on learning the semantic associations can correctly detect the mirrors but not objects that look like mirrors in the first two images, and is able to identify the mirror in a difficult case as shown in the third image in which the mirror is in an oblique angle to the camera together with some lights hanging around it.

The proposed model follows an encoder-decoder architecture. In parallel to an encoder-decoder, we add a semantic side-path for scene object discovery and semantic knowledge learning. Our model includes two novel modules to exploit semantic associations: 1) an Associations Exploration (AE) Module to explore scene object associations, and 2) a Quadruple-Graph (QG) Module to facilitate diffusion and aggregation of the semantic association knowledge. The AE module includes a scene-aware GCN to model the mutual relationships between objects, and a spatial-aware GCN for spatial relationship inference. The QG module employs self-attention [39] and reversed self-attention mechanisms to facilitate the propagation of knowledge. Both AE and QG modules are built with graph convolutions, enabling an efficient long-range semantic context aggregation. We conduct comprehensive experiments to demonstrate the effectiveness of our approach and show that our method outperforms relevant state-of-the-art methods, both quantitatively and qualitatively, on two mirror detection benchmarks, PMD [24] and MSD [47].

Our main contributions of this work are as follows:

- We present a novel mirror detection approach that learns the semantic associations between mirrors and their surrounding objects for reliable mirror detection.
- We propose a novel Associations Exploration (AE) Module to infer scene object associations with fully connected graph models, and a novel Quadruple-Graph (QG) Module to facilitate the diffusion and aggregation of semantic association knowledge using graph convolutions. We verify the effectiveness of the proposed modules with comprehensive studies.
- Extensive experiments show that our method outperforms existing state-of-the-art methods both quantitatively and qualitatively on the two popular mirror detection benchmarks, PMD [24] and MSD [47].

2. Related Work

Mirror Detection. Early mirror segmentation solutions rely on the user to specify where the mirrors are via interactions [7] or attach specialized hardware to the camera to help detect mirrors [42, 53]. Recently, Yang *et al.* [47] propose the first mirror detection benchmark (MSD) and the first deep network for automatic mirror detection by leveraging multi-scale contrasts/discontinuities around the mirror boundaries. Lin *et al.* [24] further propose another mirror detection benchmark (PMD) with a higher image diversity/complexity, along with a deep network (PMD-Net) that extracts correspondences between the mirror content and its surrounding real objects to help locate the mirror regions. Most recently, Mei *et al.* [30] propose a depth-aware mirror detection method to detect mirrors from RGBD data. Tan *et al.* [38] perform a depth refinement for 3D mirror planes on RGBD datasets.

Despite the success, these deep methods [24, 30, 47] are all based on learning a binary relation between mirror and non-mirror regions, *e.g.*, contextual contrasts [47], content correspondences [24], and depth discontinuity [30]. While these are useful features for mirror detection, they can be easily confused by objects (or distractors) with similar properties to mirrors, *e.g.*, doorways, windows and photo frames. On the other hand, [30] requires RGBD images as input, which may not always be available. In this work, we propose to learn the semantic relations between mirrors and their surrounding objects, which can help differentiate mirrors from distractors more reliably. Our focus here is on RGB image-based mirror detection.

Graph Convolutional Networks. Graph Convolutional Networks (GCNs) [19] are a kind of neural networks for handling graph data. In recent years, GCNs are becoming very popular and are widely adopted in many applications [10, 21, 22, 25, 27, 34, 41, 43, 44, 46], due to their flexibility, generality and powerful learning ability. For example, Wu *et al.* [44] employ a GCN to panoptic segmentation to bridge the information across thing-class and stuff-class. Yan *et al.* [46] employ graph convolutions to update the target similarity based on context pairs for person re-identification. Chen *et al.* [10] propose a graph convolution based unit (GloRe unit) for graph reasoning in the interaction space to model relations between distant regions.

In this work, we construct graph convolutional networks [19] as key components of our model to exploit semantic associations and facilitate the diffusion and aggregation of information for mirror detection.

Semantic Contexts. A semantic context generally refers to the co-occurrence and spatial relationships among objects. It plays an important role in many vision tasks, such as object categorization [33], semantic segmentation [50,58], and object detection [1,31]. Recently, Zhang *et al.* [51] propose a co-occurrence feature network to learn

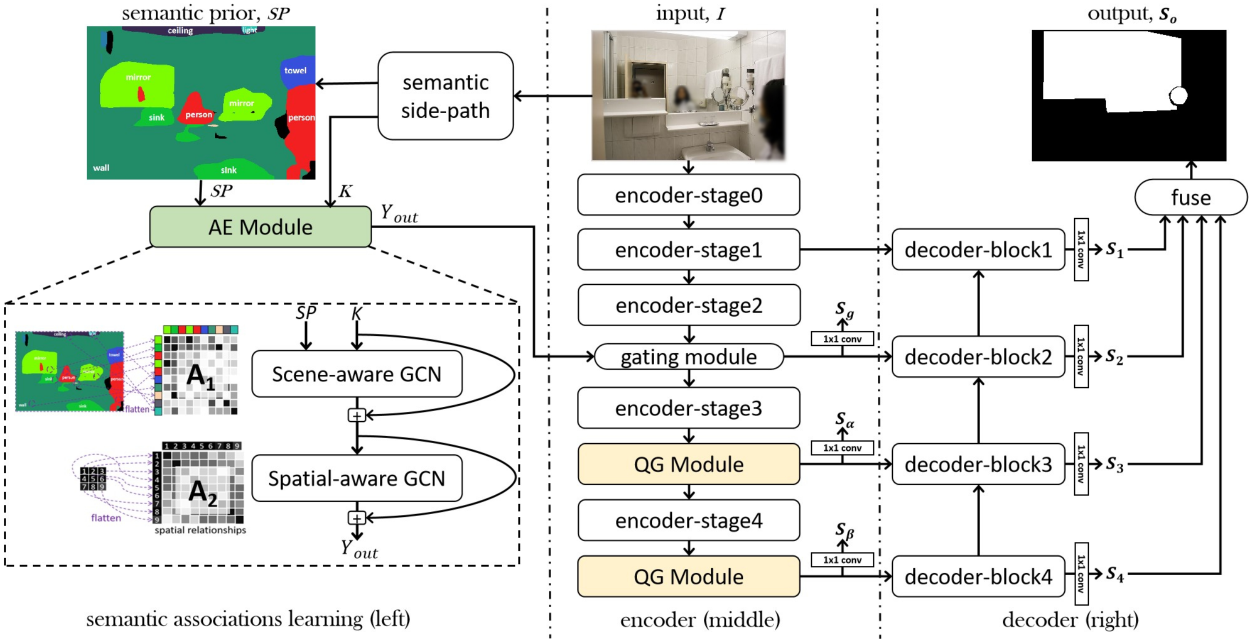

Figure 2. **Architecture of our proposed model.** We first feed the input image into the semantic side-path (upper-left) for learning the semantic prior, $\boldsymbol{SP}$, and class-specific knowledge, $\boldsymbol{K}$. $\boldsymbol{SP}$ and $\boldsymbol{K}$ are then forwarded to the AE module (light green block) for extracting semantic association knowledge, $\boldsymbol{Y_{out}}$, which is merged to the encoder (middle) via a gating module [14]. The proposed QG modules (light yellow blocks) are inserted at high-level positions of the multi-stage encoder to facilitate the propagation of semantic association knowledge. We further enhance the high-level representations with rich spatial information by joining the feature maps of the same size from the bottom-up pathway. $\boldsymbol{S_*}$ are intermediate/final score maps, where $* \in \{g, \alpha, \beta, 1, 2, 3, 4, o\}$.

object co-occurrence relations for semantic segmentation. Wan *et al.* [40] introduce a semantic prior to the crowd counting problem for eliminating the side effect of noisy false alarms presented at the density maps. Zhang *et al.* [52] enhance the semantic features by leveraging captioning as an auxiliary semantic task to facilitate salient object detection. Siris *et al.* [36] explore a scene context-aware learning approach to salient object detection.

In contrast, semantic contexts in the mirror detection problem are still unexplored, partly due to the lack of semantic annotations. To address this limitation, in this work, we first collect semantic annotations for the PMD dataset [24], which contains diverse real-world images for mirror detection, and then propose to explore semantic associations using graph convolutions to provide a long-range relation modeling.

3. Our Approach

Figure 2 shows the architecture of the proposed model. We first feed the input image into the semantic side-path for learning the semantic prior, $\boldsymbol{SP}$, and class-specific knowledge, $\boldsymbol{K}$. The semantic prior provides a scene-aware knowledge, allowing the Associations Exploration (AE) Module to model object relationships, while the class-specific knowledge contains semantic representations in high-dimensional spaces.

The AE module aims to extract semantic association knowledge by modeling the mutual relationships among the scene objects in fully connected graph structures. The extracted semantic association knowledge is merged to the encoder (the middle part of Figure 2) via a gating module [14], which serves as a fusion block here, consisting of a two-layer MLP for dynamic weights computation.

The encoder is built based on multi-stage ResNeXt [45] as in previous works [24, 47]. However, as the conventional CNNs are inefficient in performing long-range context aggregation and message passing, we propose a Quadruple-Graph (QG) Module to facilitate diffusion and aggregation of knowledge using graph convolutions. We embed the QG module in the last two stages of the multi-stage encoder. Since high-level representations contain more semantic knowledge but less spatial information while low-level representations preserve rich spatial information but lack semantic knowledge, we upsample the high-level representations and combine them with the low-level representations of the same resolution via lateral connections [26] (the right part of Figure 2) in order to recover the spatial information to produce a fine-grained prediction. Finally, we fuse intermediate score maps ($S_{1/2/3/4}$) from the decoder

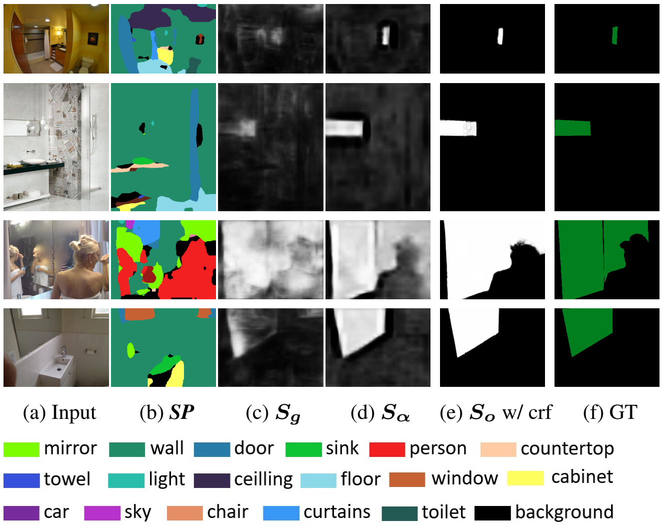

Figure 3. **Visualization of the semantic maps and score maps extracted by our model.** Given the input images, our model acquires the semantic prior $\boldsymbol{SP}$ (b) with the help of the semantic side-path. The AE module extracts semantic association knowledge, and discover the potential mirrors (c). Although the $\boldsymbol{SP}$ maps show the contents inside the mirrors instead of the mirrors, the S_g maps can roughly indicate where the mirrors are. After the first QG module, the S_α maps (d) are already much closer to the GT. Finally, we enhance the high-level features with spatial information by adding low-level features to obtain fine-grained predictions (e).

blocks to produce the output mirror score map, $\boldsymbol{S_o}$.

For the rest of this section, we first introduce the semantic side-path in Section 3.1, and then describe the AE module in Section 3.2 and QG module in Section 3.3. Finally, we present the training strategy in Section 3.4.

3.1. Semantic Side-Path

The purpose of the semantic side-path is to discover the surrounding objects and capture the class-specific knowledge. To do it, we perform a pixel-wise semantic segmentation step [12, 28, 54] with the semantic side-path. Specifically, we first collect the semantic annotations for the PMD dataset [24], which contains lots of common objects with diverse real-world scenes. Note that PMD [24] was originally collected from six public datasets: ADE20K [56, 57], NYUD-V2 [35], COCO-Stuff [6], SUNRGBD [37], Pascal-Context [31] and MINC [4]. Except for MINC [4], the other five datasets already include the semantic annotations. Hence, we simply exclude all images from MINC and all the test images from the other five datasets. We then use the remaining images in PMD, together with the corresponding semantic annotations, to form our training set (with a total of 4,746 images) for training our semantic side-path.

Our semantic side-path is based on ResNet50 [16] with the multi-grid method [8]. The dense prediction from the semantic side-path represents our semantic prior (denoted as $\boldsymbol{SP} \in \mathbb{R}^{H\times W\times C}$, where H and W are the height and width of the prediction; C indicates the number of classes), as shown in the upper left of Figure 2. The feature maps from the last stage of ResNet50 is the desired class-specific knowledge (denoted as $\boldsymbol{K} \in \mathbb{R}^{H\times W\times D}$, where D indicates the number of dimensions). The semantic prior $\boldsymbol{SP}$ and the class-specific knowledge $\boldsymbol{K}$ are then forwarded to the AE module to discover semantic association knowledge. Figure 3(b) shows that the semantic side-path can extract rich contexts from the input image for mirror detection. Figure 3(c-e) show the gradual refinement of the score maps through the exploitation of the semantic associations.

3.2. The Associations Exploration (AE) Module

We note that existing semantic segmentation methods usually fail to detect the mirrors in the input images. Rather, they would segment the mirror contents, as shown in Figure 3(b). Hence, instead of directly feeding the semantic context from the semantic side-path to the encoder for mirror detection, we propose to include the AE module to help extract the semantic associations among the scene objects and discover potential mirror signals.

When designing the AE module, we draw inspiration from our observation that mirrors, as a kind of man-made objects, are usually placed in relation to certain objects in order to achieve specific functional purposes, *e.g.*, a mirror above the sink. Hence, knowing where the sink is, we may infer where the mirror may appear. To achieve this goal, we first propose a scene-aware GCN to model the mutual relations between mirrors and their surrounding scene objects, and then a spatial-aware GCN to infer the spatial relationships, as shown in Figure 2 (bottom left).

Both GCNs in the AE module are based on the following definition of graph convolution[1] [19]:

$$Y = \sigma(AX\Theta), \tag{1}$$

where X, Y are the input and output features, respectively. Θ is a set of learnable parameters. σ is a non-linear activation function. The adjacent matrix A, which is the key component of the two GCNs, is usually approximated as normalized similarity matrix [22, 23].

For the scene-aware GCN, we define the adjacent matrix A_1 as:

$$\mathbf{SP} \in \mathbb{R}^{H\times W\times C} \mapsto \mathbf{SP} \in \mathbb{R}^{HW\times C}, \tag{2}$$

$$\mathbf{K} \in \mathbb{R}^{H\times W\times D} \mapsto \mathbf{K} \in \mathbb{R}^{HW\times D}, \tag{3}$$

$$\Sigma = W_1\,\mathbf{K}\,W_2, \tag{4}$$

$$A_1 = \mathbf{SP}\,\Sigma\,\mathbf{SP}^T, \tag{5}$$

[1] For simplicity, we have omitted some basic operations such as permutation, residual connection and upsampling in our description.

where ***SP*** is the semantic prior, and ***K*** is the class-specific knowledge. $\varSigma \in \mathbb{R}^{C\times C}$ is a learnable covariance matrix to allow the adjustment of correlations between different kinds of objects. $W_1 \in \mathbb{R}^{C\times HW}$ and $W_2 \in \mathbb{R}^{D\times C}$ are learnable parameters for data dependent covariance matrix learning.

For the spatial-aware GCN, we define the adjacent matrix A_2 as:

$$(A_2)_{ij} = distance(loc_i, loc_j), \tag{6}$$

where $distance(\cdot,\cdot)$ is a distance function to measure the 2D distance between two input pixels. We use the Manhattan distance in our implementation.

As suggested by [22], we use max pooling and bilinear interpolation in the graph projection and re-projection phases, to save computation time and preserve the original spatial relations. However, unlike [22], our scene-aware GCN is constructed based on two inputs, semantic prior (***SP***) and class-specific knowledge (***K***), to explicitly model the associations among different scene objects. The spatial-aware GCN is a simple yet efficient component to build up spatial relationships. As shown in the 1^{st} and 2^{nd} rows of Figure 3, despite the weak mirror signals in the ***SP*** maps, the S_g maps could recover distinct mirror signals, benefited by the AE module that takes the surrounding objects into account to uncover the "invisible" mirrors by inferring where the mirrors may be from the associated objects, *e.g*., the sink.

3.3. The Quadruple-Graph (QG) Module

From Figure 3(c), we can see the detected mirror signals in the S_g maps but they are not very accurate. For examples, the mirror signals in the 1^{st} and 2^{nd} rows are not very strong. It may be difficult to tell exactly where the mirrors are. Although the mirror signals in the 3^{rd} and 4^{th} rows are much stronger, they contain false positives (the lady in the 3^{rd} row and the window in the the 4^{th} row). Our analysis is that the semantic association knowledge obtained by the AE module concentrates at local regions, resulting in the incomplete mirror score maps. To address this limitation, we propose a Quadruple-Graph (QG) Module to facilitate diffusion and aggregation of semantic association knowledge. Inspired by [48], we propose to perform long-range message passing in two separated streams, to separately model intra-class and inter-class contexts. In addition, as conventional convolutions are not efficient in long-range modeling, we employ graph convolutions here to build our QG module, for better long-range relation modeling.

Figure 4 shows the structure of the proposed QG module. It consists of four graph convolutional networks: one MH**S**A GCN, one MH**R**A GCN and two spatial-aware GCNs. We build the four GCNs based on Eq. 1, but with different adjacent matrices due to their different design purposes. Specifically, the MH**S**A (multi-head **self**-attention) GCN focuses on intra-class correlations. We apply a self-attention mechanism [39] on the input features F_{in} to derive the adjacent matrix, $A_s = softmax(\frac{QK^T}{\sqrt{d}})$, where $softmax$ is a softmax function and Q, K are the query and key, respectively, of dimension d from the reshaped F_{in}. The MH**R**A (multi-head **reversed** attention) GCN focuses on inter-class correlations. We define the reversed attention matrix $A_r = 1.0 - A_s$. Following [39], we adopt a multi-head mechanism on the MH**S**A GCN and the MH**R**A GCN. The two spatial-aware GCNs are the same as that used in the AE module (see Eq 6). Finally, we merge the two streams together via a gating module [14] to produce the output.

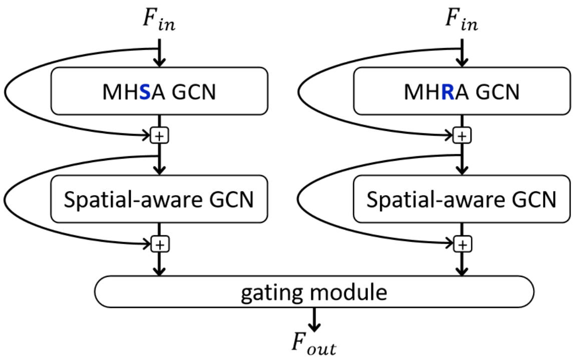

Figure 4. **Quadruple-Graph Module.** The QG module consists of four GCNs: one MH**S**A GCN, one MH**R**A GCN and two spatial-aware GCNs. MH**S**A (multi-head **self**-attention) GCN focuses on intra-class correlations based on a self-attention mechanism. MH**R**A (multi-head **reversed** attention) GCN focuses on inter-class correlations by reversing the self-attention matrix to construct the adjacent matrix. Two spatial-aware GCNs conduct spatial relationship inferences. These two information flows are merged together via a gating module [14].

As shown in Figure 3(d), we can see that the first QG module already produces a significant performance improvement. Compared to the S_g maps, the S_α maps are much better. While they cover the mirror regions more accurately, the false positive signals in the S_g maps are significantly suppressed.

3.4. The Loss Function

We use the Lovász-Softmax loss [5] to supervise the intermediate score maps, and use both Lovász-Softmax and binary cross-entropy (BCE) losses to supervise the final output, $\boldsymbol{S_o}$. The total loss is as follows:

$$Loss = \sum_{i\in\{g,\alpha,\beta,1,2,3,4,o\}} \ell_{ls}(S_i, GT) + \ell_{bce}(S_o, GT) \tag{7}$$

where ℓ_{ls} is the Lovász-Softmax loss. ℓ_{bce} is the BCE loss. S_* $(* \in \{g, \alpha, \beta, 1, 2, 3, 4, o\})$ represent the intermediate and the final score maps, as shown in Figure 2.

Table 1. Quantitative comparison between our method and ten related methods. We report max f-measure, IoU, accuracy and MAE. Best results are marked in red, and the second best results are marked in blue.

Method	PMD Dataset [24]				MSD Dataset [47]			
	f-measure↑	IoU↑	Accuracy↑	MAE↓	f-measure↑	IoU↑	Accuracy↑	MAE↓
GCPANet [11]	0.7548	58.01	95.59	0.04428	0.8477	74.76	93.11	0.06929
EGNet [55]	0.7987	60.17	96.34	0.03676	0.8238	66.68	91.54	0.08479
BDRAR [59]	0.7433	58.43	95.66	0.04346	0.8619	75.37	93.50	0.06510
DSC [17]	0.7548	59.81	95.65	0.04372	0.8479	75.36	92.82	0.07206
CPNet [48]	0.7342	56.36	94.85	0.05175	0.8314	69.86	92.44	0.07603
GloRe [10]	0.7743	61.25	95.61	0.04411	0.8600	76.10	93.07	0.06957
PSPNet [54]	0.8057	60.44	96.13	0.03920	0.8459	67.99	92.19	0.07875
DeepLabv3+ [9]	0.8096	64.08	96.43	0.04001	0.8750	77.48	94.13	0.05932
MirrorNet [47]	0.7775	62.50	96.27	0.04101	0.8597	77.41	92.75	0.07257
PMD-Net [24]	0.8276	62.40	96.80	0.08782	0.8691	76.88	93.94	0.06130
Ours	0.8437	66.84	96.82	0.04935	0.8887	79.85	94.63	0.05421

4. Experiments

4.1. Datasets and Evaluation Metrics

We experiment on the two popular 2D mirror detection benchmarks, PMD [24] and MSD [47].

PMD [24]: This dataset consists of 5,095 training images and 571 test images. It contains diverse real-world images that cover a variety of scenes and common objects, making it a very challenging dataset.

MSD [47]: This dataset consists of 3,063 training images and 955 test images. As most of the images are taken in close shots, they contain mostly large mirrors and lack contexts. In addition, a lot of images are also similar to each other. As a result, this dataset is less challenging.

To evaluate the results, we use max f-measure, Intersection over Union (IoU), accuracy, and Mean Absolute Error (MAE).

4.2. Implementation Details

We adopt ResNet50 [16] with the multi-grid method [8] as our semantic side-path. We train it on the collected semantic annotations (see Section 3.1). After training the semantic side-path, we freeze its weights during the following fine-tune stages. That is we share the same semantic side-path across the two mirror datasets, PMD [24] and MSD [47]. We use ResNeXt-101 [45] pre-trained on ImageNet-1K [13] as our encoder backbone. We use the Glorot initialization[2] [15] to initialize our GCNs. We train our model on a single NVIDIA TESLA V100 graphic card. The detail training settings are listed in Table 2. We apply CRF [20] as post-processing during inference. We use a combination of random horizontal flipping, random rotation, random center cropping, and adding random noises as the data augmentation method.

4.3. Quantitative Results

To fully evaluate our approach, we compare it with 10 different related methods, including mirror detection methods, MirrorNet [47] and PMD-Net [24], salient object detection methods [11, 55], shadow detection methods [17, 59], and semantic segmentation methods [9, 10, 48, 54].

For a fair comparison, we re-train all the methods and test them under the same dataset, *i.e.*, either PMD or MSD. Table 1 shows the quantitative results. We can see that the proposed method outperforms all the other methods on almost all the metrics on the two datasets. In particular, our f-measure/IoU scores surpass the second best scores by 1.95%/4.31% and 1.57%/3.06% on PMD [24] and MSD [47], respectively. We also observe that there are wider performance gaps between our method and the compared methods on PMD than those on MSD. There are two main reasons for this. First, since MSD contains mostly simple scenes, most methods can work well on it and the results are already close to the peak. In contrast, PMD contains more complex real-world scenes, the performances of different methods spread out much wider. Second, as PMD

Table 2. **Experiment Settings.** We show the basic settings used in our experiments. **ssp**: semantic side-path trained on the collected semantic annotations for PMD [24]. **Ours-PMD**: our model trained on PMD [24]. **Ours-MSD**: our model trained on MSD [47].

Experiment	ssp	Ours-PMD	Ours-MSD
iterations	30K	40K	40K
base lr	0.01	5e-4	8e-4
lr scheduler	poly	poly	poly
optimizer	SGD	SGD	SGD
batch size	16	10	10
weight decay	5e-4	5e-4	5e-4
momentum	0.9	0.9	0.9

[2] In Pytorch: torch.nn.init.xavier_uniform_(tensor, gain=1.0).

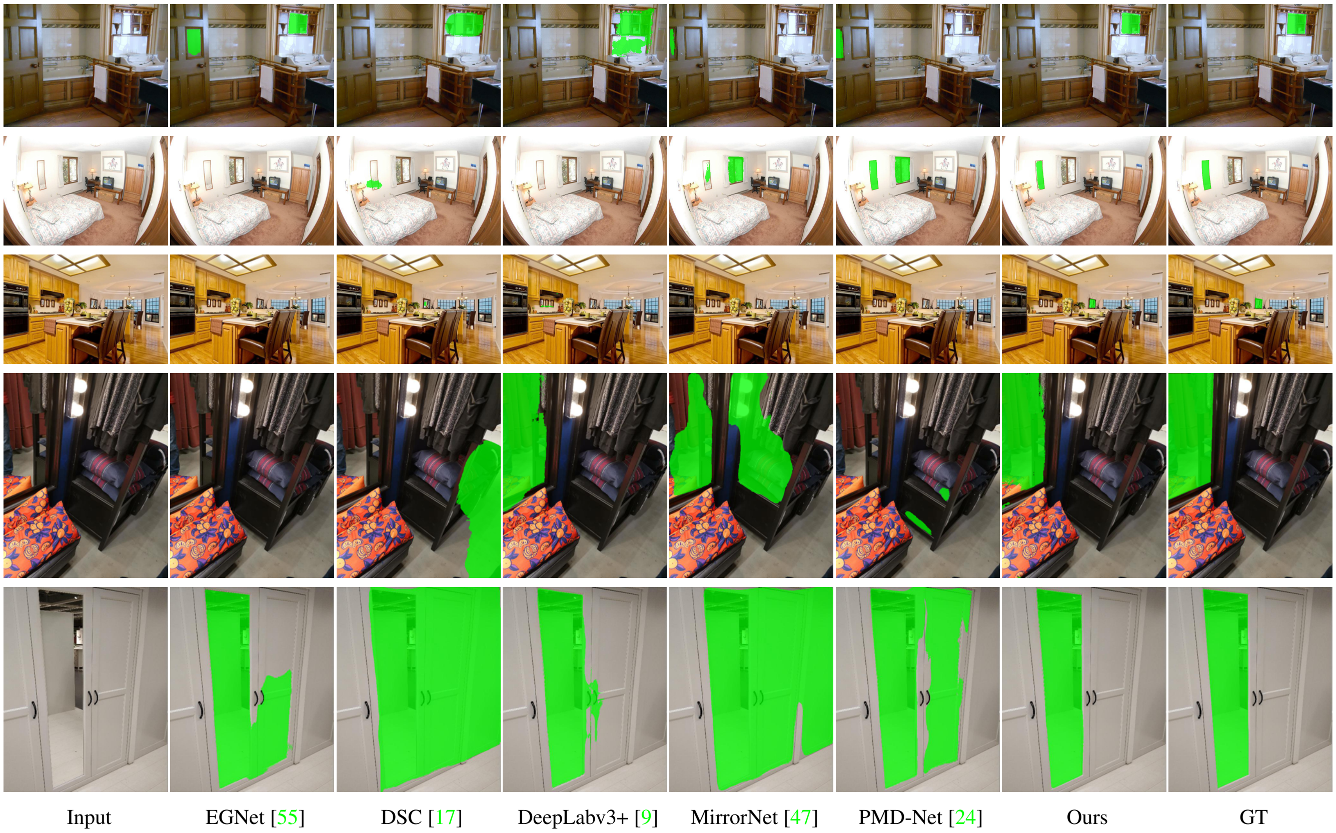

Figure 5. **Qualitative Comparison.** The detected mirrors and the GT mirrors are shown in green color. In general, our model produces much favorable results compared to the other methods.

contains more objects in each image, it can provide sufficient contexts to learn the semantic associations. In contrast, it is more difficult to learn them from MSD due to its simple image contents.

4.4. Qualitative Results

We have also evaluated our method qualitatively. Due to the limited space, we compare with the best performing methods from each group in Table 1, as shown in Figure 5. In general, our method produces more accurate results, compared with all the other methods. For example, in the first and second rows, the existing mirror detection methods, PMD-Net [24] and MirrorNet [47], both detect some distractors that look like mirrors as mirrors. By learning semantic associations, our model can identify the mirrors accurately, but not the distractors. In the third row, although the mirror is very small in a cluttered background, our method can correctly detect it while all the other methods fail. The fourth row shows a scene with a mirror inside a closet. Although it may not be easy even for humans to locate the mirror, our method can locate it well. In the last row, only our method can accurately locate the mirror, without distracted by the wardrobe doors.

4.5. Ablation Study

To analyze the importance of each component of our model, we conduct an ablation study as shown in Table 3.

Analysis on the Semantic Side-Path. We construct our baseline as an encoder-decoder (see ID1). We add the semantic side-path to our baseline to measure the perf. gain of the semantic side-path (see ID2). We can see that there is an absolute improvement of 1.15% on f-measure, indicating the importance of semantic context to mirror detection.

Analysis on the proposed modules. We then add the proposed modules one by one to reveal the effectiveness of each module, as ID3, ID4, ID5 and Ours in Table 3. We can see that both the AE and the QG modules further enhance the performance.

Analysis on the spatial-aware GCNs. Both the AE and the QG modules include spatial-aware GCNs, which help perform spatial inference. We analysis the effectiveness of these spatial-aware GCNs by removing all of them from the full model, as ID6 in Table 3. We can see that f-measure now drops to 0.8297 from 0.8437 (Ours), suggesting that the proposed modules with the spatial-aware GCNs can lead to better performances.

Table 3. **Ablation Study.** We evaluate different parts of our model, and report the max f-measure and accuracy on PMD [24]. A ✓indicates that the corresponding module/technique is selected. **ssp**: semantic side-path. **AEM**: AE module. **QGM1/QGM2**: QG module after the encoder-stage3/encoder-stage4. **Spat. GCNs**: spatial-aware GCNs in the AE and QG modules. **freeze ssp**: freeze the semantic side-path during the fine-tune stage.

Method	encoder	decoder	ssp	AEM	QGM1	QGM2	Spat. GCNs	freeze ssp	f-measure↑	Accuracy↑
ID1	ResNeXt101	✓							0.8129	96.52
ID2	ResNeXt101	✓	✓					✓	0.8244	96.51
ID3	ResNeXt101	✓	✓	✓			✓	✓	0.8284	96.91
ID4	ResNeXt101	✓	✓	✓	✓		✓	✓	0.8368	96.72
ID5	ResNeXt101	✓	✓		✓	✓	✓	✓	0.8398	96.75
ID6	ResNeXt101	✓	✓	✓	✓	✓		✓	0.8297	96.67
ID7	ResNeXt101	✓	✓	✓	✓	✓	✓		0.8373	96.78
Ours	ResNeXt101	✓	✓	✓	✓	✓	✓	✓	0.8437	96.82

Table 4. **Comparison on the RGBD dataset [30].** F_β^ω: weighted f-measure score [29]. **BER**: Balanced Error Rate. †: using depth maps during both training and inference. *: results are reported by [30]. Best performances are in red.

Method	IoU↑	F_β^ω ↑	MAE↓	BER↓
MirrorNet [47]	68.37*	0.723*	0.062*	8.66*
PMD [24]	72.27*	0.775*	0.054*	10.71*
PDNet [30]	73.57*	0.783*	0.053*	9.26*
Ours	74.99	0.800	0.048	10.56
PDNet† [30]	77.77*	0.825*	0.042*	7.77*
Ours†	78.43	0.834	0.041	8.16

Freeze semantic side-path. In the above experiments, the semantic side-path is frozen to make sure the class-specific knowledge would not be changed during the fine-tune stage. In ID7, we try **not** freezing the semantic side-path, and therefore we fine-tune the whole model. The results show that the performance of ID7 is similar to those of ID4 and ID5. We believe that without freezing the semantic side-path, it causes a semantic distortion, diverging the desired semantic association learning procedure. In contrast, our full model (with frozen semantic side-path) can produce the best performance.

4.6. Comparison on the RGBD Dataset

Although our proposed method is based on RGB input, for a more comprehensive evaluation, we would like to explore how it compares with PDNet [30], which is based on RGBD input. We conduct the experiment on their RGBD dataset, with and without using the depth maps. To utilize the depth maps, we add a depth branch in the same way as PDNet *et al.* [30] and combine the depth features after the gating module (between encoder-stage2 and encoder-stage3) using concatenation and a 1x1 convolution.

Table 4 shows the results. We can see that our method outperforms all the other methods on IoU, F_β^ω and MAE under the same data modality, *i.e.*, RGB or RGBD images.

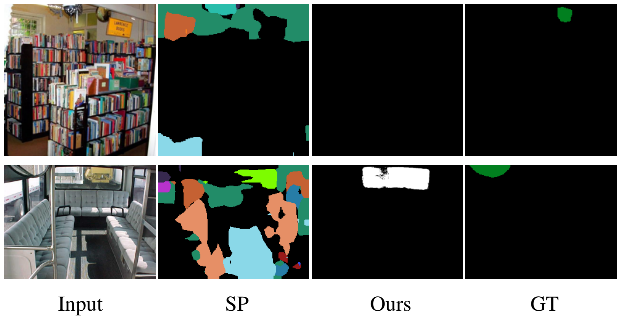

Figure 6. **Limitations.** Top row: the ***SP*** map contains a large unknown area, and our model fails to detect any mirrors. Bottom row: the ***SP*** map contains a lot of labelling errors, resulting in an incorrect detection of the mirror.

5. Conclusion

In this paper, we have proposed to learn the semantic scene contexts for the mirror detection task, and proposed a model to exploit the semantic associations between mirrors and their surrounding objects. Our experiments show that while the proposed AE module can help learn semantic associations effectively, the proposed QG module can help detect mirrors accurately with the learned semantic associations. As a result, our proposed method produces new state-of-the-art results on both RGB-based and RGBD-based mirror detection benchmarks.

Our method does have some limitations. Our approach relies on learning semantic knowledge and associations. If a scene does not contain sufficient known semantics (*e.g.*, Figure 6 top example) or if the mirror appears in an unusual place (*e.g.*, Figure 6 bottom example), our method may fail to detect the mirror correctly.

Acknowledgements

This project is in part supported by a GRF from RGC of Hong Kong (RGC Ref.: 11205620), and a SRG from City University of Hong Kong (Ref.: 7005674).

References

[1] Bogdan Alexe, Nicolas Heess, Yee Whye Teh, and Vittorio Ferrari. Searching for objects driven by context. In *NeurIPS*, 2012. 2

[2] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5:617–629, 2004. 1

[3] Moshe Bar and Elissa Aminoff. Cortical analysis of visual context. *Neuron*, 38:347–58, 2003. 1

[4] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *CVPR 2015*. 4

[5] Maxim Berman, Amal Rannen Triki, and Matthew Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR 2018*. 5

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR 2018*. 4

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *3DV 2017*. 1, 2

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 4, 6

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV 2018*. 6, 7

[10] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Shuicheng Yan, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR 2019*. 2, 6

[11] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *AAAI 2020*. 6

[12] Bowen Cheng, Maxwell Collins, Yukun Zhu, Ting Liu, Thomas Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR 2020*. 4

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*. 6

[14] Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. Frameexit: Conditional early exiting for efficient video recognition. In *CVPR 2021*. 3, 5

[15] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS 2010*. 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR 2016*. 4, 6

[17] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *CVPR 2018*. 6, 7

[18] Daniel Kaiser, Timo Stein, and Marius V Peelen. Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences*, 111(30):11217–11222, 2014. 1

[19] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR 2017*. 2, 4

[20] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *NeurIPS*, pages 109–117, 2011. 6

[21] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV 2019*. 2

[22] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *CVPR 2020*. 2, 4, 5

[23] Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. In *NeurIPS*, pages 9245–9255, 2018. 4

[24] Jiaying Lin, Guodong Wang, and Rynson W. H. Lau. Progressive mirror detection. In *CVPR 2020*. 1, 2, 3, 4, 6, 7, 8

[25] Peiwen Lin, Peng Sun, Guangliang Cheng, Sirui Xie, Xi Li, and Jianping Shi. Graph-guided architecture search for real-time semantic segmentation. In *CVPR 2020*. 2

[26] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR 2017*. 3

[27] Bang Liu, Di Niu, Haojie Wei, Jinghong Lin, Yancheng He, Kunfeng Lai, and Yu Xu. Matching article pairs with graphical decomposition and convolutions. In *ACL 2019*. 2

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR 2015*. 4

[29] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps. In *CVPR 2014*. 8

[30] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *CVPR 2021*. 1, 2, 8

[31] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR 2014*. 2, 4

[32] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007. 1

[33] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *ICCV 2017*. 2

[34] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *ICLR 2018*. 2

[35] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV 2012*. 4

[36] Avishek Siris, Jianbo Jiao, Gary Tam, Xianghua Xie, and Rynson Lau. Scene context-aware salient object detection. In *ICCV 2021*. 3

[37] Shuran Song, Samuel Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR 2015*. 4

[38] Jiaqi Tan, Weijie Lin, Angel X. Chang, and Manolis Savva. Mirror3d: Depth refinement for mirror surfaces. In *CVPR 2021*. 2

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2, 5

[40] Jia Wan, Wenhan Luo, Baoyuan Wu, Antoni Chan, and Wei Liu. Residual regression with semantic prior for crowd counting. In *CVPR 2019*. 3

[41] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV 2018*. 2

[42] Thomas Whelan, Michael Goesele, Steven Lovegrove, Julian Straub, Simon Green, Richard Szeliski, Steven Butterfield, Shobhit Verma, and Richard A. Newcombe. Reconstructing scenes with mirror and glass surfaces. *ACM Trans. Graph.*, 37(4):102:1–102:11, 2018. 2

[43] Aming Wu, Linchao Zhu, Yahong Han, and Yi Yang. Connective cognition network for directional visual commonsense reasoning. In *Advances in Neural Information Processing Systems*, 2019. 2

[44] Yangxin Wu, Gengwei Zhang, Yiming Gao, Xiajun Deng, Ke Gong, Xiaodan Liang, and Liang Lin. Bidirectional graph reasoning network for panoptic segmentation. In *CVPR 2020*. 2

[45] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR 2017*. 3, 6

[46] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *CVPR 2019*. 2

[47] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In *ICCV 2019*. 1, 2, 3, 6, 7, 8

[48] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *CVPR 2020*. 5, 6

[49] Oliver Zendel, Katrin Honauer, Markus Murschitz, Martin Humenberger, and Gustavo Fernández Domínguez. Analyzing computer vision data - the good, the bad and the ugly. In *CVPR 2017*. 1

[50] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR 2018*. 2

[51] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *CVPR 2019*. 2

[52] Lu Zhang, Jianming Zhang, Zhe Lin, Huchuan Lu, and You He. Capsal: Leveraging captioning to boost semantics for salient object detection. In *CVPR 2019*. 3

[53] Yu Zhang, Mao Ye, Dinesh Manocha, and Ruigang Yang. 3d reconstruction in the presence of glass and mirrors by acoustic and visual fusion. *IEEE TPAMI*, 40(8):1785–1798, 2018. 1, 2

[54] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR 2017*. 4, 6

[55] 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*. 6, 7

[56] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR 2017*. 4

[57] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *IJCV*, 127(3):302–321, 2019. 4

[58] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Context-reinforced semantic segmentation. In *CVPR 2019*. 2

[59] 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*. 6