Symmetry-Aware Transformer-Based Mirror Detection

Tianyu Huang[1,2], Bowen Dong[1], Jiaying Lin[2], Xiaohui Liu[1], Rynson W.H. Lau[2], Wangmeng Zuo[1, 3*]

[1] Harbin Institute of Technology
[2] City University of Hong Kong
[3] Peng Cheng Laboratory
{tyhuang0428,cndongsky,lxh720199}@gmail.com, jiayinlin5-c@my.cityu.edu.hk,
rynson.lau@cityu.edu.hk, wmzuo@hit.edu.cn

Abstract

Mirror detection aims to identify the mirror regions in the given input image. Existing works mainly focus on integrating the semantic features and structural features to mine specific relations between mirror and non-mirror regions, or introducing mirror properties like depth or chirality to help analyze the existence of mirrors. In this work, we observe that a real object typically forms a loose symmetry relationship with its corresponding reflection in the mirror, which is beneficial in distinguishing mirrors from real objects. Based on this observation, we propose a dual-path Symmetry-Aware Transformer-based mirror detection Network (SATNet), which includes two novel modules: Symmetry-Aware Attention Module (SAAM) and Contrast and Fusion Decoder Module (CFDM). Specifically, we first adopt a transformer backbone to model global information aggregation in images, extracting multi-scale features in two paths. We then feed the high-level dual-path features to SAAMs to capture the symmetry relations. Finally, we fuse the dual-path features and refine our prediction maps progressively with CFDMs to obtain the final mirror mask. Experimental results show that SATNet outperforms both RGB and RGB-D mirror detection methods on all available mirror detection datasets.

Introduction

Mirrors are common objects in the human world, and their presence can affect the performance of a range of vision tasks. For example, Zendel et al. propose a list of potential hazards within the CV domain, and the existence of mirrors is one of them. However, mirror detection can be challenging by using some general detection methods from related tasks, such as salient object detection and semantic segmentation. As such, it is necessary to treat mirror detection as an independent vision task, and previous works have managed to tackle this issue from either relation-based frameworks or property-based paradigms.

Relations between mirror and non-mirror regions are counted in most mirror detection methods. Yang et al. propose to extract contextual discontinuities among regions, but it can only be effective when mirror boundaries are clear against backgrounds. Lin, Wang, and Lau propose to perceive similarity relationships for contents inside and outside mirrors, which may easily fail when similarities come from non-mirror regions. Guan, Lin, and Lau propose to learn semantic associations in mirror scenes, while such relation quite relies on the environments nearby mirrors. It can only adapt to a few mirror cases, e.g., a mirror above a sink. Considering mirror properties, Mei et al. and Tan et al. regard depth and chirality as additional information for the detection, respectively. However, these property aggregations only focus on mirror regions, dismissing the environmental semantics related to mirrors.

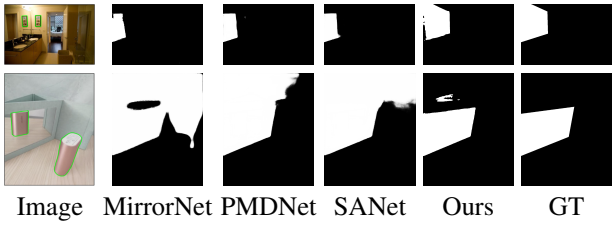


Figure 1: Comparison of mirror detection among state-of-the-art methods. MirrorNet (Yang et al. 2019) cannot handle scenes with vague mirror boundaries. Although PMDNet (Lin, Wang, and Lau 2020) considers similarity semantics, it can hardly detect the symmetry pair (1st row), and can easily count part of the similar non-mirror regions into mirrors (2nd row). SANet (Guan, Lin, and Lau 2022) only detects the mirror region above the sink (1st row), and it has worse predictions when semantic associations are lacking (2nd row). By modeling a loose symmetry relationship, SATNet succeeds in both cases.

For a general solution, we need to fully leverage the relationship between mirror and non-mirror regions based on mirror properties. Considering mirror reflection, symmetry relationships between mirror and non-mirror regions are supposed to be an essential cue for mirror detection. In Fig. 1(1st row), the right half of the mirror would not be missed if the mirror detection model could detect the mirror symmetry relationship of the two paintings. In Fig. 1(2nd row), if the model recognizes the left power bank as the mirror region, it can then classify the corresponding real power bank on the right as a non-mirror region. However, this symmetry relationship is not a strict mirror symmetry and is highly dependent on the camera viewpoint. The paint-

*Corresponding Author
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ings inside and outside mirrors form a nearly perfect reflection symmetry pair in Fig. 1(1st row), while the power banks inside and outside mirrors are from different views in Fig. 1(2nd row). We cannot adopt reflection symmetry detection methods directly. Instead, we observe that real-world object and their reflection in mirrors always maintain semantic or luminance consistency with each other, even though they may not be strictly symmetric regarding the position or orientation. That is, an object in a mirror should be a mirror reflection of an object in the real world from a certain view. We regard this kind of relationship as loose symmetry and aim to explore a new solution to model and leverage this loose symmetry relationship for mirror detection.

Taking loose symmetry into account, we present our Symmetry-Aware Transformer-based mirror detection Network (SATNet). In particular, we introduce the first transformer baseline in mirror detection, considering the long-range dependencies that loose symmetry requires. We construct a dual-path network to extract and enhance symmetric features, taking an input image as well as its corresponding horizontally flipped image as inputs. For modeling symmetry semantics, we propose a novel Symmetry-Aware Attention Module (SAAM) in high-level dual-path features. For mirror region segmentation, we propose a novel Contrast and Fusion Decoder Module (CFDM), which constructs a pyramidal decoder to progressively fuse and refine dual-path features.

To sum up, our main contributions include:

- We observe that there are typically loose symmetry relationships between mirror and non-mirror regions. Based on this observation, we propose a novel dual-path Symmetry-Aware Transformer-based mirror detection network (SATNet) to learn symmetry relations for mirror detection. This is the first transformer pipeline in mirror detection.
- We present a novel Symmetry-Aware Attention Module (SAAM) to extract high-level symmetry semantics and a novel Contrast and Fusion Decoder Module (CFDM) to refine multi-scale mirror features.
- Our network SATNet achieves state-of-the-art results on various mirror detection datasets. Experimental results clearly demonstrate the benefit of loose symmetry relationships for mirror detection.

Related Work

Mirror Detection

The mirror detection task aims to identify the mirror regions of the given input image. To tackle this problem, several methods attempt to model specific relations between mirror and non-mirror regions. Yang et al. proposed the first mirror detection network called MirrorNet, which focuses on perceiving the contrasting features between the contexts inside and outside mirrors. Lin, Wang, and Lau suggested a progressive mirror detection network PMDNet, designing a relational contextual contrasted local module to extract similarity features. Guan, Lin, and Lau proposed to learn semantic association in mirror scenes, which may imply the existence of mirrors. However, those methods can hardly adapt to general mirror detection cases as the relations they match are either too simple or too strict. Recent methods take mirror properties into account. Mei et al. introduced depth information to mirror detection as the depth value in mirror regions is irregular. In contrast, the depth input is unreliable, and the method can be easily misled by depth. Tan et al. proposed a dense visual chirality discriminator to judge the possible mirror existence, while the improvement is limited since chirality information tends to be subtle when mirror contents are clean. The leverage of mirror properties in these methods mainly depends on the semantics of mirror regions, dismissing the interaction with non-mirror regions. Unlike existing works, we aim to utilize loose symmetry relationships between real-world objects and corresponding mirror regions to enhance the overall detection ability.

Reflection Symmetry Detection

Reflection symmetry detection aims to detect symmetry axes in given images. Early works in this task can be divided into two categories: keypoint matching detection and dense heatmap detection. Loy and Eklundh adopted SIFT (Lowe 2004) to compute matched keypoints, and generated potential symmetry axes accordingly. Cornelius and Loy took a single matching pair for hypothesizing with the local affine frame. For dense heatmap, Tsogkas and Kokkinos utilized pixel-level features to predict the symmetry area densely. Funk and Liu employed CNNs to extract the symmetric features directly. Recently, Seo, Shim, and Cho proposed a novel polar matching convolution to encode the similarities among pixels. Contrary to the strict reflection symmetry, symmetry relationships in mirror cases are loosely defined. Therefore, reflection symmetry detection methods cannot be directly employed in mirror detection. To tackle this, we propose a dual-path Transformer-based structure with attention mechanisms in high-level features to model the loose symmetry relationships.

Salient Object Detection

Salient object detection (SOD) aims to detect and segment the most distinct object in an input image. Existing methods in RGB SOD are mainly based on the UNet structure (Ronneberger, Fischer, and Brox 2015), like (Wang et al. 2017; Pang et al. 2020b). Deng et al. adopt a recurrent network to refine the salient map progressively. Liu, Han, and Yang adopt attention mechanisms to learn more dependencies among features. Recently, RGB-D SOD has received considerable attention. Several methods (Song et al. 2017) treat depth as an additional dimension of the input features, while the others (Fan et al. 2020a) separately extract RGB and depth features and fuse them in the decoding process. Liu, Zhang, and Han proposed to fuse depth information with attention mechanisms. Pang et al. integrated RGB and depth through densely connected structures. Liu et al. proposed a vision transformer network, rethinking this field from an aspect of sequence-to-sequence architectures. Albeit similar to mirror detection, SOD methods can hardly have a good performance on the mirror detection task as mirrors are not salient enough to detect in most cases. SOD methods may wrongly detect some conspicuous objects inside mirrors.

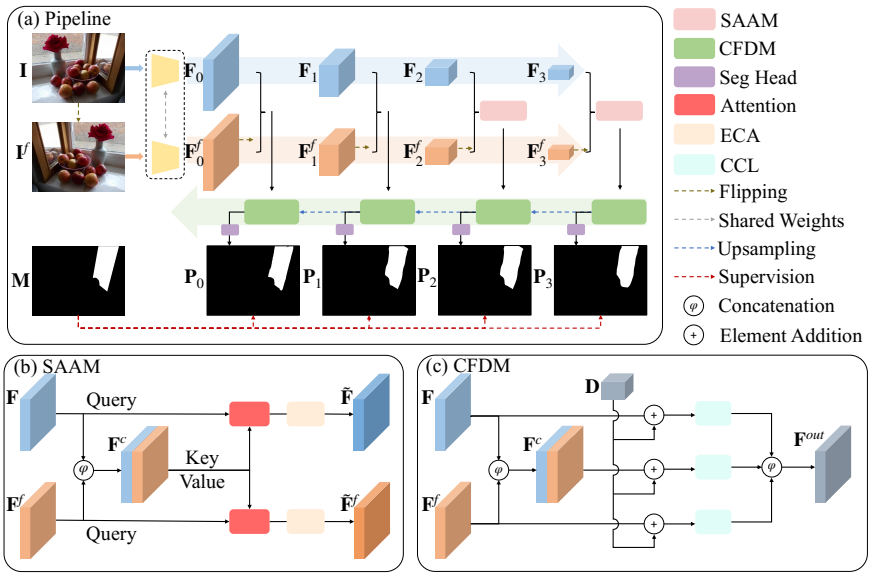


Figure 2: (a) Pipeline of our SATNet. (b) Symmetry-Aware Attention Module. (c) Contrast and Fusion Decoder Module.

Method

Based on the idea of loose symmetry relationships, we propose a dual-path Symmetry-Aware Transformer-based network for mirror detection. Loose symmetry relationships can assist the detecting process in two aspects: presences of loose symmetry relationships imply the possible existence of mirrors; differences between symmetric pairs indicate which part belongs to mirror regions. The dual-path structure and our novel Symmetry-Aware Attention Module are designed for the first aspect. Additionally, to better encode the symmetry features as well as recognize the corresponding mirror semantics, a transformer backbone and our Contrast and Fusion Decoder Module are proposed for the second aspect.

Overview

Fig. 2(a) illustrates the pipeline of our SATNet. Given an input image $\mathbf{I}$ as well as its flipped image $\mathbf{I}^f$, we feed them into a shared-weights transformer backbone to obtain multi-scale features $\{\mathbf{F}_0, ..., \mathbf{F}_3\}$ and corresponding flipped features $\{\mathbf{F}_0^f, ..., \mathbf{F}_3^f\}$, respectively. For modeling symmetry relations, we select features from the highest two levels of both paths, and feed features in the same level into our Symmetry-Aware Attention Module (SAAM), fetching joint features $\hat{\mathbf{F}}$ and $\hat{\mathbf{F}}^f$. Then, the multi-scale features $\{\mathbf{F}_i/\hat{\mathbf{F}}_i\}$ as well as the flipped features $\{\mathbf{F}_i^f/\hat{\mathbf{F}}_i^f\}$ are fed into corresponding Contrast and Fusion Decoder Module (CFDM), generating coarse output features $\mathbf{F}_i^{out}$ with different scales progressively. For each $\mathbf{F}_i^{out}$ (except $\mathbf{F}_0^{out}$), we upsample it into the next decoder as the reference features $\mathbf{D}_{i-1}$ for further prediction refinement. Meanwhile, we get the prediction map $\mathbf{P}_i$ in each decoder through a segmentation head and supervise it via the ground-truth mask $\mathbf{M}$. Finally, our prediction result $\tilde{\mathbf{M}}$ is generated by the last decoder module.

Dual-Path Structure

In most cases, loose symmetry relationships are implicit under complex semantics. Such a relationship is too hidden to be perceived by existing baselines, which has been verified in our ablation study. To better perceive the relationship, dual enhancements are suggested. As a common method of data augmentation, horizontal flipping can modify the global semantics of natural images, while symmetry relationships still exist (they just display in the opposite direction). Thus, we introduce a dual-path network to extract symmetric features: Given $\mathbf{F}$ and $\mathbf{F}^f$ from both paths, we expect they differ from each other but have features of the same loose symmetry relationship. When we concatenate them together as $\mathbf{F}^c$, symmetry semantics in the symmetric region can be enhanced. To extract the same symmetric features, input images $\mathbf{I}$ and $\mathbf{I}^f$ must be fed into the same backbone. Our fusion function is defined as follows:

$$\varphi(\mathbf{a_1}, ..., \mathbf{a_n}) = \sigma(BN(\psi_{3\times3}(\psi_{1\times1}([\mathbf{a_1}, ..., \mathbf{a_n}])))), \quad (1)$$

$$\mathbf{F}^c = \varphi(\mathbf{F}, flip(\mathbf{F}^f)), \quad (2)$$

where $[\cdot, ..., \cdot]$ denotes the concatenation operation on the channel dimension. $\psi_{w\times w}$ is a $w \times w$ convolution, BN denotes the Batch Normalization, σ is the ReLU activation function, and flip is the horizontal flipping. To align features in the spatial level during concatenation, we flip $\mathbf{F}^f$ back before SAAM or CFDM.

Symmetry-Aware Attention Module

Fig. 2(b) shows the architecture of our symmetry-aware attention module. With SAAM, we aim to perceive loose symmetry relationships in an image that indicates the possible existence of mirrors. To this end, we use the attention mechanisms (i) to enhance the feature $\mathbf{F}$ of the input image as well as (ii) to obtain the symmetry-aware feature by modeling the dependency between the input and its flipped images.

In general, the attention mechanism can model the dependencies among each position in a global manner (Vaswani et al. 2017), which can be formulated as,

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}, \quad (3)$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ denote Query, Key, and Value, respectively.

Our SAAM takes both $\mathbf{F}^c$ as well as $\mathbf{F}$ and $\mathbf{F}^f$ as the input. Among them, $\mathbf{F}^c$ aggregates the features from $\mathbf{F}$ and $\mathbf{F}^f$ from both paths and is spatially consistent with $\mathbf{F}$, and thus can be treated as an augmented representation of $\mathbf{F}$. To exploit the attention to enhance the feature $\mathbf{F}$, we treat $\mathbf{F}$ as query and $\mathbf{F}^c$ as key and value, and further apply channel transformation with Efficient Channel Attention (ECA) (Wang et al. 2020) module right after the attention module to obtain the enhanced feature $\hat{\mathbf{F}}$,

$$\hat{\mathbf{F}} = ECA(Attention(\mathbf{F}, \mathbf{F}^c, \mathbf{F}^c)), \quad (4)$$

where $ECA(\cdot)$ denotes the Efficient Channel Attention module. To obtain the symmetry-aware feature, we treat $\mathbf{F}^f$ as query and $\mathbf{F}^c$ as key and value. Note that $\mathbf{F}^f$ is extracted from the flipped image, and $\mathbf{F}^c$ is spatially consistent with the input image. Their similarity score can thus be treated as an indicator of loose mirror symmetry between parts from the input and its flipped images. And the output of the attention module can then be regarded to be symmetry-aware. Analogous to $\hat{\mathbf{F}}$, the symmetry-aware feature is obtained by,

$$\hat{\mathbf{F}}^f = ECA(Attention(\mathbf{F}^f, \mathbf{F}^c, \mathbf{F}^c)). \quad (5)$$

Contrast and Fusion Decoder Module

Since MirrorNet (Yang et al. 2019), Context Contrasted Local (CCL) decoder (Ding et al. 2018) has been widely adopted in mirror detection networks. To better refine the prediction, edge extractors are joint as an extra supervision to previous methods (Lin, Wang, and Lau 2020; Tan et al. 2022) as well. In this subsection, we further extend the CCL module to present our CFDM for handling multiple features. With no edge information, our CFDM can outline precise mirror boundaries efficiently by refining multi-level features progressively in a top-down structure.

As shown in Fig. 2(c), our CFDM takes $\mathbf{F}_i$ and $\mathbf{F}_i^f$ as the input when $i = 0, 1$, and $\hat{\mathbf{F}}_i$ and $\hat{\mathbf{F}}_i^f$ when $i = 2, 3$. Without loss of generality, we use $\mathbf{F}_i$ and $\mathbf{F}_i^f$ as an example to explain the CFDM module. To begin with, we use Eq. (2) to obtain the fused feature $\mathbf{F}_i^c$. Denote by $\mathbf{F}_{i+1}^{out}$ the $(i+1)$-scale CFDM output. We then upsample $\mathbf{F}_{i+1}^{out}$ to obtain the higher-level feature map,

$$\mathbf{D}_i = U_2(\sigma(BN(\psi_{3\times3}(\mathbf{F}_{i+1}^{out})))), \quad (6)$$

where U_2 denotes the bilinear upsampling operation. Subsequently, the reference features for $(\mathbf{F}_i^c, \mathbf{F}_i, \mathbf{F}_i^f)$ can be given by,

$$(\tilde{\mathbf{F}}_i^c, \tilde{\mathbf{F}}_i, \tilde{\mathbf{F}}_i^f) = \begin{cases} (\mathbf{F}_i^c, \mathbf{F}_i, \mathbf{F}_i^f) \oplus (\mathbf{D}_i, \mathbf{D}_i, \mathbf{D}_i), & i < 3 \\ (\mathbf{F}_i^c, \mathbf{F}_i, \mathbf{F}_i^f), & i = 3 \end{cases} \quad (7)$$

where $\oplus$ denotes the element-wise summation operator.

The three feature maps $\tilde{\mathbf{F}}_i^c, \tilde{\mathbf{F}}_i, \tilde{\mathbf{F}}_i^f$ are separately fed into the CCL module to extract contrastive semantics. Here we use $\tilde{\mathbf{F}}_i$ as an example,

$$CCL(\tilde{\mathbf{F}}_i) = \sigma(BN(f_l(\tilde{\mathbf{F}}_i) - f_{ct}(\tilde{\mathbf{F}}_i))), \quad (8)$$

where f_l is the local feature extractor which contains a 3×3 convolution with a dilation rate of 1, BN, and ReLU in turn. Considering the changes in the receptive field, we set dilation rates to $\{8, 6, 4, 2\}$ for layer $\{0, 1, 2, 3\}$, respectively. Finally, we concatenate those three CCL outputs together to get the output features $\mathbf{F}_i^{out}$ and the corresponding prediction map $\mathbf{P_i}$, which is given as follows,

$$\mathbf{F}_i^{out} = \varphi(CCL(\tilde{\mathbf{F}}_i^c), CCL(\tilde{\mathbf{F}}_i), CCL(\tilde{\mathbf{F}}_i^f)), \quad (9)$$

$$\mathbf{P}_i = f_{seg}(\mathbf{F}_i^{out}), \quad (10)$$

where f_{seg} is a segmentation head whose output has two channels. And the output of the last decoder layer $\mathbf{P}_0$ is adopted as the final prediction result $\tilde{\mathbf{M}}$ of our network.

Transformer for Mirror Detection

As for the feature extraction, loose symmetry is typically a long-range relationship, which means our network needs a large receptive field to perceive it. CNN-based methods utilize a couple of convolution kernels to fulfill local feature aggregation. However, the convolution with a small kernel size cannot construct global feature aggregation directly, which restricts the feature representation ability of those methods in complex scenarios. In contrast, the self-attention module in transformers can model the long-range interaction explicitly, making vision transformers very competitive in several complex scene understanding tasks (Zheng et al. 2021). Swin Transformer (Liu et al. 2021b) proposes regular and shifted window self-attention modules to construct local and global feature aggregation with limited computation complexity while achieving state-of-the-art performance in scene parsing. Thus, we adopt a transformer pipeline in mirror detection based on Swin Transformer.

Loss Function

Our learning objective is defined by considering all scales. For each prediction map $\mathbf{P}_i$, we calculate the binary cross-entropy (BCE) loss (De Boer et al. 2005) between $\mathbf{P}_i$ and the ground-truth $\mathbf{M}$. The overall loss function $\mathcal{L}$ is then given as the summation of BCE loss for each prediction map,

$$\mathcal{L} = \sum_{i=0}^{3} w_i \mathcal{L}_{\text{bce}}(\mathbf{P}_i, \mathbf{M}), \quad (11)$$

where w_i is the corresponding weight for the i-th layer. We empirically set the weight w_i as [1.25, 1.25, 1.0, 1.5] according to the experimental results.

Experiments

Datasets and Evaluation Metrics

Following previous works (Yang et al. 2019; Lin, Wang, and Lau 2020), we use Mirror Segmentation Dataset (MSD) and

Method	MSD			PMD		
	$IoU \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$IoU \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
CPDNet	57.58	0.743	0.115	60.04	0.733	0.041
MINet	66.39	0.823	0.087	60.83	0.798	0.037
LDF	72.88	0.843	0.068	63.31	0.796	0.037
VST	79.09	0.867	0.052	59.06	0.769	0.035
MirrorNet	78.88	0.856	0.066	58.51	0.741	0.043
PMDNet	81.54	0.892	0.047	66.05	0.792	0.032
SANet	79.85	0.879	0.054	66.84	0.837	0.032
VCNet	80.08	0.898	0.044	64.02	0.815	0.028
Ours	**85.41**	**0.922**	**0.033**	**69.38**	**0.847**	**0.025**

Table 1: Quantitative results of the state-of-the-art methods on MSD dataset and PMD dataset. Our method achieves the best performance in terms of all the evaluation metrics.

Method	w/ Depth	RGBD-Mirror		
		$IoU \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
JL-DCF	✓	69.65	0.844	0.056
DANet	✓	67.81	0.835	0.060
BBSNet	✓	74.33	0.868	0.046
VST	✓	70.20	0.851	0.052
PDNet		73.57	-	0.053
PDNet	✓	77.77	0.878	0.041
SANet		74.99	0.873	0.048
VCNet		73.01	0.849	0.052
Ours		**78.42**	**0.906**	**0.031**

Table 2: Quantitative results of the state-of-the-art methods on RGBD-Mirror dataset. w/ Depth denotes the usage of depth information in a corresponding method. Our method outperforms all the competing methods, even though we do not use depth information.

Progressive Mirror Dataset (PMD) to evaluate our method. Besides, we adopt an RGB-D dataset RGBD-Mirror to make a comparison with the state-of-the-art RGB-D mirror detection method PDNet (Mei et al. 2021). To assess mirror detection performance, we adopt three commonly used dense prediction evaluation metrics: intersection over union (IoU), F-measure F_β, and mean absolute error (MAE).

Implementation Details

We implement our network on PyTorch (Paszke et al. 2019) and use the small version of Swin Transformer (namely Swin-S) pre-trained on ImageNet-1k (Deng et al. 2009) as the backbone of our network. Note that dual-path features are fed into the same backbone and share weights. Following data augmentation methods adopted by previous works, we adopt random resize and crop as well as random horizontal flipping to augment training images. And for testing, we simply resize input images to 512×512 to evaluate our network. Our network is trained on 8 Tesla V100 GPUs with 2 images per GPU for 20K iterations. During training, we use ADAM weight decay optimizer and set β_1, β_2, and the weight decay to 0.9, 0.999, and 0.01, respectively. The learning rate is initialized to 6×10^{-4} and decayed by the *poly* strategy with the power of 1.0. It takes 6 hours to train our network, and testing on a single GPU needs 0.08s per image.

Comparison with State-of-the-Arts

To evaluate SATNet, we extensively compare it with several state-of-the-art methods. As shown in Table 1, we select 8 state-of-the-art methods for the comparison on MSD dataset and PMD dataset, including 4 RGB salient object detection methods CPDNet (Wu, Su, and Huang 2019), MINet (Pang et al. 2020c), LDF (Wei et al. 2020), and VST (Liu et al. 2021a), and 4 mirror detection methods MirrorNet (Yang et al. 2019), PMDNet (Lin, Wang, and Lau 2020), SANet (Guan, Lin, and Lau 2022), and VCNet (Tan et al. 2022). Our network outperforms other methods in terms of all the evaluation metrics. Fig. 3 provides the visualized comparison with those methods. The first two rows are examples of loose symmetry relationships. Our network can precisely distinguish real-world objects from their mirror reflections. In the first row, the cartoon toy and its reflection in mirrors cannot construct an apparent reflection symmetry, but our network can still perceive which part is in the mirrors. Albeit PMDNet (Lin, Wang, and Lau 2020) has a specific module for modeling similarity relationships, it fails in handling an easy case in the second row, in which a chalk eraser is reflected in the mirror. The last row has scenes where mirrors are similar to their surroundings. Our method can well exclude the non-mirror region, while the competing methods tend to classify a similar area as the mirror region, especially four mirror detection methods. The results show that symmetry awareness is beneficial for mirror detection, and our method can utilize symmetry information well.

Our method is also compared with 4 RGB-D salient object detection methods JL-DCF (Fu et al. 2020), DANet (Zhao et al. 2020), BBSNet (Fan et al. 2020b) and VST (Liu et al. 2021a), and 3 mirror detection methods PDNet (Mei et al. 2021), SANet (Guan, Lin, and Lau 2022) and VCNet (Tan et al. 2022) on the RGBD-Mirror dataset. As shown in Table 2, our method does not leverage depth information, and can still achieve the best performance in terms of all the evaluation metrics. Visualization results are shown in Fig. 4. In all the four examples, RGB-D methods are likely misled by depth information. Especially in the first row, they wrongly judge the depth changes as the existence of mirrors. In the second row, our method correctly detects the mirror region by exploiting the loose symmetry relationship between the television and its reflection, while some competing methods even fail to detect the correct side of the mirror. In the third row, there is a mirror that can be easily missed. All the competing methods ignore the left mirror, although the depth map has an obvious change in that area. Our method can still discover the mirror as the scene has a kind of symmetry relation with the nearby cabinet. In the last row, we note that our method does not mis-detect the glasses as a mirror region, while the competing methods can hardly tell subtle differences between mirrors and glasses. Different from mirrors, glasses can transmit most of the light, which weakens reflection effects. It shows that our method can identify corresponding reflection features from mirrors. All the cases show that symmetry information can greatly benefit the per-

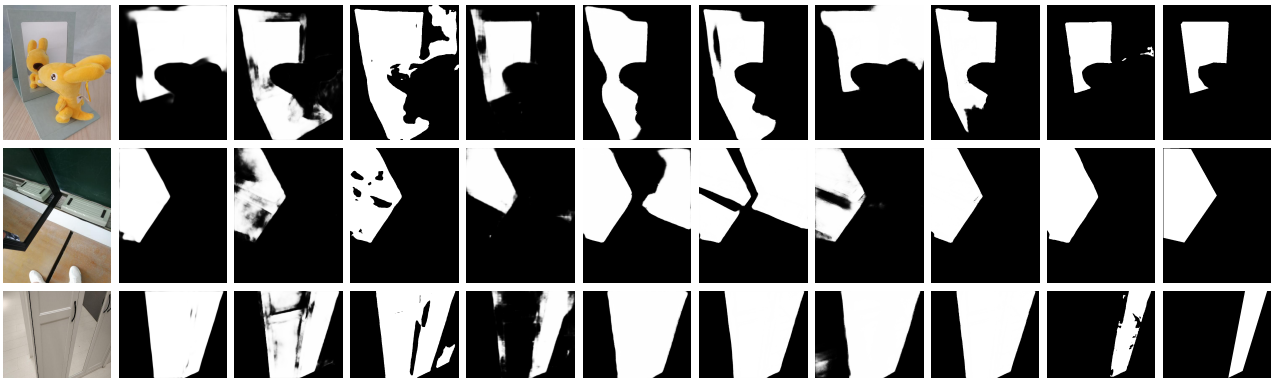


Figure 3: Visualization results on MSD and PMD datasets. The first two rows are examples of loose symmetry relationships. The last row has scenes where mirrors are similar to their surroundings.

Method	$IoU \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
Baseline	80.46	0.901	0.045
Dual-Path	79.59	0.903	0.044
Dual-Path + SAAM	80.01	0.918	0.042
Dual-Path + SAAMs	80.03	0.903	0.043
Dual-Path + CFDM	81.98	0.918	0.039
Dual-Path + SAAM + CFDM	82.96	0.911	0.039
SATNet(Ours)	**85.41**	**0.922**	**0.033**

Table 3: Ablation study results on MSD. Swin-S denotes our baseline method, which is decoded by UperNet. Dual-Path denotes the dual-path Swin Transformer. SAAM denotes our Symmetry-Aware Attention Module on a scale of 3. SAAMs denotes SAAM on both scale 2 and scale 3. CFDM denotes our Contrast and Fusion Decoder Module.

formance of mirror detection, especially in complex scenes.

Ablation Study

Benefits of Dual-Path Structure. To better analyze the benefits of our dual-path structure, we conduct two experiments: One is a pure Swin Transformer decoded by UperNet (Xiao et al. 2018) (1st row); the other is a dual-path Swin Transformer, where features are trained and supervised separately in two paths (2nd row). Results in the first two rows show that, with extra features and supervision, the second method has no clear advantage when compared against the first one. That is to say, we cannot simply attribute the improvement of our method to the extra features we extract. Albeit we introduce the dual-path structure to enhance the symmetry semantics, the extra features are more like a repeated computation of the original ones if there are no appropriate fusing and matching mechanisms for the two paths.

Effect of SAAM. To evaluate the effect of our attention module, we conduct another two experiments: One is a dual-path Swin-S with a SAAM in the highest level (3-rd row), and the other is the same structure, but with SAAMs in the highest two levels (4-th row). Comparing the third row with the second row, we discover that "Dual-Path + SAAM" gets better results in all the three metrics, which is reasonable as our SAAM models symmetry relationships in high-level features. However, F_β in the fourth row drops back to 0.903, indicating that directly applying SAAM to features in lower levels may not work well. We further visualize the attention map in SAAM. In Fig. 5(c), the mirror region (green contour in (b)) of the attention map focuses on the mirror itself. While in Fig. 5(d), the highest attention signal of the power bank region inside the mirror (red contour in (b)) is located on the real power bank in the image. This observation supports that SAAM is able to model loose symmetry relations.

Effect of CFDM. In the fifth row, we conduct an experiment based on the second row, replacing the UperNet decoder with our CFDM. Comparing results of the two rows, "Dual-Path + CFDM" have a gain of 2.39%, 1.50%, and 0.5% in IoU, F_β, and MAE, respectively. The improvement proves that our decoder module can properly fuse features in the two paths, and is more suitable for the mirror detection task.

Combination of SAAM and CFDM. To explore the best way to combine our SAAM and CFDM, we conduct two experiments: one is a dual-path Swin Transformer with a SAAM in the highest level and CFDMs as the decoder (6th row), and the other is our final network SATNet, which has SAAMs in the highest two levels (last row). Analyzing the last three rows, we conclude that applying SAAM before CFDM is effective as the three evaluation metrics have progressively improved to 85.41%. 0.922 and 0.033. On the other hand, comparing the network in the fourth row with SATNet, the improvement from "Dual-Path + SAAMs" to SATNet is even larger, which means our CFDM contributes to the fusion of dual-path features, especially for symmetry semantics in high levels.

Visualization results for the ablation study. To further analyze the effectiveness of each component, we visualize the prediction results of Swin-S, Dual-Path + SAAM, Dual-Path + CFDM, and SATNet in Fig. 6. Swin-S can provide the approximate location of the mirror but is not sensitive to symmetry relationships, which demonstrates that current baselines can hardly model loose symmetry relationships. Equipped with our attention module SAAM, the network

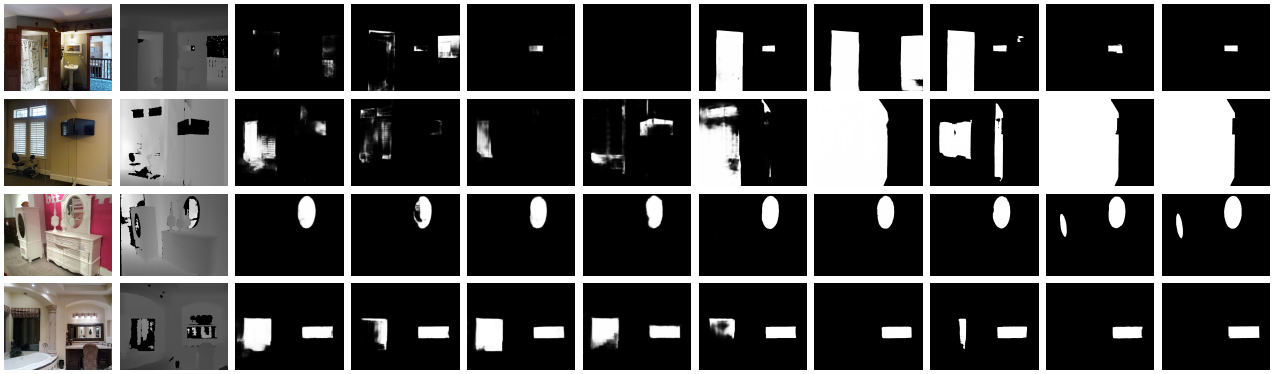


Figure 4: Visualization results on RGB-D dataset. In the first row, changes in depth can easily affect the judgement of RGB-D methods. The second row contains a pair of symmetric objects inside and outside mirrors. The third row represents mirrors that can hardly be recognized. And the last row is a scene including both glasses and mirrors.

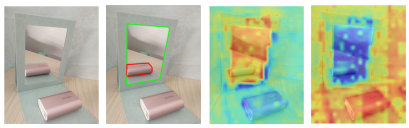


Figure 5: Visualization of attention maps in SAAM. (a)-(d) denote image, region of interest, attention of mirrors, and attention of objects in mirrors. While attention of the mirror region focuses on the mirror itself, attention of the power bank in mirrors lies in the corresponding real object.

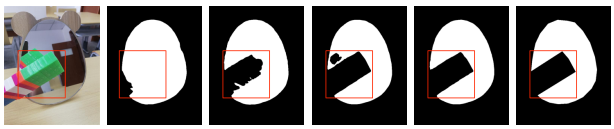


Figure 6: Visualization results for the ablation study. In this example, our baseline Swin-S cannot perceive the symmetry relationship. The network embedded with SAAM does not outline a precise boundary. And when adding CFDM, the network is still confused about the symmetry relationship. Only SATNet can correctly detect the mirror region.

can exclude the real-world object which shades the mirror from the mirror region, showing the ability of perceiving symmetry relationships. However, its prediction map is not precise enough, especially near the boundary of mirrors. In comparison to our SAAM, our decoder module CFDM refines mirror boundaries well, but it wrongly excludes the symmetry area in the mirror region. Analogous to Swin-S, it cannot handle symmetry relationships well. Only with both two modules, SATNet marks the mirror region correctly. The visualization result is basically consistent with the corresponding effects of the components we expect.

Input Size. Following Swin Transformer (Liu et al. 2021b), we train SATNet with an input image of 512×512. Nonetheless, previous networks usually adopt smaller input sizes, e.g., 384×384. To show that the superiority of our SATNet cannot be simply ascribed to larger input image size, we further train another two SATNet models respectively for input images with the sizes of 384×384 on MSD. Table 4 lists the quantitative results of our SATNet and the competing methods. From the table, one can see that: (i) increasing input image size is beneficial to the performance of our SATNet; (ii) with the same input image size, our SATNet consistently outperforms the competing methods.

Method	MACs(G)	Para.(M)	$IoU \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
MirrorNet	77.7	121.77	78.88	0.856	0.066
PMDNet	101.5	147.66	81.54	0.892	0.047
Ours-384	84.1	111.34	82.56	0.911	0.041
Ours-512	147.26	111.34	85.41	0.922	0.033

Table 4: Comparison of different mirror detection networks on MSD dataset. We report the results of our SATNet with the input image sizes of 384×384 and 512×512.

Conclusion

In this paper, we proposed a dual-path Symmetry-Aware Transformer-based mirror detection network (SATNet) for better mirror detection. We presented a new perspective on detecting mirrors by leveraging loose symmetry relationships. Then, we suggested a novel dual-path network, introducing a transformer pipeline to enhance the ability of long-range dependencies understanding for mirror detection. Furthermore, we proposed the Symmetry-Aware Attention Module (SAAM) to aggregate better feature representation of symmetry relations, while exploiting Contrast and Fusion Decoder Module (CFDM) to generate refined prediction maps progressively. Experimental results on multiple datasets demonstrate the benefit of loose symmetry relationships in mirror detection. Our network can effectively model such relationships and greatly improve the performance of mirror detection in comparison to state-of-the-arts.

Acknowledgements

This work was supported by the Major Key Project of Peng Cheng Laboratory (PCL2021A12), the National Natural Science Foundation of China (NSFC) under Grants No.s U19A2073, and two SRG grants from City University of Hong Kong (Ref: 7005674 and 7005843).

References

Cornelius, H.; and Loy, G. 2006. Detecting bilateral symmetry in perspective. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 191–191. IEEE.

De Boer, P.-T.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1): 19–67.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

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

Ding, H.; Jiang, X.; Shuai, B.; Liu, A. Q.; and Wang, G. 2018. Context Contrasted Feature and Gated Multi-Scale Aggregation for Scene Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fan, D.-P.; Lin, Z.; Zhang, Z.; Zhu, M.; and Cheng, M.-M. 2020a. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems*, 32(5): 2075–2089.

Fan, D.-P.; Zhai, Y.; Borji, A.; Yang, J.; and Shao, L. 2020b. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In *European Conference on Computer Vision*, 275–292. Springer.

Fu, K.; Fan, D.-P.; Ji, G.-P.; and Zhao, Q. 2020. JL-DCF: Joint Learning and Densely-Cooperative Fusion Framework for RGB-D Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Funk, C.; and Liu, Y. 2017. Beyond planar symmetry: Modeling human perception of reflection and rotation symmetries in the wild. In *Proceedings of the IEEE international conference on computer vision*, 793–803.

Guan, H.; Lin, J.; and Lau, R. W. 2022. Learning Semantic Associations for Mirror Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5941–5950.

Lin, J.; Wang, G.; and Lau, R. W. 2020. Progressive mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3697–3705.

Liu, N.; Han, J.; and Yang, M.-H. 2018. PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, N.; Zhang, N.; and Han, J. 2020. Learning Selective Self-Mutual Attention for RGB-D Saliency Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, N.; Zhang, N.; Wan, K.; Shao, L.; and Han, J. 2021a. Visual Saliency Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4722–4732.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2): 91–110.

Loy, G.; and Eklundh, J.-O. 2006. Detecting symmetry and symmetric constellations of features. In *European Conference on Computer Vision*, 508–521. Springer.

Mei, H.; Dong, B.; Dong, W.; Peers, P.; Yang, X.; Zhang, Q.; and Wei, X. 2021. Depth-aware mirror segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3044–3053.

Pang, Y.; Zhang, L.; Zhao, X.; and Lu, H. 2020a. Hierarchical dynamic filtering network for rgb-d salient object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 235–252. Springer.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020b. Multi-Scale Interactive Network for Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Seo, A.; Shim, W.; and Cho, M. 2021. Learning To Discover Reflection Symmetry via Polar Matching Convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1285–1294.

Song, H.; Liu, Z.; Du, H.; Sun, G.; Le Meur, O.; and Ren, T. 2017. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *IEEE Transactions on Image Processing*, 26(9): 4204–4216.

Tan, X.; Lin, J.; Xu, K.; Chen, P.; Ma, L.; and Lau, R. W. 2022. Mirror Detection With the Visual Chirality Cue. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tsogkas, S.; and Kokkinos, I. 2012. Learning-based symmetry detection in natural images. In *European Conference on Computer Vision*, 41–54. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; and Hu, Q. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, T.; Borji, A.; Zhang, L.; Zhang, P.; and Lu, H. 2017. A Stagewise Refinement Model for Detecting Salient Objects in Images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Wei, J.; Wang, S.; Wu, Z.; Su, C.; Huang, Q.; and Tian, Q. 2020. Label Decoupling Framework for Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 418–434.

Yang, X.; Mei, H.; Xu, K.; Wei, X.; Yin, B.; and Lau, R. W. 2019. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8809–8818.

Zendel, O.; Honauer, K.; Murschitz, M.; Humenberger, M.; and Fernandez Dominguez, G. 2017. Analyzing computer vision data-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1980–1990.

Zhao, X.; Zhang, L.; Pang, Y.; Lu, H.; and Zhang, L. 2020. A single stream network for robust and real-time RGB-D salient object detection. In *European Conference on Computer Vision*, 646–662. Springer.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; and Zhang, L. 2021. Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6881–6890.