Self-supervised Pre-training for Mirror Detection

Jiaying Lin* Rynson W.H. Lau*
City University of Hong Kong
{jiayinlin5, rynson.lau}@cityu.edu.hk

Abstract

Existing mirror detection methods require supervised ImageNet pre-training to obtain good general-purpose image features. However, supervised ImageNet pre-training focuses on category-level discrimination and may not be suitable for downstream tasks like mirror detection, due to the overfitting upstream tasks (e.g., supervised image classification). We observe that mirror reflection is crucial to how people perceive the presence of mirrors, and such mid-level features can be better transferred from self-supervised pre-trained models. Inspired by this observation, in this paper we aim to improve mirror detection methods by proposing a new self-supervised learning (SSL) pre-training framework for modeling the representation of mirror reflection progressively in the pre-training process. Our framework consists of three pre-training stages at different levels: 1) an image-level pre-training stage to globally incorporate mirror reflection features into the pre-trained model; 2) a patch-level pre-training stage to spatially simulate and learn local mirror reflection from image patches; and 3) a pixel-level pre-training stage to pixel-wisely capture mirror reflection via reconstructing corrupted mirror images based on the relationship between the inside and outside of mirrors. Extensive experiments show that our SSL pre-training framework significantly outperforms previous state-of-the-art CNN-based SSL pre-training frameworks and even outperforms supervised ImageNet pre-training when transferred to the mirror detection task. Code and models are available at https://jiaying.link/iccv2023-sslmirror/

1. Introduction

Mirrors are prevalent in our daily lives. As their appearances are largely determined by their surroundings, they generally lack a consistent appearance, making it difficult to separate them from their surroundings. This may affect many computer vision tasks such as object detection [2], vision-language navigation [1] and depth estimation [29]. It is therefore crucial to design high-performance mirror detection methods to facilitate computer vision applications.

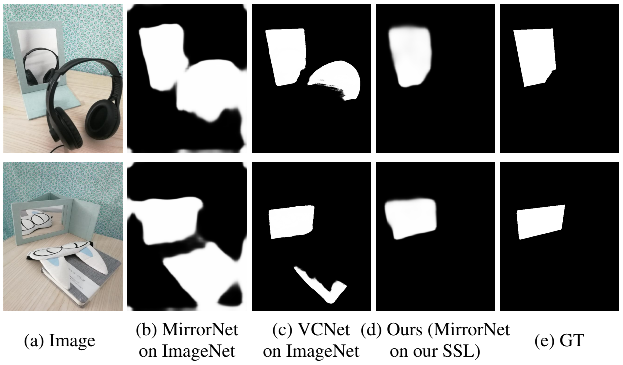

Figure 1. State-of-the-art mirror detection methods [37, 30] are based on costly supervised ImageNet pre-training. They may fail even in obvious cases, *e.g.,* (a) when the mirror clearly reflects a real object outside of the mirror. (b) and (c) are MirrorNet [37] and VCNet [30], respectively, pre-trained on ImageNet with full supervision. (c) is MirrorNet with our proposed SSL framework and without supervised ImageNet pre-training. Our SSL scheme leverages the mirror reflection cue and avoids feature redundancy in the pre-training stage. It outperforms those pre-trained on ImageNet with full supervision (*i.e.*, (b) and (c)).

Recently, a few methods [37, 23, 30] have been proposed for mirror detection, but they all require to be initialized with supervised ImageNet [8] pre-trained weights and then fine-tuned on mirror detection datasets. While such a transfer learning approach is common in general computer vision tasks (*e.g.*, object detection [24] and semantic segmentation [7]) to utilize category-oriented features from large-scale image datasets, we challenge if this strategy is necessary for the mirror detection task for two reasons. First, the supervised ImageNet pre-training process involves high labelling costs, as it is very time-consuming and labor-intensive to label such a large-scale image dataset (∼1.3M images). Second, unlike general vision tasks, mirror detection does not require category-level object understanding. Instead, it requires an understanding of the relationship between inside and outside of mirrors (*e.g.*, contrast [37], similarity [23], and visual chirality [30]). Thus,

[*] Corresponding authors.

while pre-training on ImageNet requires expensive labels, it causes feature redundancy for the mirror detection task.

Figure 1 shows that existing models pre-trained on ImageNet tend to over-detect the mirror regions. Although the two SOTA methods [37, 30] can correctly locate the mirror regions, they are unable to distinguish between the mirrored objects inside the mirrors and the real objects outside, even though these methods contain well-designed modules to learn to address this problem via the fine-tuning stage. This motivates us to investigate whether the current pre-training approach may affect the final detection performance, and whether we may incorporate some intrinsic mirror properties in the pre-training process to assist mirror detection.

An interesting observation is that humans do not need to be "trained" to recognize mirrors. Instead, we learn to recognize them implicitly from young. This aligns with the key idea of self-supervised learning (SSL). Works from neuroscience [31, 20] have shown that humans use mid-level visual cortex to recognize mirror reflection. According to [44], mid-level features are hard to be transferred from supervised pre-training on ImageNet, and SSL is suitable for learning mid-level representations, which inspires us to develop an SSL pre-training framework for modeling mirror reflection.

In this paper, we propose a new SSL pre-training framework for mirror detection, which explicitly considers mirror reflection during the pre-training process. Our framework does not require human-annotated labels from large-scale image datasets. It consists of three stages to progressively pre-train the backbone network from global to local: 1) an *image-level pre-training stage* to obtain the representation of mirror reflection globally by recognizing the geometric transformation applied to the image; 2) a *patch-level pre-training stage* to mimic patch-wise mirror reflection and then learn the spatial correlation between the original object patches and the corresponding mirrored object patches; and 3) a *pixel-level pre-training stage* to extract the pixel-to-pixel relationship of mirror reflection by image reconstruction; Under such progressive pre-training scheme, our SSL pre-training framework can learn the representation of mirror reflection, and then effectively transfer this knowledge to the subsequent mirror detection process for better detection performances. We conduct comprehensive experiments to demonstrate the effectiveness of our SSL pre-training framework. We show that it can significantly boost the performances of existing mirror detection methods, and outperform other CNN-based SSL pre-training frameworks on the mirror detection task.

To conclude, this paper makes three key contributions:

- To the best of our knowledge, we are the first to investigate how existing SSL pre-training frameworks perform on the mirror detection task, compared with supervised ImageNet pre-training.
- We propose a new SSL pre-training framework that consists of three stages at different levels to progressively learn the representation of mirror reflection. Compared with the features from a supervised ImageNet pre-trained model, our representation is better due to the reduced gap between the pre-training task and the target downstream task (*i.e.*, mirror detection).
- Extensive experiments show that our SSL pre-training framework performs the best among all state-of-the-art SSL methods, and even better than models with supervised ImageNet pre-training.

2. Related Work

2.1. Mirror Detection

Mirror detection is an essential task in computer vision, as mirrors are everywhere these days. It is also a challenging task, since mirrors do not possess a consistent appearance, but reflect those of their surroundings. Other vision tasks such as generic object detection and segmentation would tend to detect and segment objects outside as well as inside of the mirror, providing incorrect information on the mirror region, and a depth prediction task may return the depths of the reflected objects inside the mirror region instead of the depth of the mirror.

To address the mirror detection problem, Yang *et al*. [37] propose the first large-scale mirror detection dataset and the first deep-learning model, MirrorNet, for mirror detection by learning the contextual contrast between the inside and outside of mirrors. Lin *et al*. [23] further propose a new model with similarity and edge learning to progressively detect mirrors. Most recently, Guan *et al*. [15] employ graph convolutional networks to model contextual associations for mirror detection, while Tan *et al*. propose VCNet [30] to exploit the visual chirality cue at a feature level.

However, existing mirror detection methods are all fully-supervised. They also directly adopt the supervised ImageNet [8] pre-training to obtain generic image features. Unlike recent works [19, 18, 22] that propose novel network architectures for mirror detection, we propose in this paper to tackle the mirror detection problem from a learning-based perspective, offering the first self-supervised pre-training framework for mirror detection.

2.2. Self-supervised Learning (SSL)

SSL is a popular research topic in computer vision [38, 39], and has attracted a lot of research interests. It aims to learn good feature representations without needing labeled data. Early SSL methods focus on proposing novel pretext tasks, such as relative location prediction [10] and rotation prediction [21]. Recently, most methods are based on the utilization of contrastive learning [16, 4, 14, 3, 32, 5, 41,

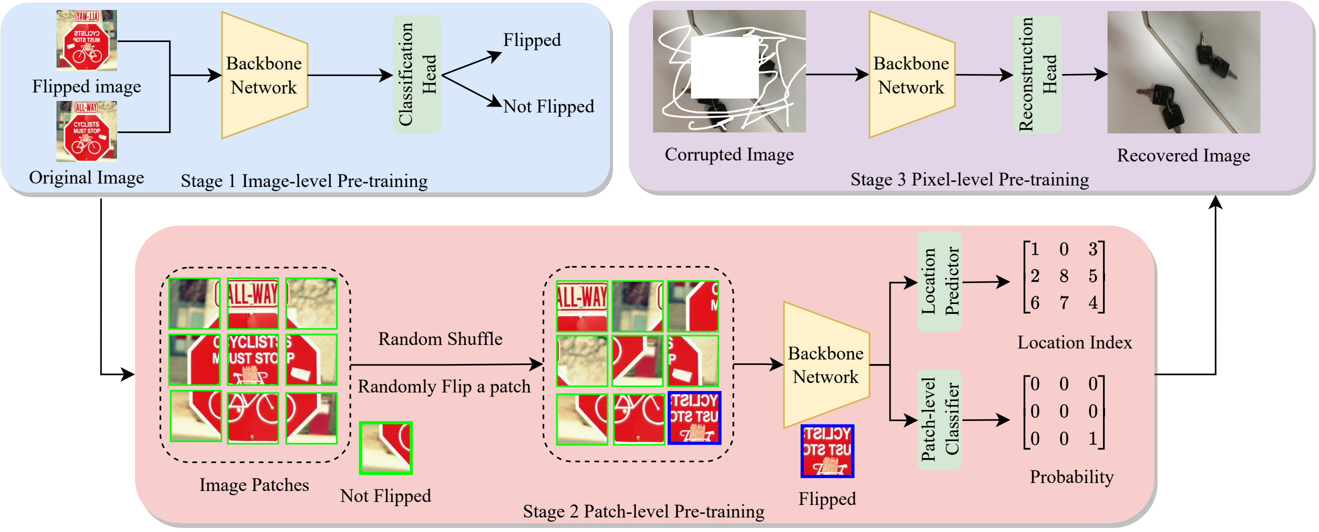

Figure 2. Our SSL pre-training framework. It consists of three pre-training stages at different levels: an image-level pre-training stage, a patch-level pre-training stage, and a pixel-level pre-training stage. It pre-trains the backbone network from global to local (image-level → patch-level → pixel-level) progressively.

40]. Most recently, [44] shows that some SSL methods perform even better than supervised ImageNet pre-training in downstream tasks, such as object detection [11, 24] and semantic segmentation [7, 34], with the help of instance discrimination during SSL pre-training. Recent transformer-based SSL frameworks [36] also involve masked image modeling to obtain better self-supervised representations.

In this paper, we observe that unlike general vision tasks, mirror detection is not benefited from supervised ImageNet pre-training since it does not require category-level object understanding. Hence, we propose a novel SSL pre-training framework specifically for the mirror detection task. The proposed framework aims to simulate and learn the relationship between mirrors and their surrounding during the pre-training stage.

3. Method

Figure 2 shows the overall architecture of the proposed SSL pre-training framework. Our SSL pre-training framework consists of three pre-training stages at different levels: an image-level pre-training, a patch-level pre-training, and a pixel-level pre-training. The pre-training at each level is conducted sequentially and independently. The progressive and sequential training strategy can help the pre-trained model obtain a global-to-local representation of mirrors.

3.1. Image-level Pre-training

Inspired by the definition of visual chirality [26] and its potential applications [46, 30] in various computer vision problems, we formulate the image-level pre-training as a binary classification problem: Given an input image and its corresponding horizontal flipped image, can the backbone network tell which one is the original image? To achieve this objective, we attach a classification head to predict the probability that the input image is flipped. The classification head consists of a convolution layer followed by a sigmoid activation function to project the output value into $[0, 1]$. In this pre-training stage, we use binary cross-entropy (BCE) as our loss function $\mathcal{L}_f$ to optimize the backbone network. Formally, we have:

$$\mathcal{L}_f = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})], \tag{1}$$

where y is the ground truth label and $\hat{y}$ is the predicted probability that the input image is horizontally flipped.

3.2. Patch-level Pre-training

The image-level pre-training pretext task alone cannot bring much help in locating the mirror inside an image since mirror detection is a pixel-level problem. However, directly connecting the image-level and the pixel-level pretext tasks is challenging due to the gap between the scales of these two pretext tasks. Thus, we design a patch-level pre-training stage as a bridge to cover the feature representations of mirrors at different scales. Besides, the image-level pre-training does not benefit relational spatial understanding, which has been proven important in mirror detection [23]. Our patch-level pre-training consists of two sub-tasks to both obtain the patch-wise relationship of mirror reflection and simulate the mirrored region with pseudo labels. We first split the input image into a grid of 3×3 image patches and randomly shuffle them. After that, we randomly apply a flipping operation to one of the image patches, to simulate a mirrored region. The first sub-task is to predict the location index from the shuffled grid by the backbone network. We append a location predictor head to the backbone network to predict the location index of input patches. We apply cross-entropy loss as $\mathcal{L}_{loc}$ to optimize this sub-task. The second sub-task is to predict the flipped patch from the nine randomly shuffled image patches. Similar to our image-level pre-training, we use BCE loss as our loss function $\mathcal{L}_{cls}$ in this sub-task.

$$\mathcal{L}_{loc} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y_i}), \tag{2}$$

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{j=1}^{N} y_j \log(\hat{y_j}) + (1 - y_j)\log(1 - \hat{y_j}), \tag{3}$$

where $\hat{y_i}$ is the predicted location index, and y_i is the corresponding ground truth label. $\hat{y}_j$ is the predicted probability that the input patch is horizontally flipped and y_j is the corresponding ground truth probability. Finally, we jointly optimize these two sub-tasks and the final loss function is:

$$\mathcal{L}_{patch} = \mathcal{L}_{loc} + \mathcal{L}_{cls}. \tag{4}$$

3.3. Pixel-level Pre-training

Extracting the pixel-level relationship of mirror reflection is a challenging task, as studied by previous related work [12] for symmetry detection. It requires point-level GT labels to represent symmetry centers. However, obtaining such labels requires huge human annotation effort and they are not applicable for SSL pre-training. Instead of generating pseudo labels for symmetry detection, our solution is that we formulate this task as an image reconstruction problem on images containing mirrors. An ideal pre-trained network should be able to reconstruct the input image containing mirrors with the pixel-level understanding of mirror reflection. We randomly mask out the input image by cutout augmentation [9] with the ratio of 0.3 and then feed it to the backbone network with a reconstruction head. The details of the reconstruction head we used are listed in Table 1. The objective of our pixel-level pre-training is defined as:

$$\mathcal{L}_{pix} = \|x - \hat{x}\|_1, \tag{5}$$

where $\mathcal{L}_{pix}$ is an L1 loss. x and $\hat{x}$ are the input image and the reconstructed image, respectively.

4. Experiments

4.1. Experimental Settings

Table 1. The architecture of the reconstruction head used in the pixel-level pre-training. Note that each "conv-IR" corresponds to a sequence of convolution layer, InstanceNorm layer and ReLU activation. K, S and P denote the number of kernels, the number of strides and the padding size, respectively, used in the convolution layer.

Layer Details
3×3 conv-IR, $K = 1024$, $S = 1$, $P = 1$
2× Upsample
3×3 conv-IR, $K = 512$, $S = 1$, $P = 1$
2× Upsample
3×3 conv-IR, $K = 256$, $S = 1$, $P = 1$
2× Upsample
3×3 conv-IR, $K = 128$, $S = 1$, $P = 1$
2× Upsample
3×3 conv-IR, $K = 128$, $S = 1$, $P = 1$
2× Upsample
1×1 conv-IR, $K = 3$, $S = 1$, $P = 1$

Dataset and Evaluation Metrics. The image-level and patch-level pre-training experiments are conducted on the training set of ImageNet-100 [28] and MS COCO [25], which contains 130K and ∼118k images respectively. [1] In the pixel-level pre-training and fine-tuning stage, our experiments are conducted on MSD [37], the first large-scale dataset for mirror detection. MSD [37] provides 4,018 mirror images, which are divided into 3,063 images for training and 955 images for testing. We adopt two popular metrics namely, F-measure (F$_\beta$) and mean absolute error (MAE) to evaluate the performance of our SSL pre-training framework. F-measure can reflect the overall model performance and is calculated by the weighted harmonic mean of the precision and recall: $F_\beta = \frac{(1+\beta^2)\cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$, where β^2 is set to 0.3. MAE can measure the average pixel-wise absolute disparity between the predicted mirror maps M and the binary ground truth G.

Implementation Details. We use Pytorch to implement our proposed SSL pre-training framework. All experiments are conducted on a GPU server with eight NVIDIA RTX 3090 GPUs. Following existing SSL pre-training frameworks, we adopt ResNet-50 [17] as our backbone network. All pre-trained model weights of existing SSL pre-training frameworks are obtained from the released models of `mmselfsup` [6] for fair comparisons. For each pretext task in the pre-training stage, all images are resized to 256×256. The batch size is set to 352. We use AdamW [27] with an initial learning rate of 1e-4 to optimize our SSL pre-training framework and stop pre-training after 20,000 iterations. We only adopt random color jittering for data augmentation. In the fine-tuning stage, we use the same hyperparameters reported in the original papers of the adopted

[1] It is worth noting that the datasets that we used (about 250K training images) are much smaller than the pre-trained dataset ImageNet (14 million images) used by other baseline SSL methods. We adopted ImageNet-100 to pre-train our network due to the limited computational resources.

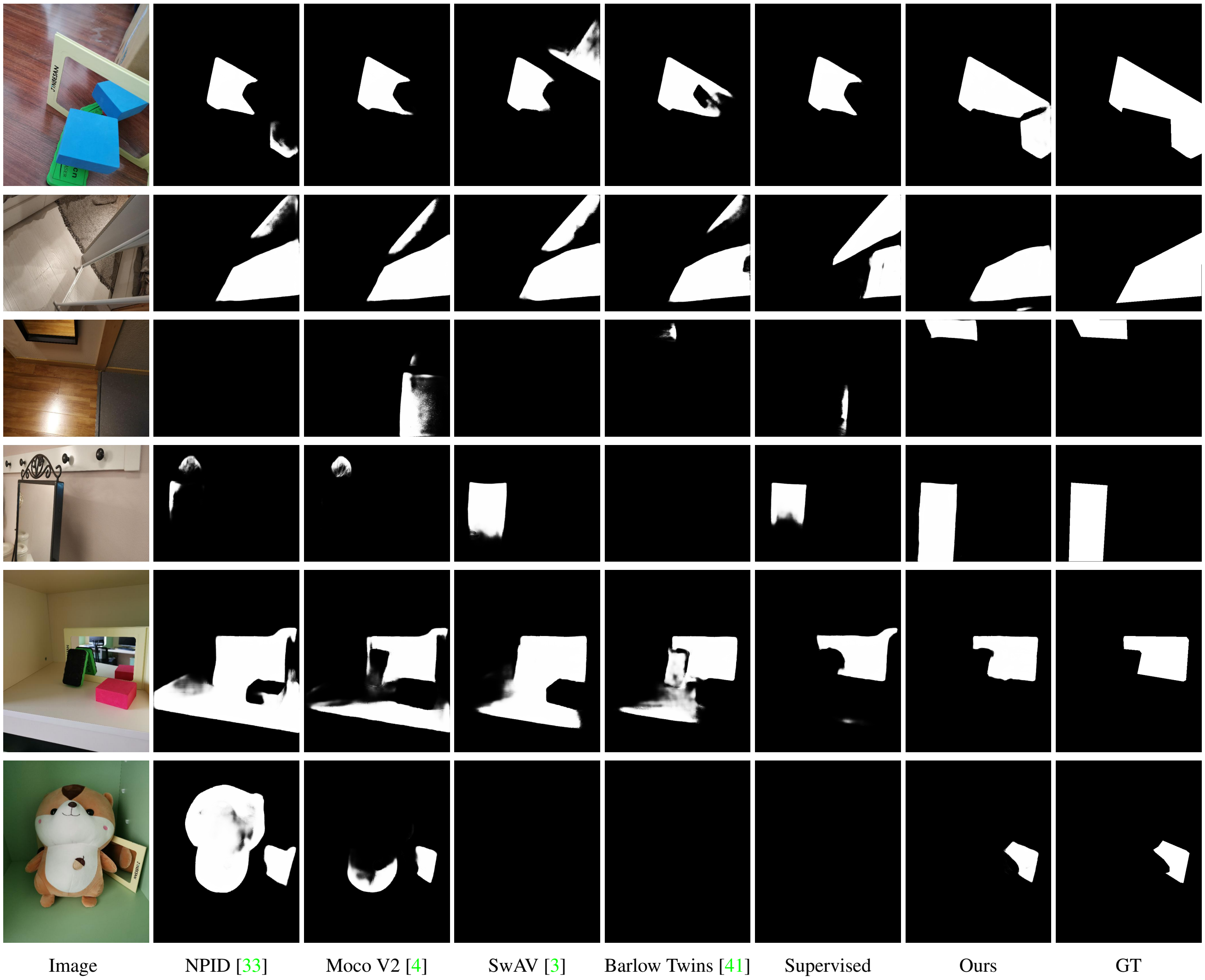

Figure 3. Qualitative Comparison of our proposed method with NPID [33], MocoV2 [4], SwAV [3], Barlow Twins [41], ImageNet pre-training (Supervised). We use the released pre-trained models of these existing SSL methods from `mmselfsup` [6] for fair comparison.

base methods MirrorNet [37] and VCNet [30].

4.2. Comparison with State-of-the-Arts

We compare against 11 state-of-the-art CNN-based frameworks in self-supervised learning, including Relative Location (Rel. Loc.) [10], Rotation Prediction (Rot. Pred.) [21], NPID [33], ODC [43], Moco V2 [16], SimCLR [4], BYOL [14], SwAV [3], DenseCL [32], SimSiam [5] and Barlow Twins (Bar. Twins) [41]. We select MirrorNet [37] and VCNet [30], which are two representative mirror detection methods, as base methods to conduct experiments in the fine-tuning stage.

Qualitative Evaluation. We further show the performance of our SSL pre-training framework in Figure 3. We compare our framework with four state-of-the-art SSL pre-training frameworks (including the two best-performing frameworks NPID [33] and Moco V2 [16] based on MirrorNet and the two best-performing frameworks SwAV [3] and Barlow Twins [41] based on VCNet, according to their performance in Table 2) and supervised ImageNet pre-training due to the limitation of space. Our framework can correctly detect the mirror regions by utilizing the relationship of mirror reflection caused by mirrors, even though the reflected regions are pure texture (*i.e.*, lack of object keypoints in the second row) and relatively small in the image (*e.g.*, the last three rows). We attribute the superior performance of our framework to the utilization of mirror reflection in the pre-training stage.

Quantitative Evaluation. Table 2 shows the experimentation results comparing the proposed SSL pre-training framework with the state-of-the-arts. Besides, we also re-

Table 2. Quantitative comparison between our SSL pre-training framework and existing CNN-based SSL methods. We also compare it with random initialization (Random) and ImageNet pre-training (Supervised).

Methods	Venue	MirrorNet		VCNet	
		$F_\beta \uparrow$	MAE↓	$F_\beta \uparrow$	MAE↓
Rel. Loc.	ICCV '15	0.582	0.254	0.842	0.072
Rot. Pred.	ICLR '18	0.537	0.212	0.851	0.069
NPID	CVPR '18	0.739	0.132	0.832	0.076
ODC	CVPR '20	0.558	0.223	0.836	0.077
Moco V2	CVPR '20	0.707	0.177	0.835	0.072
SimCLR	ICML '20	0.594	0.244	0.837	0.071
BYOL	NeurIPS '20	0.704	0.158	0.832	0.077
SwAV	NeurIPS '20	0.578	0.224	0.856	0.071
DenseCL	CVPR '21	0.694	0.240	0.844	0.072
SimSiam	CVPR '21	0.672	0.198	0.818	0.080
Bar. Twins	ICML '21	0.655	0.178	0.857	0.069
Random		0.586	0.200	0.400	0.103
Supervised [2]		0.727	0.170	0.871	0.062
Ours		**0.763**	**0.116**	**0.886**	**0.057**

port the results from models adopting Xavier initialization [13] (Random) or supervised ImageNet pre-training (Supervised). Our SSL pre-training framework (Ours) achieves the best performance across both metrics when applied on MirrorNet and VCNet and shows a good generalization of different mirror detection methods. It is worth noting that our framework even outperforms the supervised ImageNet pre-training.

4.3. Ablation Study

Table 3 demonstrates the effectiveness of each pre-training stage in our framework. We analyze its results point by point as follows.

Image-level Pre-training *vs.* Random Initialization. To evaluate the effectiveness of the proposed image-level pre-training, we compare the model only adopting image-level pre-training (B2) with the model initialized randomly without any pre-training (B1). We can see that only adopting image-level pre-training (B2) will even cause a performance drop compared with the one without pre-training (B1). This somehow supports the finding from VCNet [30] that image-level flipping cannot provide sufficient information to pixel-wisely locate the mirror inside an image.

Table 3. Ablation study of the proposed SSL pre-training framework. We use MirrorNet [37] trained on the MSD dataset as our base method to conduct the experiments in this ablation study. Best results are shown in bold.

	Rand. Init.	Image-lev. Pretrain	Patch-lev. Pretrain	Pixel-lev. Pretrain	$F_\beta \uparrow$	MAE↓
B1	✓				0.601	0.195
B2		✓			0.623	0.191
B3			✓		0.689	0.162
B4				✓	0.620	0.198
B5		✓	✓		0.723	0.166
B6			✓	✓	0.758	0.124
B7		✓		✓	0.666	0.178
Ours		✓	✓	✓	**0.763**	**0.116**

Patch-level Pre-training *vs.* Random Initialization. We compare our patch-level pre-training (B3) with random initialization (B1) to analyze if patch-level pre-training is useful in our SSL pre-training framework. Our patch-level pre-training outperforms the model without pre-training on both two metrics (14.64% improvement on F_β and 16.32% improvement on MAE), which indicates the effectiveness of our patch-level pre-training. It also demonstrates the importance of modeling the spatial relationship of mirror reflection at a patch level.

Pixel-level Pre-training *vs.* Random Initialization. Similarly, we conduct the ablation experiment on our pixel-level pre-training. The model adopting pixel-level pre-training performs better than the model with image pre-training. However, the performance gain is not as much as the model with patch-level pre-training (B3). One possible reason is that directly optimizing the backbone network for the representation of mirrors at a pixel level is challenging and may not produce satisfactory results for mirror detection.

Location Predictor in Patch-level Pre-training. The location predictor in our patch-level pre-training is designed to capture the relational spatial information, which has been proven important in mirror detection [23]. We conduct an ablation experiment to remove the location predictor and the performance of our methods drops from 0.763 to 0.731. This shows the importance of learning spatial information in the mirror detection task.

Combinations of Different Pre-training Stages. Based on the above three comparisons, we find that the patch-level pre-training contributes the most improvement. It demonstrates the importance of modeling the spatial relationship of mirror reflection at a patch level. We then analyze the results from the model with different combinations of our pre-training stages. The model with image-level pre-training and patch-level pre-training (B5) performs better than the model that only adopts patch-level pre-training (B3), which shows that the global information provided by image-level

[2] Note that the original backbone networks reported in their papers are ResNext-101 [35]. However, the model weights of ResNext-101 pre-trained by our baseline SSL frameworks are not available in `mmselfsup` [6]. Thus, we switch the backbone network of these two methods to ResNet-50 [17], which is a common backbone network used in evaluating the performances of SSL frameworks. We also directly use the released ResNet-50 SSL pre-trained weights in our experiments for fair comparisons.

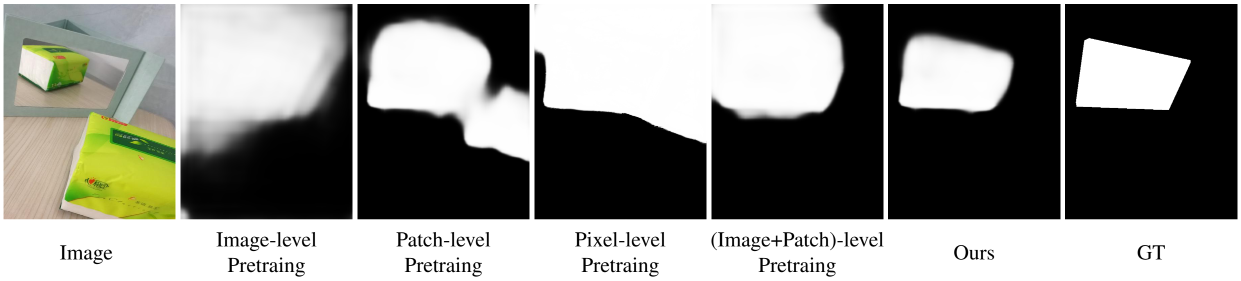

Figure 4. Visualization of the effects of the proposed pre-training stages.

pre-training can benefit the model with patch-level pre-training. Similarly, as shown in B7, B4 (i.e., only pixel-level pretraining) can boost the performance of the pre-trained model after image-level pre-training. These two comparisons indicate that image-level pre-training should be used with other pre-training stages to benefit the overall performance and it should not be solely adopted to our SSL pre-training framework. Besides, we find that combining patch-level and pixel-level pre-training (B6) has superior performance when compared with B3, especially on MAE with about 23.4% improvement. This combination also outperforms supervised ImageNet pre-training, according to the results from Table 2. We attribute its superior performance to the pixel-level representation of mirror reflection. Our final model with all pre-training stages outperforms other ablation models, showing the effectiveness of adopting pre-training stages progressively.

Visual Comparison of Our Pre-training Stages. Figure 4 shows a visual example of our ablation study. We can see that only adopting image-level pre-training on our SSL pre-training framework can approximately locate the mirror region but fails to produce a precise boundary for the predicted mask of the mirror. Patch-level pre-training can help predict mirrors more accurately with clearer boundaries but over-predict the non-mirror regions as mirrors. In comparison to predictions from all other ablation models, the prediction from the model with pixel-level pre-training has the clearest boundary but contains the largest over-predicted regions, primarily because of the insufficient global-level learning on mirrors. Our framework can significantly reduce over-predictions when image-level and patch-level pre-training are used, as opposed to when patch-level pre-training is used only, and performs the best with the help of the global-to-local pre-training process.

4.4. Discussions

Impacts of Pre-training Strategy. Some previous SSL works [42, 45], especially for those focusing on self-supervised pre-training for downstream tasks like segmentation [42] and salient object detection [45], would adopt different pretext tasks sequentially to pre-train their backbone networks. We also adopt this common strategy in our proposed SSL framework. To verify if employing different pre-training strategies may affect the performance of the proposed SSL framework, we have tried training the three stages sequentially as well as in parallel. We find that when pre-training the three stages in parallel, the backbone network fails to converge and its final fine-tuning results are much worse than those using sequential pre-training. Table 4 shows the results of this experiment. We can see that the MAE performance of pre-training the different stages in parallel is even worse than that of random initialization. A possible reason for this is that directly pre-training different pretext tasks in parallel is difficult to transfer the learned knowledge from the pre-trained network to the target downstream task (*i.e.*, pixel-level mirror detection).

Table 4. The results of MirrorNet [37] when using different pre-training strategies. Best results are shown in bold.

	$F_\beta \uparrow$	MAE$\downarrow$
Random Initialized	0.586	0.200
ImageNet Supervised	0.727	0.170
Ours (In parallel)	0.708	0.331
Ours (Sequential)	**0.763**	**0.116**

Apart from the superior performance made by the sequential pre-training strategy, we also notice that the number of pre-training iterations for each pretext task can heavily affect the performance of our SSL framework. To avoid heavy hyperparameter tuning and increase the reproducibility of our pre-training framework, we adopt the **same** number of pre-training iterations for all stages. Although this strategy of using the same number of pre-training iterations is likely not the optimal design for individual stages, we believe that it can better reflect the real performance of our framework with less stochasticity and more reproductivity. We provide a more detailed discussion on this issue next.

Impacts of Pre-training Iterations. SSL pre-training frameworks usually require huge pre-training iterations to

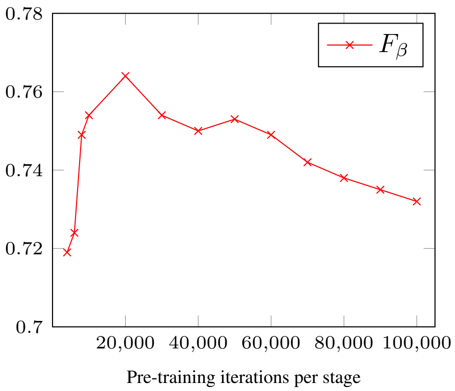

Figure 5. Ablation study on the number of pre-training iterations for our SSL pre-training framework. We find that our SSL pre-training framework continues to improve in performance until ∼20,000 pre-training iterations. After that, the performance starts to decline, mainly due to overfitting.

converge. For example, SimSiam [5] requires about 250K iterations to pre-train their model on ImageNet, which is very time-consuming. To validate how this issue may affect our SSL pre-training framework, we conduct the ablation study on the number of pre-training iterations for our SSL pre-training framework, and the results are shown in Figure 5. Note that we use the same number of iterations for all pre-training stages for convenience and to get rid of heavy hyperparameter tuning.

From the results, we find that the performance of our SSL pre-training framework improves rapidly as the number of pre-training iterations increases, up to around 20,000, which is a relatively small number compared with those used in existing SSL pre-training frameworks. The performance of our SSL pre-training framework gradually drops as the number of pre-training iterations exceeds 20,000, mainly due to overfitting. One possible explanation of this is that the model pre-trained by our SSL pre-training framework can efficiently learn the representation of mirror reflection under a short pre-training period. Another possible explanation is that unlike generic SSL pre-training frameworks that attempt to extract general-propose image representation, our SSL pre-training framework is developed specially for a specific task, *i.e.*, mirror detection. While generic SSL pre-training may cause feature redundancy for our task, having too many pre-training iterations (*i.e.*, over 20,000) causes overfitting.

Failure Cases. Our SSL pre-training framework does have limitations. It continues to struggle with the limitations of the base method [37], as the base method adopted can also significantly affect the final detection performances. For example, as shown in first row of Figure 6, MirrorNet with our SSL pre-training may still fail to handle some mirror-like regions (*e.g.*, the top-right region). It may also fail to detect mirrors in some challenging cases when the input image contains ambiguous mirrors without the relationship of mirror reflection and sufficient contextual contrast, as shown in the second row of Figure 6.

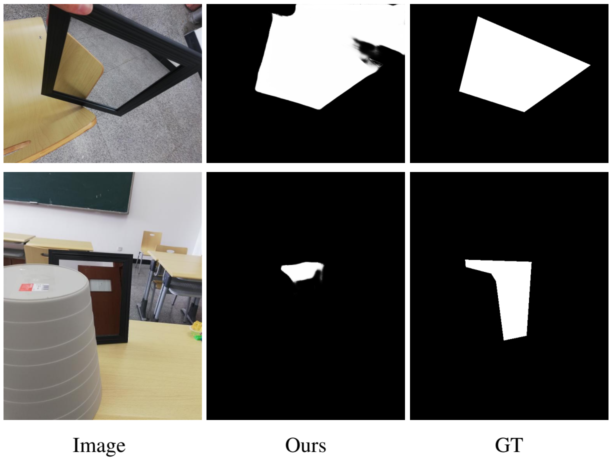

Figure 6. Failure cases of our SSL pre-training framework based on MirrorNet. Our SSL pre-training framework fails to tackle the intrinsic limitation of existing mirror detection methods. The model pre-trained by our SSL pre-training framework still has wrong predictions when the input images are challenging (*e.g.*, containing mirror-like regions in the first row, and insufficient relationship between mirror/non-mirror regions in the second row.

5. Conclusion

In this paper, we have investigated how self-supervised learning (SSL) works with the mirror detection task. To the best of our knowledge, we are the first to explore SSL pre-training frameworks applied to mirror detection. We have found that the supervised ImageNet pre-training might not be the ideal way to extract backbone image features for mirror detection due to the discrepancy between general-propose representation and mirror-specific representation. We have also proposed a new SSL pre-training framework to pre-train the backbone network for mirror detection. Our SSL pre-training framework does not require any labeled data. It progressively (image-, patch- and pixel-level pre-training) models the relationship of mirror reflection, and obtains a better representation of mirrors in the pre-training stage. Experimental results show that the proposed SSL pre-training framework outperforms CNN-based state-of-the-art SSL pre-training frameworks, and even achieves better results compared with supervised ImageNet pre-training.

Acknowledgments: The work described in this paper was partially supported by a GRF grant from the Research Grants Council of Hong Kong (Project No.: CityU 11211223).

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pages 3674–3683, 2018.

[2] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE TPAMI*, 41(8):1844–1861, Aug 2019.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[6] MMSelfSup Contributors. MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark. https://github.com/open-mmlab/mmselfsup, 2021.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] Christopher Funk and Yanxi Liu. Beyond planar symmetry: Modeling human perception of reflection and rotation symmetries in the wild. In *ICCV*, pages 793–803, 2017.

[13] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

[15] Huankang Guan, Jiaying Lin, and Rynson W.H. Lau. Learning semantic associations for mirror detection. In *CVPR*, pages 5941–5950, June 2022.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] Ruozhen He, Jiaying Lin, and Rynson WH Lau. Efficient mirror detection via multi-level heterogeneous learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 790–798, 2023.

[19] Tianyu Huang, Bowen Dong, Jiaying Lin, Xiaohui Liu, Rynson WH Lau, and Wangmeng Zuo. Symmetry-aware transformer-based mirror detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 935–943, 2023.

[20] Peter Kohler, Alexandra Yakovleva, Alasdair Clarke, Yanxi Liu, and Anthony Norcia. Parametric responses to rotation symmetry in mid-level visual cortex. *Journal of Vision*, 15(12):1122–1122, 2015.

[21] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[22] Jiaying Lin, Xin Tan, and Rynson WH Lau. Learning to detect mirrors from videos via dual correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9109–9118, 2023.

[23] Jiaying Lin, Guodong Wang, and Rynson W. H. Lau. Progressive mirror detection. In *CVPR*, 2020.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[26] Zhiqiu Lin, Jin Sun, Abe Davis, and Noah Snavely. Visual chirality. In *CVPR*, pages 12295–12303, 2020.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[29] Jiaqi Tan, Weijie Lin, Angel X Chang, and Manolis Savva. Mirror3d: Depth refinement for mirror surfaces. In *CVPR*, pages 15990–15999, 2021.

[30] Xin Tan, Jiaying Lin, Ke Xu, Chen Pan, Lizhuang Ma, and Rynson W. H. Lau. Mirror detection with the visual chirality cue. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[31] Christopher W Tyler. *Human symmetry perception and its computational analysis*. Psychology Press, 2003.

[32] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.

[33] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[34] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018.

[35] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017.

[36] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022.

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

[38] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. Learning with noisy labels for robust point cloud segmentation. In *ICCV*, pages 6443–6452, 2021.

[39] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. Robust point cloud segmentation with noisy annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7696–7710, 2022.

[40] Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, and Jing Liao. Improving commonsense in vision-language models via knowledge graph riddles. In *CVPR*, pages 2634–2645, 2023.

[41] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

[42] Xiaohang Zhan, Ziwei Liu, Ping Luo, Xiaoou Tang, and Chen Loy. Mix-and-match tuning for self-supervised semantic segmentation. In *AAAI*, volume 32, 2018.

[43] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *CVPR*, 2020.

[44] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? In *ICLR*, 2021.

[45] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, , Huchuan Lu, and Xiang Ruan. Self-supervised pretraining for rgb-d salient object detection. In *AAAI*, 2022.

[46] Ying Zheng, Yiyi Zhang, Xiaogang Xu, Jun Wang, and Hongxun Yao. Visual chirality meets freehand sketches. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1544–1548. IEEE, 2021.