Make Segment Anything Model Perfect on Shadow Detection

Xiao-Diao Chen, Wen Wu, Wenya Yang, Hongshuai Qin, Xiantao Wu, and Xiaoyang Mao, Member, IEEE

Abstract—Compared to models pretrained on ImageNet, the segment anything model (SAM) has been trained on a massive segmentation corpus, excelling in both generalization ability and boundary localization. However, these strengths are still insufficient to enhance shadow detection without additional training, and it raises the question: do we still need precise manual annotations to fine-tune SAM for high detection accuracy? This article proposes an annotation-free framework for deep unsupervised shadow detection (USD) by leveraging SAM’s capabilities. The key lies in how to exploit the abilities acquired from a large-scale corpus and utilize them to improve downstream tasks. Instead of directly fine-tuning SAM, we propose a prompt-like tuning method to inject task-specific cues into SAM in a lightweight manner, namely, ShadowSAM. This adaptation manner can ensure a good fitting when training data are limited. Moreover, considering that the pseudo labels used in our framework are generated by traditional USD approaches and may contain severe label noises, we propose an illumination and texture-guided updating (ITU) strategy to selectively boost the quality of pseudo masks. To further improve the model’s robustness, we design a mask diversity index (MDI) to establish easy-to-hard subsets for incremental curriculum learning. Extensive experiments on benchmark datasets (i.e., SBU, UCF, ISTD, and CUHK-Shadow) demonstrate that our unsupervised solution can achieve comparable performance to state-of-the-art (SOTA) fully supervised methods. Our code is available at this repository.

Index Terms—Curriculum learning, noisy label, segment anything model (SAM), shadow detection, unsupervised learning.

I. INTRODUCTION

SHADOWS are natural phenomena produced when objects obstruct light sources. Identifying shadow locations can provide valuable cues for scene understanding, including geometry [1], camera parameters [2], and light source [3]. Furthermore, detecting [4], [5], [6] and removing shadows [7], [8], [9], [10] from remote sensing images [11], [12] can greatly improve the robustness of change detection [13], semantic segmentation [14], and building extraction [15]. However, deep neural networks typically require precise and dense labeling for high performance. To address this challenge, several works have explored semisupervised [16], [17], [18] or weakly supervised [19] solutions. Pursuing coarser supervision signals has emerged as an active research direction, as it can reduce labeling costs and enable scalable models. Although some unsupervised learning methods [20], [21] for shadow detection have been proposed to tackle the cross-domain problem, most of them still depend on precise annotation, i.e., ground truths (GTs), on the source domain to attain unsupervised generalization on the target domain. This means that they are, in effect, semisupervised. Until now, no genuine deep unsupervised shadow detection (USD) framework has been introduced.

Recently, Meta AI Research released the segment anything model (SAM) [22],1 which was trained on over 11 million images and 1.1 billion segmentation masks. SAM is a Vision Transformer (ViT)-based model, consisting of three parts: an image encoder, a prompt encoder, and a lightweight mask decoder. They claim that SAM can segment any objects by providing a few prompts with three models (i.e., hover and click, box, or everything), known as the zero-shot transfer. However, similar to other vision foundation models (e.g., CLIP [23]), SAM has obvious limitations due to its training data not encompassing the entire corpus of the computer vision community [24]. As shown in Fig. 1, SAM struggles to

Fig. 1. Visualization results of SAM in different shadow scenes. (a) Inputs. (b) Hover and click. (c) Box. (d) Everything.

Authorized licensed use limited to: HANGZHOU DIANZI UNIVERSITY. Downloaded on November 29,2023 at 02:15:56 UTC from IEEE Xplore. Restrictions apply.
identify shadow regions when they coexist with other semantic objects, such as airplanes, trees, or cars. Undoubtedly, SAM excels in generalizability and boundary localization. However, directly using it for specific downstream tasks cannot always obtain satisfactory performance. Therefore, further fine-tuning is necessary.

Existing paradigms for tuning pretrained models in the computer vision community can be mainly categorized into two branches: updating all parameters in the backbone (i.e., full tuning) or only a subset of parameters, such as the classification parts (classifier head [25], [26] or decoder [27]), bias terms [28], or residual blocks [29], [30] (i.e., partial tuning). The former often struggles to fit small datasets well, particularly tuning the model with a large number of parameters. At the same time, despite the reduced dependencies on the scale of training data in the latter, they may underperform full tuning as the frozen parameters in the feature extraction stage (i.e., encoder). Recently, prompt tuning [31], [32] has been a popular strategy in the natural language processing community. It inserts task-specific vectors into each input that can be tuned for different downstream tasks. This approach can significantly reduce training costs by providing more accurate guidance and helping the model learn the rules of the downstream tasks more quickly. In computer vision, however, this technique only works when the tuning tasks are very close to pretraining task [33], e.g., salient object detection and image recognition.

Consider that SAM is essentially an interaction model where the prompt encoder is used to receive prompts from the user. If we fine-tune it to a downstream task in a supervised learning manner, the prompt encoder would no longer be useful and should be removed. Therefore, we attempt to introduce an additional network to replace the prompt encoder. Specifically, we propose a prompt-like tuning method to maintain SAM’s internal knowledge and draw task-specific cues from the shadow dataset by injecting a set of two-layered multilayer perceptron (MLP) blocks, namely, ShadowSAM. Rather than adjusting the inputs to accommodate the pretrained model in prompt tuning, our tuning method aims to fill the gap between the pretrained task and the downstream task from the perspective of feature extraction. Our approach is more suitable for computer vision tasks since visual corpus is not as easy to cover downstream tasks as language corpus. Noting that, our proposed ShadowSAM only introduces about 3.7% parameter into the SAM. This lightweight design allows us to fit well with limited training samples.

Unlike models pretrained on ImageNet [34], used for image-level classification tasks, focusing more on faster convergence and a little bit of accuracy improvement, SAM stands out with its exceptional generalization ability and boundary localization due to its training on high-quality segmentation masks. Noting that these strengths can reduce the model’s reliance on precise annotations, rather than directly training with time-consuming GTs, we opt to use inaccurate pseudo masks to fine-tune our ShadowSAM, which will potentially pave the way for an annotation-free framework for USD. For unsupervised learning, recent progress [35] in related vision task, salient object detection, transforms the class activation maps (CAMs) into high-quality pseudo masks for unsupervised learning. However, these maps tend to capture salient cues in images rather than shadow information. Consequently, we consider utilizing traditional USD approaches to generate pseudo masks. These methods identify shadow via various priors, including texture [36], color [37], gradient [38], illumination [39], and contour [40].

Considering that the quality of these pseudo masks is significantly lower than manual annotations, training directly with them would lead to potential performance degradation [41]. Inspired by self-training [42], we try to enhance the robustness of models by replacing the initial pseudo masks with predicted shadow maps. However, recent studies [43], [44], [45] on naive self-training tend to update all training targets simultaneously, which may introduce severe confirmation bias and performance degradation [46] in some challenging samples. Noting that the initial pseudo masks may exhibit superior quality due to incorporating various shadow priors, we propose an illumination and texture-guided updating (ITU) strategy for selectively refining the pseudo labels. By leveraging the illumination and texture between the initial pseudo masks and deep shadow maps, we preserve high-quality initial pseudo masks until the model generates improved pseudo labels.

Furthermore, we observed that traditional USD approaches exhibit high consistency in simple shadow scenarios but more unreliable and distinguishing predictions in complex scenes. This characteristic can help us to divide the training data into easy-to-hard subsets, facilitating curriculum learning [47]. This learning manner offers an optimized convergence direction for model learning, thus enhancing both the robustness and generalization ability of the model. Instead of simply dividing training data into two parts (e.g., easy and hard), we design a mask diversity index (MDI) to construct multilevel curricula without the need for predefined thresholds [48], [49], [50] or additional training [51]. Combining these two strategies, our unsupervised framework can achieve performance comparable to state-of-the-art (SOTA) fully supervised shadow detectors.

The four main contributions are summarized as follows.
1) A deep USD framework based on the SAM is proposed to completely eliminate the dependence on GTs. By exploiting the SAM’s powerful capability in boundary location, our framework can perform comparably to SOTA fully supervised methods.
2) To ensure a good fine-tuning performance of SAM in shadow detection with a limited number of training data, we propose a prompt-like tuning method to learn task-specific knowledge in a lightweight designing manner.
3) To alleviate the confirmation bias and performance degradation in naive self-training, we propose a selective updating strategy based on the illumination and texture priors to dynamically boost the training targets in the learning process.
4) To better match the diverse complexity in shadow images and thereby fully facilitate curriculum learning, we design an MDI for effectively splitting the training sample into multiple complexity levels.
II. RELATED WORK

Based on the utilization of human annotations, shadow detection techniques can be categorized into unsupervised and supervised ones.

A. Supervised Shadow Detection

Early supervised shadow detectors are mainly proposed by utilizing various handcrafted features extracted from limited training data to train machine-learning classifiers. For example, Huang et al. [52] and Guo et al. [53] utilized edge features and illumination characteristics, respectively, to train support vector machines for shadow detection. Vicente et al. [54] employed Markov random fields to incorporate contextual information between pairwise regions and enhance shadow detection performance. However, these methods demonstrated limited effectiveness in complex scenes, as the inherent limitations of handcrafted features hindered their ability to distinguish shadow regions from the background of the image accurately.

In recent years, significant progress has been made in shadow detection by leveraging large-scale annotated datasets and utilizing deep neural networks. Earlier deep-learning-based methods have demonstrated improved performance by incorporating efficient attention modules, such as direction-aware spatial context (DSC) [55] and recurrent attention residual (RAR) [56], or by employing adversarial learning techniques [57], [58]. Zheng et al. [59] introduced the distraction-aware shadow detection (DSD) network, which incorporates a distraction-aware shadow module to learn discriminative features. Chen et al. [16] proposed the multitask mean teacher (MTMT) network, which leverages unlabeled shadow data to enhance the network’s generalization ability. Hu et al. [60] focused on computational efficiency using lightweight networks and introduced the CUHK-Shadow benchmark dataset, which simulates the complexity of shadows in real-world scenarios. Zhu et al. [61] addressed the intensity bias issue in shadow detection with the feature decomposition and reweighting (FDR) scheme. Recently, Zhu et al. [62] explored complementary mechanisms in shadow detection. Wu et al. [18] introduced an unsupervised approach to generalize to new shadow datasets with minimal annotation costs.

While these deep-learning-based methods have shown impressive performance, they still heavily rely on manual annotations, which are time-consuming and costly. This study aims to combat this issue by introducing a deep USD framework, which uses coarse and inaccurate pseudo masks to replace precise labeling.

B. Unsupervised Shadow Detection

Before the deep-learning boom, USD methods were primarily developed by leveraging various shadow priors or physical models. For example, Tsai [63] proposed a shadow detection method that relies on different invariant color spaces such as HSV, HCV, YIQ, and YCbCr. This method effectively detects shadows by exploiting shadows’ brightness and chromaticity characteristics. Leone and Distanti [36] evaluated the compatibility of photometric properties with shadow characteristics and then improved shadow detection performance by measuring the similarity between little textured patches. Chung et al. [37] introduced a successive thresholding scheme (STS) that detects shadows coarse-to-finely. Instead of using color information, Shoaib et al. [38] presented a novel scheme for the real-time detection of cast shadows using contour-like structures of objects, which are obtained by gradient-based background subtraction. Makarau et al. [39] utilized a blackbody radiation model for shadow detection. Elbakary and Iftekharuddin [40] proposed a modified geometric active-contour model for shadow detection.

Since these methods are designed for different sources of images (e.g., aerial images [37], [63], surveillance environment [36], moving object [38], multispectral image [39], or satellite images [40]), they demonstrate weak generalizability in something else shadow scenes. In this work, we aim to develop a general USD framework by learning more shadow patterns from the large-scale benchmarks.

III. PROPOSED METHODS

Given an unlabeled shadow dataset \( U = \{u_i\}_{i=1}^{N} \), consisting of \( N \) images, this work aims to train a deep shadow detector with pseudo masks \( Y^\text{pred} = \{y_{i,1}, \ldots, y_{i,m}\}_{i=1}^{N} \) obtained from \( m \) traditional USD methods. This work attempts to address two primary challenges: 1) how to effectively tune the SAM with limited training data? and 2) how to ensure a good shadow detection performance from low-quality pseudo masks? As shown in Fig. 2, our deep USD framework consists of three main steps: 1) pseudo mask generation for annotation-free learning (Section III-A); 2) a prompt-like tuning method for injecting task-specific shadow cues into the general knowledge from SAM (Section III-B); and 3) two robust training strategies, including ITU-based selective self-training (Section III-C) and MDI-based incremental curriculum learning (Section III-D).

A. Pseudo Mask Generation

In practice, using the pseudo masks generated by deep shadow detectors can lead to better detection performance since these pseudo masks are closer to the GTs. However, these methods still need accurate GTs for model training. The core idea of this work is to completely eliminate reliance on precise shadow annotations. Therefore, we consider using the traditional USD method to generate pseudo masks at the first step, which needs no GTs and ensures that our framework is entirely unsupervised.

In theory, we can directly utilize the pseudo masks generated by the best-performing traditional USD algorithm to train the deep USD framework. However, since existing traditional methods are designed for different image scenes, including surveillance environment [36], aerial images [37], moving object [38], multispectral image [39], and satellite images [40], they cannot perform well in all types of scenes from shadow benchmarks. Therefore, using either pseudo mask in our framework will result in an obvious confirmation bias. Considering that these detectors are well-designed by different
shadow priors, they can offer complementary information to each other, thereby securing reliable supervisory signals. We take an average of the pseudo masks generated from these five USD methods, followed by a conditional random field (CRF) for edge refinement and noise suppression.

\[ y_i = \text{CRF} \left( u_i, \frac{1}{m} \sum_{t} y_{i,m} \right). \]  

(1)

Here, we can generate a corresponding pseudo mask \( y_i \) for each image \( u_i \) in the unlabeled shadow dataset \( \mathcal{U} \), denoted as \( \mathcal{D} = (\mathcal{U}, \mathcal{Y}) \), where \( \mathcal{Y} = \{ y_i \}_{i=1}^N \).

Discussion: Despite merging different shadow priors and conducting post-refinement by the CRF, the quality of these pseudo masks is still significantly lower than manual annotations. In this connection, directly learning with them on a conventional pretraining model (e.g., ResNeXt [64] or EfficientNet [65] in ImageNet [34]) cannot ensure a good detection performance.

B. ShadowSAM: Prompt-Like Tuning

As shown in Fig. 3(a), SAM consists of three parts: a prompt encoder, an image encoder, and a lightweight mask decoder. As a promptable model, SAM takes an image and prompts as inputs. Then, it extracts image features by the image encoder and encodes the given prompts into a token by the prompt encoder. Finally, they are fed into the mask decoder to generate the predicted mask. In the tuning stage, our goal is to leverage the internal knowledge from SAM and learn additional task-specific information. Note that we do not need the prompt encoder anymore since we cannot provide any prompt for training samples in our unsupervised framework. Undoubtedly, such prompts play a significant role in generalizing downstream tasks. Motivated by this, our idea is to replace it by introducing a series of learnable blocks (i.e., prompters) into the SAM, namely, ShadowSAM. In detail, we first freeze the weights of the image encoder of SAM. Then, we tune the prompters and mask decoder during the training, as shown in Fig. 3(b).

A prompter consisting of only two MLPs and a GELU activation function [66], which emphasizes the simple and efficient design. Formally, such prompters aim to learn task-specific knowledge \( F_i \) and produce the output \( P^i \), which can
be regarded as prompts
\[ P_i = \text{Up}_{\text{share}}(\text{GELU}(\text{Tune}_i(F_i))) \]  \hspace{1cm} (2)

where Tune\(_i\) are linear layers used to shadow patterns from training data, containing 32 linear layers. Up\(_{\text{share}}\) is a shared single linear layer among all prompters used to adjust the dimensions of network features, \( P_i \) will be attached to each transformer layer of the SAM model.

1) Basic USD Framework: At this point, we can train our ShadowSAM on \( D \) by the weighted cross-entropy loss \([61]\)
\[ L_{\text{wce}}(y, \hat{y}) = -\sum_{i=1}^{M} \frac{S}{M} y_i \log \hat{y}_i + \frac{NS}{M} \log(1 - y_i) \]  \hspace{1cm} (3)

where \( i \) represents the pixel index; \( \hat{y} \) is the predicted mask by the network; and \( S, NS, \) and \( M \) are the number of shadow pixels, nonshadow pixels, and total pixels in the image, respectively.

2) Discussion: By inheriting the powerful boundary location from SAM and learning some shadow patterns from pseudo masks, our basic USD framework can help us reduce the dependence on the high-quality training mask, thereby completing the needs of shadow detection. However, the severe label noises in the initial pseudo mask will be overfitted in the latter training phase, leading to a potential performance degradation \([41]\).

C. ITU-Based Selective Self-Training

In self-training \([42]\), deep-learning-based shadow detectors have the potential to capture a wider range of shadow patterns during the training phase. As a result, the shadow maps inferred from these detectors are expected to demonstrate superior qualities compared to the initial pseudo labels. However, regular practice \([43]\), \([44]\), \([45]\) in self-training tends to update all training targets at one time, resulting in a severe confirmation bias \([46]\) in some challenging samples. Unlike them, we found that the initial pseudo masks considered by various shadow priors have better quality than almost early deep shadow maps as the underfitting of the deep model. Therefore, we have the idea to maintain initial pseudo masks until the model can generate better ones.

Formally, for an unlabeled training image \( u_i \), we first view the initial pseudo mask \( y_i \) and the deep shadow map \( \hat{y}_i \) as two candidates at the beginning of training. Then, we design a criterion to choose a better one from them: the more accurate shadow regions should contain lower brightness and weaker texture information. Therefore, we compute the score of illumination and texture for these two candidates as follows:
\[ \text{score}(y) = \alpha \cdot L(y) + (1 - \alpha) \cdot T(y) \]  \hspace{1cm} (4)

where \( \alpha \) is the balance factor, \( L(y) = \text{Mean}(L \odot y) \), \( T(y) = \text{Mean}(T \odot y) \), \( L \) is the L channel in the lab color space, \( T \) is the texture map extracted by the HOG operator, \( \text{Mean}(\#) \) represents the mean operation, and \( \odot \) indicates elementwise multiplication. As shown in Fig. 4, we use deep shadow map \( \hat{y}_i \) to replace initial pseudo mask \( y_i \), since \( \text{score}(\hat{y}_i) < \text{score}(y_i) \).

Fig. 4. Comparing the covered illumination and texture cues for two pseudo masks. (a) Shadow image. (b) Illumination cues. (c) Texture cues. (d) Initial pseudo mask. (e) Deep shadow map.

**ITU-Based Selective Self-Training:** Based on the basic USD framework (Section III-B), we further introduce the illumination and texture-guided pseudo updating for safely selective self-training. In practice, most initial pseudo masks will not be replaced immediately in the early training stage, and some replaced pseudo masks may also be replaced again. In this manner, we can progressively alleviate the impact of noise labels in the “training targets.”

D. MDI-Based Incremental Curriculum Learning

Curriculum learning \([47]\) imitates human learning procedures from simple to hard, providing a better direction for model convergence. This approach, which has been proven, can also boost the robustness and generalizability of deep networks. The crux of this learning manner is effectively splitting the training data from easy to hard. Previous work mainly distinguishes between easy and difficult samples by predicting the reliability or uncertainty of pixel-level samples, such as taking the final softmax output as the confidence distribution and then filtering hard pixel samples by a predefined threshold \([48]\), \([49]\), \([50]\), as well as training two models with different parameter initialization to predict the same image and then identify their disagreements as hard parts \([51]\). Rather than the widely used pixel-level sample selection, we propose to conduct data splitting in an image-level manner, enabling the model to learn more holistic contextual patterns.

Specifically, we found that most traditional USD methods provide accurate and highly consistent predictions in simple shadow scenes. However, in more complex scenarios, the prediction results may contain much unreliable information and exhibit poor consistency. Therefore, we can sort the unlabeled images by prediction consistency. Formally, for an unlabeled image \( u_i \), with the results of \( m \) different traditional USD methods \( \langle y_{i,j} \rangle_{m=1}^{m} \), we introduce an MDI to obtain the consistency value reflecting the level of difficulty
\[ \text{MDI}_i = \sum_{m=1}^{K} \sum_{n=m+1}^{K} \text{meanIoU}(y_{i,m}, y_{i,n}) \]  \hspace{1cm} (5)

where meanIoU denotes the mean intersection over union. When \( \text{MDI}_i = 1 \), this indicates that the \( m \) different
prediction results are entirely consistent, which also means that \( u \) is a simplest sample. Our MDI allows us to split all unlabeled training data without additional training and thresholds. Benefiting from this, instead of only two categories in previous works (i.e., easy and hard), we set \( C \) complexity levels \( \{ D_c \}_{c = 1, \ldots, C} \) for better matching the diverse complexity in shadow image, as shown in Fig. 5.

**Algorithm 1** Training Details of Our Deep USD Framework

```plaintext
U \(_{usd}\) : Unlabeled shadow dataset
\( Y_{usd} \) : Pseudo masks from traditional USD
Input: ShadowSAM : Model with pre-trained parameters
C : Pre-defined number of curriculums
Output: Fine-tuned ShadowSAM
Generating initial pseudo masks: Refining \( Y_{usd} \) to be \( Y \) and creating training set \( D = \{ U, Y \} \) by Eq. 1
Curriculum construction: Splitting \( D \) into \( C \) curriculums \( D_0 = \{ U^0, Y^0 \}_c \) by mask diversity index (MDI) in Eq. 5
for \( c = 1 \rightarrow C \) do
    Training ShadowSAM\(_c\) using \( \{ D_1, \ldots, D_c \} \)
    if \( c < C \) then
        Inferencing on \( \{ U^1, \ldots, U^c \} \) and obtaining deep shadow map \( \{ Y^1, \ldots, Y^c \} \) using ShadowSAM\(_c\).
        for \( i = 1 \rightarrow \text{Number}(D_1, \ldots, D_c) \) do
            if score\(_i\) \( \leq \) score\(_y\) then
                \( y_i \leftarrow \hat{y}_i \) and obtaining updated \( \{ Y^1, \ldots, Y^c \} \)
            end if
        end for
    end if
end for
return ShadowSAM\(_C\).
```

**MDI-Based Incremental Curriculum Learning:** After splitting sample subsets via our MDI, rather than gradually replacing easy samples with harder ones for replacement retraining [Fig. 6(a)], this scheme is prone to result in model overfitting as the limited samples at each curriculum. Inspired by incremental learning [67], which can preserve learned knowledge when the model continually learns from additional datasets, we introduce an incremental curriculum learning to safely exploit shadow patterns inside a dataset from easy to hard [Fig. 6(b)]. Based on the proposed selective self-training (Section III-C), our further applied incremental curriculum learning for better robustness, see more details in Algorithm 1.

**IV. EXPERIMENTS**

In this section, we first describe the benchmark dataset and evaluation metric for shadow detection. Then, we demonstrate how to use the vanilla SAM for shadow detection. Next, we compare our proposed deep USD framework with SOTA supervised- and unsupervised-shadow detectors and then conduct extensive ablation studies for each component in our framework. Finally, we discuss the applications and limitations of our proposed methods.

**A. Datasets and Metrics**

1) **Datasets:** SBU [68],2 UCF [69],3 ISTD [70],4 and CUHK-Shadow [60]5 are four benchmark datasets widely used in shadow detection. The SBU shadow dataset consists of 4089 pairs of shadow images/masks for training and 638 for testing. The UCF dataset, primarily collected for early machine-learning-based algorithms, comprises 145 and 78 training and testing pairs, respectively. The ISTD dataset can be used for both shadow detection and removal, containing 1870 triplets of shadow images, shadow masks, and shadow-free images. It employs 1330 of such triplets for training and 540 for testing. CUHK-Shadow [60] is the largest shadow detection dataset, consisting of four subsets, comprising 7350 training images, 1050 validation images, and 2100 testing images. Following existing deep-learning-based shadow detectors, UCF is only used for testing.

2) **Evaluation Metrics:** We evaluate the performance of shadow detection using the balanced error rate (BER), which is defined as

\[
\text{BER} = \left( 1 - \frac{1}{2} \left( \frac{TP}{S} + \frac{TN}{NS} \right) \right) \times 100
\]

2https://www3.cs.stonybrook.edu/ cvl/projects/shadow_noisy_label
3https://drive.google.com/file/d/12DOmIMvMVe-oNuJXmkBrzkfBvuDd007ON
4https://github.com/DeepInsight-PCALab/ST-CGAN
5https://github.com/xw-hu/CUHK-Shadow

---

**Fig. 5.** Easy to hard sample splitting. (a) Inputs. (b) Leone and Distante [36]. (c) Chung et al. [37]. (d) Shoaib et al. [38]. (e) Makarau et al. [39]. (f) Elbakary and Iftekharuddin [40].
CHEN et al.: MAKE SAM PERFECT ON SHADOW DETECTION

Fig. 7. Visualization results of vanilla SAM for shadow detection. (a) Overview of our Bottom-N strategy for shadow detection. (b) Good cases. (c) Failure cases.

**TABLE I**

<table>
<thead>
<tr>
<th>Method</th>
<th>SBU</th>
<th>UCP</th>
<th>ISTD</th>
<th>CUTIK</th>
<th>Shadow</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDTR+ [18]</td>
<td>2.95</td>
<td>6.35</td>
<td>1.51</td>
<td>7.15</td>
<td></td>
</tr>
<tr>
<td>Bottom-1</td>
<td>13.51</td>
<td>14.13</td>
<td><strong>4.38</strong></td>
<td>16.19</td>
<td></td>
</tr>
<tr>
<td>Bottom-3</td>
<td>11.28</td>
<td>12.46</td>
<td>6.62</td>
<td>14.55</td>
<td></td>
</tr>
<tr>
<td>Bottom-5</td>
<td>10.21</td>
<td>11.22</td>
<td>7.89</td>
<td>14.71</td>
<td></td>
</tr>
<tr>
<td>Bottom-7</td>
<td>12.32</td>
<td>13.48</td>
<td>9.17</td>
<td><strong>13.82</strong></td>
<td></td>
</tr>
</tbody>
</table>

where TP, TN, S, and NS represent the true positive, true negative, shadow, and nonshadow pixel numbers in the shadow image, respectively. Considering the class imbalance problem, BER considers both false positive and false negative rates, providing a more balanced assessment of classifier performance. A lower BER value indicates better shadow detection performance.

**B. Implementation Details**

We use PyTorch 1.10 to implement all our experiments on four NVIDIA 4090 GPUs. We employ the ViT-B [71] as the backbone of our ShadowSAM with AdamW optimizer. The initial learning rate is set to $2e^{-4}$ with cosine decay. We set the balance factor $\alpha$ to 0.6 and the number of curricula $C$ to 5, and each curriculum requires the model to train for ten epochs.

**C. Performance of Vanilla SAM in Shadow Detection**

When no prompts are fed into SAM, it will generate the corresponding binary masks for most of the objects in the input image, known as the “everything model.” Therefore, we should select the most matching shadow candidates as the shadow detection results.

**Bottom-N Strategy:** Given an input image, SAM first generates binary masks for it. Then, we calculate the BER value between each predicted and GT shadow mask. Next, we merge the masks with the Bottom $N$ scores as the final shadow detection result, as shown in Fig. 7(a).

Table I also shows the quantitative comparison results with an SOTA shadow detector. One can observe that: 1) directly applying SAM for shadow detection leads to a significant performance gap between it and the SOTA shadow detector, i.e., SDTR; 2) SAM demonstrates better detection performance on the ISTD than other benchmarks, as its scenes are relatively simpler, while in the other three datasets, the performance of Bottom-5 or Bottom-7 is superior to Bottom-1 and Bottom-3, as they often contain more shadow regions; and 3) a larger $N$ does not always result in better performance, as it may incur more incorrect predictions. We also provide some visual results in Fig. 7(b) and (c). The backgrounds in these successful cases are relatively simple, and the shadows have high contrast intensities. However, in complex scenes, most predictions from SAM are not satisfactory.
TABLE II

<table>
<thead>
<tr>
<th>Method</th>
<th>Year</th>
<th>Type</th>
<th>SBU [68]</th>
<th>UCF [69]</th>
<th>ISTD [70]</th>
<th>CUHK-Shadow [60]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chuang et al. [37]</td>
<td>2008</td>
<td></td>
<td>14.91</td>
<td>16.32</td>
<td>11.21</td>
<td>19.78</td>
</tr>
<tr>
<td>Shoaib et al. [38]</td>
<td>2009</td>
<td></td>
<td>17.53</td>
<td>25.49</td>
<td>9.15</td>
<td>25.44</td>
</tr>
<tr>
<td>Makarau et al. [39]</td>
<td>2011</td>
<td></td>
<td>12.38</td>
<td>20.48</td>
<td>8.78</td>
<td>22.61</td>
</tr>
<tr>
<td>Elbakary et al. [40]</td>
<td>2013</td>
<td></td>
<td>11.56</td>
<td>14.83</td>
<td>6.32</td>
<td>18.75</td>
</tr>
<tr>
<td>SegNeXt-B [72]</td>
<td>2022</td>
<td></td>
<td>6.16</td>
<td>7.65</td>
<td>2.46</td>
<td>13.14</td>
</tr>
<tr>
<td>ICON-R [73]</td>
<td>2023</td>
<td></td>
<td>6.31</td>
<td>7.79</td>
<td>2.32</td>
<td>14.11</td>
</tr>
<tr>
<td>SAM-Max [74]</td>
<td>2023</td>
<td></td>
<td>25.35</td>
<td>24.07</td>
<td>26.37</td>
<td>30.23</td>
</tr>
<tr>
<td>ShadowSAM-U (Ours)</td>
<td>-</td>
<td></td>
<td>3.12</td>
<td>7.27</td>
<td>1.84</td>
<td>11.39</td>
</tr>
</tbody>
</table>

ShadowSAM-S (Ours) -

<table>
<thead>
<tr>
<th>Year</th>
<th>Type</th>
<th>SBU [68]</th>
<th>UCF [69]</th>
<th>ISTD [70]</th>
<th>CUHK-Shadow [60]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2023</td>
<td>Supervised</td>
<td>2.67</td>
<td>7.57</td>
<td>1.36</td>
<td>6.31</td>
</tr>
</tbody>
</table>

D. Comparison With SOTA Methods

We first compare our method with 17 shadow detectors, including five traditional unsupervised methods (i.e., Leone and Distanta [36], Chung et al. [37], Shoaib et al. [38], Makarau et al. [39], and Elbakary and Iftekharuddin [40]) and ten SOTA shadow detectors, namely, SDTR+ [18], SDCM [62], FDRNet [61], FSDNet [60], MTMT-Net [16], DSD [59], A + D Net [58], DSC [55], BDRAR [56], and scGAN [57]. We adopt their reported BER values in their published papers for a fair comparison.

Considering that shadow detection is a kind of pixel-level classification problem, it is related to saliency object detection and semantic segmentation, and we select an SOTA saliency object detection method, ICON-R [73], and one semantic segmentation method, SegNeXt-B [72], and train them using pseudo masks generated by traditional USD methods to achieve deep unsupervised setting. For a more comprehensive comparison, we also compare our method with two SAM-based methods, SAM-Max [74], designed for shadow detection, and SAMed [75], proposed for medical semantic segmentation.

1) Quantitative Comparison: Table II summarizes the quantitative results of different methods on four benchmarks. One can observe that our ShadowSAM-S achieves the SOTA performance across all datasets in a supervised manner. Specifically, compared to the second-best performed method (SDTR+ [18] or SDCM [62]), it improves by 9.18%, 9.45%, 5.56%, and 11.75% in terms of BER on SBU, UCF, ISTD, and CUHK-Shadow datasets, respectively. At the same time, the ShadowSAM-U significantly outperforms traditional USD methods, which is contributed by our proposed prompt-tuning and robust training strategies (i.e., selective self-training and incremental curriculum learning). Compared to the second-best performed traditional shadow detection method, it improves by 49.35%, 4.97%, 20.69%, and 13.32% in terms of BER on SBU, UCF, ISTD, and CUHK-Shadow datasets, respectively. Its performance can even be on par with most fully supervised shadow detectors.

Although SegNeXt-B [72] and ICON-R [73] also trained with pseudo masks, they cannot inherit the boundary localization capabilities of vision foundation model and dynamically refine imprecise initial pseudo masks like our proposed ShadowSAM-U, resulting in unsatisfactory results. Moreover, SAM-Max [74] directly applies the vanilla SAM to detect shadows, identifying the matchest segmented region with GT as the final shadow detection result. It only performs well in simple shadow scenes with one single shadow region. On the other hand, SAMed [75] fine-tunes SAM to adopt downstream task by inserting additional trainable layers in each Transformer block, which retains the prompt encoder and requires updating default embedding for each input. Despite the promising performance on relatively larger datasets, i.e., SBU, it shows a suboptimal performance on the smaller dataset, i.e., ISTD.

2) Qualitative Comparison: We also provide some visualization results in Fig. 8. One can see that our ShadowSAM-U and ShadowSAM-S can both detect shadows accurately. In particular, the ShadowSAM-U outperforms most of the supervised methods. Our unsupervised model can accurately detect shadow areas, while those unsupervised shadow detectors cannot successfully distinguish shadows from shadow-like nonshadow regions (e.g., words, clothes, and plants). In the fifth, sixth, and seventh rows, where the scenes are particularly complex, our method still successfully distinguishes them, while the unsupervised shadow detector and some supervised shadow detectors erroneously identify trees, roofs, and walls as shadows.
In this section, we first conduct extensive ablation studies on the SBU [68] and ISTD [70] datasets to comprehensively evaluate the effectiveness of each component in our deep USD framework.

1) Effect of Tuning Strategy: To intuitively demonstrate the impact of different tuning strategies, in Table III, we carry out numerous fully supervised training with three types of tuning strategies, i.e., fully tuning, partial tuning, and our prompt-like tuning for SAM and ShadowSAM. One can observe that the following conditions hold.

1) Retraining the SAM without using any pretrained weights can also handle the shadow detection task, contributing to the ViT-based network architecture in SAM. However, the performance is still unsatisfactory as it lacks task-related design.

2) The performance of full tuning incurs a slight improvement on the SBU dataset but drops significantly on the ISTD dataset. The reason is the scale of the ISTD dataset (1350 pairs) is insufficient to adjust SAM’s internal knowledge (91.2 M parameters) for adapting shadow detection, which is much better in the SBU dataset (4096 pairs).

3) When only tuning the mask decoder of SAM, the performance significantly increases on ISTD but slightly decreases on SBU. This indicates that although partial tuning has a lower requirement on the data scale, it cannot fully learn task-specific knowledge as the parameters in the feature extraction stage (i.e., encoder) are completely frozen.

4) Compared to the partial tuning, our proposed adaptation methods introduce only 3.52 M extra parameters and perform impressively on SBU and ISTD datasets. This suggests that our prompt-like tuning can better inherit the learned knowledge and sufficiently learn task-specific knowledge even with smaller data.

2) Effect of Pretrained Dataset: To validate the effectiveness of pretrained datasets for shadow detection, i.e., ImageNet and SA-1B. We first replace the mask decoder of SAM with two fully connected layers for retraining on ImageNet and then save the encoder weights for subsequent fine-tuning. One can observe from Table IV: compared to not pretraining on any data, pretrained on ImageNet can slightly improve model performance. However, pretraining on the SA-1B dataset performs better than ImageNet, as SA-1B is collected for image segmentation tasks, and its vast corpus is beneficial for identifying shadows and boundaries, while models pretrained on ImageNet will focus more on extracting objects’ salient cues, which may not help us to recognize the shadow and its boundary.

3) Pruning Experiment: To verify the effectiveness of different components in our proposed deep USD framework, we compare the entire framework with its three variants.

TABLE III
FULLY SUPERVISED SHADOW DETECTION PERFORMANCE WITH DIFFERENT TUNING STRATEGIES. “TUNING PARA.” INDICATES THE NUMBER OF PARAMETERS THAT NEED TO BE FINE-TUNED

<table>
<thead>
<tr>
<th>Network</th>
<th>Training</th>
<th>Tuning Para.</th>
<th>BER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>SBU</td>
</tr>
<tr>
<td>SAM</td>
<td>without pre-training</td>
<td>-</td>
<td>4.88</td>
</tr>
<tr>
<td></td>
<td>full tuning</td>
<td>91.2 M</td>
<td>4.54</td>
</tr>
<tr>
<td></td>
<td>partial tuning</td>
<td>3.87 M</td>
<td>4.65</td>
</tr>
<tr>
<td>ShadowSAM</td>
<td>prompt-like tuning</td>
<td>3.52 M+3.87 M</td>
<td>2.67</td>
</tr>
</tbody>
</table>

E. Ablation Study

In this section, we first conduct extensive ablation studies on the SBU [68] and ISTD [70] datasets to comprehensively evaluate the effectiveness of each component in our deep USD framework.

1) Effect of Tuning Strategy: To intuitively demonstrate the impact of different tuning strategies, in Table III, we carry out numerous fully supervised training with three types of tuning strategies, i.e., fully tuning, partial tuning, and our prompt-like tuning for SAM and ShadowSAM. One can observe that the following conditions hold.

1) Retraining the SAM without using any pretrained weights can also handle the shadow detection task, contributing to the ViT-based network architecture in SAM. However, the performance is still unsatisfactory as it lacks task-related design.

2) The performance of full tuning incurs a slight improvement on the SBU dataset but drops significantly on the ISTD dataset. The reason is the scale of the ISTD dataset (1350 pairs) is insufficient to adjust SAM’s internal knowledge (91.2 M parameters) for adapting shadow detection, which is much better in the SBU dataset (4096 pairs).

3) When only tuning the mask decoder of SAM, the performance significantly increases on ISTD but slightly decreases on SBU. This indicates that although partial tuning has a lower requirement on the data scale, it cannot fully learn task-specific knowledge as the parameters in the feature extraction stage (i.e., encoder) are completely frozen.

4) Compared to the partial tuning, our proposed adaptation methods introduce only 3.52 M extra parameters and perform impressively on SBU and ISTD datasets. This suggests that our prompt-like tuning can better inherit the learned knowledge and sufficiently learn task-specific knowledge even with smaller data.

2) Effect of Pretrained Dataset: To validate the effectiveness of pretrained datasets for shadow detection, i.e., ImageNet and SA-1B. We first replace the mask decoder of SAM with two fully connected layers for retraining on ImageNet and then save the encoder weights for subsequent fine-tuning. One can observe from Table IV: compared to not pretraining on any data, pretrained on ImageNet can slightly improve model performance. However, pretraining on the SA-1B dataset performs better than ImageNet, as SA-1B is collected for image segmentation tasks, and its vast corpus is beneficial for identifying shadows and boundaries, while models pretrained on ImageNet will focus more on extracting objects’ salient cues, which may not help us to recognize the shadow and its boundary.

3) Pruning Experiment: To verify the effectiveness of different components in our proposed deep USD framework, we compare the entire framework with its three variants.
<table>
<thead>
<tr>
<th>BackBone</th>
<th>Dataset</th>
<th>BER SBU</th>
<th>BER ISTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>ImageNet</td>
<td>6.08</td>
<td>2.97</td>
</tr>
<tr>
<td>ShadowSAM</td>
<td>ImageNet</td>
<td>3.12</td>
<td>1.84</td>
</tr>
</tbody>
</table>

### Table V

**Performance of Different USD Frameworks.** Here, “PMG” represents pseudo mask generation, “SS” refers to selective self-training, and “ICL” is incremental curriculum learning.

<table>
<thead>
<tr>
<th>Framework</th>
<th>PMG</th>
<th>SS</th>
<th>ICL</th>
<th>BER SBU</th>
<th>BER ISTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD1</td>
<td>✓</td>
<td></td>
<td></td>
<td>5.94</td>
<td>2.37</td>
</tr>
<tr>
<td>USD2</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>3.97</td>
<td>2.03</td>
</tr>
<tr>
<td>USD3</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>3.44</td>
<td>2.12</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.12</td>
<td>1.84</td>
</tr>
</tbody>
</table>

4) **Single Supervision:** In Fig. 9, we also evaluate the performance of our USD framework when using single-source supervision. One can see that compared to the multisource supervision in Table II (3.12 and 1.84 in terms of BER in SBU and ISTD datasets, respectively), single-source supervision is difficult for the deep model to learn enough shadow patterns due to lacking sufficient shadow prior knowledge. However, their superior performance compared to traditional USD methods also demonstrates the effectiveness of our deep USD framework.

5) **Number of Curriculums:** In the stage of curriculum construction, we evenly divide the training set from easy to difficult into $C$ subsets. To explore the impact of the number of $C$ in our USD framework, we set $C$ to 2, 5, 8, and 10. One can observe from Table VI: 1) a larger number of $C$ seems to have more advantages and this is because a smaller $C$ is hard to represent the multilevel complexity of training samples, which limits the advantage of curriculum learning and 2) too large $C$ could lead to a limited number of training samples in each subset, resulting in overfitting during model training. In this connection, the obvious differences in data scale and scene complexity between ISTD and SBU datasets result in a different set of $C$.

### F. Applications

As shown in Fig. 10, the presence of shadows will bring more challenges for the current remote sensing semantic segmentation method (e.g., DBFNet [76]). For example, the following hold: 1) shadows will interfere with image semantics leading to the misclassification between the low vegetable (blue color) and tree (green color) in regions “1” and “3” and 2) the cast shadows occluded by the car will be misrecognized as the car (yellow) in region “2.” However, by combining our accurate shadow detection and an effective shadow removal technique [77], we can reduce these errors. Similarly, the shadows in the remote sensing change detection, widely used in urban expansion, will often be regarded as new changes. As shown in Fig. 11, detecting the shadows by our ShadowSAM and removing them can significantly improve the performance.

**Deep Unsupervised Salient Object Detection (USOD):** Our deep USD framework can also be applied for high-relevant vision tasks, i.e., salient object detection. Specifically, we use...
four traditional salient object detection methods [78], [79], [80], [81] to produce initial pseudo masks, while in the stage of selective self-training, we use the color variance and gradient information to guide the updating of the pseudo mask. We argue that more accurate salient masks contain more gradient information and smaller color variance due to the clearer boundaries of the salient objects and relatively consistent color within them. As shown in Fig. 12, the modified deep USOD framework can also achieve impressive detection results.

G. Limitations

Our proposed deep USD framework has several limitations: 1) compared with conventional training, our incremental curriculum learning will incur 3–4 times the training cost and 2) the performance of our framework mainly depends on the SAM pretrained on large-scale corpora. Even though we use the ViT-B, the smallest SAM backbone, it still has a 94.7 M parameter, resulting in weak real-time performance, as shown in Table VII.

V. Conclusion

This work presents ShadowSAM, which adapts the pretrained SAM model to USD and effectively reduces annotation reliance. By introducing a series of prompters, ShadowSAM integrates task-specific knowledge with the general knowledge gained by SAM. We address the challenge of robust training under noisy labels through our ITU for selective self-training and MDI for incremental curriculum learning. Our USD framework exhibits comparable performance to fully supervised methods, significantly enhancing applications such as remote sensing shadow semantic segmentation and salient object detection. However, we also encounter longer training costs and lower efficiency. This study elucidates the potential of large-scale pretrained models in annotation-free task-specific adaptation, opening up exciting opportunities in deep unsupervised learning.

REFERENCES


