SILT: Shadow-aware Iterative Label Tuning for
Learning to Detect Shadows from Noisy Labels

Han Yang[1,*], Tianyu Wang[1,2*], Xiaowei Hu[3,1,†] and Chi-Wing Fu[1,2]

[1] The Chinese University of Hong Kong [2] The Shun Hing Institute of Advanced Engineering
[3] Shanghai Artificial Intelligence Laboratory

Abstract

Existing shadow detection datasets often contain missing or mislabeled shadows, which can hinder the performance of deep learning models trained directly on such data. To address this issue, we propose SILT, the Shadow-aware Iterative Label Tuning framework, which explicitly considers noise in shadow labels and trains the deep model in a self-training manner. Specifically, we incorporate strong data augmentations with shadow counterfeiting to help the network better recognize non-shadow regions and alleviate overfitting. We also devise a simple yet effective label tuning strategy with global-local fusion and shadow-aware filtering to encourage the network to make significant refinements on the noisy labels. We evaluate the performance of SILT by relabeling the test set of the SBU [55] dataset and conducting various experiments. Our results show that even a simple U-Net [42] trained with SILT can outperform all state-of-the-art methods by a large margin. When trained on SBU / UCF [78] / ISTD [56], our network can successfully reduce the Balanced Error Rate by 25.2% / 36.9% / 21.3% over the best state-of-the-art method.

1. Introduction

Detecting shadows is very challenging, since shadows have no specific shapes, colors, or textures, and their intensity may just be slightly lower than the surroundings. Thanks to the advances in deep learning, many works [81, 80, 3, 73] have been developed and they show great progress in detecting shadows. They mostly propose new and delicate network architectures and train the network directly on shadow datasets with labeled shadow regions.

However, it can be observed that labels in existing datasets [78, 13] may not be accurate. For example, as Fig. 1 shows, the training samples may lack details (row 1); some shadows could be incomplete (row 2); some self

(a) Training images in existing datasets (b) Original labels in existing datasets (c) Our refined labels
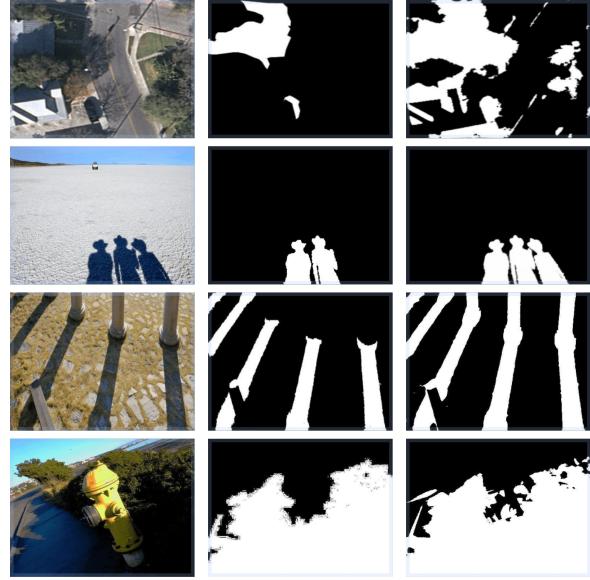
Figure 1. Column (b): Labels in existing datasets may not be accurate. Column (c): Our automatically-refined annotations. Note that row 1 is from UCF [78] and rows 2-4 are from SBU [55].

shadows may be missed (row 3); and the annotations could be rough (row 4). Fundamentally, there are two main reasons. First, the annotations of the SBU [55] dataset are generated from a manually lazy-labeled dataset using an LSSVM-based method, so the resulting annotations can be noisy, and some detailed and background shadows may be ignored. Second, the perception of shadow can be subjective, especially for self shadow and soft shadow. Since datasets are typically prepared by several human annotators, shadow and non-shadow regions could be labeled inconsistently in the same dataset. Hence, existing methods trained on such noisy datasets could be easily biased by the noisy labels, which hinder them to achieve better performance.

To learn from data with noisy labels, it is intuitive to adopt a self-training framework [65], *i.e.*, train a network

| Input image | Original Label | Input image | Original Label |

Refined Labels

Round 2

Round 4

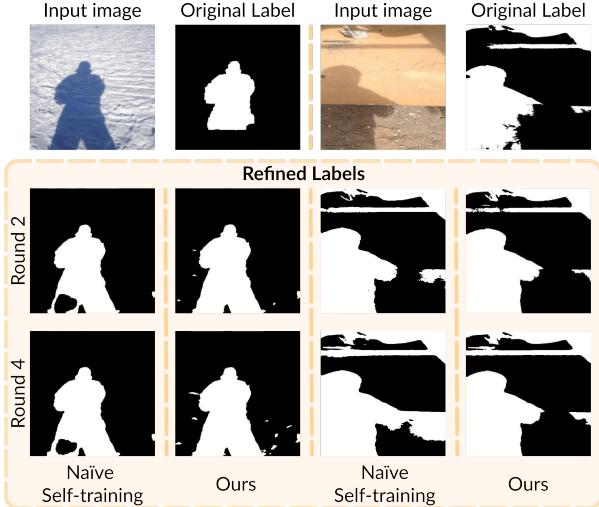| Naïve Self-training | Ours | Naïve Self-training | Ours |

Figure 2. Comparison of label quality produced by a naïve self-training framework and our SILT framework. Our SILT framework generates more accurate labels with finer details.

on noisy data; use the trained network to relabel the training data; and repeat these two steps alternately to refine the labels. Yet, directly adopting self-training to detect shadows may not work well for the following reasons:

- Overfitting the training data. Considering the limited size of existing datasets, the trained network is prone to remember every single sample. So, if we directly use it to relabel the training data, the network may be too conservative to try large refinements on the noisy labels. As Fig. 2 column 1 shows, the network could fail to label some obvious shadows due to overfitting.

- Error accumulation. Existing methods often falsely predict dark objects as shadows. So, the iterative self-training process may easily accumulate such prediction errors and wrongly encourage the network to label all dark objects as shadows; see Fig. 2 column 3.

In this paper, we present the Shadow-aware Iterative Label Tuning (SILT) framework by formulating the following strategies to address the above challenges. Firstly, we adopt a shadow-aware data augmentation strategy that involves shadow counterfeiting and incorporating dark regions and noise into input images. This enhances the ability of the deep network to recognize non-shadow regions by teaching it to identify these regions amidst the added noise. Secondly, we propose a global-local fusion approach that involves splitting the input image into multiple patches and using the network to predict masks for each patch and the entire image. This approach helps alleviate overfitting. We then perform shadow-aware filtering that considers both the image brightness information and the previous shadow masks to select accurate shadow masks while filter-

ing out inaccurate predictions. Thirdly, we collect a set of zero-labeled non-shadow images with dark objects to train the network to better identify non-shadow regions. Using the above techniques, we can effectively train a simple U-Net [42] in SILT to iteratively refine the noisy labels in the original datasets. Examples demonstrating the effectiveness of the proposed approach are shown in Fig 1 (c).

For quantitative evaluation, we relabel the test set of SBU [55] to obtain high-quality and accurate shadow masks, due to the existing issue of noisy labels. With this carefully relabeled test set, we conduct various experiments and demonstrate the superiority of our approach in producing more precise shadow detection results compared to existing state-of-the-art methods. Furthermore, we find that our refined training set can significantly improve the performance of these state-of-the-art methods. The code, pretrained model, and dataset are publicly available at `https://github.com/Cralence/SILT`.

2. Related Works

Shadow detection. Single-image shadow detection has been studied for a long time. Earlier methods detect shadows mainly based on physical features such as geometrical properties [43, 39], spectrum ratios [49], color [22, 10, 52, 54], texture [78, 10, 52, 54], edge [22, 78, 17], etc. However, due to the physical nature of the shadow, it is hard to quantify shadow by a few situation-invariant features. Therefore, these methods work well mainly on simple cases and are less effective on complex real-world shadows.

With deep convolutional neural networks (CNN), features can be automatically learned instead of hand-crafted. Since then, the performance of shadow detection improves significantly. Khan et al. [20] first use a CNN followed by a conditional random field to achieve pixel-wise shadow detection. After that, many architectures [44, 55, 36, 16, 14, 56, 23, 79, 73, 6, 3, 80, 15, 81, 7, 74, 62, 18, 53, 59, 57, 58] are proposed to adopt deep neural networks for shadow detection. Besides, various high-level shadow features are designed, e.g., direction-aware spatial context [16, 14], distraction-aware [73], shadow edges [44, 3], shadow count [3], etc. Further, several datasets are built, e.g., SBU [55] and ISTD [56]. Trained on these data, deep neural networks can learn to detect shadows in more situations.

However, prior works mostly do not consider the noisy label problem. One relevant work [55] is an LSSVM-based noisy label recovery method, which adopts image clustering to recover the annotations from a lazily labeled dataset. However, the SBU dataset refined by this method still contains many noises. In this work, we reformulate shadow detection as a noisy label problem by explicitly considering the noisy labels in the training data and present a new shadow-aware iterative label tuning framework to learn to detect shadows from noisy labels.
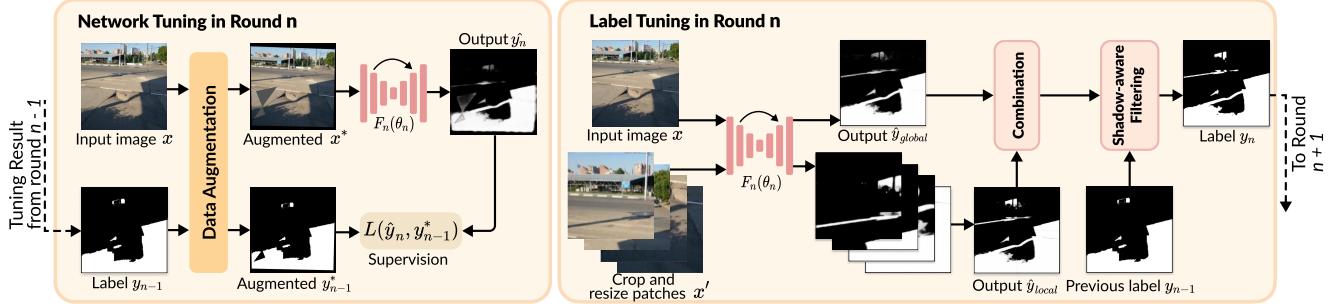
Figure 3. The architecture of Shadow-aware Iterative Label Tuning (SILT) for learning to detect shadows from data with noisy labels.

Learning with noisy labels. Existing works for the classification task have widely studied the noisy label issue. Various aspects have been exploited, *e.g.*, network architectures [2, 64, 9, 11], losses [8, 31, 61, 72, 76, 77], regularizations [33, 48, 63], *etc*. On the other hand, some methods aim to first detect [41, 75] noises then re-weight [40] or correct [47, 50, 51, 68, 12] the detected noises. To perform image segmentation with noisy labels, some methods improve the robustness of the network to noise by designing meta-structures [30] or leveraging reliability of annotators [70] and noisy gradient [34]. Others correct errors by exploiting early-learning phase [28], using local visual cues [45] and formulating spatial label smoothing regularization [71].

Self-training. Self-training methods [67, 37, 24] typically first train a teacher network on a small labeled dataset. Then, we can use it to generate pseudo-labels for a large unlabeled dataset and take the pseudo-labels to train a student network; by iterate this process, the quality of the pseudo-labels can be gradually improved. Recently, self-training methods achieve state-of-the-art performance in tasks like image classification [25, 65] and segmentation [32, 38, 26].

Our method differs from previous works in self-training and noisy labels in two aspects. First, *we only have small but noisy datasets*, while previous self-training methods typically adopt a huge unlabeled or badly-labeled dataset [24, 65, 47, 12, 64], sometimes together with a small and clean dataset [24, 65]. Thus, we propose global-local fusion in label tuning to mitigate the over-fitting issue. Second, *in shadow detection, errors appear more with a common pattern*, compared with general image classification and segmentation, so it is hard to tackle the error accumulation issue. Thus, we propose shadow counterfeiting in network tuning to enhance the discriminating ability of the deep network to recognize non-shadow regions and shadow-aware filtering in label tuning to select accurate shadow masks while filtering out inaccurate predictions.

3. Methodology

Fig. 3 shows the overall architecture of the proposed Shadow-aware Iterative Label Tuning (SILT) framework,

which automatically tunes the labels in the noisy dataset in a shadow-aware manner. For each round $n \in \{1, 2, ..., N\}$, where N is the total number of rounds, we denote the input images as X, and its corresponding shadow masks as Y_{n-1}. When $n = 1$, the corresponding Y_0 is the original noisy label. There are two stages in each round of SILT: *Network Tuning* and *Label Tuning*. In network tuning, we train a shadow detection network supervised by previously tuned labels Y_{n-1}, during which we introduce strong data augmentation with shadow counterfeiting; see details in Sec. 3.1. In label tuning, we tune the shadow masks using the previous freeze-weight network with global-local fusion and shadow-aware filtering strategies; see details in Sec. 3.2. By iteratively performing these two stages, the labels gradually become more accurate and contain more fine details.

3.1. Network Tuning with Shadow Counterfeiting

Dark objects are easily misclassified as shadow regions and the errors could be accumulated in the self-training frameworks [65, 12]. In order to let the network learn to distinguish the shadow and dark non-shadow objects, we introduce the shadow counterfeiting in network tuning. This strategy is implemented by two data augmentations, namely Distraction and RandomNoise, which add different types of dark regions (counterfeited shadows) to train the network to recognize the dark regions as non-shadows.

Specifically, in Distraction (Fig. 4 (a)), we randomly choose an area in the training image and fill it with dark color. The area is a randomly generated polygon, and we multiply the value of the pixels in that polygon by a factor of 0.3. In RandomNoise (Fig. 4 (b)), we randomly add black points in the training images. We label both augmentations' newly added regions/pixels as **non-shadow regions**. These two augmentations introduce "counterfeited" shadows with various sizes and numbers to the training data, thereby improving its ability to distinguish between shadows and dark non-shadows. To mitigate overfitting, we add other common data augmentations following previous works [69, 4].
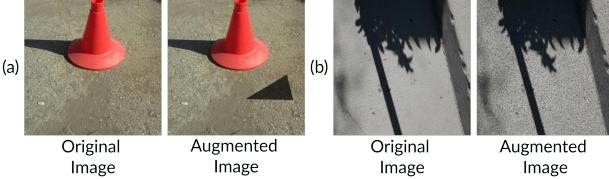
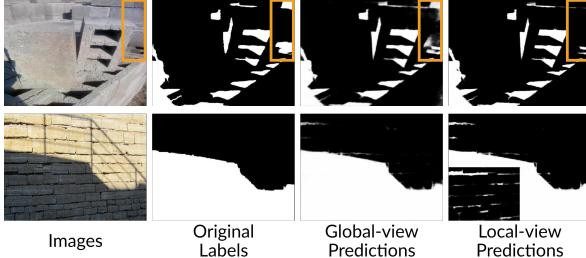Figure 4. An illustration of shadow counterfeiting.

Figure 5. Visualizations of global-view and local-view prediction.

3.2. Label Tuning

In label tuning, we aim to employ the network to correct wrong labels and add missing labels in the training data. However, the network may easily overfit the wrong labels since the shadow detection datasets [78, 55] are relatively small compared with other datasets used for general computer vision tasks, e.g., ImageNet [5] for image classification and COCO [27] for object detection. To solve the above issues, we present the label tuning strategy, which first adopts global-local fusion to enhance details as well as alleviate overfitting and then uses shadow-aware filtering to filter out inaccurate and vague predictions.

3.2.1 Global-Local Fusion

As shown in the right part of Fig. 3, unlike the previous works [25, 65] that use the network to relabel their training data directly, we first split the input image into four parts, resize them to the size of the original input image, and adopt the network to predict the shadow mask for each part. Lastly, we obtain a unified output $\hat{y}_{local}$, referred to as local-view prediction, by combining different local parts.

The benefits of such operations are two-fold. First, inspired by the concept of multi-scale inference [35], a higher resolution input image helps the network to predict more detailed shadow masks. More importantly, taking crop-and-resized patches as input avoids directly using the raw training data, which mitigates overfitting and encourages noisy label correction, as depicted in the first row of Fig. 5.

Besides the local-view prediction, we further take the whole image as input to generate the global-view prediction $\hat{y}_{global}$, which helps to distinguish the large shadow region by considering the global image context, as shown in the second row in Fig. 5. After obtaining the predictions from the global view and the local view, we combine them into the final prediction $\hat{y}_n$ by

$$\hat{y}_n = \begin{cases} \max(\hat{y}_{global}, \hat{y}_{local}), & \hat{y}_{global} > R_{filt} \\ \hat{y}_{global}, & \hat{y}_{global} \leqslant R_{filt} \end{cases}, \quad (1)$$

where R_{filt} is the threshold to ensure the priority of $\hat{y}_{global}$, since the global-view result is more reliable by considering the global image context. We empirically set the R_{filt} as 0.1 during the experiments.

3.2.2 Shadow-aware Filtering

The prediction $\hat{y}_n$ contains the continuous values, and some ambiguous regions, e.g. dark non-shadow objects, usually have low confidence. If we keep these ambiguous regions, the confidence value will accumulate stage by stage and finally mislead the network to recognize these regions as shadow regions with high confidence. To avoid this situation, we pass the prediction $\hat{y}_n$ through a threshold map to generate a binary mask.

We observe that the network often has lower confidence in predicting shadows in bright regions while has higher confidence in dark regions. But the dark regions are easily to be mislabeled. Therefore, we prefer to adopt a lower threshold for bright regions, thus making the network easier to refine the wrong labels in those regions. Meanwhile, we prefer a higher threshold for dark regions, which helps to filter out the wrong labels of dark non-shadow objects predicted by the network. For this purpose, we construct a shadow-aware filter map to perform binarization with different thresholds according to the image brightness.

Specifically, to obtain the illumination intensity, we transform the RGB image into YC_bC_r color space [19] and take the first channel as brightness map Y. Then, we construct a shadow-aware filter map F, which gives a smaller threshold when the brightness is high and vice versa. In detail, we set R_{min} and R_{max} as two hyper-parameters denoting the minimum and maximum value of the threshold map, respectively, and for each pixel i, the corresponding threshold F_i is defined as:

$$F_i = \frac{R_{max}Y_{max} - R_{min}Y_{min}}{Y_{max} - Y_{min}} - \frac{R_{max} - R_{min}}{Y_{max} - Y_{min}} \cdot Y_i, \quad (2)$$

where Y_{max} and Y_{min} denote the maximum and minimum brightness value of the image, respectively, and Y_i is the brightness at the pixel i. Note that F_i is defined as R_{max} when Y_i takes Y_{min} and as R_{min} when Y_i takes Y_{max}. During the experiments, we empirically set R_{min} as 0.5 and R_{max} as 0.6. Through the above formulation, we adopt a higher threshold to make a strict selection in dark regions and use a lower threshold to avoid filtering out correct masks in bright regions.

After we obtain the newly generated shadow mask $\hat{y}_n$, we compare it with the previous shadow mask y_{n-1} and use it to replace the previous shadow mask when $\hat{y}_n$ contains at least R_{corr} percentage of the previous shadow mask. Otherwise, we will keep the previous shadow mask as the final result. This is because the major part of the original/previous shadow mask is usually reliable, especially in the larger regions. We empirically set R_{corr} as 0.95 in the following experiments.

4. Experiments

4.1. Datasets and Evaluation Metrics

We employ three mostly-used datasets. The first one is SBU [55], which contains 4089 training images and 638 testing images. The second is UCF Shadow Dataset [78], which includes 360 images. The third is ISTD [56], which has 1330 training images and 540 testing images. SBU and UCF contain more noisy labels, whereas ISTD is relatively cleaner. For evaluation metrics, we choose the most widely-used Balanced Error Rate (BER), Shadow Error Rate (BER_S), and Non-shadow Error Rate (BER_{NS}).

4.1.1 Relabeled SBU test set

Considering that the test sets in existing shadow datasets also contain wrong labels, we decide to relabel the test set of SBU [55] to better evaluate the performance of various shadow detection networks. Specifically, we hired three experts to do the relabeling work. Before relabeling, we showed them some examples to clarify the definitions and identifications of shadow, especially the self shadow. Then, they used Affinity Photo [1] on the iPad with Apple Pencil to draw shadow labels image by image. Specifically, they were required to zoom in to label the details. It took on average five minutes to relabel each image. Finally, we cross-validated the refined labels from three experts and integrated them to obtain the final ground truth masks.

In total, our relabeled test set contains 638 test images and masks with fine details. In the original test set, 11.34% of the mask pixels have been changed. Among them, 11.05% are incomplete masks (new=1, old=0), while only 0.29% are wrongly-labeled masks (new=0, old=1). Therefore, incomplete masks contribute the most to inaccurate annotations. Meanwhile, out of the 638 masks in our relabeled test set, 517 of them have modifications larger than 5%, and 374 of them have modifications larger than 20%. Figure 6 shows comparison examples between our relabeled test set and the original SBU test set. Ours features more detailed shadow masks and exhibits more consistent labeling regarding self shadows and background shadows.
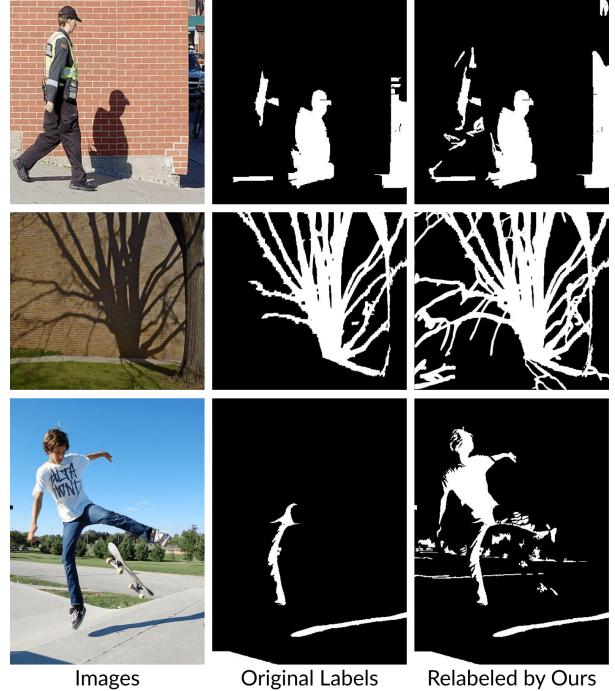

Figure 6. Comparison of our relabeled SBU test set and original SBU test set.

Figure 7. Examples of our additional dataset.

4.1.2 Additional Non-shadow Training Data

We collect some images from the Internet to further help train the network to distinguish shadows and dark objects and alleviate error accumulation problem. Specifically, we search for some images that contain dark objects but no shadows, see Figure 7. Therefore, their shadow masks should be zeros, and no hand labeling is needed. Considering that too many training images with zeros as labels may deteriorate training, we only choose 89 images in total as our additional training data, which is 2% of the total number of the SBU. Note that we only use this additional training data when training on SBU [55], as the size of the training set of UCF [78] and ISTD [56] is very small.

4.2. Experiment Details

We adopt a simple U-Net structure [42] as our shadow detection network. For a fair comparison, we use differ-

Table 1. Quantitative Comparison of our method with recent state-of-the-art methods on three benchmark datasets. (a) and (b) are evaluated on our relabeled SBU test set; (c) is evaluated on original ISTD test set.

Training set		Param.(M)	SBU (a)			UCF (b)			ISTD (c)		
Method	Year		BER↓	BER_S	BER_{NS}	BER↓	BER_S	BER_{NS}	BER↓	BER_S	BER_{NS}
BDRAR	ECCV2018	42.46	6.49	9.68	3.29	11.48	18.81	4.15	2.69	0.50	4.87
DSC	CVPR2018	79.03	8.08	12.25	3.91	14.15	24.86	3.44	3.42	3.85	3.00
DSD	CVPR2019	58.16	5.60	7.86	3.34	-	-	-	2.17	1.36	2.98
MTMT	ICCV2021	44.13	7.41	13.68	0.97	-	-	-	1.72	1.36	2.08
FSD	TIP2021	4.40	10.87	20.39	1.34	17.66	32.38	2.93	2.68	3.69	1.66
FDRnet	ICCV2021	10.77	5.93	10.93	1.71	12.91	22.50	3.32	1.55	1.22	1.88
SDCM	ACM MM22	10.95	5.71	9.07	2.36	11.45	20.69	2.22	1.41	1.19	1.69
ours (EfficientNet-B3)		12.18	5.23	6.22	4.23	9.18	12.32	6.04	2.00	1.62	2.37
ours (ConvNeXt-B)		100.68	5.11	7.07	3.15	8.62	11.54	5.70	1.15	0.82	1.48
ours (ResNeXt-101)		90.50	5.08	4.86	5.30	9.27	12.71	5.82	1.53	1.20	1.86
ours (EfficientNet-B7)		67.80	4.62	**4.24**	4.90	7.97	9.41	6.54	1.46	1.01	1.90
ours (PVT v2-B3)		49.42	4.36	5.29	3.43	7.25	**7.39**	7.12	**1.11**	0.79	**1.44**
ours (PVT v2-B5)		86.14	**4.19**	4.28	4.09	**7.23**	7.78	6.69	1.16	0.85	1.47

Table 2. Comparison results of the recent state-of-the-art methods trained on the original dataset and our refined dataset. Evaluations are done on our relabeled SBU test set.

	Method	Year	Trained on original dataset	Trained on our refined dataset			% Reduction	
			BER↓	BER↓	$BER_S↓$	$BER_{NS}↓$	BER	BER_S
	BDRAR [79]	ECCV2018	6.49	5.35	5.57	5.12	17.6%	42.5%
	DSC [16]	CVPR2018	8.08	5.62	5.38	5.86	30.4%	56.1%
SBU [55]	FSD [15]	TIP2021	10.87	6.13	8.19	4.08	43.6%	59.8%
	FDRnet [80]	ICCV2021	5.93	5.48	7.44	3.52	7.5%	31.9%
	SDCM [81]	ACM MM22	5.71	4.87	4.32	5.41	14.4%	55.7%
	BDRAR [79]	ECCV2018	11.45	9.59	9.55	9.64	16.2%	49.2%
	DSC [16]	CVPR2018	14.15	10.71	12.49	8.93	24.3%	49.8%
UCF [78]	FSD [15]	TIP2021	17.66	13.39	21.21	5.57	24.2%	34.5%
	FDRnet [80]	ICCV2021	12.91	9.12	11.97	6.27	29.4%	72.1%
	SDCM [81]	ACM MM22	11.45	8.37	8.23	8.51	26.9%	60.2%

ent encoder networks with similar parameter size as previous SOTA networks, *i.e.*, ResNeXt-101 [66], ConvNeXt-B [29], EfficientNet-B3 [46], EfficientNet-B7 [46], PVT v2-B3 [60], and PVT v2-B5 [60]. Following previous works, we initialize the backbones with the weights pretrained on ImageNet [5]. We set the total number of self-training rounds N as seven, and in each round, we train the network for 20 epochs with a batch size of six. We set the learning rate to 1×10^{-4} for PVTs and 5×10^{-4} for other networks, and use Adamax [21] as the optimizer. We resize each image to 512×512 while training, refining, and inferring, and calculate the BERs on the original image size. For models trained on SBU and UCF, we test the models on our relabeled SBU test set. Besides, we employ the original ISTD test set to evaluate the models trained on ISTD, as the test set in ISTD is better labeled. Note that we obtain the best result in round 6, 5, and 3 for models trained on SBU, UCF, and ISTD, respectively.

4.3. Comparison with the State-of-the-art Methods

We compare our training framework with seven previous SOTA networks, namely BDRAR [79], DSC [16, 14],

DSD [73], MTMT [3], FSD [15], FDRnet [80], SDCM [81]. We use their public pre-trained models on the SBU and ISTD, and re-trained models on the UCF for evaluation.

Table 1 shows the results. We can see that our U-Net with various backbones outperforms all the previous SOTA networks by a large margin. On SBU, the PVTv2-based variant achieves the best performance, with a 25.2% and 45.5% decrease in BER and BER_S compared with the best previous work. On UCF, our PVTv2-based variant achieves a 36.9% and 60.7% decrease in BER and BER_S compared with the best previous work. Note that we do not re-train MTMT and DSD, which require additional training data. On ISTD, our PVTv2-based variant also reduces the BER and BER_{NS} by 21.3% and 13.3%, respectively.

Fig. 8 shows the qualitative comparisons of our method with the recent state-of-the-art methods. From the results, we can see that our method is able to detect more fine shadow details with the help of the proposed SILT.

To show the effectiveness of our SILT, we further retrained the previous state-of-the-art networks on our tuned SBU and UCF dataset. The comparison results are shown in Table 2, where all the previous networks achieve signifi-
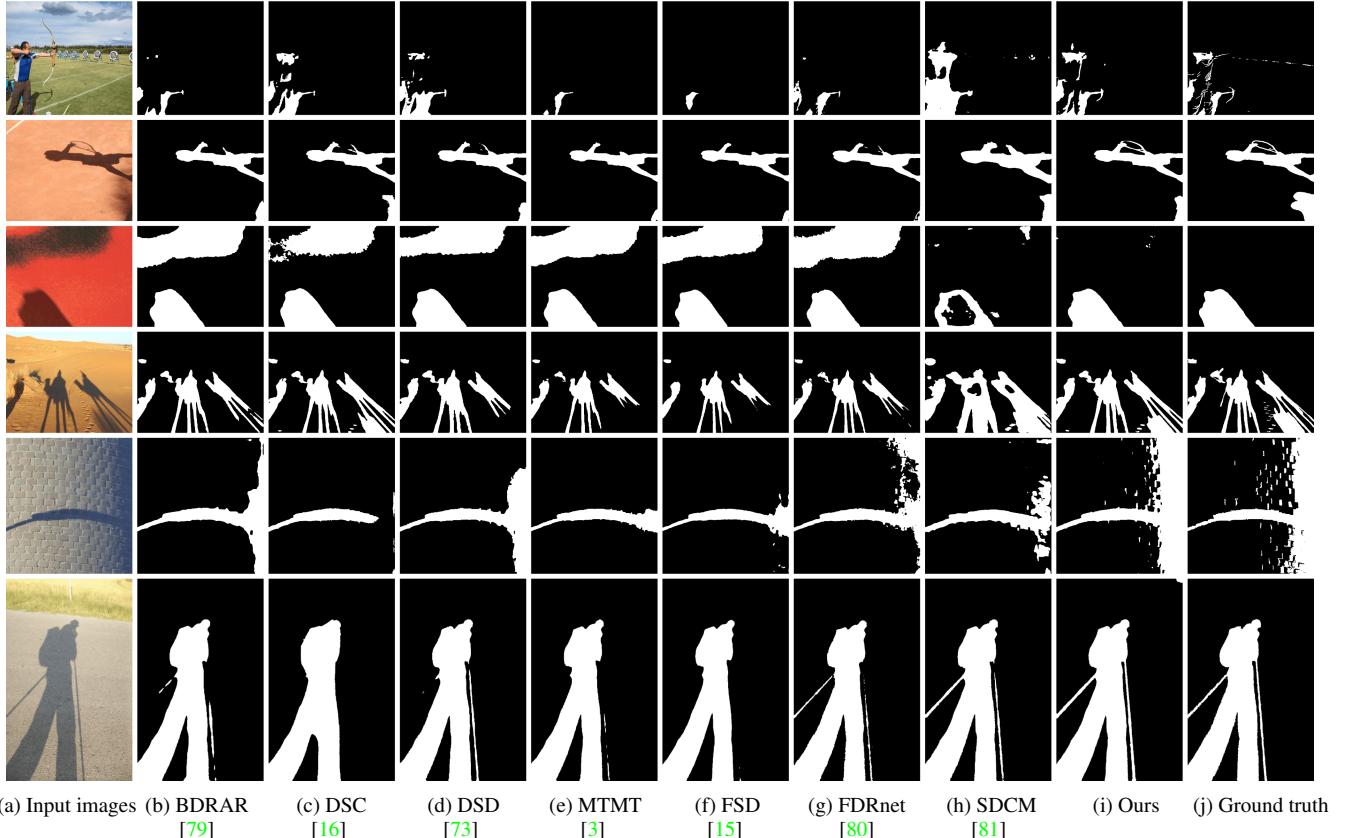
| (a) Input images | (b) BDRAR [79] | (c) DSC [16] | (d) DSD [73] | (e) MTMT [3] | (f) FSD [15] | (g) FDRnet [80] | (h) SDCM [81] | (i) Ours | (j) Ground truth |

Figure 8. Qualitative comparison of our method with recent state-of-the-art methods.

cant improvements, especially in terms of BER_S. It proves that the performance of existing networks is limited by the noisy labels and our tuned labels help to train a better deep model than using the original labels. It also shows that our SILT framework is applicable to various kinds of network architectures, not limited to the U-Net.

4.4. Ablation Study
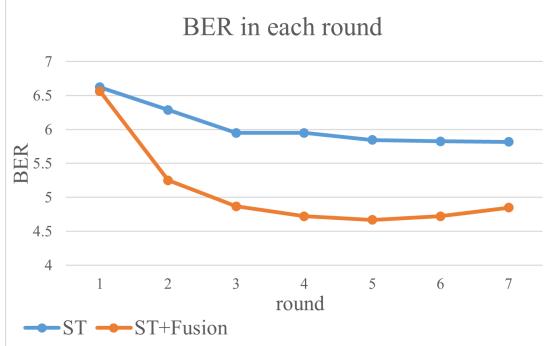
4.4.1 Component Analysis

We conduct the ablation study by applying our SILT on SBU [55] dataset and testing it on our refined SBU test set. In the experiments, we use a U-Net [42] with PVT v2-B5 [60] as the backbone. In total, we consider the following baseline networks: 1) Base: directly train a U-Net [42] on the noisy dataset. 2) ST: directly apply a self-training framework, where we train a network, use the trained network to relabel the training data, and repeat these two stages alternately. 3) ST + Filter: add shadow-aware filtering in label tuning, but without global-local fusion. 4) ST + Fusion: add global-local fusion in label tuning, but without shadow-aware filtering. 5) ST + LT: full label tuning but without shadow counterfeiting. 6) ST + LT + SDA: add the common strong data augmentations, *e.g.*, RandomPerspective,
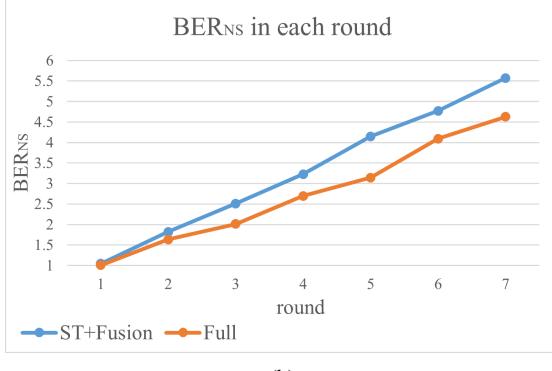
Table 3. Component analysis on the proposed SILT.

	BER↓
Base	6.62
ST	5.82
ST+Filter	5.44
ST+Fusion	4.67
ST+LT	4.52
ST+LT+SDA	4.46
ST+LT+SDA+SC	4.37
Full	**4.19**

GaussianBlur, etc. 7) ST + LT + SDA + SC: further use shadow counterfeiting in the network tuning stage. 8) Full: further add an additional unlabeled non-shadow dataset to the training data.

Table 3 shows the BER of each variant evaluated on our relabeled SBU test set, where we can see that (i) simply applying a self-training framework (ST) to refine noisy labels gives a limited improvement; (ii) each component in our framework design improves the quality of shadow masks in training data, thus promoting the performance of shadow detection; and (iii) our full pipeline with the additional non-shadow training set achieves the best performance, showing the effectiveness of the proposed method.

(a)

(b)

Figure 9. Comparison of the baseline networks in terms of error rates in each round.

Table 3 shows that global-local fusion (ST + Fusion) brings the largest improvement. Fig. 9 (a) shows the BER in each round of the self-training framework with and without global-local fusion. Compared with ST, the ST + Fusion effectively alleviates overfitting in the first few training rounds. Fig. 9 (b) shows that our full pipeline with shadow-aware filtering and shadow counterfeiting reliefs error accumulation in shadow regions.

4.4.2 Architecture Analysis

Round number. In Fig. 10, we show the results where we use label tuning for different rounds in the training stage. We can observe that (i) after the first round, we are able to relabel the large shadow regions in the images; (ii) fine details are gradually added in the shadow masks during the training process; and (iii) if using a large number of training rounds (the last column), the shadow masks become noisy again due to the error accumulation. Hence, we empirically set the round number as seven.

Hyper-parameters: R_{min} **and** R_{max}**.** R_{min} and R_{max} are two hyper-parameters in equation (2), which control the threshold to split the non-confident predictions into shadow and non-shadow regions. Table 4 shows the BER of different combinations, where when the difference between R_{min}
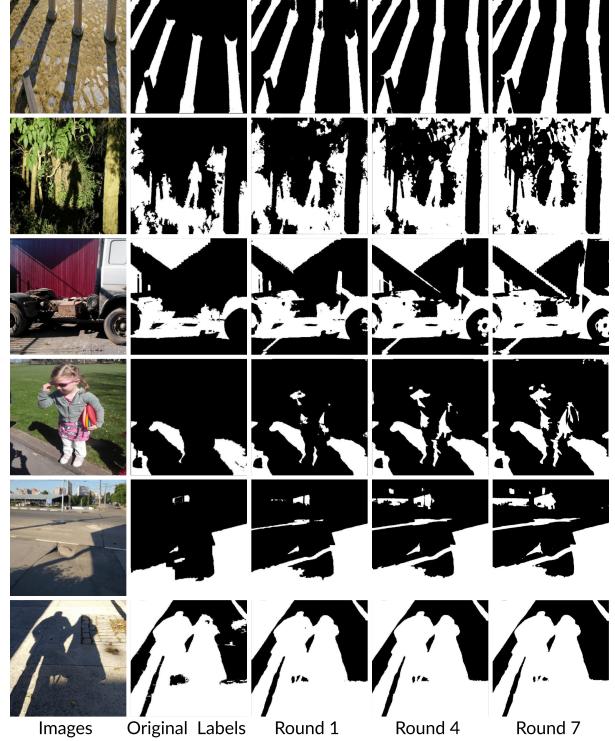

Figure 10. Shadow masks refined by different rounds of SILT.

Table 4. BER values of different R_{min} and R_{max}.

		R_{max}				
		0.4	0.5	0.6	0.7	0.8
R_{min}	0.4	4.77	4.44	4.40	4.41	-
	0.5	-	4.53	**4.19**	4.75	-
	0.6	-	-	4.45	4.36	4.80

Table 5. BER values of different R_{corr}

R_{corr}	0.91	0.93	0.95	0.97	0.99
BER	4.59	4.32	**4.19**	4.51	4.52

Table 6. BER values of different R_{filt}

R_{corr}	0.05	0.075	0.1	0.125	0.15
BER	4.70	4.47	**4.19**	4.42	4.62

and R_{max} is neither 0 nor too large, the shadow-aware filtering gives the best results, and we set R_{min} and R_{max} as 0.5 and 0.6 in experiments.

Hyper-parameters: R_{corr}**.** R_{corr} controls the least ratio of the previous mask that is contained in the new mask. A larger R_{corr} leads to fewer label corrections, while a smaller R_{corr} introduces more wrong labels. As shown in Table 5, we empirically set R_{corr} as 0.95.

Hyper-parameters: R_{filt}**.** R_{filt} determines how much we take the local-view prediction into account. A smaller R_{filt} keeps more local-view prediction, and vice versa. As shown in Table 6, we empirically set R_{filt} as 0.1.

4.4.3 Analysis on Additional Data

Dataset Name	SBU	UCF	ISTD
with additional dataset	4.19	7.60	1.18
w/o additional dataset	4.37	7.23	1.16

Table 7. The BER values of PVT v2-B5 based model trained on different datasets with and without additional dataset.

For a fair comparison, we conduct the ablation study on the additional non-shadow data to test its effectiveness. The results of our PVT v2-B5 based model trained on the three datasets are listed in Table 7. As the results of UCF and ISTD show, data with excessive hard negative cases would deteriorate the training, leading to a performance degradation. Meanwhile, the results on all the three datasets show that, without the additional dataset, SILT is still, or even more, competitive.

4.5. Discussion on Non-Shadow Error Rate

From Table 1, we can observe that our method has a higher Non-shadow Error Rate (BER_{NS}) than the prior ones. We hypothesize that a model trained on a well-labeled dataset tends to produce balanced BER_S and BER_{NS} values. In contrast, the previous training datasets with lots of missing masks lead the networks to predict fewer shadow regions, resulting in BER_S much higher than BER_{NS}.

To validate our assumption, we randomly split the relabeled SBU test set into a new training set of 538 images and a new test set of 100 images. Then, we train the SOTA method SDCM [81] on this new training set [denoted as SDCM(c)] and test it on the new test set. We also used this same test set to evaluate the performance of our SILT and SDCM [81] that trained on the original SBU training set [denoted as SDCM(a)] and our SILT relabeled training set [denoted as SDCM(b)]. We conduct the experiments three times, and each time, we randomly select a new test set of the size of 100.

The results are shown in Table 8, where we can observe that (i) from (c), SDCM achieves lower BER and balanced BER_S and BER_{NS} with a smaller well-labeled training set, supporting our hypothesis; (ii) the comparison among SDCM (a-c) shows that our SILT-relabeled training set improves shadow detection performance and indicates the high quality of our SILT-relabeled dataset;

5. Conclusion

We revisit the shadow detection task by considering noise in the shadow labels, and design SILT, a novel Shadow-aware Iterative Label Tuning framework, to enable effective model training on data with noisy labels. Technically, we present a label tuning strategy to encourage the

Table 8. Comparison results of the state-of-the-art (SOTA) method [81] trained on three different training sets: (a) the original SBU training set, (b) our SILT relabeled training set, and (c) a subset of manually relabeled SBU test set (538 images). The evaluation was conducted on the rest of the relabeled SBU test set.

Exp. Time	SDCM(a)			SDCM(b)		
	BER	BER_S	BER_{NS}	BER	BER_S	BER_{NS}
1	5.29	8.95	1.63	4.43	3.95	4.91
2	5.44	7.33	3.55	5.27	4.48	6.06
3	5.18	7.85	2.51	4.42	4.27	4.57
AVG.	5.30	8.04	2.56	4.70	4.23	5.18
	SDCM(c)			Ours		
	BER	BER_S	BER_{NS}	BER	BER_S	BER_{NS}
1	4.39	4.68	4.10	3.99	4.38	3.60
2	5.73	5.21	6.25	4.34	3.54	5.15
3	4.26	4.34	4.18	3.61	3.79	3.44
AVG.	4.79	4.74	4.84	3.98	3.90	4.06

network to get rid of overfitting and try large refinement on the shadow labels. Also, we design two shadow-specific data augmentation strategies, which add "fake" shadows into the training images to improve the ability of the network to distinguish the shadows and dark objects. Considering the noise in existing datasets, we carefully relabel the test set of SBU [55] for evaluation and conduct various experiments. The experimental results show that with SLIT, even a simple network can achieve state-of-the-art performance on shadow detection.

Limitations. Our SILT may fail to distinguish the shadow regions and dark non-shadow regions that have very similar colors and textures to the shadow regions. This is a very challenging issue in the shadow detection task, which detects shadow regions from single images. In the future, we aim to explore depth information to more accurately detect shadows in 3D space, since the shadows are usually projected on the 2D background.

Acknowledgements

References

[1] Affinity photo for iPad. https://affinity.serif.com/en-us/photo/ipad/. 5

[2] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *IEEE International Conference on Computer Vision*, pages 1431–1439, 2015. 3

[3] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5611–5620, 2020. 1, 2, 6, 7

[4] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624, 2020. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4, 6

[6] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. ARGAN: Attentive recurrent generative adversarial network for shadow detection and removal. In *IEEE International Conference on Computer Vision*, pages 10213–10222, 2019. 2

[7] Xianyong Fang, Xiaohao He, Linbo Wang, and Jianbing Shen. Robust shadow detection by exploring effective shadow contexts. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2927–2935, 2021. 2

[8] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI Conference on Artificial Intelligence*, volume 31, 2017. 3

[9] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations*, 2016. 3

[10] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Single-image shadow detection and removal using paired regions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2033–2040, 2011. 2

[11] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[12] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *IEEE International Conference on Computer Vision*, pages 5138–5147, 2019. 3

[13] Le Hou, Tomás F. Yago Vicente, Minh Hoai, and Dimitris Samaras. Large scale shadow annotation and detection using lazy annotation and stacked CNNs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1337–1351, 2021. 1

[14] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2795–2808, 2019. 2, 6

[15] Xiaowei Hu, Tianyu Wang, Chi-Wing Fu, Yitong Jiang, Qiong Wang, and Pheng-Ann Heng. Revisiting shadow detection: A new benchmark dataset for complex world. *IEEE Transactions on Image Processing*, 30:1925–1934, 2021. 2, 6, 7

[16] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7454–7462, 2018. 2, 6, 7

[17] Xiang Huang, Gang Hua, Jack Tumblin, and Lance Williams. What characterizes a shadow boundary under the sun and sky? In *IEEE International Conference on Computer Vision*, pages 898–905, 2011. 2

[18] Leiping Jie and Hui Zhang. RMLANet: Random multi-level attention network for shadow detection. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2022. 2

[19] Hang Jin and Yanming Feng. Towards an automatic road lane marks extraction based on isodata segmentation and shadow detection from large-scale aerial images. In *Proceedings of the 24th International Federation of Surveyors*, pages 1–12, 2010. 4

[20] Salman Hameed Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic feature learning for robust shadow detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1939–1946, 2014. 2

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[22] Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *European Conference on Computer Vision*, pages 322–335, 2010. 2

[23] Hieu Le, Tomás F. Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. A+D Net: Training a shadow detector with adversarial shadow attenuation. In *European Conference on Computer Vision*, pages 662–678, 2018. 2

[24] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference on Machine Learning*, volume 3, page 896. Atlanta, 2013. 3

[25] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. *Advances in Neural Information Processing Systems*, 32, 2019. 3, 4

[26] Yinlin Li, Lihao Jia, Zidong Wang, Yang Qian, and Hong Qiao. Un-supervised and semi-supervised hand segmentation in egocentric images with noisy label learning. *Neurocomputing*, 334:11–24, 2019. 3

[27] 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 *European Conference on Computer Vision*, pages 740–755, 2014. 4

[28] Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2606–2616, 2022. 3

[29] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 6

[30] Yaoru Luo, Guole Liu, Yuanhao Guo, and Ge Yang. Deep neural networks learn meta-structures from noisy labels in semantic segmentation. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 1908–1916, 2022. 3

[31] Yueming Lyu and Ivor W Tsang. Curriculum loss: Robust learning and generalization against label corruption. *arXiv preprint arXiv:1905.10045*, 2019. 3

[32] Robert Mendel, Luis Antonio de Souza, David Rauber, Joao Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *European Conference on Computer Vision*, pages 141–157. Springer, 2020. 3

[33] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2019. 3

[34] Shaobo Min, Xuejin Chen, Zheng-Jun Zha, Feng Wu, and Yongdong Zhang. A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 4578–4585, 2019. 3

[35] Mahyar Najibi, Bharat Singh, and Larry S Davis. Autofocus: Efficient multi-scale inference. In *IEEE International Conference on Computer Vision*, pages 9745–9755, 2019. 4

[36] Vu Nguyen, Tomás F. Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *IEEE International Conference on Computer Vision*, pages 4510–4518, 2017. 2

[37] Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 86–93, 2000. 3

[38] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 3

[39] Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras, and Nikos Paragios. Illumination estimation and cast shadow detection through a higher-order graphical model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 673–680, 2011. 2

[40] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017. 3

[41] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343, 2018. 3

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 1, 2, 5, 7

[43] Elena Salvador, Andrea Cavallaro, and Touradj Ebrahimi. Cast shadow segmentation using invariant color features. *Computer Vision and Image Understanding*, 95(2):238–259, 2004. 2

[44] Li Shen, Teck Wee Chua, and Karianto Leman. Shadow optimization from structured deep edge detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2067–2074, 2015. 2

[45] Yucheng Shu, Xiao Wu, and Weisheng Li. LVC-Net: Medical image segmentation with noisy label based on local visual cues. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 558–566. Springer, 2019. 3

[46] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019. 6

[47] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5552–5560, 2018. 3

[48] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2019. 3

[49] Jiandong Tian, Xiaojun Qi, Liangqiong Qu, and Yandong Tang. New spectrum ratio properties and features for shadow detection. *Pattern Recognition*, 51:85–96, 2016. 2

[50] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[51] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 839–847, 2017. 3

[52] Tomás F. Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection. In *IEEE International Conference on Computer Vision*, pages 3388–3396, 2015. 2

[53] Tomás F. Yago Vicente, Minh Hoai, and Dimitris Samaras. Noisy label recovery for shadow detection in unfamiliar domains. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3783–3792, 2016. 2

[54] Tomás F. Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):682–695, 2018. 2

[55] Tomás F. Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *European Conference on Computer Vision*, pages 816–832, 2016. 1, 2, 4, 5, 6, 7, 9

[56] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 1, 2, 5

[57] Tianyu Wang, Xiaowei Hu, Chi-Wing Fu, and Pheng-Ann Heng. Single-stage instance shadow detection with bidirectional relation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–11, June 2021. 2

[58] Tianyu Wang, Xiaowei Hu, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection with a single-stage detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2022. 2

[59] Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 2

[60] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 6, 7

[61] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *IEEE International Conference on Computer Vision*, pages 322–330, 2019. 3

[62] Wen Wu, Kai Zhou, Xiao-Diao Chen, and Jun-Hai Yong. Light-weight shadow detection via gcn-based annotation strategy and knowledge distillation. *Computer Vision and Image Understanding*, 216:103341, 2022. 2

[63] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*, 2020. 3

[64] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015. 3

[65] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 1, 3, 4

[66] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 6

[67] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, 1995. 3

[68] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7025, 2019. 3

[69] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *IEEE International Conference on Computer Vision*, pages 8229–8238, 2021. 3

[70] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Cicarrelli, Frederik Barkhof, and Daniel Alexander. Disentangling human error from ground truth in segmentation of medical images. *Advances in Neural Information Processing Systems*, 33:15750–15762, 2020. 3

[71] Minqing Zhang, Jiantao Gao, Zhen Lyu, Weibing Zhao, Qin Wang, Weizhen Ding, Sheng Wang, Zhen Li, and Shuguang Cui. Characterizing label errors: confident learning for noisy-labeled image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–730. Springer, 2020. 3

[72] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[73] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W.H. Lau. Distraction-aware shadow detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2019. 1, 2, 6, 7

[74] Kai Zhou, Wen Wu, Yan-Li Shao, Jing-Long Fang, Xing-Qi Wang, and Dan Wei. Shadow detection via multi-scale feature fusion and unsupervised domain adaptation. *Journal of Visual Communication and Image Representation*, 88:103596, 2022. 2

[75] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Robust curriculum learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*, 2020. 3

[76] Xiong Zhou, Xianming Liu, Junjun Jiang, Xin Gao, and Xiangyang Ji. Asymmetric loss functions for learning with noisy labels. In *International Conference on Machine Learning*, pages 12846–12856, 2021. 3

[77] Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Learning with noisy labels via sparse regularization. In *IEEE International Conference on Computer Vision*, pages 72–81, 2021. 3

[78] Jiejie Zhu, Kegan G.G. Samuel, Syed Z. Masood, and Marshall F. Tappen. Learning to recognize shadows in monochromatic natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 223–230, 2010. 1, 2, 4, 5, 6

[79] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *European Conference on Computer Vision*, pages 121–136, 2018. 2, 6, 7

[80] Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson WH Lau. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *IEEE International Conference on Computer Vision*, pages 4702–4711, 2021. 1, 2, 6, 7

[81] Yurui Zhu, Xueyang Fu, Chengzhi Cao, Xi Wang, Qibin Sun, and Zheng-Jun Zha. Single image shadow detection via complementary mechanism. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6717–6726, 2022. 1, 2, 6, 7, 9