SDDNet: Style-guided Dual-layer Disentanglement Network for Shadow Detection

Runmin Cong*
rmcong@sdu.edu.cn
Shandong University
Jinan, Shandong, China

Yuchen Guan
gyc23@mails.tsinghua.edu.cn
Beijing Jiaotong University
Beijing, China

Jinpeng Chen†
jinpechen2-c@my.cityu.edu.hk
City University of Hong Kong
Hong Kong SAR, China

Wei Zhang*
davidzhang@sdu.edu.cn
Shandong University
Jinan, Shandong, China

Yao Zhao
yzhao@bjtu.edu.cn
Beijing Jiaotong University
Beijing, China

Sam Kwong
cssamk@cityu.edu.hk
City University of Hong Kong &
Lingnan University
Hong Kong SAR, China

ABSTRACT

Despite significant progress in shadow detection, current methods still struggle with the adverse impact of background color, which may lead to errors when shadows are present on complex backgrounds. Drawing inspiration from the human visual system, we treat the input shadow image as a composition of a background layer and a shadow layer, and design a Style-guided Dual-layer Disentanglement Network (SDDNet) to model these layers independently. To achieve this, we devise a Feature Separation and Recombination (FSR) module that decomposes multi-level features into shadow-related and background-related components by offering specialized supervision for each component, while preserving information integrity and avoiding redundancy through the reconstruction constraint. Moreover, we propose a Shadow Style Filter (SSF) module to guide the feature disentanglement by focusing on style differentiation and uniformization. With these two modules and our overall pipeline, our model effectively minimizes the detrimental effects of background color, yielding superior performance on three public datasets with a real-time inference speed of 32 FPS. Our code is publicly available at: *https://github.com/rmcong/SDDNet_ACMMM23*.

CCS CONCEPTS

• **Computing methodologies** → **Image segmentation**.

KEYWORDS

shadow detection, feature disentanglement, style constraint

---

*Runmin Cong and Wei Zhang are also affiliated with the Key Laboratory of Machine Intelligence and System Control, Ministry of Education, Jinan, Shandong, China.
†Corresponding author.

1 INTRODUCTION

Shadows are a ubiquitous illumination phenomenon resulting from the linear propagation of light, which can negatively affect tasks such as object detection [12, 59]. Accurate shadow detection can provide valuable insights into scene geometry [31, 45] and light source positioning [34, 46], thereby enhancing scene understanding. As a result, shadow detection has become a foundational task in computer vision, attracting increasing attention in recent years.

Early shadow detection efforts mainly relied on physical methods, constructing physical illumination models [19, 20] and employing hand-crafted features, *e.g.*, illumination cues [15, 47] and texture [51, 64]. However, with the remarkable performance of convolutional neural networks (CNNs) [1, 4–11, 23, 28, 38, 39, 44, 57, 58, 60, 61, 63] in computer vision tasks, deep learning-based methods [2, 3, 14, 26, 40, 46, 54, 62, 66] have progressively become the mainstream of shadow detection. Previous deep learning-based shadow detection methods predominantly focused on guidance from location, semantics, and context perspectives, without attaching importance to the detrimental influence of background color. Consequently, these detectors tend to associate shadow features with dark colors, leading to incorrect detection results in some complex scenes. Specifically, errors can be divided into two categories: 1) weak shadow regions on light-colored backgrounds (*e.g.*, the first example in Figure 1), which are wrongly detected as non-shadow regions, and 2) dark-colored background areas (*e.g.*, the second example in Figure 1), which are often misclassified as shadows. In summary, detection results are greatly affected by background color, and similar shadows in different backgrounds may yield completely disparate detection outcomes. For instance, detecting a shadow is relatively simple when it forms on a ground composed entirely of light-colored bricks, but becomes challenging when the shadow appears on a ground with alternating dark and light bricks, as shown in the third and fourth examples of Figure 1. In these scenarios, the shadow on the light-colored brick is difficult to detect, while the dark brick is prone to being misidentified as a shadow, irrespective of whether it is in shadow or not.

Figure 1: Some difficult cases in shadow detection. (a) The input images. (b) The ground truth shadow maps. (c) The predicted results of ECA [14]. (d) The predicted results of MTMT-Net [3]. (e) The predicted results of our SDDNet.

Since the shadows are created by light being blocked, they are inherently colorless. Based on this, humans can discern shadows on complex backgrounds through a three-step process: first, recognizing background attributes; second, identifying shadow attributes by observing confident shadow regions; and finally, detecting all shadow regions based on our understanding of background and shadow attributes. Inspired by this process, we propose treating shadow images as a composition of background and shadow layers, and modeling them separately to effectively reduce the impact of background color on detection performance. This objective can be accomplished through the strategy of feature disentanglement, which is proved to be effective in many computer vision works. For instance, [56] conducted an early trial to incorporate task-feature co-disentanglement regularizations for multi-task learning and achieved satisfactory performance.

In this paper, we introduce the concept of feature disentanglement into shadow detection to realize the separated modeling of background and shadow layers. We present a novel Style-guided Dual-layer Disentanglement Network (SDDNet) featuring two innovative modules, *i.e.*, the Feature Separation and Recombination (FSR) module and the Shadow Style Filter (SSF) module. The FSR module effectively decomposes multi-level features into background-related and shadow-related components, which is explicitly achieved by providing distinct supervision for each set of components. The shadow-related component receives supervision from the ground-truth shadow map, while the background-related component is guided to generate a shadow-free background image. Furthermore, to ensure information integrity, the recombined features merged from both components are supervised to reconstruct the input image. During prediction, only the shadow-related component is utilized to generate the final shadow map, effectively eliminating

the adverse influence of the background. Additionally, to further constrain the FSR module on feature disentanglement, particularly for background-related component that lacks a background ground-truth, we propose a Shadow Style Filter (SSF) module to extract and constrain style attributes of the shadow-related component, background-related component, and recombined features. Specifically, we regard the presence or absence of shadows as a style. From this perspective, the recombined features and shadow-related component should have consistent styles, while the background-related component should exhibit a different style from them. Based on this principle, we can generate background images in an indirect style-guided manner, thereby facilitating feature disentanglement within the FSR module.

In summary, our contributions are primarily three-fold:

- We model the shadow image as a superposition of shadow layers on background layers, and then propose a Style-guided Dual-layer Disentanglement Network (SDDNet) for shadow detection. Extensive experiments on three public datasets demonstrate that our proposed method outperforms all state-of-the-art shadow detection methods with a real-time inference speed of 32 FPS.
- We design a Feature Separation and Recombination (FSR) module to decompose image features into shadow-related and background-related components, thereby preventing predictions from being misled by background information.
- We devise a Shadow Style Filter (SSF) module that assists feature separation through style differentiation and uniformization, especially to help generate the background-related component in an indirect style-guided manner.

2 RELATED WORKS

2.1 Shadow Detection
Related works on shadow detection can be broadly categorized into traditional methods and deep learning-based methods.

Early efforts in shadow detection primarily focused on constructing physical illumination models [15, 16, 19, 20, 27, 33, 35, 47, 48] to analyze the shadow formation process. Based on these models, shadows were detected either by using physical models [15, 16] or by employing traditional machine learning-based detectors with hand-crafted features, such as illumination cues [15, 19, 20, 47], texture [20, 64], and edge [27, 33]. Although these methods led to improvements, most of them relied on assumptions (*e.g.*, fixed background classes, uniform illumination, *etc.*) that are difficult to satisfy in complex situations. Additionally, the hand-crafted features may not be discriminative enough for detecting intricate shadow regions.

Inspired by the outstanding performance of CNN in computer vision tasks, deep learning-based methods [2, 3, 14, 22, 25, 26, 40–42, 46, 54, 62, 62, 66] have gained popularity in shadow detection. With their ability to extract and select discriminative features, CNNs are more robust than traditional methods that use hand-crafted features. Khan *et al.* [46] were the first to apply CNNs to shadow detection, extracting features from superpixels using a 7-layer CNN and feeding these features to a conditional random field (CRF) model to refine the detection results. Zheng *et al.* [62] integrated the semantics of distraction regions to extend CNNs for robust shadow detection. Some researchers [22, 36, 42, 43] employed generative

adversarial networks (GAN) [18] for shadow detection. Recently, Chen *et al.* [3] proposed a semi-supervised teacher-student framework to detect shadow regions, edges, and count under consistency constraints. Zhu *et al.* [66] introduced a feature reweighting method to balance the intensity-variant and intensity-invariant features obtained by self-supervised decomposition. Liao *et al.* [40] incorporated confidence maps into shadow detection and combined the prediction results of multiple methods for shadow detection.

Despite the significant improvements offered by these methods, they still suffer from background color interference. This interference causes confusion between dark background areas and shadow regions, as well as between light background areas and weak shadow regions. In this study, we disentangle background-related and shadow-related components, utilizing only the shadow-related component to predict the final results. This approach enhances the robustness of shadow detection in complex scenes.

2.2 Style Transfer

In the domain of neural style transfer, research has been conducted to comprehend the content and style of image features. Gatys *et al.* [17] proposed utilizing the Gram matrix of image features as a means to encapsulate the distinctive style of an image. Subsequent studies [30, 37] have further corroborated the efficacy of the Gram matrix in capturing and representing image styles.

In this paper, we employ the Gram matrix to extract style attributes and regulate the consistency or diversity of these attributes across various features and components of the input shadow image. This approach serves to bolster our feature disentanglement process, ultimately leading to enhanced outcomes.

3 PROPOSED METHOD

3.1 Overview

In Figure 2, we present the overall framework of our SDDNet, which adopts an encoder-decoder architecture. During training, SDDNet generates the shadow map, background image, and reconstructed input image; however, only the shadow map is predicted during the inference stage. The generation of reconstructed and background images constitutes our joint training strategy with the main aim of improving the quality of feature disentanglement.

To elaborate, we initially input the image into the backbone network to extract multi-level features $\{F_k\}_{k=1}^N$, where N represents the number of levels. To fully exploit the detail and global semantics, we divide the features into two groups: the low-level group $F_{low} = \{F_k\}_{k=1}^{N_{low}}$ and the high-level group $F_{high} = \{F_k\}_{k=N_{low}+1}^N$. We process these two groups in two paths with the same structure, omitting the subscripts *low* and *high* for simplicity. The features in each group are concatenated together after upsampling to unify the spatial sizes, generating merged features $\hat{F}$. Subsequently, the FSR module is fed $\hat{F}$ and outputs the shadow-related component $\hat{F}^{sd}$, the background-related component $\hat{F}^{bg}$, and the recombined features $\hat{F}^{re}$. The SSF module then extracts style attributes from $\hat{F}^{sd}$, $\hat{F}^{bg}$, and $\hat{F}^{re}$, and constrains the consistency or diversity of specific style attribute pairs to guide the upstream feature separation. Finally, in the parallel decoder, the shadow-related component, background-related component, and recombined features from both

paths are fused separately, and then generate the shadow map, the background image, and the reconstructed input image.

3.2 Feature Separation and Recombination Module

From the perspective of the human visual system, shadow images can be considered as shadows of other objects superimposed on the background image. This kind of dual-layer separation is not a difficult task for humans, but it is not a simple matter for computers. Therefore, we aim to emulate this way of perceiving shadow images in a bio-inspired manner, and thereby achieve the disentanglement of shadow image content/features. Effective feature disentanglement can promote the focus on more informative components for shadow detection.

One of the main challenges lies in the complex coupling between background and shadow images, making it extremely difficult to model the relationship with complete accuracy. To simplify this process, we model it as a straightforward linear model, which is also a relatively intuitive modeling approach. Nevertheless, to achieve accurate disentanglement within this simple linear model, we design comprehensive strategies based on differentiated supervision. However, differentiated supervision presents its own challenge. Specifically, we only have ground truth shadow maps and lack labels for shadow-free background images, which means that we cannot accomplish our goal solely through direct supervision. Instead, we must find ingenious indirect supervision methods. To this end, in addition to shadow image supervision, we also incorporate joint supervision and style supervision (introduced in the following section). In this manner, the generation of background images can be supervised indirectly, thereby improving the overall feature disentanglement process.

To accomplish this objective, we design the FSR module to achieve feature disentanglement and reorganization through a shadow branch and a background branch. Each branch consists of a residual block [21], which comprises two convolutional layers and a skip connection. Given $\hat{F}$, the shadow branch produces the shadow-related component $\hat{F}^{sd}$, and the background branch generates the background-related component $\hat{F}^{bg}$ as follows:

$$\hat{F}^{sd} = Conv\left(Conv\left(\hat{F}\right)\right) + \hat{F}, \tag{1}$$

$$\hat{F}^{bg} = Conv\left(Conv\left(\hat{F}\right)\right) + \hat{F}, \tag{2}$$

where $Conv$ denotes a convolutional layer. Additionally, we combine them to obtain recombined features $\hat{F}^{re}$:

$$\hat{F}^{re} = \hat{F}^{sd} \oplus \hat{F}^{bg}, \tag{3}$$

where $\oplus$ signifies element-wise addition.

Upon obtaining the outputs of the FSR modules in the low- and high-level paths, $\hat{F}^{sd}_{low}, \hat{F}^{bg}_{low}, \hat{F}^{re}_{low}$ and $\hat{F}^{sd}_{high}, \hat{F}^{bg}_{high}, \hat{F}^{re}_{high}$ (with subscripts restored), they are individually fused in the parallel decoder:

$$F^{sd} = Conv\left(CA\left(concat(\hat{F}^{sd}_{low}, \hat{F}^{sd}_{high})\right)\right), \tag{4}$$

$$F^{bg} = Conv\left(CA\left(concat(\hat{F}^{bg}_{low}, \hat{F}^{bg}_{high})\right)\right), \tag{5}$$

$$F^{re} = Conv\left(CA\left(concat(\hat{F}^{re}_{low}, \hat{F}^{re}_{high})\right)\right), \tag{6}$$

Figure 2: Architecture of the proposed SDDNet. Given an input image, SDDNet outputs the shadow map, background image, and reconstructed image in an end-to-end manner. Firstly, the backbone extracts integrated low-level and high-level features. Then, the proposed FSR module decomposes the features and produce shadow-related component, background-related component, and recombined features. In addition, the SSF module extracts style attributes and guide the feature disentanglement process. Finally, the low-level and high-level features are fused through the parallel decoder to generate three outputs (*i.e.,* background image, shadow map, and reconstructed input image).

where *concat* represents a concatenation operation, and *CA* denotes channel attention [24]. By applying channel attention, the network can automatically select informative channels from both low-level and high-level features while suppressing non-informative channels. Finally, the F^{sd}, F^{bg}, and F^{re} are use to generate the shadow map P^{sd}, the background image P^{bg}, and the reconstructed input image P^{re}, respectively, after passing through a convolutional layer. These processes can be expressed as:

$$P^{sd} = Conv\left(F^{sd}\right), \tag{7}$$

$$P^{bg} = Conv\left(F^{bg}\right), \tag{8}$$

$$P^{re} = Conv\left(F^{re}\right). \tag{9}$$

Although the process for generating these three outputs does not involve distinct operations tailored to their specific targets, we can apply differentiated supervision to enable the network to autonomously learn the optimal way to decompose features. In particular, the supervision for P^{sd} is provided by the ground truth shadow map, while the supervision for P^{re} is derived from the input image. The two losses can be calculated as follows:

$$\mathcal{L}_{sd} = BBCE\left(P^{sd},\ G^{sd}\right), \tag{10}$$

$$\mathcal{L}_{re} = MAE(P^{re},\ I), \tag{11}$$

where G^{sd} represents the ground-truth shadow map, I denotes the input image, and *BBCE* and *MAE* signify the balanced binary cross

entropy and mean absolute error, respectively. Here, we employ the same balanced binary cross entropy as in [66], formulated by:

$$BBCE(P^{sd}, G^{sd}) =$$
$$- \sum_i \left[\frac{N_n}{N} G^{sd} log(P_i^{sd}) + \frac{N_p}{N} (1 - G^{sd}) log(1 - P_i^{sd}) \right], \tag{12}$$

where i denotes the index of spatial locations, N_p and N_n represent the number of shadow and non-shadow pixels, and N corresponds to the total number of pixels. The mean absolute error is given by:

$$MAE(P^{re}, I) = \frac{1}{N} \sum_i |P_i^{re} - I_i|. \tag{13}$$

The supervision for these two outputs is relatively straightforward. However, for P^{bg}, the problem becomes more complex due to the absence of a ground-truth background image. As a result, we use some indirect manners to guide the network learning. In areas without shadows, the input image and the background image are identical, enabling us to directly use the input image to supervise these regions. This can be expressed as:

$$\mathcal{L}_{bg} = MAE\left(P^{bg} \otimes \left(1 - P^{sd}\right),\ I \otimes \left(1 - G^{sd}\right)\right), \tag{14}$$

where $\otimes$ denotes element-wise multiplication. The two terms in this loss correspond to the ground-truth shadow-free regions of the input image and the predicted shadow-free regions of the generated background image. As we employ the predicted shadow-free map $1 - P^{sd}$, $\mathcal{L}_{bg}$ has the advantage of constraining the generation of the

Figure 3: Structure of the SSF module. The Gram matrix is used to extract style attributes of the background-related component, the shadow-related component, and the recombined features. Based on the presence or absence of shadows, we aim to bring the style of the shadow-related component closer to that of the recombined features, while differentiating it with that of the background-related component.

background image while simultaneously aiding the prediction of the shadow map. Through $\mathcal{L}_{bg}$, we offer guidance to the network for predicting the shadow-free region of the background image. Nevertheless, to predict the complete background image and thereby enhance the quality of disentangling the background-related component, we also need to provide guidance for the shadowed areas. This aspect is accomplished through our SSF module, which will be discussed in Section 3.3.

3.3 Shadow Style Filter Module

In Section 3.2, we decompose the integrated features into background-related and shadow-related components using the proposed FSR module with a differentiated supervision strategy. However, there are two imperfections: 1) The supervision of the background-related component is insufficient, as it only involves shadow-free regions, leading to a lack of guidance for generating shadowed regions. 2) It does not further emphasize the differences between the shadows and the background, which may results in unclear boundaries for isolating different components, making them less pure.

To address these issues, we consider incorporating style guidance into our method, as the presence or absence of shadows inherently represents a common style attribute. Following this idea, we design the SSF module, as depicted in Figure 3. It extracts style attributes from each of the three outputs from the FSR module (*i.e.*, $\hat{F}^{sd}$, $\hat{F}^{bg}$, and $\hat{F}^{re}$), and then constrains the consistency and diversity between different style pairs in a contrastive learning fashion.

For the style attribute extraction, we adopt the Gram matrix [17] of the feature map as the style representation. For the input features $\hat{F} \in \mathbb{R}^{C \times H \times W}$, the Gram matrix $M \in \mathbb{R}^{C \times C}$ captures correlations between its channels, which can be computed as follows:

$$M_{x,y} = \hat{F}_x^T \hat{F}_y, \tag{15}$$

where $M_{x,y}$ denotes the (x, y) element of Gram matrix M, and $\hat{F}_x$ and $\hat{F}y$ represent the xth and y^{th} channels of $\hat{F}$, respectively. Subsequently, we employ two consecutive linear layers to further

extract the style attribute $\rho \in \mathbb{R}^{C^2}$ from $M_{x,y}$:

$$\rho = Linear(Linear(Flatten(M))), \tag{16}$$

where *Linear* signifies a linear layer, and *Flatten* indicates a flatten operation. For the input components and features, $\hat{F}^{sd}$, $\hat{F}^{bg}$, and $\hat{F}^{re}$, the extracted style attribute vectors are denoted as ρ^{sd}, ρ^{bg}, and ρ^{re}, respectively.

In our approach, the primary style consideration is the presence or absence of shadows. From this perspective, the style of the shadow-related component and recombined features should be consistent, as they collectively represent the existence of shadows. To achieve this, we employ the following loss function to enhance their consistency:

$$\mathcal{L}^{con} = 1 - cos\left(\rho^{sd}, \rho^{re}\right) = 1 - \frac{\rho^{sd} \cdot \rho^{re}}{|\rho^{sd}||\rho^{re}|}, \tag{17}$$

in which *cos* denotes the cosine similarity. Reducing $\mathcal{L}^{con}$ is equivalent to increasing the cosine similarity, which in turn improves the consistency between ρ^{sd} and ρ^{re}.

Conversely, the styles of the shadow-related component and background-related component ought to be distinct, as the latter embodies a shadow-free style. To augment their difference, we employ the subsequent differentiate loss:

$$\mathcal{L}^{diff} = \frac{(\rho^{re} \cdot \rho^{bg})^2}{C^2}, \tag{18}$$

A smaller $\mathcal{L}^{diff}$ signifies that the two vectors are more orthogonal, meaning the difference between them is larger.

The comprehensive style constraint loss, denoted as $\mathcal{L}_{style}$, encompasses the similarity and diversity losses from both low-level and high-level pathways. This loss can be computed using the following equation:

$$\mathcal{L}_{style} = \mathcal{L}_{low}^{con} + \mathcal{L}_{low}^{diff} + \mathcal{L}_{high}^{con} + \mathcal{L}_{high}^{diff}. \tag{19}$$

The two constraints in the SSF module enable the two linear layers to extract the style related to the presence or absence of shadows from the Gram matrix more effectively. As the presence or absence of shadows serves as the decisive factor for the diversity or consistency of the two style attribute pairs, if the linear layers were to focus on other styles, the diversity or consistency would not be adequately captured. Thus, the process of back-propagation encourages the linear layers to concentrate on the shadow aspect. With this premise, the constraint that differentiates background-related and shadow-related components fosters the formation of distinctly different characteristics between them. This ensures that the information they contain is not easily duplicated, supporting the feasibility of our dual-layer modeling approach. More importantly, when combined with the shadow-free region constraint described in Section 3.2, the network gains the ability to separate background component without requiring a ground-truth background image, which in turn refines the shadow-related component.

3.4 Overall Loss Function

The overall loss function of our method is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{sd} + \alpha(\mathcal{L}_{re} + \mathcal{L}_{bg}) + \beta\mathcal{L}_{style}, \tag{20}$$

where α and β are two balancing hyperparameters, which are empirically set to $\alpha = 0.2$ and $\beta = 0.1$, respectively.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metric

4.1.1 Datasets. We evaluate our method on three public datasets: SBU [52], ISTD [53], and UCF [64]. The SBU dataset comprises 4,089 training images and 638 testing images. The ISTD dataset contains 1,330 training images and 540 testing images. Although it provides ground truths for both shadow maps and shadow-free images, we only use the ground truths for shadow maps in our task. The UCF dataset consists of 135 training images and 110 testing images. Following previous shadow detection works [3, 25, 40, 62, 65, 66], we evaluate our method on both the SBU and UCF test sets using the model trained on the SBU training images, and on the ISTD testing set using the model trained on its own training set.

4.1.2 Evaluation Metrics. We following previous shadow detection works [43, 66, 67] to adopt the widely-used metric, balanced error rate (BER), to quantitatively evaluate performance:

$$BER = \left(1 - \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)\right) \times 100, \quad (21)$$

where TP, TN, FP, and FN represent the numbers of true positive, true negative, false positive, and false negative pixels, respectively. BER considers error rates for both shadow and non-shadow regions, with lower values indicating better performance. Additionally, we also report the error rate for the shadow region, $1 - \frac{TP}{TP+FN}$, and the error rate for the non-shadow region, $1 - \frac{TN}{TN+FP}$.

4.2 Implementation Details

For the backbone, we adopt the lightweight EfficientNet-B3 [49] as in [66, 67] and initialize it with pre-trained parameters from ImageNet [13]. EfficientNet-B3 comprises 25 consecutive blocks, with the output of the first 6 layers serving as our low-level features and the output of the remaining layers as the high-level features. Our code is implemented with PyTorch and accelerated by a single NVIDIA RTX2080Ti. We also implement our network by using the MindSpore Lite tool[1].

During both training and inference, we resize the input image to 512×512. In the training stage, we optimize the entire network for 20 epochs with a batch size of 4, using the Adam optimizer. The initial learning rate is set to 0.0005. The learning rate is adjusted using the exponential decay strategy, with a decay rate of 0.7. During testing stage, we employ a fully connected conditional random field (CRF) [32] to further refine our predicted shadow map, following the approaches in [3, 25, 40, 62, 65, 66]. Our proposed model has a real-time inference speed of 32 FPS for processing an image with the size of 512×512.

4.3 Comparisons

We compare our method with 15 previous state-of-the-art shadow detection methods both quantitatively and qualitatively, the methods we select include Unary-Pariwise [19], scGAN [43], ST-CGAN [53], DC-DSPF [55], A+D Net [36], BDRAR [65], DSDNet [62], DSC

[25], MTMTNet [3], ECA [14], RCMPNet [40], FDRNet [66], CM-Net [67], TransShadow [29], and R2D [50]. Among them, Unary-Pariwise is based on hand-crafted features, while all the others are CNN-based methods. For a fair comparison, all the results are provided directly by the authors or generated by the source codes under the default parameter settings in the corresponding models.

4.3.1 Quantitative Comparison. We present the quantitative comparison results between our SDDNet and other models in Table 1. It is clear that our model is highly competitive among all these methods, securing either the first place or a tie for first place in terms of BER across all three datasets. This achievement demonstrates our model's ability to handle data with diverse characteristics and deliver satisfactory outcomes. In comparison to the previously best-performing CM-Net [67], our model exhibits an equal BER on the SBU dataset, while outperforming it by 11.81% and 2.08% on the ISTD and UCF datasets, respectively. Additionally, in the comparison of error rates within shadow and non-shadow regions, our model exhibits a consistently stable performance, ranking among the top positions across all three datasets.

4.3.2 Qualitative Comparison. We also qualitatively compare the results of our model with those of previous models, as illustrated in Figure 4. It can be observed that the results of our model exhibit advantages, particularly in scenes with confusing background colors. For instance, in the first example, the dark eyes and hair of the cartoon character might be misclassified as shadows by other models; however, our SDDNet can effectively mitigate this interference due to its dual-layer modeling. Likewise, in the second and third examples, dark objects or dark ground may be erroneously identified as shadows by other networks, whereas our model prevents this error from occurring. Moreover, in the fourth and fifth examples, other models may misidentify shadows on light-colored backgrounds as non-shadows due to the varying background colors covered by the shadows. In contrast, our model avoids this interference as the shadow-related component utilized for prediction do not incorporate any background information.

4.4 Ablation Study

To verify the effect of each part in our model, we conduct ablation studies on the SBU dataset with the following configurations:

- *Baseline*: Compared with the full model introduced in Section 3, we remove the FSR module and the SSF module.
- *Baseline+FSR*: Compared with *Baseline*, we add the FSR module.
- *Baseline+FSR**: Compared with *Baseline+FSR*, we remove the joint training of generating the background image and reconstructing the input image, namely remove $\mathcal{L}_{joint}$.
- *Baseline+FSR+SSF*: Compared with *Baseline+FSR*, we add the SSF module.

The quantitative results with all these different configurations are reported in Table 2. We also present the quantitative results for several different configurations in Figure 5.

4.4.1 Effectiveness of the FSR module. In this part, we showcase the effectiveness of our proposed FSR module by comparing its performance to the results obtained without its implementation. The

---

[1]https://www.mindspore.cn/

Table 1: Quantitative comparison results between our SDDNet and existing state-of-the-art methods. "Shad." and "No Shad." denote the error rates of shadow and non-shadow regions, respectively. Bold indicates the best performances, and <u>underline</u> indicates the second best performances.

Model	Source	ISTD [53]			SBU [52]			UCF [64]		
		BER↓	Shad.↓	No Shad.↓	BER↓	Shad.↓	No Shad.↓	BER↓	Shad.↓	No Shad.↓
Unary-Pariwise [19]	CVPR'11	-	-	-	25.03	36.26	13.80	-	-	-
scGAN [43]	ICCV'17	4.70	3.22	6.18	9.10	8.39	9.69	11.50	7.74	15.30
ST-CGAN [53]	CVPR'18	3.85	2.14	5.55	8.14	3.75	12.53	11.23	4.94	17.52
DC-DSPF [55]	IJCAI'18	-	-	-	4.00	4.70	5.10	7.90	6.50	9.30
A+D Net [36]	ECCV'18	-	-	-	5.37	4.45	6.30	9.25	8.37	10.14
BDRAR [65]	ECCV'18	2.69	0.50	4.87	3.64	3.40	3.89	7.81	9.69	5.44
DSDNet [62]	CVPR'19	2.17	1.36	2.98	3.45	3.33	3.58	7.59	9.74	5.44
DSC [25]	TPAMI'19	3.42	3.85	3.00	5.59	9.76	1.42	10.54	18.08	3.00
MTMT-Net [3]	CVPR'20	1.72	1.36	2.08	3.15	3.73	2.57	7.47	10.31	4.63
ECA [14]	MM'21	2.03	2.88	1.19	5.93	10.82	1.03	10.71	18.59	2.83
RCMPNet [40]	MM'21	1.61	1.22	2.00	2.98	3.26	2.69	6.75	8.36	5.75
FDRNet [66]	ICCV'21	1.55	1.22	1.88	3.04	2.91	3.18	7.28	8.31	6.26
CM-Net [67]	MM'22	<u>1.44</u>	-	-	**2.94**	-	-	<u>6.73</u>	-	-
TransShadow [29]	ICASSP'22	1.73	-	-	3.17	-	-	6.95	-	-
R2D [50]	WACV'23	1.69	0.59	2.79	3.15	2.74	3.56	6.96	8.32	5.60
Ours	/	**1.27**	1.01	1.52	**2.94**	3.23	2.64	**6.59**	7.89	5.29

Figure 4: Qualitative comparison between our SDDNet and existing state-of-the-art methods. (a) Input images. (b) Ground-truths. (c) The prediction of BDRAR [65]. (d) The prediction of DSDNet [62]. (e) The prediction of MTMT-Net [3]. (f) The prediction of FDRNet [66]. (g) The prediction of ECA [14]. (h) The prediction of CM-Net [67]. (i) The prediction of our SDDNet.

FSR module allows for independent modeling of shadow and background layers, efficiently reducing the adverse effects of confounding background colors. By comparing the performance of *Baseline* and *Baseline+FSR*, it is evident that the FSR module improves the

Table 2: Ablation study results for our SDDNet. Bold indicates the best performances.

Configuration	FSR	$\mathcal{L}_{joint}$	SSF	BER↓
Baseline				3.39
Baseline+FSR	✓	✓		3.29
Baseline+FSR*	✓			3.32
Baseline+FSR+SSF	✓	✓	✓	**2.94**

prediction accuracy. Compared to the scenario without the FSR module, the BER score improves from 3.39 to 3.29, with the percentage gain of 2.9%. As illustrated in Figure 5, when the backgrounds in shadowed areas (the first example) or non-shadowed areas (the second example) display various distinct characteristics, *Baseline* struggles to eliminate such interference. It predicts the light yellow line in shadows as non-shadow and misclassifies the non-shadow dark part of the red clay court as shadows. In contrast, *Baseline+FSR* performs better, as its isolated shadow-related components can mitigate the impact of background colors to some extent, achieving improved predictions. However, there are still noticeable discrepancies between the result and the ground truth, indicating that the absence of clear guidance for disentangling background-related components hinders the feature disentanglement from achieving complete success.

Additionally, we conduct experiments focusing on the joint training strategy in the FSR module, specifically the $\mathcal{L}_{joint}$ term in the loss function. This joint training serves two purposes. Firstly, it constrains the reconstruction of the input image, ensuring that the information within the two isolated components is neither omitted nor redundant. Secondly, it constrains the generation of the background image, encouraging the production of the background-related component, thereby more effectively eliminating the interference of background information from shadow-related components. Compared with *Baseline+FSR* and *Baseline+FSR**, *Baseline+FSR* that incorporates joint training yields superior results, with a BER improvement of 0.03. This observation demonstrates the significance of joint training, and both of its functions are essential for achieving high-quality feature disentanglement.

4.4.2 Effectiveness of the SSF module. Furthermore, we compare the performance of our model with and without the proposed SSF module. This module constrains feature disentanglement, taking into account both style diversity and consistency. By examining the results of *Baseline+FSR* and *Baseline+FSR+SSF* in Table 2, it is evident that adding the SSF module yields considerably improved results, with a 0.35 higher BER. This suggests that the style constraints within the SSF module indeed enhance the ability to more effectively separate the two feature components, thus simplifying the prediction of shadow maps. In the absence of the SSF module, disentangling background-related component proves challenging, as ground truth background images are unavailable. However, the SSF module ingeniously addresses this issue by diversifying the styles of background features and shadow-related components in a weakly supervised manner.

In Figure 5, we can also observe the superiority brought by the SSF module. The predictions of *Baseline+FSR+SSF* demonstrate an

Figure 5: The qualitative results of the ablation study. (a) Input images. (b) Ground-truths. (c) The prediction of *Baseline*. (d) The prediction of *Baseline+FSR*. (e) The prediction of *Baseline+FSR+SSF*.

obvious advantage over *Baseline+FSR*, exhibiting a clear improvement in handling complex backgrounds and fully mitigating the impact of confounding background colors. Consequently, both the FSR and SSF modules are indispensable for obtaining stable and robust prediction results. They need to coordinate with each other in order to maximize their effectiveness.

5 CONCLUSION

In this paper, we present a novel Style-guided Dual-layer Disentanglement Network (SDDNet) for shadow detection. Our central idea is to separate the shadow and background layers of the input image to reduce the impact of background color. To achieve this goal, we introduce two novel modules. The first one is the Feature Separation and Recombination (FSR) module that separates complete features into shadow-related and background-related components using differentiated supervisions. Simultaneously, the joint training strategy of reconstructing the input image and generating the background image ensures the reliability of the separation process. Furthermore, we consider the presence and absence of shadows as a type of style and introduce style constraints to our model through a Shadow Style Filter (SSF) module, further enhancing the quality of feature disentanglement. Experimental results on three datasets demonstrate that our SDDNet achieves state-of-the-art performance, proving the effectiveness of our approach.

REFERENCES

[1] Zuyao Chen, Runmin Cong, Qianqian Xu, and Qingming Huang. 2021. DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection. *IEEE Transactions on Image Processing* 30 (2021), 7012–7024.

[2] Zhihao Chen, Liang Wan, Lei Zhu, Jia Shen, Huazhu Fu, Wennan Liu, and Jing Qin. 2021. Triple-cooperative video shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2715–2724.

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

[4] Runmin Cong, Ke Huang, Jianjun Lei, Yao Zhao, Qingming Huang, and Sam Kwong. 2023. Multi-projection fusion and refinement network for salient object detection in 360° omnidirectional image. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[5] Runmin Cong, Qinwei Lin, Chen Zhang, Chongyi Li, Xiaochun Cao, Qingming Huang, and Yao Zhao. 2022. CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection. *IEEE Transactions on Image Processing* 31 (2022), 6800–6815.

[6] Runmin Cong, Qi Qin, Chen Zhang, Qiuping Jiang, Shiqi Wang, Yao Zhao, and Sam Kwong. 2022. A weakly supervised learning framework for salient object detection via hybrid labels. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 2 (2022), 534–548.

[7] Runmin Cong, Weiyu Song, Jianjun Lei, Guanghui Yue, Yao Zhao, and Sam Kwong. 2023. PSNet: Parallel symmetric network for video salient object detection. *IEEE Transactions on Emerging Topics in Computational Intelligence* 7, 2 (2023), 402–414.

[8] Runmin Cong, Haowei Yang, Qiuping Jiang, Wei Gao, Hai-Sheng Li, Cong Wang, Yao Zhao, and Sam Kwong. 2022. BCS-Net: Boundary, context, and semantic for automatic COVID-19 lung infection segmentation from CT images. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–11.

[9] Runmin Cong, Ning Yang, Chongyi Li, Huazhu Fu, Yao Zhao, Qingming Huang, and Sam Kwong. 2022. Global-and-local collaborative learning for co-salient object detection. *IEEE Transactions on Cybernetics* 53, 3 (2022), 1920–1931.

[10] Runmin Cong, Wenyu Yang, Wei Zhang, Chongyi Li, Chun-Le Guo, Qingming Huang, and Sam Kwong. 2023. PUGAN: Physical model-guided underwater image enhancement using GAN with dual-discriminators. *IEEE Transactions on Image Processing* 32 (2023), 4472–4485.

[11] Runmin Cong, Kepu Zhang, Chen Zhang, Feng Zheng, Yao Zhao, Qingming Huang, and Sam Kwong. 2022. Does thermal really always matter for RGB-T salient object detection? *IEEE Transactions on Multimedia* (2022).

[12] Runmin Cong, Yumo Zhang, Leyuan Fang, Jun Li, Yao Zhao, and Sam Kwong. 2022. RRNet: Relational reasoning network with parallel multi-scale attention for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1558–0644.

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

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

[15] Graham D Finlayson, Mark S Drew, and Cheng Lu. 2009. Entropy minimization for shadow removal. *International Journal of Computer Vision* 85, 1 (2009), 35–57.

[16] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. 2005. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 1 (2005), 59–68.

[17] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2414–2423.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

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

[20] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. 2012. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 12 (2012), 2956–2967.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 770–778.

[22] Yingqing He, Yazhou Xing, Tianjia Zhang, and Qifeng Chen. 2021. Unsupervised portrait shadow removal via generative priors. In *Proceedings of the 29th ACM International Conference on Multimedia*. 236–244.

[23] Junkang Hu, Qiuping Jiang, Runmin Cong, Wei Gao, and Feng Shao. 2021. Two-Branch Deep Neural Network for Underwater Image Enhancement in HSV Color Space. *IEEE Signal Process. Lett.* 28 (2021), 2152–2156.

[24] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7132–7141.

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

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

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

[28] Yawen Huang, Feng Zheng, Runmin Cong, Weilin Huang, Matthew R. Scott, and Ling Shao. 2020. MCMT-GAN: Multi-Task Coherent Modality Transferable GAN for 3D Brain Image Synthesis. *IEEE Trans. Image Process.* 29 (2020), 8187–8198.

[29] Leiping Jie and Hui Zhang. 2022. A fast and efficient network for single image shadow detection. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. 2634–2638.

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*. 694–711.

[31] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. 2011. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics* 30, 6 (2011), 1–12.

[32] Philipp Krähenbühl and Vladlen Koltun. 2011. Efficient inference in fully connected CRFs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, Vol. 24.

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

[34] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. 2012. Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision* 98 (2012), 123–145.

[35] Hieu Le and Dimitris Samaras. 2019. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8578–8587.

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

[37] Sohyun Lee, Taeyoung Son, and Suha Kwak. 2022. Fifo: Learning fog-invariant features for foggy scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18911–18921.

[38] Chongyi Li, Runmin Cong, Sam Kwong, Junhui Hou, Huazhu Fu, Guopu Zhu, Dingwen Zhang, and Qingming Huang. 2021. ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. *IEEE Transactions on Cybernetics* 50, 1 (2021), 88–100.

[39] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. 2020. RGB-D salient object detection with cross-modality modulation and selection. In *Proceedings of the European Conference on Computer Vision*. 225–241.

[40] Jingwei Liao, Yanli Liu, Guanyu Xing, Housheng Wei, Jueyu Chen, and Songhua Xu. 2021. Shadow detection via predicting the confidence maps of shadow detection methods. In *Proceedings of the 29th ACM International Conference on Multimedia*. 704–712.

[41] Xiao Lu, Yihong Cao, Sheng Liu, Chengjiang Long, Zipei Chen, Xuanyu Zhou, Yimin Yang, and Chunxia Xiao. 2022. Video shadow detection via spatio-temporal interpolation consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3116–3125.

[42] Junfeng Lyu, Zhibo Wang, and Feng Xu. 2022. Portrait eyeglasses and shadow removal by leveraging 3D synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3429–3439.

[43] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. 2017. Shadow detection with conditional generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4510–4518.

[44] Min Ni, Jianjun Lei, Runmin Cong, Kaifu Zheng, Bo Peng, and Xiaoting Fan. 2017. Color-Guided Depth Map Super Resolution Using Convolutional Neural Network. *IEEE Access* 5 (2017), 26666–26672.

[45] Takahiro Okabe, Imari Sato, and Yoichi Sato. 2009. Attached shadow coding: Estimating surface normals from shadows under unknown reflectance and lighting conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1693–1700.

[46] Alexandros Panagopoulos, Dimitris Samaras, and Nikos Paragios. 2009. Robust shadow and illumination estimation using a mixture model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 651–658.

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

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

[49] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*. 6105–6114.

[50] Jeya Maria Jose Valanarasu and Vishal M Patel. 2023. Fine-Context shadow detection using shadow removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1705–1714.

[51] Tomas F Yago Vicente, Minh Hoai, and Dimitris Samaras. 2017. Leave-one-out kernel optimization for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 3 (2017), 682–695.

[52] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. 2016. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *Proceedings of the European Conference on Computer Vision*. 816–832.

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

[54] Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. 2020. Instance shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1880–1889.

[55] Yupei Wang, Xin Zhao, Yin Li, Xuecai Hu, Kaiqi Huang, and NLPR CRIPAC. 2018. Densely cascaded shadow detection network via deeply supervised parallel fusion. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*. 1007–1013.

[56] Zhiyong Yang, Qianqian Xu, Xiaochun Cao, and Qingming Huang. 2020. Task-Feature collaborative learning with application to personalized attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (2020), 4094–4110.

[57] Guanghui Yue, Wanwan Han, Bin Jiang, Tianwei Zhou, Runmin Cong, and Tianfu Wang. 2022. Boundary constraint network with cross layer feature integration for polyp segmentation. *IEEE Journal of Biomedical and Health Informatics* 26, 8 (2022), 4090–4099.

[58] Chen Zhang, Runmin Cong, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, and Sam Kwong. 2021. Cross-modality discrepant interaction network for RGB-D salient object detection. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2094–2102.

[59] Qijian Zhang, Runmin Cong, Chongyi Li, Ming-Ming Cheng, Yuming Fang, Xiaochun Cao, Yao Zhao, and Sam Kwong. 2021. Dense attention fluid network for salient object detection in optical remote sensing images. *IEEE Transactions on Image Processing* 30 (2021), 1305–1317.

[60] Qi Zhang, Jingyu Xiao, Chunwei Tian, Jerry Chun-Wei Lin, and Shichao Zhang. 2023. A robust deformed convolutional neural network (CNN) for image denoising. *CAAI Transactions on Intelligence Technology* 8, 2 (2023), 331–342.

[61] Guangzhe Zhao, Yimeng Zhang, Maoning Ge, and Min Yu. 2023. Bilateral U-Net semantic segmentation with spatial attention mechanism. *CAAI Transactions on Intelligence Technology* 8, 2 (2023), 297–307.

[62] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. 2019. Distraction-aware shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5167–5176.

[63] Wujie Zhou, Fan Sun, Qiuping Jiang, Runmin Cong, and Jenq-Neng Hwang. 2023. WaveNet: Wavelet network with knowledge distillation for RGB-T salient object detection. *IEEE Transactions on Image Processing* 32 (2023), 3027–3039.

[64] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. 2010. Learning to recognize shadows in monochromatic natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitionn*. 223–230.

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

[66] Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson WH Lau. 2021. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4702–4711.

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