Two-Stage Video Shadow Detection via Temporal-Spatial Adaption

Xin Duan[1], Yu Cao[1], Lei Zhu[2,3], Gang Fu[1], Xin Wang[1], Renjie Zhang[1], and Ping Li[1](✉)

[1] The Hong Kong Polytechnic University, Kowloon, Hong Kong
p.li@polyu.edu.hk

[2] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

[3] The Hong Kong University of Science and Technology, Kowloon, Hong Kong

Abstract. Video Shadow Detection (VSD) is an important computer vision task focusing on detecting and segmenting shadows throughout the entire video sequence. Despite their remarkable performance, existing VSD methods and datasets mainly focus on the dominant and isolated shadows. Consequently, VSD under complex scenes is still an unexplored challenge. To address this issue, we built a new dataset, Complex Video Shadow Dataset (CVSD), which contains 196 video clips including 19,757 frames with complex shadow patterns, to enhance the practical applicability of VSD. We propose a two-stage training paradigm and a novel network to handle complex dynamic shadow scenarios. Regarding the complex video shadow detection as conditioned feature adaption, we propose temporal- and spatial-adaption blocks for incorporating temporal information and attaining high-quality shadow detection, respectively. To the best of our knowledge, we are the first to construct the dataset and model tailored for the complex VSD task. Experimental results show the superiority of our model over state-of-the-art VSD methods. Our project will be publicly available at: https://hizuka590.github.io/CVSD.

Keywords: Video Shadow Detection · Video Understanding · Large-Scale Complex Video Shadow Dataset · Conditioned Feature Adaption

1 Introduction

Shadows can be commonly observed in digital images or videos, which can provide a variety of visual properties, including depth relations [1,2,41], object shapes [32,54], light direction [6,30,51], and spatial layout [36,49]. On the other hand, misunderstanding shadows may fail many computer vision tasks such as object detection [8,45,59], tracking [29], and interpretation of visual

X. Duan and Y. Cao—Contributed equally to this work.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15106, pp. 196–214, 2025.
https://doi.org/10.1007/978-3-031-73195-2_12

data [14,17,33]. Hence, detecting shadows in static or dynamic scenes accurately [12,13,16,23,37,38,54] is a fundamental and challenging task.

Several shadow detection methods have been developed, including estimating shadow masks from a single image [15,18,21,23,27,47,55,58,65] and from temporal-related video sequences [4,11,24,25,40,42,53,69], while the state-of-the-art learning-based Video Shadow Detection (VSD) methods [4,11,35,40,66] have demonstrated their remarkable performance primarily on the ViSha dataset [4]. Nonetheless, it is important to acknowledge that placing exclusive emphasis on achieving a numerical score on a single shadow detection dataset may lack meaningfulness in practical applications. This leads us to question whether machines can truly perceive shadows like humans in real-world scenarios.

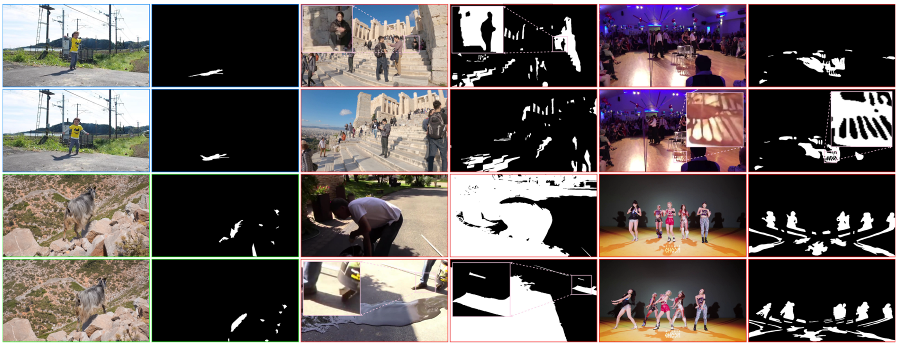

(a) ViSha (blue) (b) VISAD (green) (c) Our CVSD Dataset (red)

Fig. 1. Visual comparison between ViSha [4], VISAD [40] and our CVSD with magnified local details. Compared with isolated and dominant shadows in others, shadow patterns in our CVSD are much more complex and diverse.

To answer this question, we first revisit existing shadow detection datasets. While the majority of these datasets are single image-based [22,52,58,60,62,65], Chen et al. [4] construct the first video shadow dataset and Lu et al. [40] introduce another partially annotated video shadow dataset for semi-supervised learning. Nevertheless, existing video shadow datasets primarily consist of isolated and dominant shadow instances, which do not represent real-world and complex scenes. To facilitate the exploration of video shadow detection in the wild, we build a new dataset named Complex Video Shadow Dataset (CVSD). It comprises 196 video clips featuring diverse scenarios encompassing various shadow patterns across 149 categories, resulting in a wide range of challenging cases and shadow characteristics. Within the dataset, we carefully annotate 309,183 disjoint shadow areas, yielding a collection of 19,757 frames with high-quality shadow masks for training and evaluating video shadow detection methods in real-world and complex scenarios. Figure 1 illustrates the visual comparison between existing ViSha dataset [4] (Fig. 1a), VISAD dataset [40] (Fig. 1b) and our CVSD (Fig. 1c). Our CVSD enjoys notable features including complex and diverse shadow patterns, which pose more opportunities and challenges for video shadow detection.

Detecting shadows in complex dynamic scenes presents two main challenges. The first is temporal consistency modeling since shadows are deformative making them even harder to track over time. The second is the detection of shadows with intricate local details further complicating the task. **We consider our complex video shadow detection task as a conditioned feature adaptation problem in both temporal and spatial domains.** We introduce a novel two-stage training paradigm that enables the direct transfer from a pre-trained image-level model to a detail-preserving video-level model. Specifically, to address the temporally correlated shadow modeling, we propose a *Temporal-Adaption Block*, which emphasizes deformable parts of the shadows while preserving consistent features. With this design, we condition the extracted feature embedding of other reference frames with the given main frame feature in the temporal domain, resulting in a high correlation with the main frame feature while preserving temporal disengagement. For accurately localizing intricate details of shadows in each frame, we utilize low-resolution context features as guidance and spatially condition high-resolution local shadow details on it. We develop a *Spatial-Adaption Block* that obtains a high-quality mask by integrating high-resolution local patch information into global low-resolution context features.

Our main contributions can be summarized as follows:

- We build a new dataset named Complex Video Shadow Dataset (CVSD) which features more diverse shadow patterns and challenging cases in real-world scenarios.
- We design a two-stage paradigm for video shadow detection which adapts the image-level model to the video-level with local detail concentration using our proposed Temporal-Adaption Block and Spatial-Adaption Block.
- We conduct comprehensive comparisons with SOTA shadow detection methods on the ViSha [4] and our CVSD dataset, demonstrating the complexity of our CVSD and the superiority of our proposed model.

2 Related Work

Shadow Dataset. Existing datasets for image shadow detection [19,22,52,60–62,65] have been extensively used recently. Among them, one notable example [60] was delivered for high-resolution image shadow detection. Another dataset [22] was specially crafted for complex real-world scenes. We should note that there is currently a lack of emphasis on high-resolution or real-world settings in video shadow datasets. The most extensively used and fully annotated dataset in the video domain was ViSha [4], while Lu et al. [40] introduced another partially annotated video shadow dataset VISAD for a semi-supervised setting. The existing video shadow detection methods commonly underperform when facing complex real-world scenes due to the data limitations of current datasets.

One notable drawback is the low resolution of the current dataset. This reduction in resolution might lead to erroneous results and affect the reliability of the trained models. Furthermore, the current datasets only cover a limited range of scenes, which may not adequately represent the diverse circumstances encountered in real-world shadow detection scenarios. This narrow scope restricts the ability of the models to generalize and perform well in different contexts. Most importantly, current datasets primarily address shadows in dominant and isolated patterns, which are not as applicable to real-world scenes. To overcome these limitations, it is crucial to create larger, more diverse, and more challenging datasets. Our complex Video Shadow Dataset (CVSD) aims to establish a more practical foundation and advance research in the field of video shadow detection.

Video Shadow Detection. Research on shadow detection of single images [5,16,20,60,63,73,75] has been prevailing. However, shadow detection in dynamic scenes was less successful. The primary challenge lies in temporal modeling, which involves managing redundant and complementary information across multiple frames. Previous methods typically employed a shared encoder [4,11,40,48,66] to extract multi-frame information. These methods often required an auxiliary feature fusion module, such as optical flow [11], LSTM [66], or attention [48] to assist feature extraction. So it necessitates careful design in the feature fusion part, and performance can easily be compromised if the extracted features are not properly managed [50]. To mitigate this issue, some methods introduced implicit loss, such as contrastive loss [4] or consistency regularization [11] to help manage complementary information across multiple frames. An alternative approach was to directly extract spatio-temporal features [35]. Different from the above two types of solutions, we introduce a novel two-stage paradigm that enables direct adaption from a pre-trained image-level model to a detailed-preserving video-level model.

Other Video Processing Tasks. Video object segmentation aims to segment a video into semantically meaningful objects [7,34,67,68,72], while video saliency detection aims to identify primary foreground objects from their background in all frames of a video. These tasks are closely related to our work on video shadow detection, as they involve the extraction of meaningful information from video data. Unlike single-image segmentation, video objects are correlated in the temporal domain, which poses a greater challenge for detection. Some researchers addressed temporal consistency by propagating information from neighboring frames using optical flow [31] or feature matching modules [39,56]. Another approach involved using external memory to retain past information [48]. However, such methods focused on designing auxiliary modules or feature fusion strategies to achieve temporal consistency, which may result in error accumulation due to inaccurate motion estimation or feature fusion. Additionally, they often neglect the rich constraints provided by image-level information.

3 Complex Video Shadow Dataset

In this section, we introduce our Complex Video Shadow Dataset (CVSD). We first present the video collection and annotation process in Sect. 3.1 and then give the dataset statistics and comparisons in Sect. 3.2.

3.1 Building the Dataset

Video Collection. To ensure the representativeness and robustness of the collected video shadow dataset, we source video data from various origins, encompassing diverse lighting conditions, scenes, and viewpoints. To accurately reflect the typical distribution of real-world shadows, we select videos with crowded objects to ensure they represent complex scenarios. For this purpose, we manually select intricate shadow videos from well-known video processing benchmarks such as DAVIS [3], UAV123 [44], VSPW [43], MOT [10], TAO [9], Kinetics [26], Inter4k [57], and REDS [46]. Originally, these video datasets are intended for purposes other than shadow detection, such as object tracking, motion recognition, and video deblurring. We curate diverse video samples, highlighting various items, scenarios, artifacts, and shadows of different scales. This comprehensive selection accurately represents the complexities of the real world and the interactions between shadows and objects.

Shadow Annotation. Once all the videos for our CVSD are collected, we follow existing preprocessing algorithms [4] to generate video sequences. In complex scenes, shadows can be numerous and ambiguous, particularly when objects interact with each other. Therefore, we only annotate cast shadows in our dataset which is also in line with previous works [4,40]. We employ an annotation team consisting of well-trained annotators to label all the cast shadows. Subsequently, we carefully inspect and validate the labeled results one by one, by overlaying them onto the original RGB image. During the review process, we focus on the labeling quality of shadow boundary. We return labels with poor quality or mislabeled self-shadows to the annotation team for refinement. After three rounds of annotation and refinement, our dataset comprises a total of 196 video sequences, consisting of 19,757 frames. Our CVSD is the first dataset that holds the title of being the most extensive collection of video shadow data and it encompasses the most diverse range of scenarios encountered in real-world situations.

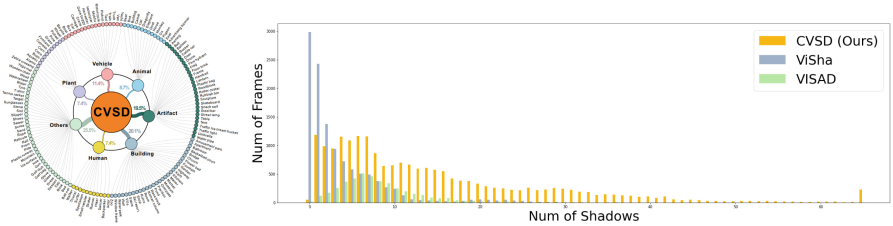
3.2 Statistics and Comparisons

Diverse Shadow Pattern. Unlike prior datasets that offer a uniform representation of shadow patterns, our dataset encompasses a broad spectrum of shadow features. This diversity is attributed to factors such as different types of motion, changes in viewpoint, and a wide range of objects and scene types.

As demonstrated in Table 1 and Fig. 2, our dataset not only includes more shadow instances but also incorporates shadows generated by multiple light sources. This extends the original illumination scenarios beyond the conventional indoor, outdoor, day, and night categories to more diverse types, such as stage lighting, overcast lighting, and dusk lighting. For instance, overcast scenarios depict scenes with uniform and diffuse lighting conditions caused by heavy cloud cover, resulting in soft shadows. In contrast, stage lighting produces sharp shadows and induces rapid changes in illumination color. Furthermore, we expand the original 60 shadow categories in ViSha [4] to 149 types, as illustrated in Fig. 2a. Unlike previous datasets, which primarily consist of footage from handheld or standard cameras, our collection includes examples captured using a diverse array of camera types, thereby providing dynamic viewpoints, vivid motion patterns, and motion blur.

Table 1. Comparison of dataset characteristics, including shadow instances, shadow motion, viewpoint, motion blur, object type, and scene type. This table illustrates how our CVSD dataset surpasses ViSha [4] and VISAD [40] in diversity and complexity.

Dataset	Shadow Instances				Shadow Motion		Viewpoint	Motion Blur	Object Type	Scene Type
	1–4	5–10	11–15	16+	Stable	Moving	-	-	-	-
ViSha [4]	6,792	3,625	743	525	16.7%	83.3%	Single	No	60	4
VISAD [40]	308	2,381	1,076	423	19.4%	80.6%	Single	No	33	2
CVSD	**2,818**	**5,374**	**4,583**	**6,982**	**11.2%**	**88.8%**	**Multiple**	**Yes**	**149**	**12**



(a) Shadow Categories **(b)** Count of Disjointed Shadow Areas

Fig. 2. Detailed statistics of our proposed CVSD. The CVSD comprises 7 main classes with 149 sub-classes. (a) provides a detailed breakdown of these categories. (b) presents a distribution disjointed shadow areas across three datasets.

Improved Resolution. According to the comparison presented in Table 2, it is evident that available datasets suffer from poor resolution, which makes it challenging to identify small or distant shadows. In contrast, our dataset is specifically designed to facilitate precise and accurate shadow identification, primarily due to its significantly higher resolution. To ensure superior quality, we handpicked high-resolution base videos from the most recent video benchmark dataset, achieving an average resolution of $(1,358 \times 2,412)$ pixels in our dataset. This deliberate choice guarantees that our dataset matches the resolution of modern cameras, thus enabling high-quality shadow detection. The utilization of high-resolution datasets provides several advantages. Firstly, it leads to improved visualization of shadow textures and details, enhancing the model quality based on it. Additionally, our dataset is more compatible with contemporary cameras and screens, making its integration into real-world applications much more seamless. Lastly, the availability of high-resolution datasets opens up greater potential for future research and development in the field of shadow detection.

Table 2. Comparison of training and testing sets on ViSha [4], VISAD [40], and CVSD datasets. Our CVSD dataset exceeds the others in the number of frames, meanwhile boasts a significantly higher average of disconnected shadow area and mean resolution.

Dataset	Training Set				Testing Set			
	Frames	Avg Num	Avg Ratio	Mean Resolution	Frames	Avg Num	Avg Ratio	Mean Resolution
ViSha [4]	4,788	4.7	15.1%	681×825	6,897	6.2	9.6%	701×905
VISAD [40]	1,125	8.1	22.0%	$700 \times 1,208$	3,063	8.8	19.1%	$820 \times 1,458$
CVSD	**8,026**	**13.6**	7.5%	$\mathbf{1,405 \times 2,493}$	**11,731**	**16.5**	7.7%	$\mathbf{1,334 \times 2,358}$

Count of Disjointed Shadow Areas. We measure the number of distinct shadow regions by counting the disjoint shadow areas in each frame. We disregard any shadow region that occupies less than 0.01% of the total image and set the maximum number to 66. As shown in Table 2, our average number of shadows significantly exceeds that of other datasets. Reflecting real-world complex shadow patterns where shadows are present but not dominating, our average ratio of shadows is 8.3% for training and 7.3% for testing. The distribution of shadow instances (Fig. 2b) further supports our findings. The majority of frames in the ViSha and VISAD datasets contain fewer than 10 dominant shadow instances per frame. In contrast, our dataset exhibits a much higher number of shadow instances, demonstrating our superior ability to represent real-world complex shadow patterns.

4 Method

4.1 Overview

In this work, we formulate the task of complex video shadow detection as a conditioned feature adaption problem in both temporal and spatial domains. We propose a novel two-stage training paradigm including the first stage (Fig. 3) involves training an image-level model, and the second stage (Fig. 4) transfers the pre-trained image-level model to dynamic video-level with detail concentration.

4.2 First Pre-training Stage

In the first pre-training stage (Fig. 3), by following [70], the network takes a single image X as input. Similar to traditional image-level models, it generates hierarchical feature maps F_i and learns to create a shadow mask Y.

$$Y = \mathcal{I}(X), F_i = \mathcal{E}\mathcal{T}(X). \tag{1}$$

Here, encoder $\mathcal{E}$ (marked gray) which consists of Conv2D, flatten, and layer norm operations, together with transformer block $\mathcal{T}$ (marked blue) of the image-level model $\mathcal{I}$, transform the input image into multi-level feature maps F_i, $i \in 1, 2, 3, 4$ represents feature at the level i.

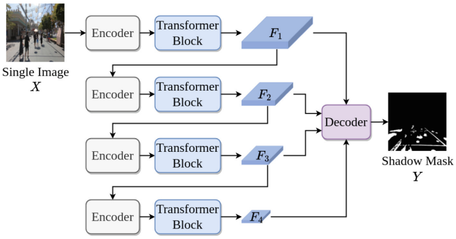


Fig. 3. Image-domain pre-training stage for image shadow detection. Once trained on individual images, the encoders acquire the basic shadow feature extraction ability.

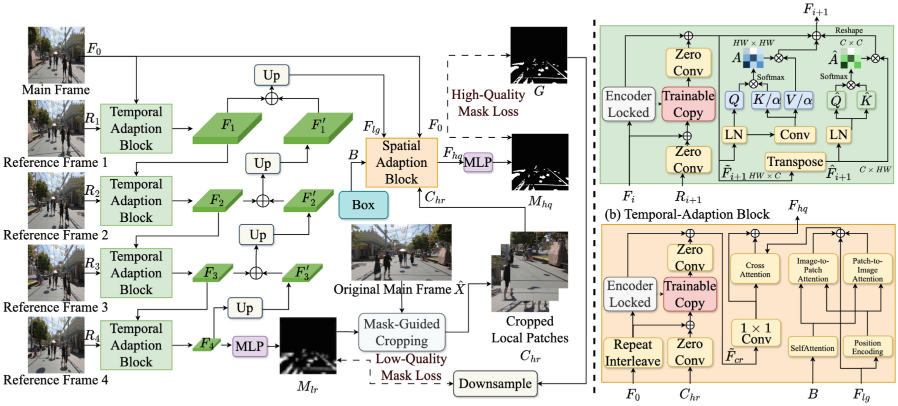
4.3 Second Feature Adaption Stage

The target of the second training stage (Fig. 4a) is to transfer the image-level model $\mathcal{I}$ to the video-level model $\mathcal{V}$, which is capable of reasoning about temporal sequences in videos and spatially detailed information. We apply the encoder $\mathcal{E}$ of the pre-trained image-level model $\mathcal{I}$ to incorporate both temporal and spatial auxiliary information. The pre-trained encoder $\mathcal{E}$ (marked gray) and its trainable copy $\mathcal{E}'$ (marked red) are embedded in our proposed Temporal-Adaption Block (Fig. 4b) and Spatial-Adaption Block (Fig. 4c).

$$Y = \mathcal{I}(X) \rightarrow Y = \mathcal{V}(X, R, C_{hr}), \tag{2}$$

where X is main frame, R is reference frame and C_{hr} is the cropped high-resolution local patch.

As depicted in Fig. 4, reference frames are progressively fused using our Temporal-Adaption Block during the encoding phase. A low-resolution mask M_{lr} is generated using the intermediate feature F_4, which serves as a guide to pinpoint potential shadow locations. With this position hint, the original main frame $\hat{X}$ is locally cropped into several high-resolution patches C_{hr}. These patches are then integrated into the low-resolution global feature F_{lg} using our



(a) Video Shadow Detection Via Temporal-Spatial Domain Adaption

(b) Temporal-Adaption Block

(c) Spatial-Adaption Block

Fig. 4. We transfer the pre-trained image-level model to video-level model (a) using Temporal-Adaption Block (b) which emphasizes deformable parts of the shadows while preserving consistent features and Spatial-Adaption Block (c) that obtains a high-quality mask by integrating high-resolution local patch information into global low-resolution context features. (Color figure online)

proposed Spatial-Adaption Block during the decoding phase. By incorporating temporal information from neighboring frames and integrating high-resolution local patches, our model is capable of generating a high-quality mask M_{hq}. The optimization goal is given by

$$\mathcal{L} = \mathcal{L}_{BCE}(M_{hq}, G) + \mathcal{L}_{Hinge}(M_{hq}, G) + \lambda_1 \mathcal{L}_{BCE}(M_{lr}, \hat{G}) + \lambda_2 \mathcal{L}_{Hinge}(M_{lr}, \hat{G}), \tag{3}$$

where $\mathcal{L}_{BCE}(\cdot)$ and $\mathcal{L}_{Hinge}(\cdot)$ denote the BCE loss and the lovász-hinge loss. M_{lr} and M_{hq} denote low-resolution mask and high-quality mask. G is a ground-truth mask and $\hat{G}$ is a down-sampled ground-truth mask.

Temporal-Adaption Block. We employ a pre-trained image-level encoder $\mathcal{E}$ to estimate cross-frame information progressively. It takes input as reference frame R_{i+1} in level $i+1$ and context feature F_i in level i, to output a fused global feature $\tilde{F}_{i+1}$. Inspired by ControlNet [74], we freeze the parameters Θ of the $\mathcal{E}$ and concurrently create a trainable copy $\mathcal{E}'$. The frozen parameter helps to preserve the original redundant context of the main frame. And $\mathcal{E}'$ helps to capture discriminative features across different reference frames. We formulate the cross-frame feature fusion process of our Temporal-Adaption Block as:

$$\tilde{F}_{i+1} = \mathcal{E}(F_i; \Theta) + \mathcal{Z}(\mathcal{E}'(F_i + \mathcal{Z}(R_{i+1}); \Theta')). \tag{4}$$

In this step, we follow a progressive fusion protocol, where frames near/farther away the main frame contribute more/less. Therefore, reference frames are not fed in the same resolution. When the reference image undergoes encoding, it is scaled into different sizes to match the dimensions of F_i with resolutions of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$. Context feature is faithfully preserved using the frozen encoder $\mathcal{E}$. The trainable copy $\mathcal{E}'$ outputs a temporal discriminative feature. Here, $\mathcal{Z}$ represents zero convolution. Finally, the context feature and temporal discriminative feature are added together to obtain the fused global feature $\tilde{F}_{i+1}$.

As shadows deform across time, directly applying self-attention leads to suboptimal solutions and heavy computation. Therefore, we incorporate modified self-attention with channel-wise attention to achieve deformable feature matching. This is done by generating query Q, downscaled key K/α, and downscaled value V/α mapped from the fused global feature. Additionally, we compute the channel-wise attention by reshaping the queries and key projections, resulting in a transposed-attention map $\hat{A}$ of size $\mathbb{R}^{C\times C}$. The final output F_{i+1} of Temporal-Adaption Block is given by the sum of fused global feature $\tilde{F}_{i+1}$, self-attention output, and channel-wise attention output:

$$F_{i+1} = \tilde{F}_{i+1} + \mathrm{SA}(Q, K/\alpha, V/\alpha) + \mathrm{CA}(\hat{Q}, \hat{K}, \hat{F}_{i+1}). \tag{5}$$

Mask-Guided Cropping. Shadows in complex real-world settings are intricate. A simple upscale from a downgraded feature may result in the loss of local details. Therefore, it is crucial for a model to integrate non-downscaled information. However, directly inputting all non-downscaled patches into the network is computationally demanding and unfeasible. To maximize the preservation of local details, we propose a mask-guided cropping strategy that selectively crops areas of the original main frame $\hat{X}$ that truly contain shadows. With the low-resolution shadow mask M_{lr} generated using the intermediate feature F_4, we identify potential shadows and calculate their bounding boxes. By cropping the original resolution image $\hat{X}$ accordingly and resize operation We obtain high-resolution local patches $C_{hr} \in \mathbb{R}^{N\times512\times512\times3}$. N represents the patch number. Please note those box crops are directly taken from the original size (e.g., $3,840 \times 2,160$), preserving high-resolution local details.

Spatial-Adaption Block. We resort to a pre-trained image encoder $\mathcal{E}$ with the ability to process high-resolution local patch information. Similar to the Temporal-Adaption Block, we freeze the parameters Θ of the pre-trained image-domain encoder and simultaneously create a trainable copy $\mathcal{E}'$. The cross-resolution feature fusion of our Spatial-Adaption Block can be formally expressed as:

$$\tilde{F}_{cr} = \mathcal{E}(F_0; \Theta) + \mathcal{Z}(\mathcal{E}'(F_0 + \mathcal{Z}(C_{hr}); \Theta')), \tag{6}$$

where $\tilde{F}_{cr}$ is cross-resolution feature and C_{hr} represents high-resolution local patches. By simply repeating the main frame F_0 along a batch dimension, we can match the dimension of local patches $C_{hr} \in \mathbb{R}^{N\times512\times512\times3}$. The frozen parameter Θ is used to preserve the original low-resolution context of the main frame. While the trainable copy $\mathcal{E}'$ captures high-resolution local features.

The positions of cropped patches are represented with $B \in \mathbb{R}^{128\times128\times64}$, following the prompt encoding strategy [28]. We add position encoding (PE) to the low-resolution global feature F_{lg}. To embed F_{lg} with the position of high-resolution local details, we enable localization ability by:

$$F_{lg} = F_{lg} + \mathrm{Cross}(\mathrm{PE}(F_{lg}), \mathrm{SA}(B)) + \mathrm{Cross}(\mathrm{SA}(B), \mathrm{PE}(F_{lg})). \tag{7}$$

We then direct the model's attention to critical local regions with the cross-attention to get a high-quality main frame feature F_{hq}, which can be written as:

$$F_{hq} = \mathrm{Conv}(\tilde{F}_{cr}) + \mathrm{Cross}(\mathrm{Conv}(\tilde{F}_{cr}), F_{lg}). \tag{8}$$

5 Experimental Results

5.1 Implementation Details

We utilize the PyTorch and PyTorch-Lighting frameworks to construct our network. In the first stage, we employ Segformer [70] as our image-level model and pre-train it on both training set separately. It is important to note that each sample is treated as an independent image, disregarding any temporal information in the first stage. In the second stage, we transfer the pre-trained image-domain model to incorporate temporal correlation and local high-resolution information. Each image is fed empirically in a resolution of 512 × 512 along with 4 neighboring frames. For optimization, we employ the Adan [71] optimizer with an initial learning rate of 1.5×10^{-5} and a weight decay of 0.02. All experiments are trained on a single NVIDIA GeForce RTX 3090 GPU.

5.2 Datasets and Evaluation Metrics

We conduct evaluations on both the ViSha [4] dataset and our CVSD dataset. We adopt the same data preprocessing strategy as outlined in [4]. To provide a quantitative comparison of the effectiveness of various video shadow detection methods, we utilize four commonly used evaluation metrics: MAE, F_β, IoU, and BER [4,36]. A lower MAE and BER score, along with higher F_β and IoU scores, indicate superior video shadow detection performance.

5.3 Comparison with SOTA Methods

Comparative Methods. We compare our method with existing Video Object Segmentation (VOS), Image Shadow Detection (ISD), and Video Shadow Detection (VSD) methods. We directly download available results on the ViSha dataset from [4,35]. We retrain the public codes of FDRNet [35] and SDCM [4] for a fair comparison, as there are no reported results on the ViSha dataset. We should note that the reason we report a lower result of the DAS [66] is because it requires the first-frame ground-truth bounding box as input. To ensure a fair comparison, we replace the first-frame ground-truth bounding box with the bounding box generated by an image shadow detection model. We also train all these methods on public codes with the same settings on our CVSD dataset. This adjustment ensures consistency in the evaluation process and allows for a reliable comparison between our method and the existing state-of-the-art methods.

Table 3. Quantitative comparison between our method with the existing SOTA methods on the ViSha dataset and our CVSD dataset.

Tasks	Methods	ViSha [4]				CVSD			
		MAE ↓	F_{β} ↑	IoU ↑	BER ↓	MAE ↓	F_{β} ↑	IoU ↑	BER ↓
VOS	STM [48]	0.068	0.597	0.408	25.69	0.085	0.519	0.401	20.74
	COSNet [39]	0.040	0.705	0.514	20.50	0.089	0.554	0.417	22.52
	FEELVOS [64]	0.043	0.710	0.512	19.76	0.081	0.496	0.385	23.13
ISD	BDRAR [76]	0.050	0.695	0.484	21.29	0.071	0.538	0.416	29.26
	DSD [75]	0.043	0.702	0.518	19.88	0.086	0.502	0.408	22.16
	MTMT [5]	0.043	0.729	0.517	20.28	0.074	0.550	0.402	20.17
	FSDNet [22]	0.057	0.671	0.486	20.57	0.091	0.475	0.334	24.99
	FDRNet [77]	0.044	0.612	0.497	20.23	0.079	0.531	0.376	28.14
	SDCM [78]	0.053	0.643	0.485	21.20	0.098	0.497	0.355	30.31
VSD	TVSD [4]	0.033	0.757	0.567	17.70	0.099	0.539	0.369	27.28
	STICT [40]	0.046	0.702	0.545	16.60	0.073	0.608	0.447	23.27
	SC-Cor [11]	0.042	0.762	0.615	13.61	0.070	0.573	0.476	19.94
	SCOTCH and SODA [35]	0.029	0.793	0.640	9.07	0.082	0.585	0.426	23.27
	DAS [66]	0.032	0.788	0.613	16.47	0.087	0.561	0.435	19.15
	⋆ Ours	**0.027**	**0.801**	**0.684**	**8.96**	**0.046**	**0.638**	**0.515**	**18.32**

Quantitative Comparison. Table 3 reports a quantitative comparison between our proposed method and SOTA methods on the ViSha dataset [4] and our CVSD dataset. Our approach outperforms all other SOTA methods across all evaluation metrics on both datasets, demonstrating its superior performance. However, it is crucial to acknowledge that despite our significant achievements in video shadow detection on previous benchmarks, there are still unresolved challenges when dealing with complex scenes. For instance, the performance on our CVSD dataset still has a lot of room for improvement, which calls for more efficient and real-world applicable models. Moreover, it is noteworthy that all video shadow detection methods exhibit relatively better performance compared to their image-domain counterparts on both datasets. This emphasizes the importance of considering the temporal aspect in video shadow detection tasks.

Qualitative Comparison. Figure 5 presents a qualitative comparison of the video shadow detection masks generated by our method and other SOTA methods. The first four rows of the comparison are derived from the ViSha dataset [4] (marked blue), while the remaining rows are obtained from our CVSD dataset (marked red). It is worth noting that we choose results of DSD [75] that show the best performance as a representative of ISD methods. The results demonstrate that our method, shown in the 3rd column, outperforms the compared methods in accurately identifying shadow pixels. Our method effectively detects various types of shadows against different backgrounds and successfully identifies shadows in video frames containing crowded objects, as observed in the fifth and sixth rows. In contrast, the compared methods tend to miss or incorrectly classify small shadow regions. Additionally, the compared methods often fail to obtain local shadow details, whereas our method performs better in such scenarios. This is evident in row 7, where our method produces highly accurate detection results with clear shadow boundaries. This indicates that our method is well-suited for complex real-world video shadow detection task.

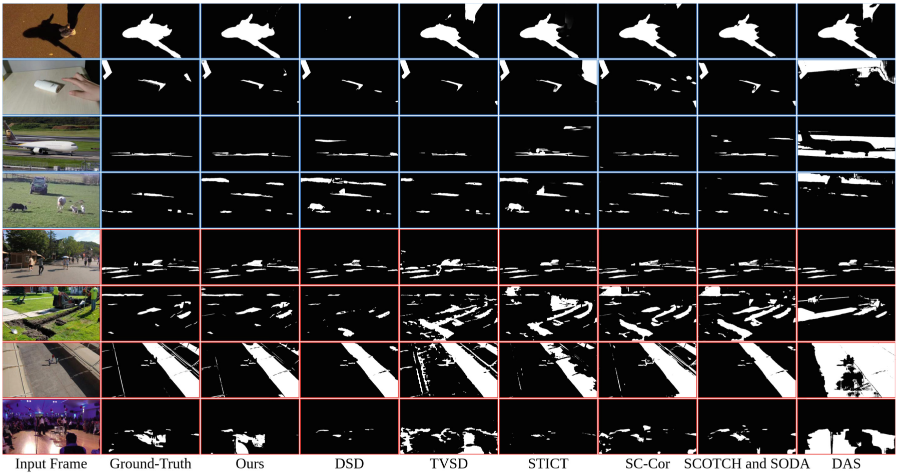


Fig. 5. Qualitative comparison of shadow masks generated by our method and other SOTA methods. The top four rows (labeled in blue) display generated masks using the ViSha [4] dataset, while the remaining rows (labeled in red) display results using our CVSD dataset. (Color figure online)

5.4 Ablation Study

Two-Stage Training Paradigm. We present the results of an ablation study aimed at evaluating the effectiveness of our proposed two-stage training paradigm. Initially, we obtain the performance of the baseline network, referred to as the Baseline, which is a stage-1 model trained on individual images. We examine our paradigm through two different approaches. First, we experiment with a randomly initialized encoder and train the stage-2 model directly (referred to as direct train, i.e., the model extracts shadow features without the guidance of a pre-trained encoder.), resulting in model collapse as shown in the second row of Table 4.

Table 4. Ablations on two-stage training paradigm.

Variants of Ours	T	S	ViSha [4]				CVSD			
-	-	-	MAE ↓	F_β ↑	IoU ↑	BER ↓	MAE ↓	F_β ↑	IoU ↑	BER ↓
Baseline	✗	✗	0.044	0.660	0.532	18.94	0.074	0.445	0.375	24.58
directly train	✗	✗	0.175	0.231	0.216	38.38	0.189	0.197	0.183	40.14
directly transfer	✗	✗	0.053	0.620	0.487	23.83	0.101	0.429	0.355	28.31

Next, we examine the results of directly transferring the pre-trained image-level model to the video-level, without introducing our proposed Temporal-Adaption Block and Spatial-Adaption Block (referred to as direct transfer), as shown in the third row of Table 4. We feed reference frames and local patches into a shared encoder following the two-stage training paradigm with weights initialized from stage-1. This direct transfer can work to some extent, but the performance is compromised with the introduction of auxiliary information, as it does not contribute positively but rather acts as noise.

Table 5. Ablation study on T-Block and S-Block. 'T' denotes Temporal-Adaption Block, and 'S' denotes Spatial-Adaption Block. The baseline is the first-stage image-domain model. ⋆ shows only with both our proposed T and S block, the model can achieve domain transfer from the image domain to the detail-preserving video domain.

Variants of Ours	T	S	ViSha [4]				CVSD			
-	-	-	MAE ↓	F_β ↑	IoU ↑	BER ↓	MAE ↓	F_β ↑	IoU ↑	BER ↓
Baseline	✗	✗	0.044	0.660	0.532	18.94	0.074	0.445	0.375	24.58
Ours w/o T	✗	✓	0.029	0.789	0.675	11.98	0.056	0.629	0.509	22.01
Ours w/o S	✓	✗	0.031	0.767	0.636	12.36	0.070	0.541	0.463	19.03
⋆ Our Method	✓	✓	**0.027**	**0.801**	**0.684**	**8.96**	**0.046**	**0.638**	**0.515**	**18.32**

T-Block and S-Block. Subsequently, we investigate the impact of solely integrating our S block into the network (referred to as "Ours w/o T"), as is shown in Table 5. We observe that the S block provides more benefits when dealing with complex shadow structures. This is evident from the greater performance gains achieved on our CVSD dataset compared to the ViSha dataset. Furthermore, we explore the effects of adding the T block to our network (referred to as "Ours w/o S"). It is observed that the T block successfully integrates reference information, resulting in a performance boost. Finally, we present the full version of our second-stage model (referred to as "⋆ Our method"), which incorporates both the T and S blocks. This model achieves the highest performance on both datasets, demonstrating the effectiveness of combining our proposed blocks.

Number of Reference Images. We add extra ablations by removing all/3/2/1 reference frames, denoted as 0,1,2,3 to check the integration of reference frames indeed contributes to the final performance. As is shown in Table 6.

Adding more reference frames generally enhances overall performance, with the inclusion of the most neighboring frame resulting in the highest performance gain, which is in line with our progressive fusion strategy.

Table 6. Ablation study for varying numbers of the reference image.

Num of Ref	ViSha				CVSD			
	MAE	F_β	IoU	BER	MAE	F_β	IoU	BER
0	0.045	0.657	0.537	17.63	0.072	0.428	0.395	23.58
1	0.033	0.764	0.643	11.76	0.053	0.618	0.481	19.17
2	0.031	0.766	0.645	11.07	0.054	0.633	0.488	18.58
3	0.029	0.787	0.673	10.22	0.051	0.632	0.491	18.47
4(Original)	**0.027**	**0.801**	**0.684**	**8.96**	**0.046**	**0.638**	**0.515**	**18.32**

6 Conclusion

In this paper, we introduce **CVSD**, the first large-scale dataset specifically designed for complex video shadow detection. This dataset encompasses a wide range of challenging shadow patterns, aiming to stimulate further research in detecting complex real-world video shadows. High-quality, complex real-world video shadow datasets are essential for exploring and improving detection methods in real-world settings. Our contribution with **CVSD** is intended to facilitate progress in this important research area.

We also introduce a novel two-stage paradigm, equipped with the Temporal-Adaption Block and Spatial-Adaption Block, for complex video shadow detection. By considering the complex video shadow detection task as a conditioned feature adaptation problem, we tackle temporally correlated shadow modeling using the Temporal-Adaption Block, which emphasizes deformable parts of the shadows while preserving consistent features. Subsequently, we accurately localize intricate shadow details using the Spatial-Adaption Block, which fuses high-resolution local patch information with global context features. The experimental comparisons showcase the superior performances of our proposed method on public and our constructed datasets.

Acknowledgments. This work was supported in part by The Hong Kong Polytechnic University under Grants P0042740, P0044520, P0043906, and P0050657, and in part by the Guangzhou-HKUST(GZ) Joint Funding Program under Grant 2023A03J0671.

References

1. Abrams, A., Schillebeeckx, I., Pless, R.: Structure from shadow motion. In: IEEE International Conference on Computational Photography, pp. 1–8 (2014)
2. Adams, H., Stefanucci, J., Creem-Regehr, S., Bodenheimer, B.: Depth perception in augmented reality: the effects of display, shadow, and position. In: IEEE Conference on Virtual Reality and 3D User Interfaces, pp. 792–801 (2022)
3. Caelles, S., Pont-Tuset, J., Perazzi, F., Montes, A., Maninis, K.K., Gool, L.V.: The 2019 DAVIS challenge on VOS: unsupervised multi-object segmentation. arXiv:1905.00737, pp. 1–4 (2019)
4. Chen, Z., et al.: Triple-cooperative video shadow detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2714–2723 (2021)
5. Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., Heng, P.A.: A multi-task mean teacher for semi-supervised shadow detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5610–5619 (2020)
6. Chen, Z., Gao, T., Sheng, B., Li, P., Chen, C.L.P.: Outdoor shadow estimating using multiclass geometric decomposition based on BLS. IEEE Trans. Cybern. **50**(5), 2152–2165 (2020)
7. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3151–3161 (2024)
8. Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: Detecting moving objects, ghosts, and shadows in video streams. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1337–1342 (2003)

9. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: TAO: a large-scale benchmark for tracking any object. In: European Conference on Computer Vision, pp. 436–454 (2020)
10. Dendorfer, P., et al.: MOT20: a benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003, pp. 1–7 (2020)
11. Ding, X., Yang, J., Hu, X., Li, X.: Learning shadow correspondence for video shadow detection. In: European Conference on Computer Vision, pp. 705–722 (2022)
12. Finlayson, G.D., Drew, M.S., Lu, C.: Entropy minimization for shadow removal. Int. J. Comput. Vision **85**, 35–57 (2009)
13. Finlayson, G.D., Hordley, S.D., Lu, C., Drew, M.S.: On the removal of shadows from images. IEEE Trans. Pattern Anal. Mach. Intell. **28**(1), 59–68 (2006)
14. Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., Garcia-Rodriguez, J.: A survey on deep learning techniques for image and video semantic segmentation. Appl. Soft Comput. **70**, 41–65 (2018)
15. Guan, H., Xu, K., Lau, R.W.H.: Delving into dark regions for robust shadow detection. arXiv preprint arXiv:2402.13631, pp. 1–14 (2024)
16. Guo, R., Dai, Q., Hoiem, D.: Single-image shadow detection and removal using paired regions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2033–2040 (2011)
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
18. Hosseinzadeh, S., Shakeri, M., Zhang, H.: Fast shadow detection from a single image using a patched convolutional neural network. In: IEEE International Conference on Intelligent Robots and Systems, pp. 3124–3129 (2018)
19. Hou, L., Vicente, T.F.Y., Hoai, M., Samaras, D.: Large scale shadow annotation and detection using lazy annotation and stacked CNNs. IEEE Trans. Pattern Anal. Mach. Intell. **43**(4), 1337–1351 (2021)
20. Hou, Y., Zheng, L.: Multiview detection with shadow transformer (and view-coherent data augmentation). In: ACM International Conference on Multimedia, pp. 1673–1682 (2021)
21. Hu, X., Jiang, Y., Fu, C.W., Heng, P.A.: Mask-ShadowGAN: learning to remove shadows from unpaired data. In: IEEE International Conference on Computer Vision, pp. 2472–2481 (2019)
22. Hu, X., Wang, T., Fu, C.W., Jiang, Y., Wang, Q., Heng, P.A.: Revisiting shadow detection: a new benchmark dataset for complex world. IEEE Trans. Image Process. **30**, 1925–1934 (2021)
23. Hu, X., Zhu, L., Fu, C.W., Qin, J., Heng, P.A.: Direction-aware spatial context features for shadow detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7454–7462 (2018)
24. Huang, J.B., Chen, C.S.: Moving cast shadow detection using physics-based features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2310–2317 (2009)
25. Jung, C.R.: Efficient background subtraction and shadow removal for monochromatic video sequences. IEEE Trans. Multimedia **11**(3), 571–577 (2009)
26. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, pp. 1–22 (2017)
27. Khan, S.H., Bennamoun, M., Sohel, F., Togneri, R.: Automatic feature learning for robust shadow detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1939–1946 (2014)

28. Kirillov, A., et al.: Segment anything. In: IEEE International Conference on Computer Vision, pp. 3992–4003 (2023)
29. Kotera, J., Rozumnyi, D., Šroubek, F., Matas, J.: Intra-frame object tracking by deblatting. In: IEEE International Conference on Computer Vision Workshop, pp. 2300–2309 (2019)
30. Lalonde, J.F., Efros, A.A., Narasimhan, S.G.: Estimating the natural illumination conditions from a single outdoor image. Int. J. Comput. Vision **98**, 123–145 (2012)
31. Li, H., Chen, G., Li, G., Yu, Y.: Motion guided attention for video salient object detection. In: IEEE International Conference on Computer Vision, pp. 7273–7282 (2019)
32. Li, J., Li, H.: Neural reflectance for shape recovery with shadow handling. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 16200–16209 (2022)
33. Liu, F., et al.: Referring image segmentation using text supervision. In: IEEE International Conference on Computer Vision, pp. 22124–22134 (2023)
34. Liu, F., Liu, Y., Lin, J., Xu, K., Lau, R.W.: Multi-view dynamic reflection prior for video glass surface detection. In: AAAI Conference on Artificial Intelligence, pp. 3594–3602 (2024)
35. Liu, L., et al.: SCOTCH and SODA: a transformer video shadow detection framework. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10449–10458 (2023)
36. Liu, R., Menon, S., Mao, C., Park, D., Stent, S., Vondrick, C.: Shadows shed light on 3D objects. arXiv preprint arXiv:2206.08990, pp. 1–19 (2022)
37. Liu, Y., et al.: Structure-informed shadow removal networks. IEEE Trans. Image Process. **32**, 5823–5836 (2023)
38. Liu, Y., Ke, Z., Xu, K., Liu, F., Wang, Z., Lau, R.W.: Recasting regional lighting for shadow removal. In: AAAI Conference on Artificial Intelligence, pp. 3810–3818 (2024)
39. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: unsupervised video object segmentation with co-attention siamese networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3618–3627 (2019)
40. Lu, X., et al.: Video shadow detection via spatio-temporal interpolation consistency training. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3106–3115 (2022)
41. Madison, C., Thompson, W., Kersten, D., Shirley, P., Smits, B.: Use of interreflection and shadow for surface contact. Percept. Psychophys. **63**(2), 187–194 (2001)
42. Martel-Brisson, N., Zaccarin, A.: Learning and removing cast shadows through a multidistribution approach. IEEE Trans. Pattern Anal. Mach. Intell. **29**(7), 1133–1146 (2007)
43. Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: VSPW: a large-scale dataset for video scene parsing in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4131–4141 (2021)
44. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: European Conference on Computer Vision, pp. 445–461 (2016)
45. Nadimi, S., Bhanu, B.: Physical models for moving shadow and object detection in video. IEEE Trans. Pattern Anal. Mach. Intell. **26**(8), 1079–1087 (2004)
46. Nah, S., et al.: NTIRE 2019 challenge on video deblurring and super-resolution: dataset and study. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1996–2005 (2019)

47. Nguyen, V., Vicente, T.F.Y., Zhao, M., Hoai, M., Samaras, D.: Shadow detection with conditional generative adversarial networks. In: IEEE International Conference on Computer Vision, pp. 4520–4528 (2017)
48. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: IEEE International Conference on Computer Vision, pp. 9225–9234 (2019)
49. Okabe, T., Sato, I., Sato, Y.: Attached shadow coding: estimating surface normals from shadows under unknown reflectance and lighting conditions. In: IEEE International Conference on Computer Vision, pp. 1693–1700 (2009)
50. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: ST-Adapter: parameter-efficient image-to-video transfer learning. Adv. Neural. Inf. Process. Syst. **35**, 26462–26477 (2022)
51. Panagopoulos, A., Wang, C., Samaras, D., Paragios, N.: Illumination estimation and cast shadow detection through a higher-order graphical model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 673–680 (2011)
52. Qu, L., Tian, J., He, S., Tang, Y., Lau, R.W.: DeshadowNet: a multi-context embedding deep network for shadow removal. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2308–2316 (2017)
53. Sanin, A., Sanderson, C., Lovell, B.C.: Shadow detection: a survey and comparative evaluation of recent methods. Pattern Recogn. **45**(4), 1684–1695 (2012)
54. Shao, Y., Taff, G.N., Walsh, S.J.: Shadow detection and building-height estimation using IKONOS data. Int. J. Remote Sens. **32**(22), 6929–6944 (2011)
55. Shen, L., Chua, T.W., Leman, K.: Shadow optimization from structured deep edge detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2067–2074 (2015)
56. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: European Conference on Computer Vision, pp. 744–760 (2018)
57. Stergiou, A., Poppe, R.: AdaPool: exponential adaptive pooling for information-retaining downsampling. IEEE Trans. Image Process. **32**, 251–266 (2022)
58. Sun, J., et al.: Adaptive illumination mapping for shadow detection in raw images. In: IEEE International Conference on Computer Vision, pp. 12663–12672 (2023)
59. Tian, X., Xu, K., Lau, R.: Unsupervised salient instance detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2702–2712 (2024)
60. Vasluianu, F.A., Seizinger, T., Timofte, R.: WSRD: a novel benchmark for high resolution image shadow removal. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1826–1835 (2023)
61. Vicente, T.F.Y., Hoai, M., Samaras, D.: Noisy label recovery for shadow detection in unfamiliar domains. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3783–3792 (2016)
62. Vicente, T.F.Y., Hou, L., Yu, C.P., Hoai, M., Samaras, D.: Large-scale training of shadow detectors with noisily-annotated shadow examples. In: European Conference on Computer Vision, pp. 816–832 (2016)
63. Vicente, T.F.Y., Samaras, D.: Single image shadow removal via neighbor-based region relighting. In: European Conference on Computer Vision Workshops, pp. 309–320 (2014)
64. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.C.: FEELVOS: fast end-to-end embedding learning for video object segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9473–9482 (2019)
65. Wang, J., Li, X., Yang, J.: Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1788–1797 (2018)

66. Wang, Y., Zhou, W., Mao, Y., Li, H.: Detect any shadow: segment anything for video shadow detection. IEEE Trans. Circuits Syst. Video Technol. 1–13 (2023)
67. Warren, A., Xu, K., Lin, J., Tam, G.K., Lau, R.W.: Effective video mirror detection with inconsistent motion cues. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 17244–17252 (2024)
68. Wu, Q., Yang, T., Wu, W., Chan, A.B.: Scalable video object segmentation with simplified framework. In: IEEE International Conference on Computer Vision, pp. 13879–13889 (2023)
69. Wu, W., Zhou, K., Chen, X.D., Yong, J.H.: Light-weight shadow detection via GCN-based annotation strategy and knowledge distillation. Comput. Vis. Image Underst. **216**, 1–12 (2022)
70. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems, vol. 34, pp. 12077–12090 (2021)
71. Xie, X., Zhou, P., Li, H., Lin, Z., Yan, S.: Adan: adaptive nesterov momentum algorithm for faster optimizing deep models. arXiv preprint arXiv:2208.06677, pp. 1–34 (2022)
72. Xu, K., Siu, T.W., Lau, R.W.: ZOOM: learning video mirror detection with extremely-weak supervision. In: AAAI Conference on Artificial Intelligence, pp. 6315–6323 (2024)
73. Yang, H., Wang, T., Hu, X., Fu, C.W.: SILT: shadow-aware iterative label tuning for learning to detect shadows from noisy labels. In: IEEE International Conference on Computer Vision, pp. 12641–12652 (2023)
74. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE International Conference on Computer Vision, pp. 3813–3824 (2023)
75. Zheng, Q., Qiao, X., Cao, Y., Lau, R.W.: Distraction-aware shadow detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5162–5171 (2019)
76. Zhu, L., et al.: Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In: European Conference on Computer Vision, pp. 122–137 (2018)
77. Zhu, L., Xu, K., Ke, Z., Lau, R.W.: Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In: IEEE International Conference on Computer Vision, pp. 4682–4691 (2021)
78. Zhu, Y., Fu, X., Cao, C., Wang, X., Sun, Q., Zha, Z.J.: Single image shadow detection via complementary mechanism. In: ACM International Conference on Multimedia, pp. 6717–6726 (2022)