SCOTCH and SODA: A Transformer Video Shadow Detection Framework

Lihao Liu[1], Jean Prost[2], Lei Zhu[3,4], Nicolas Papadakis[2], Pietro Liò[1],
Carola-Bibiane Schönlieb[1], Angelica I Aviles-Rivero[1]
[1] University of Cambridge, United Kingdom
[2] Univ. Bordeaux, CNRS, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France
[3] The Hong Kong University of Science and Technology (Guangzhou), China
[4] The Hong Kong University of Science and Technology, HK SAR, China

Abstract

Shadows in videos are difficult to detect because of the large shadow deformation between frames. In this work, we argue that accounting for shadow deformation is essential when designing a video shadow detection method. To this end, we introduce the shadow deformation attention trajectory (SODA), a new type of video self-attention module, specially designed to handle the large shadow deformations in videos. Moreover, we present a new shadow contrastive learning mechanism (SCOTCH) which aims at guiding the network to learn a unified shadow representation from massive positive shadow pairs across different videos. We demonstrate empirically the effectiveness of our two contributions in an ablation study. Furthermore, we show that SCOTCH and SODA significantly outperforms existing techniques for video shadow detection. Code is available at the project page: https://lihaoliu-cambridge.github.io/scotch_and_soda/

1. Introduction

Shadow is an inherent part of videos, and they have an adverse effect on a wide variety of video vision tasks. Therefore, the development of robust video shadow detection techniques, to alleviate those negative effects, is of great interest for the community. Video shadow detection is usually formulated as a segmentation problem for videos, however and due to the nature of the problem, shadow detection greatly differs from other segmentation tasks such as object segmentation. For inferring the presence of shadows in an image, one has to account for the global content information such as light source orientation, and the presence of objects casting shadows. Importantly, in a given video, shadows considerably change appearance (deformation) from frame to frame due to light variation and object motion. Finally, shadows can span over different backgrounds over different frames, making approaches relying on texture information unreliable.

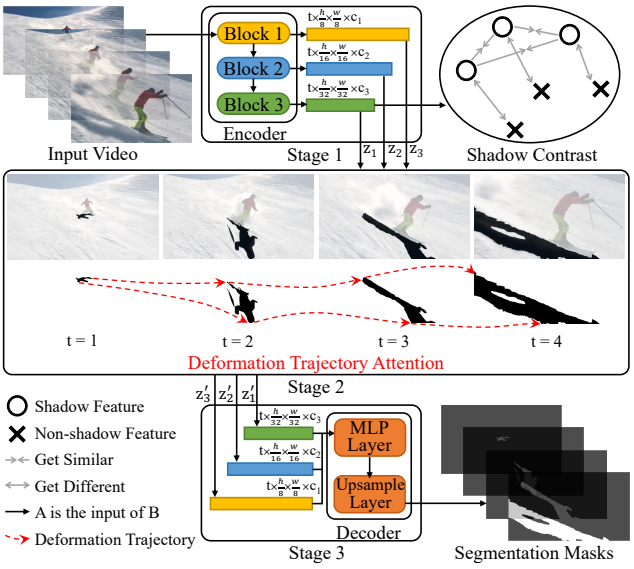


Figure 1. Overview of our SCOTCH and SODA framework. A MiT encoder extracts multi-scale features for each frame of the video (stage 1). Then, our deformation attention trajectory is applied to features individually to incorporate temporal information (stage 2). Finally, an MLP layer combines the multi-scale information to generate the segmentation masks (stage 3). The model is trained to contrast shadow and non-shadow features, by minimising our shadow contrastive loss with massive positive shadow pairs.

Particularly, video shadow detection methods can be broadly divided into two main categories. The first category refers to image shadow detection (ISD) [9,15,35,36,43,46]. This family of techniques computes the shadow detection frame by frame. Although computationally saving, these methods are incapable of handling temporal information. The second category refers to video shadow detection (VSD) [6, 9, 14, 16, 25]. These methods offer higher performance as the analysis involves spatial-temporal information. Hence, our main focus is video shadow detection.

State-of-the-art video shadow detection methods rely on deep neural networks, which are trained on large annotated datasets. Specifically, those methods are composed of three parts: (i) a feature extraction network that extracts spatial features for each frame of the video: (ii) a temporal aggregation mechanism [6, 14] enriching spatial features with information from different frames; and (iii) a decoder, that maps video features to segmentation masks. Additionally, some works enforce consistency between frames prediction by using additional training criterion [9,25]. We retain from these studies that the design of the temporal aggregation mechanism and the temporal consistency loss is crucial to the performance of a video shadow detection network, and we will investigate both of those aspects in this work.

The current temporal aggregation mechanisms available in the literature were typically designed for video tasks such as video action recognition, or video object segmentation. Currently, the most widely used temporal aggregation mechanism is based on a variant of the self-attention mechanism [1, 29, 32, 40, 41]. Recently, trajectory attention [29] has been shown to provide state-of-the-art results on video processing. Intuitively, trajectory attention aggregates information along the object's moving trajectory, while ignoring the context information, deemed as irrelevant. However, shadows in videos are subject to strong deformations, making them difficult to track, and thus they might cause the trajectory attention to fail.

In this work, we first introduce the ShadOw Deformation Attention trajectory (SODA), a spatial-temporal aggregation mechanism designed to better handle the large shadow deformations that occur in videos. SODA operates in two steps. First, for each spatial location, an associated token is computed between the given spatial location and the video, which contains information in every time-step for the given spatial location. Second, by aggregating every associated spatial token, a new token is yielded with enriched spatial deformation information. Aggregating spatial-location-to-video information along the spatial dimension helps the network to detect shape changes in videos.

Besides, we introduce the Shadow COnTrastive meCHanism (SCOTCH), a supervised contrastive loss with massive positive shadow pairs aiming to drive our network to learn more discriminative features for the shadow regions in different videos. Specifically, in training, we add a contrastive loss at the coarsest layer of the encoder, driving the features from shadow regions close together, and far from the features from the non-shadow region. Intuitively, this contrastive mechanism drives the encoder to learn high-level representations of shadow, invariant to all the various factors of shadow variations, such as shape and illumination.

In summary, our contributions are as follows:

- We introduce a new video shadow detection framework, in which we highlight:
  - SODA, a new type of trajectory attention that harmonise the features of the different video frames at each resolution.
  - SCOTCH, a contrastive loss that highlights a massive positive shadow pairs strategy in order to make our encoder learn more robust high-level representations of shadows.
- We evaluate our proposed framework on the video shadow benchmark dataset ViSha [6], and compare with the state-of-the-art methods. Numerical and visual experimental results demonstrate that our approach outperforms, by a large margin, existing ones on video shadow detection. Furthermore, we provide an ablation study to further support the effectiveness of the technical contributions.

2. Related Work

The task of video shadow detection has been extensively investigated in the community, in which solutions largely rely on analysing single frames (image shadow detection) or continuous multiple frames (video shadow detection). In this section, we review the existing techniques in turn, and then summarize the recent achievements in the video processing area to better illustrate the difference between existing work and our work.

2.1. Image Shadow Detection

Image shadow detection (ISD) can be cast as a semantic segmentation problem, where image object segmentation (IOS) methods can be used to solve this problem [8, 13, 19, 21, 22, 42]. However, IOS methods are not specifically designed for shadow detection. Hence, when re-training these methods directly for shadow detection, the performance is unsatisfactory due to the data bias.

Techniques focused on image shadow detection incorporate problem-specific shadow knowledge into the model architecture and the training criterion [7,9,15,16,35,36,43, 46]. For example, BDRAR [46] introduces a bidirectional pyramidal architecture for shadow detection. DSD [43] presents a distraction-aware shadow-module to reduce false positives. Chen *et al.* [7] make use of non-labelled data, during training, with a task-specific semi-supervised learning mechanism called MTMT. Wang *et al.* investigate the detection of shadows along with their corresponding objects [35, 36]. Finally, FSDNet [16] proposes a compact image shadow detection network. Whilst ISD techniques have demonstrated potential results, their performance on videos is limited by the lack of temporal information.

2.2. Video Shadow Detection

Another body of researchers has explored the task of shadow detection from the lens of video analysis. The work

of [6] proposes the TVSD model. It relies on a dual-gated co-attention module, to aggregate features from different frames, and uses a contrastive learning mechanism to drive the encoder to discriminate frames from different videos. Hu *et al.* [14] introduce an optical flow warping module to aggregate features from different frames. STICT [25] uses transfer learning, to transfer the knowledge of a supervised image shadow detection network to a video shadow detection network, without labelled videos, by training the network prediction to be consistent with respect to temporal interpolation [33]. Moreover, the technique called SC-Cor [9] presents a weakly supervised correspondence learning mechanism to enhance the temporal similarity of features corresponding to shadow region across frames.

2.3. Progresses in Video Processing

The specific nature of videos, containing spatial and temporal information, has motivated the design of deep neural network architectures for different video processing applications. Models based on 3D CNN [17, 31] process videos by sequentially aggregating the spatio-temporal local information using 3D convolutional filters, but fail to effectively capture long-range temporal dependencies. To alleviate this limitation, architectures using recurrent networks were introduced in [2, 30]. Moreover, another set of works uses spatio-temporal memory bank mechanism [27] or spatio-temporal attention mechanism [32] into the coarse layers of 3D CNN architectures [18, 26, 37] to better integrate the temporal information for video processing.

The success of the transformer network architecture, on a wide variety of vision tasks [4, 10], has motivated the use of transformer for video tasks. While the self-attention mechanism in transformers appears to be well suited to capture the long-range dependencies in videos, applying transformers to videos raises many challenges, such as the quadratic complexity in the input sequence length, and the large data requirement induced by the lack of problem-specific inductive bias. The works of [1, 41] propose to separate spatial and temporal attention to reduce computational complexity, and the authors of [40] propose to apply multiple encoders on multiple views of the video. Recently, trajectory attention [29] was introduced as a way to incorporate an inductive bias in the self-attention operation to capture objects moving trajectories for better video recognition tasks.

2.4. Existing Works & Comparison to Our Work

All precedent works on VSD [6, 14, 25, 33] rely on convolutional neural network architectures. *To the best of our knowledge, our work is the first video shadow detection approach based on transformers.*

Moreover, whilst the work of [29] also considers a type of attention trajectory, their modelling hypothesis is that the video objects do not change shape over time. This is a strong assumption to fulfill for several vision applications such as shadow detection; as shadows significantly change shape from frame to frame. Our work first mitigates this issue by modelling in the trajectories the inherent deformation of the shadows. Notice that TVSD [6] uses an image-level contrastive loss between frames from different videos, and SC-Cor [9] uses weakly-supervised correspondence learning for driving similar features from shadow-region close together. Unlike these works, we introduce a supervised contrastive strategy to contrast shadow features from non-shadow features. We underline that existing contrastive-based techniques assume that supervised contrastive learning is not performing better than its counterpart. In this work, we show that a well-designed supervised contrastive strategy indeed improves over existing works.

3. Methodology

In this section, we introduce all the components of the video shadow detection method presented in this work. After a global description of our framework (i), we introduce `SODA`, our new video self-attention module (ii), and we describe the training criterion of our model, which includes `SCOTCH`, our shadow contrastive mechanism (iii).

3.1. Framework Architecture Overview

Our proposed framework is composed of three main stages. The overall workflow is illustrated in Figure 1, and details are provided next.

Stage 1: Feature Extraction. In this stage, Mix Transformer (MiT) [39] is adopted as the encoder. MiT takes video clips as input, and it outputs a set of different-resolution spatial feature maps. Unlike ViT [10], which only generates single-resolution feature maps, the hierarchical structure of MiT can generate CNN-like multi-level multi-scale feature maps. These feature maps contain from high-resolution detailed information to low-resolution semantic information. By incorporating different levels of resolutions, the performance of tasks such as semantic segmentation can be boosted [39]. In this stage, MiT only encodes the spatial information of the given video clips (input), that is, the temporal information is not yet incorporated.

Stage 2: Harmonising Spatial and Temporal Information. The multi-scale feature maps from Stage 1 are then processed by the newly introduced shadow deformation attention trajectory module (`SODA`). The goal of this module is to capture the shadow's deformation trajectory along the frames in the video clips. Our deformation attention trajectory module processes independently each feature map with different-resolution. These processed feature maps are used as input to the decoder in the next stage.

Stage 3: Mask Shadow Generation. In this stage, a light-weighted decoder, with only MLP layers, is adopted to

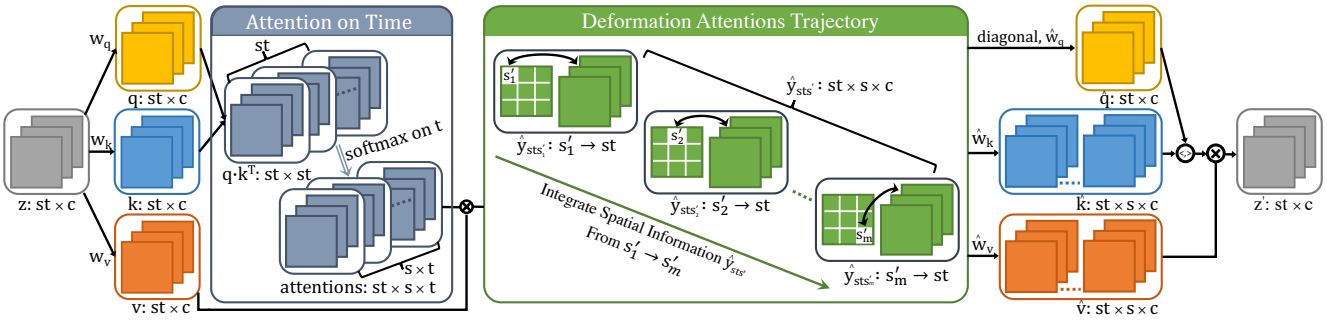


Figure 2. Deformation attention trajectory module. The input feature maps z is used to generate q, k, v, respectively. The q and k are first used to calculate the pointwise st-to-st similarity, followed by a softmax on t to get the time attention. The time attention aggregated v can generate s-to-st (spatial-location-to-video) attention, named deformation attention. Then, the spatial-location-to-video attention is integrated along the spatial dimension to capture the deformation attention trajectory. Lastly, the deformation attention trajectory is used to generate the final feature map z' with a second self-attention ($\hat{q}$, $\hat{k}$, $\hat{v}$). In this figure, each square represents a spatial feature map. (The channel dimension is not represented for simplicity).

reconstruct the shadow's segmentation masks. The decoder aims to mix the high- & low- resolution information, from processed feature maps, for better semantic segmentation. The output of the decoder is a segmented video that marks out the shadows in each frame of the video.

3.2. SODA: ShadOw Deformation Attention trajectory

Transformers have revolutionised several tasks in the computer vision area. The key is the self-attention mechanism that can accommodate with any given type of data and domain. However, for videos, the standard self-attention does not differentiate the spatial dimensions from the temporal dimension. This can lead the attention to focus on the redundant spatial information while neglecting the informative temporal variations in the videos. Nonetheless, video analysis is inherent to such temporality. Most recently, trajectory attention [29] has proposed to accommodate somehow with such issues. However and even though trajectory attention has demonstrated potential results, it has a major limitation – *the objects are assumed not to change over time*. This is a major constraint in several video tasks including shadow detection; as shadows significantly undergo deformation from one frame to another. In this subsection, we introduce a new scheme called **Sh**ad**O**w **D**eformation **A**ttention trajectory (SODA) to mitigate current drawbacks of the literature.

Like in the classical self-attention setting, it begins with an input feature map generated from the encoder. Specifically, let us denote $f_d \in \mathbb{R}^{t \times \frac{h}{d} \times \frac{w}{d} \times c}$, the generated feature map from the encoder, where d is the spatial downsampling ratio, c is the number of feature channels, and t is the number of time frames in the video. f_d is reshaped to a sequence of 1D token embedding denoted as $z \in \mathbb{R}^{n \times c}$, where $n = t \times \frac{h}{d} \times \frac{w}{d}$. As shown in the left part of Figure 2, z is then mapped to a set of query-key-value vectors $q, k, v \in \mathbb{R}^{n \times c}$ using linear projections $q = w_q \cdot z$, $k = w_k \cdot z$, $v = w_v \cdot z$, with projection matrices $w_q, w_k, w_v \in \mathbb{R}^{n \times n}$.

Our scheme considers two main parts: (i) temporal attention between the spatial location and video (Attention on Time in Figure 2), (ii) intra-space attention to capture deformation statues within a spatial scene (Deformation Attention Trajectory in Figure 2). For the first part, given a space-time position in the video $st \in \{1, \cdots, n\}$, and a spatial location $s' \in \{1, \cdots, m\}$, where $m = \frac{n}{t}$, the *temporal attention* (deformation) between the space-time position st and the spatial locations s' is computed as:

$$\hat{y}_{sts'} = \sum_{t'} v_{s't'} \cdot \frac{\exp\langle q_{st}, k_{s't'} \rangle}{\sum_{\bar{t}} \exp\langle q_{st}, k_{s'\bar{t}} \rangle} \tag{1}$$

For brevity, the notation is slightly abused by omitting the "softmax operation on time dimension" in the Fig. 2 applied to the fraction, as well as the scaling parameter $\sqrt{n}$ (we will keep this notation convention throughout the paper). The deformation encodes the connection between one space-time position and one spatial location, which indicates how the content of the space-time position st is presented in spatial location s'.

Once the temporal attentions are computed, the intra-space attention is then estimated to aggregate the spatial-location-to-video responses to space-level deformation. To do this, the computed deformation tokens are projected to a new set of query-key-value vectors using linear projections:

$$\hat{q}_{st} = \hat{w}_q \cdot \hat{y}_{sts}, \ \hat{k}_{sts'} = \hat{w}_k \cdot \hat{y}_{sts'}, \ \hat{v}_{sts'} = \hat{w}_v \cdot \hat{y}_{sts'} \tag{2}$$

where $\hat{y}_{sts}$ is the temporal connection from st to the same spatial location s, and $\hat{q}_{st}$ corresponds to the deformation reference point st that is used to aggregate the location-to-video connection:

$$\hat{y}_{st} = \sum_{s'} \hat{v}_{sts'} \cdot \frac{\exp\langle \hat{q}_{st}, \hat{k}_{sts'} \rangle}{\sum_{\bar{s}} \exp\langle \hat{q}_{st}, \hat{k}_{st\bar{s}} \rangle} \tag{3}$$

where $\hat{y}_{st}$ is the deformation attention output. The meaningful location-to-video tokens are pooled out to form the full space-level deformation status. By computing the intra-space attention, the attended feature map can capture the deformation status in different frames of the video, thus boosting the video shadow detection performance.

3.3. SCOTCH: Shadow COnTrastive meCHanism

Contrastive learning [5, 20, 23] has been proven to be an effective mechanism for learning distinctive features. By contrasting the positive pairs with high similarity and negative pairs with low similarity, the learned feature maps can be more discriminative in downstream tasks including classification and segmentation. In previous video shadow detection task [6], positive and negative pairs are sampled from frames from the same video and from two different videos respectively. Since the frames from one video have high similarity image content, the contrastive mechanism can help to discriminate different video content.

However, the key element in video shadow detection is the shadow itself instead of the video content. With the goal of boosting the detection performance, we introduce `SCOTCH`, a **S**hadow **CO**n**T**rastive me**CH**anism. `SCOTCH` seeks to better guide the segmentation process for shadows and non-shadows regions in the videos. Specifically, to learn a unified shadow feature for different videos, positive pairs are sampled from the shadow regions from different frames in different videos, whilst negative pairs are sampled as shadow and non-shadow regions on the frames in different videos. *We underline that unlike the classical contrastive loss used for unsupervised learning [5, 28], where there is only a small number of positive pairs, we proposed a massive positive shadow paired contrastive loss.* The key idea behind our loss is that – we seek to not only maximise the difference between shadow and non-shadow features, but also maximise the similarity between features of shadows presented in different videos. All shadow and non-shadow features are cropped from the last layer of the encoder presented in Section 3.1, with the supervision of the segmentation masks. The contrastive loss reads:

$$\ell_{contrast}\left(\boldsymbol{v}, \boldsymbol{v}^{+}, \boldsymbol{v}^{-}\right) = -\log\left[\frac{\sum_{n=1}^{N}\exp\left(\boldsymbol{v}\cdot\boldsymbol{v}_{n}^{+}/\tau\right)}{\sum_{n=1}^{N}\exp\left(\boldsymbol{v}\cdot\boldsymbol{v}_{n}^{+}/\tau\right)+\sum_{n=1}^{N}\exp\left(\boldsymbol{v}\cdot\boldsymbol{v}_{n}^{-}/\tau\right)}\right] \tag{4}$$

where $\boldsymbol{v} \in \mathbb{R}^{c}$ is the query shadow feature. Moreover, $\boldsymbol{v}_{n}^{+}, \boldsymbol{v}_{n}^{-} \in \mathbb{R}^{n\times c}$ are the positive and negative groups respectively, and τ is a temperature hyperparameter. The final loss is computed across all frames in a mini-batch fashion.

Optimisation Scheme for Shadow Detection. Finally, to compute the shadow segmentation loss, we follow the default setting in [6]. We use the binary cross entropy (BCE) loss with a lovasz-hinge loss [3]. These two terms are added to define the shadow segmentation loss as follows:

$$\ell_{seg} = \ell_{bce} + \lambda_1 \ell_{hinge} \tag{5}$$

Our optimisation scheme is then given by (4) and (5) as:

$$\ell_{final} = \ell_{bce} + \lambda_1 \ell_{hinge} + \lambda_2 \ell_{contrast} \tag{6}$$

where λ_1 and λ_2 are two hyper-parameters weighting the relative effect of the hinge loss and the contrastive loss in the final loss. In the following experiments, λ_1, λ_2 were empirically set to a value of 1 and 0.1, respectively.

4. Experimental Results

This section details all experiments performed to validate our proposed framework.

4.1. Dataset and Evaluation Metrics

Data Description. We utilise the largest and latest **Vi**deo **Sha**dow dataset (**ViSha**) [6] to evaluate the effectiveness of our proposed VSD method. The ViSha dataset has 120 videos, and each video contains between 29 and 101 frames. ViSha is composed of a total of 11,685 frames corresponding to a total duration of 390 seconds of video.

Data Pre-processing. We follow the setting introduced in ViSha [6]. That is, we use the same train-test split, with 50 videos for training and 70 videos for testing. During training, we also use the same data augmentation strategy as [6] to enrich the variety of the dataset. Specifically, during training, images are re-scaled to size 512×512, and are randomly flipped horizontally. In testing, only re-scaling to the unified size 512×512 is used.

Evaluation Metrics. Following the evaluation protocol used in [6, 9, 25], we employ four common evaluation metrics to measure the shadow detection accuracy: MAE, F_β, IoU, and BER. Lower MAE and BER scores, and higher F_β and IoU scores indicate a better video shadow detection result. Moreover, we also provide the shadow BER (S-BER) and the non-shadow BER scores (N-BER) to further compare different VSD methods.

Implementation Details. Our proposed segmentation architecture is built using the PyTorch-lightning [11] deep-learning framework. The parameters of the feature extraction encoder are initialised using the weights from the MiT-B3 model pre-trained for image segmentation on ADE20K dataset [44, 45], publicly available on HuggingFace [38]. The remaining parameters (attention modules and the MLP decoder) are randomly initialised using "Xavier" methods [12]. During training, AdamW optimizer [24] is used with an initial learning rate of 1×10^{-6} without decay. All experiments and ablation studies are trained for 36 epochs, for a training time of approximately 12 hours on NVIDIA A100 GPU with 80G RAM with batch size of 8.

Methods		Evaluation Metrics					
Tasks	Techniques	MAE ↓	F_β ↑	IoU ↑	BER ↓	S-BER ↓	N-BER ↓
IOS	⋆ FPN [19]	0.044	0.707	0.512	19.49	36.59	2.40
	PSPNet [42]	0.051	0.642	0.476	19.75	36.44	3.07
	DSS [13]	0.045	0.696	0.502	19.77	36.96	2.59
	R^3Net [8]	0.044	0.710	0.502	20.40	37.37	3.55
ISD	BDRAR [46]	0.050	0.695	0.484	21.29	40.28	2.31
	⋆ DSD [43]	0.043	0.702	0.518	19.88	37.89	1.88
	MTMT [7]	0.043	0.729	0.517	20.28	38.71	1.86
	FSDNet [16]	0.057	0.671	0.486	20.57	38.06	3.06
VOS	PDBM [30]	0.066	0.623	0.466	19.73	34.32	5.16
	COSNet [26]	0.040	0.705	0.514	20.50	39.22	1.79
	⋆ FEELVOS [34]	0.043	0.710	0.512	19.76	37.27	2.26
	STM [27]	0.068	0.597	0.408	25.69	47.44	3.96
VSD	TVSD [6]	0.033	0.757	0.567	17.70	33.97	1.45
	STICT [25]	0.046	0.702	0.545	16.60	29.58	3.59
	SC-Cor [9]	0.042	0.762	0.615	13.61	24.31	2.91
	⋆ SCOTCH and SODA	**0.029**	**0.793**	**0.640**	**9.066**	**16.26**	**1.44**

Table 1. Comparisons between our proposed technique and SOTA techniques on the ViSha dataset. "MAE" denotes mean absolute error, "F_β" denotes F-measure score, "IoU" denotes intersection over union, "BER" denotes balance error rate, and "S-BER" means shadow BER, "N-BER" means non-shadow BER. The ↑ denotes the higher the value is the better the performance is, whilst the ↓ means the opposite. ⋆ indicates the best performed network in each category.

4.2. Comparison to SOTA Techniques

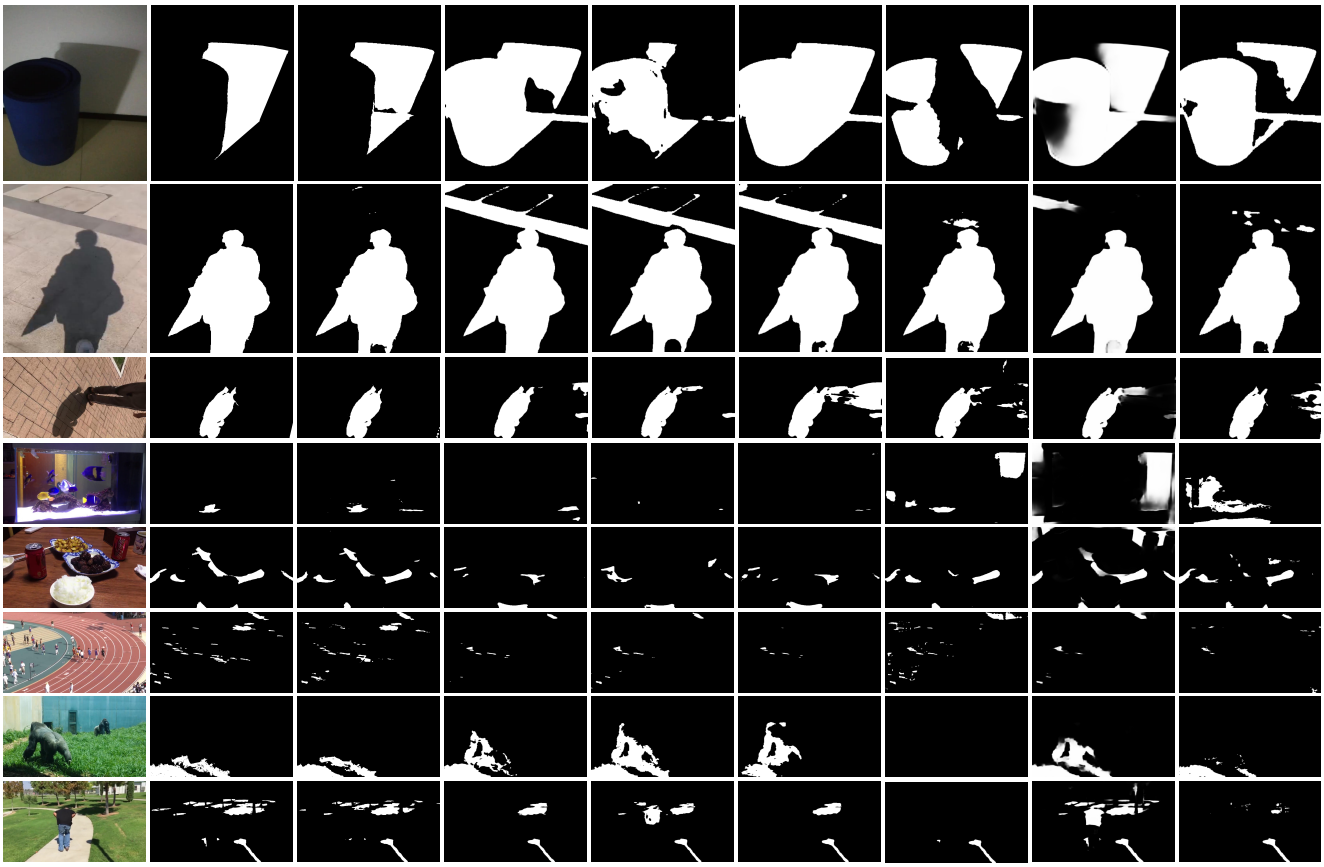
Compared Methods. Video shadow detection is a relatively recent topic, and there are only three directly-related methods designed for this task. Hence, following existing VSD methods, we make comparisons against 4 different kinds of methods, including image object segmentation (IOS), image shadow detection (ISD), video object segmentation (VOS), and video shadow detection (VSD). For IOS, they are FPN [19], PSPNet [42], DSS [13], and R^3-Net [8], while the ISD methods are BDRAR [46], DSD [43], MTMT [7], and FSDNet [16]. The compared VOS methods include PDBM [30], COSNet [26], FEELVOS [34], and STM [27], while compared VSD methods are TVSD [6], STICT [25], and SC-Cor [9]. We obtain the results by re-training their network parameters with unified training parameters or by downloading the results from the TVSD [6] repository.

Quantitative Comparisons. Table 1 summarises MAE, F_β, IoU, and BER scores of our network against SOTA techniques. For each category, we use ⋆ to mark out the method with the best performance. From these quantitative results, we observe that the IOS and ISD report readily competing results. These results are expected as the modelling hypothesis for both families of techniques relies on only the image level analysis. Whilst VOS techniques consider also temporal information, these techniques are customised as a general framework for video object segmentation. However, shadow detection is more complex due to the fast change in appearance between frames. Notably, we observe that VSD techniques indeed provide a substantial performance improvement compared to other methods; as they are designed considering the complexity of shadows.

More importantly, our method outperforms all other techniques by a significant margin for all evaluation metrics – further supporting the superiority of our approach. In particular, our method yields to a balanced error rate of 9.066 which is more than 4 points below SC-Cor, the latest SOTA technique, in terms of BER. The error rate for the shadow label (denoted as S-BER in Table 1) of our method is 8 points below SC-Cor. We also underline that our significant improvement in performance comes with a negligible computational cost. We report the test time in Table 2, where we observe that our SCOTCH and SODA framework only requires a fraction of time than compared methods.

Visual Comparison. Figure 3 visually compares the shadow segmentation masks from our method and the compared methods on different input video frames. For video frames with black objects at the first three rows, we find that compared methods tend to wrongly identify those black objects as shadow ones, while our method can still accurately detect shadow pixels under the corruption of black objects. Moreover, compared methods tend to miss some shadow regions when the input video frames contain multiple shadow detection, as shown in the 4-th row to the 8-th row of Figure 3. On the contrary, our method can identify all these shadow regions of input video frames, and our detection results are most consistent with the ground truth in the 2nd column of Figure 3. We also provide the video segmentation masks on the project page to demonstrate the temporal coherence of the results provided by our method.

Model size and inference time. Table 2 further



(a) Image (b) GT (c) Ours (d) FPN (e) DSD (f) FEELVOS (g) TVSD (h) STICT (i) SC-Cor

Figure 3. Visual comparisons of video shadow detection results produced by our network (SCOTCH and SODA) and compared methods. Apparently, our method can more accurately identify shadow pixels and our results are more consistent with the ground truth in the 2nd column. The video segmentation results can be found on the project page. (d-e) are the methods with the highest performance in IOS, ISD, and VOS in Table 1, whilst (g-i) are all from the VSD area.

Networks	Params	GMACs	Time	BER
TVSD [6]	243.32	158.89	32.4	17.70
STCIT [25]	**104.68**	**40.99**	<u>13.5</u>	16.60
SC-Cor [9]	232.63	218.4	21.8	<u>13.61</u>
SCOTCH and SODA	<u>211.79</u>	<u>122.46</u>	**9.15**	**9.006**

Table 2. Comparison of model size (Params), computational complexity (GMACs), inference time (Time), and segmentation accuracy (BER). Specifically, the units for Params and Time are (MB) and (Mins), respectively. We denote the **best** and <u>second best</u> in **bold** and <u>underline</u> font.

compares our network and three state-of-the-art video shadow detection methods in terms of the model size (Params), computational complexity (GMACs), inference time (Time), and segmentation accuracy (BER). Apparently, among the three compared VSD methods, STICT has the smallest testing time (13.5 minutes). Compared to STICT, our method further reduces the inference time from 13.5 minutes to 9.15 minutes to test all 70 testing videos with 6,897 images. Although our method takes 2nd rank in terms of the model size and computational complexity, they are only larger than STICT. This is because we only use a light-weighted MLP layer as the decoder to integrate the multi-resolution feature maps, which is computational-saving. In terms of performance, our method has a superior BER performance than STICT by reducing the BER score from 16.60 to 9.006, which indicates that with a minor compromise on the model size, our method can more accurately identify video shadows than STICT, TVSD, and SC-Cor.

4.3. Ablation Study

In Table 3, we perform ablation studies on our main contributions to evaluate the effectiveness of each component.

Baseline with MiT backbone. In order to evaluate the role of the MiT backbone on the final performance of our method, we define the segmentation network with MiT backbone as the baseline, which is trained by using the classical segmentation loss (5), but *without* the deformation attention trajectory and shadow level contrastive mechanism.

Components			Evaluation Metrics			
Backbone	Attention	Contrast	MAE ↓	F_β ↑	IoU ↑	BER ↓
MiT	✗	✗	0.048	0.755	0.584	13.18
MiT	Trajectory	✗	0.048	0.760	0.593	12.35
MiT	SODA	✗	0.041	0.791	0.613	10.55
MiT	✗	Image	0.048	0.761	0.588	12.85
MiT	✗	Shadow	0.041	0.767	0.592	12.02
MiT	✗	SCOTCH	0.034	0.771	0.606	11.29
† MiT	SODA	SCOTCH	**0.029**	**0.793**	**0.640**	**9.066**

Table 3. Ablation study on different components of our proposed methods on the ViSha dataset. The ↑ denotes the higher the value is the better the performance is, whilst the ↓ means the opposite. "†" denotes our final methods with the highest performance in all evaluation metrics. Notations : **best**, second best, third best.

This baseline already provides results on par or better than the previous three works in the VSD area (see Table 1 for comparison), even without considering any kind of temporal information. This illustrates the superior performance of the MiT transformer-based architecture over convolutional architectures on the task of video shadow detection.

Attention mechanisms. We then evaluate the effectiveness of different attention mechanisms, including trajectory attention [29] and our newly introduced shadow deformation attention. Both types of attention modules provide an improvement over the baseline, which was to be expected as those modules give the ability to consider the temporal information within the videos. Our deformation attention trajectory module also appears to provide better results than the trajectory attention, indicating the importance of considering the deformation in the design of the attention module.

Contrastive losses. Next, we compare the effect of two types of contrastive criterion, the image level contrastive loss used in TVSD [6], and our feature-level shadow contrastive loss. The inter-frame contrastive learning slightly improves the baseline, whereas our shadow contrastive loss provides a significant improvement over the baseline. Those results illustrate the superiority of the features-level contrastive loss over the frame-level contrastive loss.

Final model. SCOTCH reduces the variance of the spatial shadow features, while SODA processes the spatial features at different time-step to consider the temporal information in the video. Thus, SCOTCH and SODA have a complementary effect, as feeding SODA with more robust spatial features from SCOTCH, we are able to reach the best performance, outperforming all the previous settings.

4.4. Attention Map Visualisation

In Figure 4, we visually compare the attention maps from the baseline model (second row) with our proposed SCOTCH and SODA (third row). We can observe that the baseline model highlights the object part, whilst ours focuses more on the shadow region with the help of contrastive learning. We can also observe that the baseline model might lose track of shape in different frames (third column), whilst our approach provides consistent shape information at different times with the help of the deformation attention trajectory. Hence, the model using the shadow deformation attention trajectory and the shadow contrastive loss is better at tracking the shadow region during different frames, while ignoring the non-shadow part of the image.

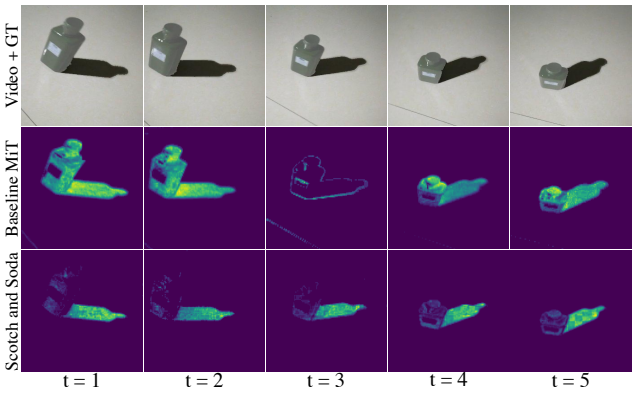


Figure 4. Attention maps visualisation. The low-resolution layers of the MiT encoder are selected for the visualization, with the query from the center of the ground truth shadow mask. From top to bottom row are the input video blended with segmentation mask, attention maps for baseline MiT methods, and attention maps from ours SCOTCH and SODA with deformation attention trajectory and shadow contrastive mechanism.

5. Conclusion

In this paper, we introduced SCOTCH and SODA, a new transformer video shadow detection framework. We developed shadow deformation attention trajectory (SODA), a self-attention module specially designed to handle the shadow deformation in videos, and we introduced a shadow contrastive mechanism (SCOTCH) to guide our network to better discriminate between shadow and non-shadow features. We demonstrate the effectiveness of the contributions with ablation studies. Finally, we show that our proposed method outperforms by a large margin concurrent video shadow segmentation works on the ViSha dataset.

Acknowledgements. This work is supported by Girton Postgraduate Research Scholarships, GSK Ph.D. Scholarship, CMIH, CCIMI, Philip Leverhulme Prize, Royal Society Wolfson Fellowship, EPSRC Advanced Career Fellowship EP/V029428/1, EPSRC grants EP/S026045/1, EP/T003553/1, EP/N014588/1, EP/T017961/1, Wellcome Innovator Awards 215733/Z/19/Z and 221633/Z/20/Z, EU Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 777826 NoMADS, and Guangzhou Municipal Science and Technology Project Grant No. 2023A03J0671.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 2, 3

[2] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015. 3

[3] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018. 5

[4] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *Advances in Neural Information Processing Systems*, 34:19594–19607, 2021. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[6] Zhihao Chen, Liang Wan, Lei Zhu, Jia Shen, Huazhu Fu, Wennan Liu, and Jing Qin. Triple-cooperative video shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2715–2724, 2021. 1, 2, 3, 5, 6, 7, 8

[7] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 5611–5620, 2020. 2, 6

[8] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 684–690. AAAI Press Menlo Park, CA, USA, 2018. 2, 6

[9] Xinpeng Ding, Jingweng Yang, Xiaowei Hu, and Xiaomeng Li. Learning shadow correspondence for video shadow detection. *arXiv preprint arXiv:2208.00150*, 2022. 1, 2, 3, 5, 6, 7

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] William Falcon et al. Pytorch lightning. *GitHub. Note: https://github. com/PyTorchLightning/pytorch-lightning*, 3(6), 2019. 5

[12] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 5

[13] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3203–3212, 2017. 2, 6

[14] Shilin Hu, Hieu Le, and Dimitris Samaras. Temporal feature warping for video shadow detection. *arXiv preprint arXiv:2107.14287*, 2021. 1, 2, 3

[15] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2795–2808, 2019. 1, 2

[16] Xiaowei Hu, Tianyu Wang, Chi-Wing Fu, Yitong Jiang, Qiong Wang, and Pheng-Ann Heng. Revisiting shadow detection: A new benchmark dataset for complex world. *IEEE Transactions on Image Processing*, 30:1925–1934, 2021. 1, 2, 6

[17] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 3

[18] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7274–7283, 2019. 3

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2, 6

[20] Lihao Liu, Angelica I Avilés-Rivero, and Carola-Bibiane Schönlieb. Contrastive registration for unsupervised medical image segmentation. *arXiv preprint arXiv:2011.08894*, 2020. 5

[21] Lihao Liu, Chenyang Hong, Angelica I Aviles-Rivero, and Carola-Bibiane Schönlieb. Simultaneous semantic and instance segmentation for colon nuclei identification and counting. *arXiv preprint arXiv:2203.00157*, 2022. 2

[22] Lihao Liu, Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. ψ-net: Stacking densely convolutional lstms for sub-cortical brain structure segmentation. *IEEE transactions on medical imaging*, 39(9):2806–2817, 2020. 2

[23] Lihao Liu, Zhening Huang, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Pc-swinmorph: Patch representation for unsupervised medical image registration and segmentation. *arXiv preprint arXiv:2203.05684*, 2022. 5

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[25] Xiao Lu, Yihong Cao, Sheng Liu, Chengjiang Long, Zipei Chen, Xuanyu Zhou, Yimin Yang, and Chunxia Xiao. Video shadow detection via spatio-temporal interpolation consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3116–3125, 2022. 1, 2, 3, 5, 6, 7

[26] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3623–3632, 2019. 3, 6

[27] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 3, 6

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[29] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021. 2, 3, 4, 8

[30] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 715–731, 2018. 3, 6

[31] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[33] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022. 3

[34] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 6

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

[36] Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1880–1889, 2020. 1, 2

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 5

[39] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3

[40] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022. 2, 3

[41] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13577–13587, 2021. 2, 3

[42] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2, 6

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

[44] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[45] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 5

[46] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018. 1, 2, 6