Structure-Aware Spatial-Temporal Interaction Network for Video Shadow Detection

Housheng Wei[1†], **Guanyu Xing**[1†], **Jingwei Liao**[2], **Yanci Zhang**[3] and **Yanli Liu**[3*]
[1]National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University
[2]Department of Information Sciences and Technology, George Mason University
[3]College of Computer Science, Sichuan University
2020326040004@stu.scu.edu.cn, xingguanyu@scu.edu.cn, jliao2@gmu.edu, {yczhang, yanliliu}@scu.edu.cn

Abstract

Video shadow detection faces significant challenges due to ambiguous semantics and variable shapes. Existing video shadow detection algorithms typically overlook the fine shadow details, resulting in inconsistent detection between consecutive frames in complex real-world video scenarios. To address this issue, we propose a spatial-temporal feature interaction strategy, which refines and enhances global shadow semantics with local prior features in the modeling of shadow relations between frames. Moreover, a structure-aware shadow prediction module is proposed, which focuses on modeling the distance relation between local shadow edges and regions. Quantitative experimental results demonstrate that our approach significantly outperforms the state-of-the-art methods, providing stable and consistent shadow detection results in complex video shadow scenarios.

1 Introduction

Video shadow detection has long served as a valuable cue for various tasks in computer vision and computer graphics, such as lighting estimation [Adams *et al.*, 2022], occlusion relationship estimation [Hao *et al.*, 2021], and scene geometry reconstruction [Karsch *et al.*, 2011]. In Augmented Reality (AR) applications, modeling the harmonized shadow casting between real scenes and virtual objects is important to enhance the visual realism of synthetic scenes [Liu *et al.*, 2022; Adams *et al.*, 2022]. A prerequisite of shadow cast modeling is obtaining consistent shadow detection results from background videos.

Recently, deep learning-based video shadow detection algorithms[Chen *et al.*, 2021; Ding *et al.*, 2022; Lu *et al.*, 2022; Hu *et al.*, 2021a; Lin and Wang, 2022; Liu *et al.*, 2023] eliminate the performance dependency of traditional algorithms on manually set parameters, and demonstrate powerful generalization capabilities to various scenes. These methods can be broadly categorized into two categories: the first one adapts image-based shadow detection methods to meet the demand of consistency across video frames, by using interpolation consistency training [Lu *et al.*, 2022] and contrastive learning [Ding *et al.*, 2022] to ensure similar shadow features extraction; the second category explicitly model the temporal shadow feature transferring, and simultaneously enhance the similarity of features after information fusion [Chen *et al.*, 2021; Hu *et al.*, 2021a; Lin and Wang, 2022; Liu *et al.*, 2023]. By increasing the sharing of shadow features in the temporal domain, methods in the second category demonstrate superior detection performance compared to those in the first.

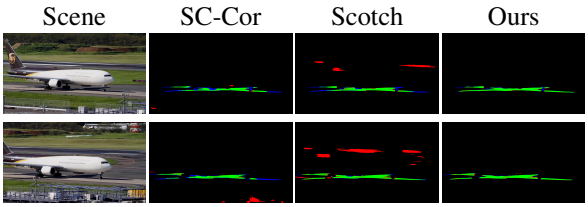

Figure 1: Visual comparison of detection errors from our method and the state-of-the-art shadow detection approaches. From left to right, the shadow scene image, the shadow detection errors maps (with TP, FP, FN) of SC-Cor [Ding *et al.*, 2022], Scotch [Liu *et al.*, 2023], and Ours. For better comparison, the true positives, false positives, and false negatives of shadow results are labeled with green, red, and blue, respectively.

While focusing on modeling temporal relationships of shadow features across video sequences, previous algorithms tend to overlook the details of shadows. For instance, TVSD [Chen *et al.*, 2021] conducts inter-frame information propagation only on low-resolution high-level shadow features, lacking frame-to-frame sharing of local details. Hu et al. [Hu *et al.*, 2021a] employ the optical flow maps to register shadow features, nonetheless the accuracy and consistency of shadow detection is limited by the short-term perspective and accuracy of optical flow maps. Recently, transformer-based video shadow detection algorithms [Liu *et al.*, 2023] have shown excellent capabilities in modeling global shadow semantics, such as shadow deformations. Nevertheless, transformers are unable to effectively capture local semantics [Chen *et al.*, 2022a], leading to inconsistent outcomes, particularly evident in complex video shadow scenes. Besides, existing algorithms use per-pixel independent labels to supervise shadow learning, neglecting the structural relationships between local shadow boundaries and the areas of shadow and non-shadow.

The defective designs of the above methods lead to discontinuous predictions for adjacent pixels, especially in nonshadow areas with significant intensity fluctuations. As shown in Figure 1, in the face of complex shadow situations in real-world scenarios, existing video shadow detection algorithms are prone to errors such as unstable shadow boundaries between consecutive frames, misidentifying dark objects, or overlooking light-colored shadows. These errors significantly impact the applications that demand high shadow detection precision. For instance, in AR, these errors usually cause flickering, overlap, or hollow areas in the integration between real and virtual shadows, thereby adversely affecting the realistic experience of AR.

An intuitive solution is to find inspiration from video object detection and segmentation tasks. However, the video shadow detection task presents significantly different characteristics from general object segmentation. Firstly, shadows in videos often exhibit vague textures in shadow areas due to reduced illumination. Secondly, the semantic content within shadow regions is ambiguous, lacking specific identifiers such as unique IDs. Consequently, existing object detection and segmentation algorithms [Huang *et al.*, 2022; Li *et al.*, 2023; Zhou *et al.*, 2022], which predominantly rely on abstract semantic modeling of object query embedding, struggle to effectively extract features from shadows.

To tackle the above challenges, we propose a new video shadow detection network. We adopt a transformer and lightweight CNN to extract video shadow features with consistent global semantics and fine local details, respectively. It's common and straightforward to directly add or stack two features in and between frames. However, the local and global semantics are not only complement each other. Under the assumption that the lighting conditions between two adjacent frames of a scene do not change drastically [Chen *et al.*, 2021], the similarity exists due to the sharing of lighting conditions in adjacent frames, and inconsistencies arise due to the shadow movement or deformation between frames. Therefore, we consider the inherent connection between the local and global semantics of multi-frame shadows. We first develop a spatial-temporal interaction strategy that enhances and refines these features in a bidirectional way and shares similar features between frames. Then, we propose a structure-aware shadow prediction head to decode the enhanced multi-scale features with local boundary details and shadow body semantics.

As for the spatial-temporal interaction strategy, we propose an enhance-exchange-refinement workflow to obtain shadow features with fine details. First, a spatial feature injection module (SFIM) is designed to enhance global shadow semantics with local shadow detail; Second, a temporal feature interaction module (TFIM) is proposed to share the enhanced features between frames. We integrate the SFIM and TFIM into the iterations of the transformer encoder, progressively aggregating and optimizing these temporally varying shadow features with different receptive fields. Between each encoder stage, a spatial-temporal feature refinement module (STFRM) is proposed to refine the spatial-temporal shadow features and output multi-scale shadow features for further prediction. As for the structure-aware shadow detection module, we decouple the shadow scene with the structure of the nonshadow region, shadow edge, and shadow region. Different from the methods that learn the pixel-level shadow boundaries directly, we manage to understand the structure relation between shadow regions and boundaries, and use distance transformation [Wei *et al.*, 2020] to represent the structural relation. Then, two complement semantic decoders are proposed to learn and refine local and global structure information, which are merged to predict the shadows. By deconstructing the relationship between shadow regions and boundaries, the proposed network achieves stable detection in scenes with significant intensity variations.

In summary, our contributions are as follows:

- We propose a novel video shadow detection network with a spatial-temporal shadow feature interaction strategy that enhances, exchanges, and refines global semantics and local details in multi-stage video shadow feature extraction.
- We present a structure-aware shadow prediction module, decoupling multi-scale shadow features into shadow regions and shadow edges with distance relation, and ensuring the robustness of shadow detection.
- We evaluate the performance of the proposed algorithm on common video shadow detection datasets and compare it with state-of-the-art (SOTA) methods. Quantitative analysis and visual results demonstrate that our framework significantly outperforms current SOTA algorithms.

2 Related Work

2.1 Video Shadow Detection

The video shadow detection task is dedicated to stably and accurately detecting shadow regions in continuous video frames. In earlier years, video shadow detection algorithms based on traditional methods were proposed [Gomes *et al.*, 2017; Shi and Liu, 2020], relying heavily on manual parameter tuning, which made them impractical for scenarios with dramatic scene changes.

Recently, with the advent of large-scale video shadow detection datasets [Chen *et al.*, 2021], researchers have begun to explore video shadow detection algorithms based on deep learning [Liu *et al.*, 2023; Lin and Wang, 2022; Chen *et al.*, 2022b; Hu *et al.*, 2021a; Ding *et al.*, 2022; Lu *et al.*, 2022]. These algorithms achieve consistent shadow detection results by considering and modeling the temporal relationships of shadows. For the temporal aggregation mechanisms in video shadow detection algorithms, the mainstream approaches include direct feature-level conditional constraints and modeling frame-to-frame dependency relations. The former improves the consistency of shadow detection results by reducing the discrepancy between shadow features of adjacent frames [Ding *et al.*, 2022; Lu *et al.*, 2022], while the latter achieves consistent shadow detection by modeling long-term dependencies [Chen *et al.*, 2021; Liu *et al.*, 2023; Lin and Wang, 2022] or short-term dependencies [Hu *et al.*, 2021a]. Among these, the schemes based on Video Transformers [Liu *et al.*, 2023; Lin and Wang,

2022] offer greater flexibility and modeling capacity compared to CNN-based methods [Chen *et al.*, 2021]. However, the self-attention mechanism of Video Transformers, which focuses on both spatial and temporal dimensions, introduces a large number of parameters, posing optimization challenges for video shadow detection tasks with limited data. Different from the existing transformer-based methods, we use two adjacent frames to realize shadow relation modeling, which is more effective than the joint spatial-temporal attention method.

2.2 Vision Transformer in Dense Prediction

In dense prediction tasks such as video shadow detection, a common architecture comprises a feature extraction backbone followed by a decoder. Recently, owing to the transformer's capability to capture and model long-term dependencies of semantics, many works adapt ViT as an encoder and design task-specific decoders. However, due to the lack of specific inductive biases inherent in CNNs, transformers struggle to capture repeated detail semantics. Recently, PVT [Wang *et al.*, 2021], Swin [Liu *et al.*, 2021], and Mit [Xie *et al.*, 2021] incorporate more vision-specific inductive biases by merging pyramid structures from CNNs. Especially, Chen et al. [Chen *et al.*, 2022a] develop a CNN adapter on top of the latest ViT models, achieving state-of-the-art results. Differing from these algorithms, this paper considers the application of inductive biases in video shadow detection. In particular, we design a specialized temporal interaction module to enhance the sharing of these spatial priors over time, thereby increasing detection consistency in video scenarios.

2.3 Boundary Attention in Shadow Detection

Inspired by the semantic segmentation work of FCN, per-pixel classification has remained the primary shadow segmentation paradigm in neural network-based video shadow detection [Liu *et al.*, 2023; Chen *et al.*, 2021; Ding *et al.*, 2022; Lu *et al.*, 2022]. These approaches consider the classification of each pixel independently, with an inability to perceive each other's prediction scores. Recently, image-level shadow detection and removal algorithms achieve better results by focusing on shadow structures. MTMT [Chen *et al.*, 2020] utilizes a separate branch to predict binary shadow edge masks, enhancing the network's perception of shadow boundaries. Le et al. [Le and Samaras, 2022] divides the shadow edge region into inner and outer shadow edge areas and supervise the shadow prediction results of these areas separately.

In the field of saliency and semantic segmentation, many works have been proposed to focus on detecting edges and regions simultaneously [Wu *et al.*, 2019; Qin *et al.*, 2019; Ding *et al.*, 2019; Li *et al.*, 2020], in which the image boundaries are represented with hard edges, as the ordinary object often has clear shapes. However, since natural light sources are basically area lights, the boundary brightness of shadows usually has a slow transition. Therefore, representing a soft shadow boundary with a hard edge easily leads to inaccurate shadow localization. We decouple a shadow into the body and distanced boundary, which enables the network to emphasize the transition in brightness and distance changes along shadow boundaries.

3 Method

3.1 Network Overview

As shown in Figure 2, we propose a new **S**tructure-Aware **S**patial-**T**emporal **I**nteraction **Net**work (**SSTINet**) for video shadow detection. SSTINet adopts the siamese structure in the temporal dimension, which takes the t-th and the $(t-1)$-th frames as inputs, and the two branches share the same structure and parameters to keep consistent shadow prediction results in the temporal domain. Each branch consists of two stages: feature extraction and shadow prediction. During the feature extraction stage, we employ a transformer and multi-scale CNN to extract global and local shadow features. The proposed extractor consists of M-stage encoders to output hierarchical shadow features, $M = 4$ in our experiments. For the i-th encoder, SFIM is employed to inject local features into the transformer attention blocks. Subsequently, TFIM is adopted to exchange global-local features from two video frames. Finally, STFRM is proposed to refine and output temporal and global-local features. In the shadow prediction phase, we introduce SASPM to learn the structure information of shadow and bolster the network's perception of local edge semantics.

3.2 Spatial-Temporal Shadow Feature Extraction

It has been proved that convolutional neural networks (CNNs) can enhance the transformer's ability to discern more semantic features [Wang *et al.*, 2022; Fang *et al.*, 2022; Wu *et al.*, 2021]. Inspired by this idea, we propose a novel spatial-temporal feature interaction strategy to synergy and enhance shadow features extracted by transformer and multi-scale CNN, which ensures accurate shadow detection across spatial dimensions by enabling the network to acquire both global and local perspectives.

As illustrated in Figure 2, a multi-layer CNN is utilized to obtain multi-scale shadow features from input image $I \in \mathbb{R}^{B\times C\times H\times W}$. Specifically, the multi-layer CNN employs the Stem structure of ResNet , and it is sequentially stacked with three convolutional blocks with kernel size of 3, stride of 2, and padding of 1 to multi-scale shadow features. Subsequently, three embedding layers consisting of $Conv$ layer with the kernel size of 1 are employed to encode the multi-scale features. Denoting the three output features as f_2, f_3, f_4, we have $f_2 \in \mathbb{R}^{B\times C\times H/8\times W/8}$, $f_3 \in \mathbb{R}^{B\times C\times H/16\times W/16}$, $f_4 \in \mathbb{R}^{B\times C\times H/32\times W/32}$. These features are then reshaped and stacked together to form a local semantic feature set, denoted as $F_l^1 = [f_2, f_3, f_4]$. Subsequently, local feature F_l^1 and global backbone features F_g^1 are fed into the SFIM for feature interaction.

Spatial Shadow Feature Injection

Diverging from the conventional approach of directly stacking two types of features, the proposed SFIM employs an attention mechanism to inject local shadow features F_l into the global features F_g.

As shown in Figure 2.a and 2.b, for the i-th SFIM, the local detail features F_l^i serve as Key and $Value$, while the global features F_g^i act as $Query$. These components are fused through Cross Attention interaction to estimate spatial

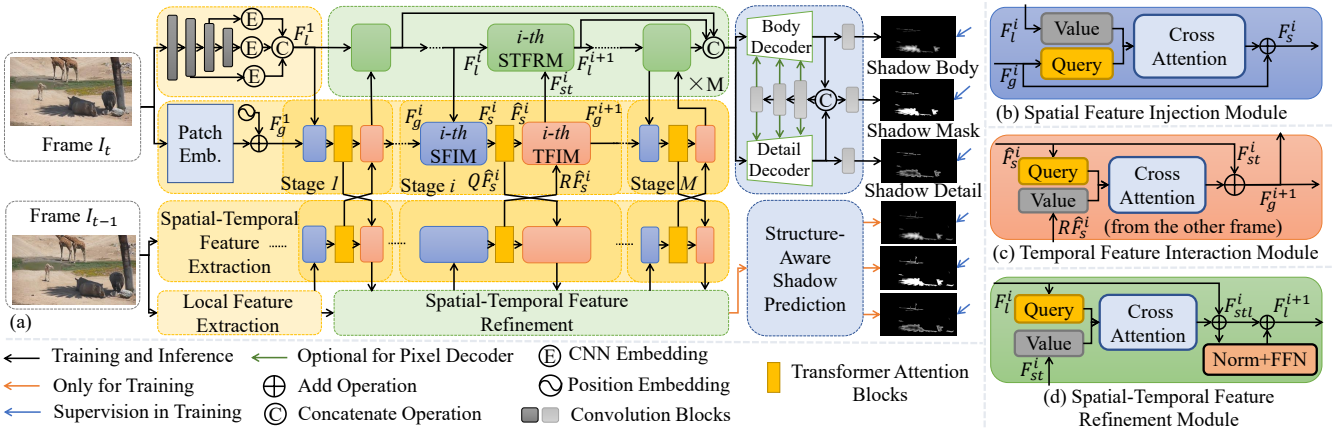

Figure 2: Overview of the proposed **SSTINet** for video shadow detection. SSTINet consists of two parts: video shadow feature extraction and structure-aware shadow prediction. Every stage in feature extraction shares the same structure. (b), (c), and (d) describe three interactive modules in these stages (the colors are consistent with the extraction stages in (a)), and the three lines in the lower left part are used as legends.

shadow features F_s^i.

$$F_s^i = F_g^i + \lambda_1^i CrossAtt(N(F_g^i), N(F_l^i)), \quad (1)$$

where $N(\cdot)$ denotes layer normalization. $CrossAtt$ represents a cross attention layer which employs the multi-scale deformable attention mechanism from Deformable DETR[Zhu *et al.*, 2020]. λ_1^i is a learnable weight that balances F_l^i and F_g^i. The spatial shadow features F_s^i are then fed into the transformer attention blocks to extract further processed features which are denoted as $\hat{F}_s^i$.

Temporal Shadow Feature Interaction

In current video shadow detection methods, simulating the spatial-temporal relationships of shadows in different frames remains a challenging task. We propose an efficient method to propagate shadow temporal information between two consecutive frames. A cross-attention machine is adopted for delivering similar shadow features between adjacent frames, which helps shadow detection results keep consistency.

Different from previous methods that model the relationship of shadows between two frames in the last layer of the backbone network [Chen *et al.*, 2021], the proposed method performs the exchange of temporal information at multiple stages. As shown in Figure 2.c, denoting the shadow features extracted from the t-th frame and the $(t-1)$-th frame as $Q\hat{F}_s$ and $R\hat{F}_s$, bidirectional interaction of $Q\hat{F}_s$ and $R\hat{F}_s$ is performed for feature integration. Similar to the SFIM, we utilize multi-scale deformable attention to facilitate the interaction. Specifically, for feature flow in the i-th TFIM of the t-th frame, $Q\hat{F}_s^i$ serves as $Query$, and $R\hat{F}_s^i$ serves as Key and $Value$. In the feature extraction of the $(t-1)$-th frame, their roles are reversed. The process is described as follows:

$$QF_{st}^i = Q\hat{F}_s^i + \lambda_2^i CrossAtt(N(Q\hat{F}_s^i), N(R\hat{F}_s^i)), \quad (2)$$

where $CrossAtt$ denotes the multi-scale deformable attention, λ_2^i represents a learnable weight that balances query features and reference features. The output features QF_{st}^i and RF_{st}^i are sent to the next stage for feature refinement. Especially, F_{st}^i sent to the $i+1$-th SFIM in t-th or $(t-1)$-th frame is served as global feature:$F_g^{i+1} = F_{st}^i$.

Spatial-Temporal Feature Refinement

We incorporate a Spatial-Temporal Feature Refinement Module (STFRM) to refine and enhance the spatial details of shadow features derived from TFIM. As shown in Figure 2.d, the output of the TFIM F_{st}^i serves as Key and $Value$, and F_l^i acts as $Query$ in multi-scale deformable attention module of the STFRM. The refined shadow features F_{stl}^i are calculated as follow:

$$F_{stl}^i = F_l^i + CrossAtt(N(F_l^i), N(F_{st}^i)), \quad (3)$$

then, F_{stl}^i is processed by a feed-forward network following the formula below:

$$F_l^{i+1} = F_{stl}^i + FFN(N(F_{stl}^i)) \quad (4)$$

where FFN consists of Fully Connected Layer (FC), $Conv$, GELU Activation Layer($GELU$), FC and $GELU$. The output F_l^{i+1} is used as the input of SFIM and STFRM in $(i+1)$-th stage.

3.3 Structure-Aware Shadow Prediction

As shown in Figure 1, even the state-of-the-art video shadow detection algorithms still have false and miss detection when detecting shadows of complex semantics, especially for areas with high-contrast and soft shadow edges. To address this problem, we work separately on the inner areas and edges of shadows to enhance segmentation performance. Different from directly generating shadow edges in the scenes[Chen *et al.*, 2020], we introduce distance transformation[Wei *et al.*, 2020] to decouple ground truth shadow mask M_s to shadow body M_b and shadow detail map M_d, which can be denoted as: $M_b = DistanceTrans(M_s)$, and $M_d = M_s - M_b$. The implementation details of $DistanceTrans$ are described in [Wei *et al.*, 2020]. The values of pixels in M_b and M_d correspond to the distance of the pixels from the shadow edges. A

large pixel value in M_b means that the pixel is farther away from the edge, while it means closer to the edge in M_d.

We propose a structure-aware shadow detection module to perform shadow mask prediction. As shown in Figure 2.a, the proposed shadow detection head comprises three primary components: a shadow body decoder, a shadow detail decoder, and a fusion module. In the first stage, the multi-scale features are sent to the body decoder and the detail decoder, each of them consisting of four $Conv + BN + Relu$ blocks. Then, the output features are sent to the output block to get coarse results P_b, P_d, and P_s. The output block consists of $Conv + BN + Relu + Conv$ layers. In the second stage, the body feature F_b and detail feature F_d are concat and fused by f_{fusion}, $\hat{F}_{fb}, \hat{F}_{fd} = f_{fusion}(cat(F_b, F_d))$, where f_{fusion} consists of four $ConvBlocks$ and three $MaxPooling$ layers to get features in four different scales, and two groups of $ConvBlocks$ are employed to generate body and detail features. The fused features are sent back to two decoders to get fine results.

3.4 Loss Function

The loss function in this paper is primarily composed of three parts: constraints on the shadow map, constraints on the shadow body, and constraints on the shadow details. The shadow map loss follows the setting in [Liu *et al.*, 2023], and Binary Cross-Entropy with Logits (BCEL) and IoU loss are adopted to constrain the prediction of shadow map. For the supervision of shadow body and details, we employ the Mean Squared Error (MSE) loss to get fine soft results.

$$L_s = L_{bcel}(P_s, M_s) + L_{iou}(P_s, M_s), \quad (5)$$

$$L_{sd} = MSE(P_b, M_b), L_{sb} = MSE(P_d, M_d), \quad (6)$$

These three losses are then combined to serve as the union constraints in the training of SSTINet:

$$L = L_s + \lambda_3 L_{sb} + \lambda_4 L_{sd}, \quad (7)$$

where λ_3 and λ_4 are hyper-parameters that trade off the losses of shadow mask, shadow body, and detail mask. We empirically set λ_3, λ_4 as 5 and 5, respectively.

4 Experiments

4.1 Dataset and Evaluation Metrics

Data Description. We utilize the currently popular **Vi**deo **Sha**dow Detection Dataset (**Visha**) [Chen *et al.*, 2021] to demonstrate the effectiveness of our study. This dataset comprises 120 scenes, with 50 used for training and 70 for testing. Each scene contains approximately 100 images, although some scenes have fewer than 30 images. In total, there are 4,838 images used for training and 6,967 images for testing.

Data Pre-processing. We conduct training and testing under the dataset setting in Visha [Chen *et al.*, 2021]. During the training phase, we employ random cropping and flipping to augment the dataset. In the testing phase, a multi-scale flipping strategy is used to enhance the robustness of the algorithm. After the augmentation in both the training and testing phases, the final input size of the SSTINet is 512×512.

Evaluation Metrics. To better evaluate the effectiveness of SSTINet, we adopt the popular accuracy metrics applied by the current methods [Chen *et al.*, 2021; Liu *et al.*, 2023; Chen *et al.*, 2022b]. Besides, we also adopt the consistency assessment [Ding *et al.*, 2022], which has been overlooked in recent works. The accuracy metrics include Mean Absolute Error (MAE), Intersection over Union (IoU), F-measures, and Balanced Error Rate (BER). Lower MAE and BER, along with higher IoU and F-measures, indicate more accurate shadow detection results. Consistency assessment primarily measures the temporal stability of prediction results. Specifically, optical flow is obtained based on the ground truth between adjacent frames. Then, the detection results from the previous frame are mapped to the next frame using this optical flow. Finally, IoU is used to measure the degree of consistency between them. For a detailed calculation method, please refer to [Ding *et al.*, 2022].

Implementation Details. The experiments in this paper are conducted using the MMSegmentation [Contributors, 2020] segmentation framework and the Pytorch framework. For parameter optimization, the backbone network of the proposed method is initialized using the pre-trained parameters of Beitv2 [Peng *et al.*, 2022] on the COCO-Stuff segmentation dataset [Caesar *et al.*, 2018]. The rest of the parameters are randomly initialized using the Xavier method [Glorot and Bengio, 2010] and optimized during training. We use the AdamW optimizer [Loshchilov and Hutter, 2017], with an initial learning rate of 2e-5, a weight decay of 0.05, and employ a poly learning rate decay. The experiments are conducted with a batch size of 2, a total of 20,000 iterations, and are trained on a single NVIDIA RTX 3090ti with 24GB of VRAM, taking approximately 10 hours and 30 minutes.

4.2 Comparison with SOTA Techniques

Compared Methods. We employ the latest image-based and video-based shadow detection methodologies, along with pixel-level detection algorithms, to demonstrate the effectiveness of SSTINet. This includes three types of methods: two image object segmentation methods, FPN [Lin *et al.*, 2017], DSS [Hou *et al.*, 2017]; three image shadow detection methods, DSD [Zheng *et al.*, 2019], MTMT [Chen *et al.*, 2020], FSDNet [Hu *et al.*, 2021b]; five video object/instance segmentation methods, PDBM [Song *et al.*, 2018], COSNet [Lu *et al.*, 2019], FeelVOS [Voigtlaender *et al.*, 2019], MinVIS [Huang *et al.*, 2022], Tube-L [Li *et al.*, 2023]; and four video shadow detection techniques, TVSD [Chen *et al.*, 2021], STICT [Lu *et al.*, 2022], SC-Cor [Ding *et al.*, 2022], Scotch [Liu *et al.*, 2023].

Quantitative Comparisons. In existing algorithms, image-based algorithms such as IOS and ISD only consider the features of individual frames and lack the concern of temporal consistency. Thus, as shown in Table 1, although these algorithms exhibit favorable metrics in Frame-Level assessments, they fall short in terms of temporal consistency(TS) compared to video-based algorithms. In video-based shadow detection algorithms, VSD, due to its consideration of shadow inherent characteristics, outperforms methods based on VOS and VIS in overall metrics, including accuracy (MAE), IOU, and BER. Especially, the algorithm proposed in this paper surpasses existing algorithms in both accuracy and consistency metrics. Specifically, our algorithm

METHODS		Frame Level						Temp. Level	
Tasks	Techniques	MAE ↓	F_β ↑	IoU ↑	BER ↓	S ↓	N ↓	TS ↑	Avg ↑
IOS	⋆ FPN (2017)	0.044	0.707	0.512	19.49	36.59	2.40	0.743	0.628
	DSS (2017)	0.045	0.696	0.502	19.77	36.96	2.59	0.750	0.626
ISD	⋆ DSD (2019)	0.043	0.702	0.518	19.88	37.89	1.88	0.747	0.633
	MTMT (2020)	0.043	0.729	0.517	20.28	38.71	1.86	0.744	0.631
	FSDNet (2021)	0.057	0.671	0.486	20.57	38.06	3.06	0.748	0.617
VS	⋆ PDBM (2018)	0.066	0.623	0.466	19.73	34.32	5.16	0.800	0.633
	COSNet (2019)	0.040	0.705	0.514	20.50	39.22	1.79	0.740	0.627
	FeelVOS (2019)	0.043	0.710	0.512	19.76	37.27	2.26	0.749	0.631
	MinVIS (2022)	0.072	0.565	0.438	12.09	18.43	5.75	0.716	0.577
	Tube-L (2023)	0.035	0.801	0.576	13.79	26.24	1.34	0.682	0.629
VSD	TVSD (2020)	0.033	0.757	0.567	17.70	33.97	1.45	0.783	0.674
	STICT (2022)	0.046	0.702	0.545	16.60	29.58	3.59	0.770	0.657
	SC-Cor (2022)	0.042	0.762	0.615	13.61	24.31	2.91	0.814	0.715
	Scotch (2023)	0.029	0.793	0.640	9.066	16.26	1.44	0.640	0.721
	SSTINet	**0.017**	**0.866**	**0.746**	**6.484**	**12.32**	**0.65**	**0.853**	**0.793**

Table 1: Comparisons between our SSTINet and SOTA techniques on the ViSha dataset. "Temp." denotes temporal, "IOS" denotes image object segmentation, "ISD" denotes image shadow detection, "VS" denotes video object/instance segmentation, "VSD" denotes video shadow detection, "MAE" denotes mean absolute error, "F_β" denotes F-measure score, "IoU" denotes intersection over union, "BER" denotes balance error rate, and "S" means shadow error rate, "N" means non-shadow error rate. 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.

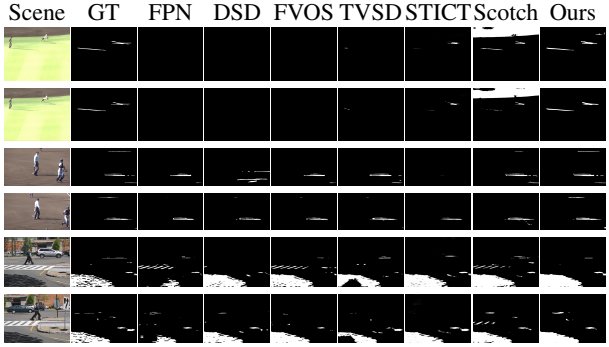

Figure 3: Visual comparison of video shadow detection results by our method (SSTINet) and the SOTA approaches. For each video shadow scene, we show the results of two adjacent frames. From left to right, the shadow scene image, the shadow ground truth, the detection results of FPN [Lin *et al.*, 2017], DSD [Zheng *et al.*, 2019], FeelVOS [Voigtlaender *et al.*, 2019], TVSD [Chen *et al.*, 2021], STICT [Lu *et al.*, 2022], Scotch [Liu *et al.*, 2023] and Ours.

exceeds the current best algorithm[Liu *et al.*, 2023] by 41.3% in the MAE metric, 9.2% in F_β, by 16.5% in IOU, and 28.48% in BER, 33.2% in TS. On the AVG metric, which considers both TS and IOU, the method proposed in this paper shows a 9.98% improvement compared to existing methods. This demonstrates that our SSTINet can maintain accurate shadow detection results while also ensuring stable and consistent detection.

Visual Comparison. Figure 3 illustrates the visual performance of current popular shadow detection and semantic segmentation algorithms on scenes from the Visha dataset [Chen *et al.*, 2021]. Compared methods include the image segmentation algorithm FPN [Lin *et al.*, 2017], image shadow detection algorithm DSD [Zheng *et al.*, 2019], video semantic segmentation algorithm FeelVOS [Voigtlaender *et al.*, 2019], and video shadow detection algorithms TVSD [Chen *et al.*, 2021], STICT [Lu *et al.*, 2022], and Soda [Liu *et al.*, 2023].

In the presented scenarios, common errors in VSD can be observed, primarily including missed detection of fragmented small shadows, such as the shadows of individuals in Scene 1; missed detection in areas with relatively weak intensity contrast, such as the shadow cast by individuals on the ground in Scene 1 and the task's shadow on the ground in Scene 2; false detection of non-shadow areas with significant intensity contrast, for instance, the gray ground in Scene 1, the sidewalk in Scene 3. The random combinations of these scenarios further increase the detection difficulty for existing algorithms. From the comparison of existing scenarios, we observe that conventional image and video segmentation algorithms exhibit less effective semantic capture of shadows compared to algorithms specifically designed for shadow detection. For instance, in Scene 3, FeelVOS and FPN demonstrate lower accuracy compared to DSD, the latter demonstrating a relatively accurate detection of the shadow cast by the tree. In the realm of VSD algorithms, both accuracy and consistency in detection have been notable improved. However, TVSD and STICT may fail to exhibit complete detection when confronted with scenarios involving minute shadows (Scenes 1 and 2), and Scotch may fail when faced with scenes featuring strong intensity contrast (Scenes 1 and 3). Compared with these methods, SSTINet captures consistent shadows with greater accuracy and delineates shadow edges with enhanced precision.

Components					Evaluation Metrics				
ind.	SFIM	TFIM	STFRM	head	MAE ↓	F_β ↑	IoU↑	BER ↓	TS↑
1				PP	0.0244	0.795	0.630	9.25	0.803
2	✓			PP	0.0217	0.822	0.673	7.881	0.827
3	✓		✓	PP	0.0226	0.831	0.695	7.256	0.828
4		✓		PP	0.0199	0.847	0.703	8.254	0.828
5	✓	✓	✓	PP	0.0174	0.856	0.739	6.227	0.841
6	✓	✓	✓	SA	0.0171	0.866	0.746	6.484	0.853

Table 2: Ablation study on different components of our SSTINet on the ViSha dataset. All architectures utilize the BEiT [Peng *et al.*, 2022] backbone, which is not depicted here due to table size limitations. "ind." denotes the index of ablation structures, and "head" signifies the shadow detection prediction head. "PP" indicates the plain per-pixel shadow mask prediction module, while "SA" denotes the use of the SASPM presented in Sec.3.3.

4.3 Ablation Study

We conduct ablation studies to validate the effectiveness of the spatial-temporal interaction strategy and the structure-aware segmentation head proposed in this paper. Specifically, we establish five ablation structures. For these structures, we utilize the pre-trained transformer BEiT-V2 [Peng *et al.*, 2022] as the feature extractor, and follow the training and inference strategies described in Sec.4.1.

Baseline. The baseline network consists of a feature extraction backbone and a per-pixel shadow mask prediction module(PP module). The PP module consists of a

Networks	Params↓	GFLOPs↓	Speed↑	Avg ↑
Sc-Cor	232.63	436.80	5.27	0.715
Scotch	<u>211.79</u>	**244.92**	**12.56**	0.721
SSTINet(Ours.)	338.28	658.33	6.06	**0.793**
SSTINet-L(Ours.)	**106.32**	<u>250.91</u>	<u>12.53</u>	<u>0.757</u>

Table 3: Analysis of model complexity and performance. Specifically, the unit of Params and Speed are MB and item/s, respectively, Avg is the average of temporal consistency(TS) and IOU, **Bold** and <u>Under</u> represent the best and second best respectively.

$ConvBlock$, which receives concatenated multi-stage features from the backbone and outputs a shadow mask. As presented in the 1st row of Table 2, "Baseline" achieves favorable outcomes when compared to current VSD algorithms. This can be attributed to the shadow semantics with multiple receptive fields from the pre-trained BEiT backbone.

Ablation on Spatial-Temporal Interaction. As shown in Table 2, to evaluate the effectiveness of the proposed SFIM and STFRM, we add SFIM to the "Baseline" as the proposed method. As the result presented in the 1st and 2nd rows, the SFIM reduces MAE and BER while improving IOU and F_{beta}. Additionally, the increasing TS score also shows the effectiveness of SFIM. To evaluate the effectiveness of the proposed TFIM, we add TFIM to the "Baseline". As the results presented in the 1st and 4-th rows, TFIM also reduces MAE and BER, with a more pronounced decrease in MAE and greater improvements in F-beta and IOU compared to SFIM. This indicates that inter-frame interaction is more critical than single-frame semantic attention in VSD. Our structure effectively augment the flow of information between frames. As the results presented in the 3rd, 4-th, and 5-th rows, "Baseline+SFIM+TFIM+STFRM" further improves accuracy and consistency metrics on top of the individual additions of temporal or spatial elements, demonstrating that our combined spatial-temporal strategy allows shadow feature spatial details to interact over time, thereby mutually enhancing each other.

Ablation on Structure-Aware Shadow Prediction. To evaluate the effectiveness of the proposed shadow prediction head, we replace the PPM with SASPM, and the last ablation structure is the SSTINet. As the result presented in the 5-th and 6-th rows of Table 2, the SASPM further reduces error rates and improves inter-frame accuracy, which also achieves optimal performance.

4.4 Complexity Analysis

The proposed SSTINet consists of three components: the transformer backbone, the spatial-temporal attention modules, and the shadow prediction head, which account for 89.5%, 10%, and 0.5% of the total computational load, respectively. While the backbone accounts for the largest share, the attention modules and prediction head proposed in the paper constitute only a small fraction of the overall network. To further optimize the potential performance, we replace the backbone in SSTINet with a lighter pre-trained BEiT-Base model and obtain a light network called SSTINet-L. As shown in Table 3, SSTINet-L reduces 61.8% of GFLOPs and maintains similar inference accuracy with SSTINet. Compared with the SOTA algorithm $Scotch$, our proposed SSTINet achieves higher accuracy and SSTINet-L achieves superior detection performance with the same computational costs.

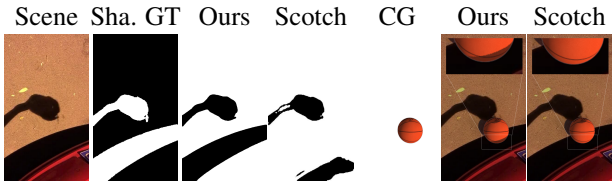

Figure 4: The application of VSD in AR shadow interaction[Liu *et al.*, 2022] and a visual comparison of synthetic results. From left to right: scene image, shadow detection ground truth, our shadow detection result(with shadow label inversion), Scotch's[Liu *et al.*, 2023] shadow detection result(with shadow label inversion), CG object inserted in the scene, synthetic result using our shadow detection mask, synthetic result using Scotch's shadow mask.

4.5 Application in AR

To validate the efficacy of our algorithm in downstream tasks, we investigate the application of VSD in AR. To achieve realistic AR interaction scenes, techniques based on shadow texture mapping [Xing *et al.*, 2013; Liu *et al.*, 2022] and shadow volume rendering are employed for shadow interaction within AR scenarios. Following the workflow of AR algorithms based on texture mapping [Liu *et al.*, 2022], we initiate the process by conducting geometric and illumination reconstruction of the scene; We then place a virtual basketball into the reconstructed 3D scene; and the shadow masks are projected onto the basketball using texture mapping technology; Finally, the 3D scenes are rendered and synthesized. As depicted in Figure 4, comparing our shadow mask with Scotch, the SOTA VSD algorithm, our method generate more accurate and complete shadows. Consequently, Scotch's fusion result on the right lacks sufficient shadow interaction, giving the impression that the basketball is suspended in mid-air. In contrast, our method provides a shadow mask that seamlessly integrates with the lower part of the basketball, effectively blending it into the scene.

5 Conclusion

This paper introduces SSTINet, an innovative video shadow detection network designed to seamlessly integrate local shadow details with global shadow semantics across video frames. We propose a spatial-temporal interaction strategy that effectively captures consistent and comprehensive shadow features through spatial feature injection, temporal feature interaction, and spatial-temporal feature refinement. Additionally, we introduce a structure-aware shadow detection module that enhances understanding of the distance relationship between shadow edges and the shadow body. This significantly improves the accuracy in determining uncertain shadow boundaries. The effectiveness of our proposed modules is rigorously demonstrated, establishing that our approach outperforms current video shadow detection algorithms across various metrics.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Project(62172290, 61972271), Sichuan Science and Technology Program under Project (2023YFS0454). We would like to thank the anonymous referees for their helpful comments. Housheng Wei and Guanyu Xing contributed equally to this work and should be considered co-first authors. Yanli Liu is the corresponding author.

References

[Adams *et al.*, 2022] Haley Adams, Jeanine Stefanucci, Sarah Creem-Regehr, Grant Pointon, William Thompson, and Bobby Bodenheimer. Shedding light on cast shadows: An investigation of perceived ground contact in ar and vr. *IEEE Transactions on Visualization and Computer Graphics*, page 4624–4639, Dec 2022.

[Caesar *et al.*, 2018] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context, 2018.

[Chen *et al.*, 2020] 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.

[Chen *et al.*, 2021] 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.

[Chen *et al.*, 2022a] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[Chen *et al.*, 2022b] Zipei Chen, Xiao Lu, Ling Zhang, and Chunxia Xiao. Semi-supervised video shadow detection via image-assisted pseudo-label generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2700–2708, 2022.

[Contributors, 2020] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[Ding *et al.*, 2019] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6819–6829, 2019.

[Ding *et al.*, 2022] Xinpeng Ding, Jingwen Yang, Xiaowei Hu, and Xiaomeng Li. Learning shadow correspondence for video shadow detection. In *European Conference on Computer Vision*, pages 705–722. Springer, 2022.

[Fang *et al.*, 2022] Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, and Xinggang Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. Apr 2022.

[Glorot and Bengio, 2010] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, Jan 2010.

[Gomes *et al.*, 2017] Vitor Gomes, Pablo Barcellos, and Jacob Scharcanski. Stochastic shadow detection using a hypergraph partitioning approach. *Pattern Recognition*, 63:30–44, 2017.

[Hao *et al.*, 2021] Hanxiang Hao, Sriram Baireddy, Emily Bartusiak, Mridul Gupta, Kevin LaTourette, Latisha Konz, Moses Chan, Mary L. Comer, and Edward J. Delp. Building height estimation via satellite metadata and shadow instance detection. In *Automatic Target Recognition XXXI*, Apr 2021.

[Hou *et al.*, 2017] 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.

[Hu *et al.*, 2021a] Shilin Hu, Hieu Le, and Dimitris Samaras. Temporal feature warping for video shadow detection. *arXiv preprint arXiv:2107.14287*, 2021.

[Hu *et al.*, 2021b] 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.

[Huang *et al.*, 2022] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems*, 35:31265–31277, 2022.

[Karsch *et al.*, 2011] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, Dec 2011.

[Le and Samaras, 2022] Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 9088–9101, Dec 2022.

[Li *et al.*, 2020] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 435–452. Springer, 2020.

[Li *et al.*, 2023] Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, and Chen Change Loy. Tube-link: A flexible cross tube baseline for universal video segmentation. *arXiv preprint arXiv:2303.12782*, 2023.

[Lin and Wang, 2022] Junhao Lin and Liansheng Wang. Spatial-temporal fusion network for fast video shadow detection. In *Proceedings of the 18th ACM SIGGRAPH In-*

ternational Conference on Virtual-Reality Continuum and its Applications in Industry, pages 1–5, 2022.

[Lin *et al.*, 2017] 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.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2022] Yanli Liu, Xingming Zou, Songhua Xu, Guanyu Xing, Housheng Wei, and Yanci Zhang. Real-time shadow detection from live outdoor videos for augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(7):2748–2763, 2022.

[Liu *et al.*, 2023] Lihao Liu, Jean Prost, Lei Zhu, Nicolas Papadakis, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Scotch and soda: A transformer video shadow detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10449–10458, 2023.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Learning,Learning*, Nov 2017.

[Lu *et al.*, 2019] 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.

[Lu *et al.*, 2022] 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.

[Peng *et al.*, 2022] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. 2022.

[Qin *et al.*, 2019] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[Shi and Liu, 2020] Hang Shi and Chengjun Liu. A new cast shadow detection method for traffic surveillance video analysis using color and statistical modeling. *Image and Vision Computing*, 94:103863, 2020.

[Song *et al.*, 2018] 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.

[Voigtlaender *et al.*, 2019] 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.

[Wang *et al.*, 2021] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

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

[Wei *et al.*, 2020] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020.

[Wu *et al.*, 2019] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019.

[Wu *et al.*, 2021] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021.

[Xie *et al.*, 2021] 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.

[Xing *et al.*, 2013] Guanyu Xing, Xuehong Zhou, Qunsheng Peng, Yanli Liu, and Xueying Qin. Lighting simulation of augmented outdoor scene based on a legacy photograph. In *Computer Graphics Forum*, volume 32, pages 101–110. Wiley Online Library, 2013.

[Zheng *et al.*, 2019] 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.

[Zhou *et al.*, 2022] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: end-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Oct 2020.