Language-Driven Interactive Shadow Detection

Anonymous Authors

ABSTRACT

Traditional shadow detectors often identify all shadow regions of static images or video sequences. This work presents the Referring Video Shadow Detection (RVSD), which is an innovative task that rejuvenates the classic paradigm by facilitating the segmentation of particular shadows in videos based on descriptive natural language prompts. This novel RVSD not only achieves segmentation of arbitrary shadow areas of interest based on descriptions (***flexibility***) but also allows users to interact with visual content more directly and naturally by using natural language prompts (***interactivity***), paving the way for abundant applications ranging from advanced video editing to virtual reality experiences. To pioneer the RVSD research, we curated a well-annotated RVSD dataset, which encompasses 86 videos and a rich set of 15,011 paired textual descriptions with corresponding shadows. To the best of our knowledge, this dataset is the first one for addressing RVSD. Based on this dataset, we propose a Referring Shadow-Track Memory Network (RSM-Net) for addressing the RVSD task. In our RSM-Net, we devise a Twin-Track Synergistic Memory (TSM) to store intra-clip memory features and hierarchical inter-clip memory features, and then pass these memory features into a memory read module to refine features of the current video frame for referring shadow detection. We also develop a Mixed-Prior Shadow Attention (MSA) to utilize physical priors to obtain a coarse shadow map for learning more visual features by weighting it with the input video frame. Experimental results show that our RSM-Net achieves state-of-the-art performance for RVSD with a notable Overall IOU increase of 4.4%. We shall release our code and dataset for future research.

CCS CONCEPTS

• **Computing methodologies** → **Video segmentation**.

KEYWORDS

Video shadow detection, referring segmentation, dataset

1 INTRODUCTION

As a cornerstone in computer vision, shadow detection has seen decades of research, and plays a pivotal role in interpreting the geometry and depth of a scene. Accurate shadow detection can significantly enhance downstream computer vision tasks, including object detection, tracking, and scene reconstruction [14, 33, 47]. Previous shadow detection methods mainly focus on discerning all shadows in static images or video sequences and have achieved promising results [6, 10, 15, 17, 24, 53]. However, these methods do not allow for the segmentation of specific shadows described by users, due to their lack of interactivity and flexibility which is essential in the era of advanced multimedia interaction.

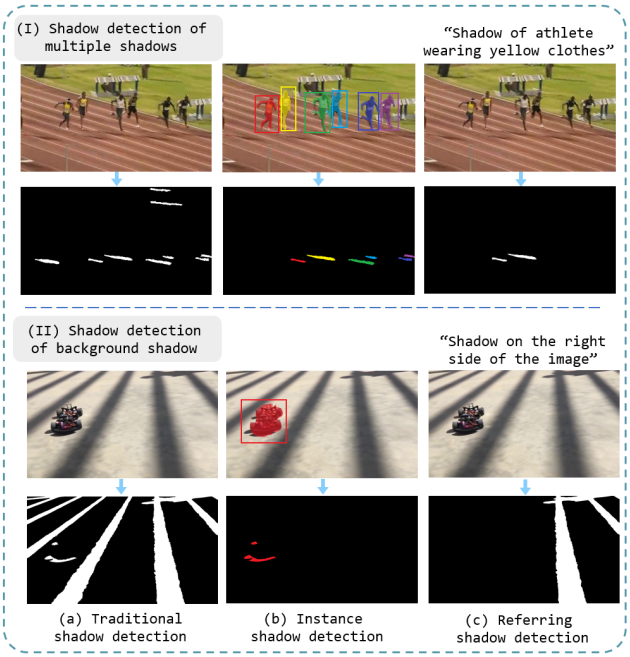

Figure 1: Comparisons of task settings for video shadow detection, instance shadow detection, and our RVSD. Traditional shadow detection (a) segments all shadows, current instance shadow detection (b) detects the foreground objects and segments the associated shadows, while our RVSD (c) can flexibly segment any shadow of interest referred by the text description, including I. those of multiple objects and II. background shadows (objects are invisible in the figure).

Therefore, we propose Referring Video Shadow Detection (RVSD) as a novel task dedicated to meeting this demand by interactively segmenting the specific shadows in videos based on descriptive natural language prompts, as illustrated in Fig. 1. The significance of RVSD can be envisaged in numerous applications. Advanced video editing tools can benefit from this RVSD task by allowing editors to precisely manipulate shadows based on verbal instructions [5, 22]. In augmented and virtual reality environments, understanding and reacting to user-specified shadows can enhance immersion and realism [1, 31]. Furthermore, remote sensing may also benefit from this RVSD task to achieve more precise results [29].

In this work, we introduce the first dataset dedicated to the RVSD task. This dataset collected 86 videos with 15,011 paired rich textual descriptions and corresponding shadow mask annotations, with some examples showcased in Fig. 2. It has been meticulously curated to cover a wide array of scenarios, and shadow dynamics, which can provide a comprehensive benchmark for evaluating future RVSD

Unpublished working draft. Not for distribution.

ACM MM, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

methods. To the best of our knowledge, our dataset is the first one for RVSD with high-quality annotation.

Compared to classical referring video object segmentation [2, 39], achieving an accurate RVSD is more challenging. Firstly, shadows lack rich appearance features and are easily confused with the dark regions of the input video [3]. Moreover, shadows are naturally influenced by ever-changing environmental factors, demonstrating severe temporal shape transformations. Hence, RVSD requires richer temporal information and physical prior information to accurately identify the specific shadow referred by users' descriptions.

Based on our annotated RVSD dataset, we develop a novel Referring Shadow-Track Memory Network (RSM-Net) for addressing the RVSD task. In our RSM-Net, we devise a Twin-Track Synergistic Memory (TSM) to store hierarchical inter-clip features and intra-clip features, and then propagate these two memory features via a memory read module to refine the current video frame for video shadow detection. Moreover, we propose a Mixed-Prior Shadow Attention (MSA) module that leverages physical knowledge to generate a preliminary shadow map, guiding the network to focus on potential shadow regions attentively.

Our main contributions are summarized as follows:

- This work is the first one to explore the RVSD task, which presents ***a fresh paradigm for language-driven interactive and flexible shadow detection*** with potential benefits for numerous downstream tasks.
- To address the RVSD task, we collect and annotate the first dataset for RVSD consisting of 86 videos with 15,011 paired video frames and the corresponding text descriptions. This is the first dataset dedicated for the RVSD task.
- We propose an RSM-Net for RVSD. Here, we devise a TSM module to learn intra-clip and inter-clip features and store both of them in a memory to refine the current video frame for video shadow detection. Moreover, an MSA module is developed to generate the coarse shadow map for focusing on potential shadow areas for RVSD.
- Extensive experimental results show that our RSM-Net clearly outperforms state-of-the-art methods for the RVSD task.

2 RELATED WORK

2.1 Shadow detection.

Shadow detection is crucial in computer vision, aiming to generate binary masks for all shadows [6, 10, 15, 17, 24, 45, 53]. Early techniques employed physical illumination and color models to analyze spectral and geometrical shadow properties [36, 38, 41]. As the machine learning development, subsequent approaches built models using handcrafted attributes like texture [42, 51], color [12, 21], and edges [18, 21]. These models were then combined with classifiers like decision trees [21, 51] and support vector machines [12, 18, 42] to differentiate shadow. However, the limited representational ability often restricted their effectiveness in various scenarios.

Recent deep-learning methods have made significant strides in shadow detection by effectively learning from shadow images. Khan *et al.* [20] first propose a framework where relevant features are automatically learned through multiple convolutional deep neural networks. Shen *et al.* [40] employ structured CNNs to analyze shadow edges' local features. Vicente *et al.* [43] recommend utilizing quickly obtained, partially accurate image labels, refining them automatically for enhanced performance. Hu *et al.* [17] develop a network using spatial RNNs for direction-aware context analysis in shadow detection. Chen *et al.* [4] propose a semi-supervised, multi-task model combining consistency loss from unlabeled data with supervised loss from labeled data.

All the above investigations primarily focus on shadow detection for images, while there have been recent endeavors in video-based shadow detection as well. Chen *et al.* [3] curate a new video shadow detection dataset and develop a triple-cooperative network for enhanced accuracy. Liu *et al.* [27] introduce shadow deformation attention trajectory, a new video self-attention module meticulously crafted to tackle substantial shadow deformations within videos. Differing from conventional video shadow detection that generates a universal binary shadow mask, we delve into a pioneering realm of referring video shadow detection. This approach facilitates the segmentation of each specific shadow through associated language expressions, enhancing user-friendly and personalized applications.

2.2 Referring segmentation.

The aim of referring segmentation is to precisely outline the particular object referred to in a natural language expression within an image. This task combines computer vision and natural language processing, similar to our RVSD approach. Hu *et al.* [16] first tackle the novel challenge of image segmentation guided by natural language expressions. They utilize a hybrid recurrent-convolutional network, encoding expressions with recurrent neural networks and generating response maps using fully convolutional networks from images. Liu *et al.* [26] present a convolutional multimodal LSTM to encode interactions among words, visual, and spatial information components, enforcing a more effective word-to-image interaction.

2.3 Instance shadow detection.

Instance shadow detection is a subtask of shadow detection, and the goal of the current instance shadow detection is to detect the object and segment the corresponding shadow region. Wang *et al.* [46] pioneer instance shadow detection and built an image-based dataset for finer segmentation. Xing *et al.* [50] expand the scope of instance shadow detection from static images to dynamic videos, introducing a new framework to extract shadow-object associations in videos with paired tracking.

However, our RVSD stands out from previous approaches in two distinctions. Firstly, RVSD is a novel ***interactive*** technique for shadow segmentation, which is more user-friendly. RVSD receives the user's linguistic description and the corresponding video frames as input to segment the relevant shadows, whereas instance shadow detection only inputs image data. Secondly, in contrast with previous approaches that focus on detecting objects in the foreground and segmenting their associated shadows separately, RVSD offers greater ***flexibility*** by enabling the segmentation of any shadow of interest, including those of multiple objects or background shadows (objects are invisible), as illustrated in Fig. 1.

3 RVSD DATASET

In this section, we introduce the newly established RVSD dataset. We begin by detailing the video collection, annotation and validation process in Section 3.1, followed by an examination of dataset statistics and analysis in Section 3.2.

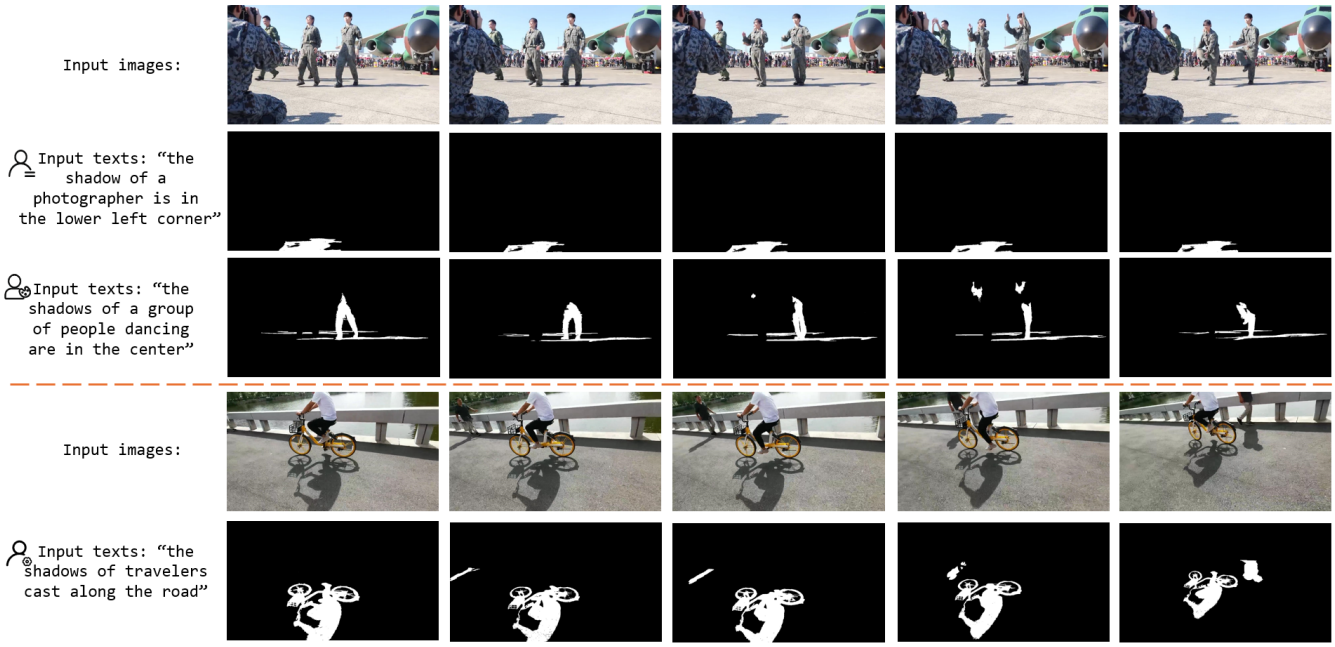

Figure 2: Sample frames from the RVSD dataset showcase pixel-level shadow annotations paired with textual descriptions that guide the corresponding shadow segmentation. These examples demonstrate that our RVSD not only facilitates the flexible segmentation of a specific shadow but also effectively segments shadows cast by groups of objects.

3.1 Constructing the RVSD dataset

3.1.1 Video Collection. We perform data collection and re-annotation using the most extensive publicly available video shadow dataset to date, which comprises 120 videos, known as ViSha [3]. After reviewing 120 potential candidates, we carefully exclude shadows that are blended together (hard to differentiate them through textual descriptions). We also remove highly fragmented shadows. Meanwhile, we re-mask the original binary shadow mask to label the shadows of different instances separately. Eventually, with an emphasis on quality over quantity, we meticulously select 86 videos to construct a benchmark dataset that encompasses diversity and represents a broad spectrum of real-world video scenarios. We also conduct an analysis of various scenarios where shadows are present in the video. Further details can be found in Table 1. According to the tag combinations for different shadow scenarios, we give reference suggestions for the corresponding language descriptions. The following section will provide a more detailed explanation of the language expression annotation and validation process.

3.1.2 Language Description Annotation. The fundamental methodology and procedure for language annotation in RVSD are following previous works [8, 19, 39]. It employs an interactive approach involving multiple annotators taking turns to annotate and validate. We invited multiple people with no computer vision-related background to participate in labeling with the guidance in our Table 1. The annotator is required to select one or multiple shadows from the video and generate corresponding referring descriptions according to the guidance for annotating language expressions. It is worth noting that our guidance in Table 1 contains three elements, the recommended ✎, optional ○, and not required – ones. In this setup, there is a distinct advantage. While ensuring the accurate depiction of various shadows, annotators have the freedom to select the description's content, thus maximizing the richness and flexibility of the sentences.

3.1.3 Language Description Validation. After the initial annotation, we conduct validation tasks for all annotations. The validation process commences by presenting the video along with the corresponding expression, prompting the validator to identify the shadow referred to in the expression. The validator is required to independently find the target shadow and record it. The shadow chosen by the validator is subsequently cross-referenced with the annotations provided by the annotator. If they align, the annotation passes the validation; otherwise, it undergoes re-labeling and subsequent validation. By implementing the validation procedure, we strive to ensure the accurate language representation of shadows within our dataset, while maintaining sentence diversity.

3.2 Analysis and Statistics of the RVSD dataset

3.2.1 Video and Shadow. The selection of video and shadow is based on diversity, which makes the dataset both comprehensive and challenging. Firstly, the number of shadows is a factor to consider, since RVSD in a scene with multiple shadows generally is more challenging than one with a single shadow. Multiple shadows, often close in location and similar in shape, can lead to confusion during model interpretation, thereby making precise segmentation of a specific shadow challenging. In the RVSD dataset, there are 53 videos containing multiple shadows, and 33 videos containing only one shadow. Next, we consider the type of shadow, namely hard and soft shadow. A hard shadow has a distinct boundary, whereas a

Table 1: Scene analysis of video shadows and corresponding linguistic description approaches. ✎ represents what we recommend a description, ○ represents the optional description, and – represents no description.

Tag combination				shadow	object	motion	shape	position	shadow type
Only shadow	One shadow	Stable	soft or hard	✎	–	–	✎	○	○
		Moving	soft or hard	✎	–	✎	✎	○	○
	Multiple shadows	Stable	soft or hard	✎	–	–	○	✎	○
			soft and hard	✎	–	–	○	✎	✎
		Moving	soft or hard	✎	–	✎	○	✎	○
			soft and hard	✎	–	✎	○	✎	✎
		Stable and Moving	soft or hard	✎	–	✎	○	✎	○
			soft and hard	✎	–	✎	○	✎	✎
Shadow and object	One shadow	Stable	soft or hard	✎	✎	–	✎	○	○
		Moving	soft or hard	✎	✎	✎	✎	○	○
	Multiple shadows	Stable	soft or hard	✎	✎	–	○	✎	○
			soft and hard	✎	✎	–	○	✎	✎
		Moving	soft or hard	✎	✎	✎	○	✎	○
			soft and hard	✎	✎	✎	○	✎	✎
		Stable and Moving	soft or hard	✎	✎	✎	○	✎	○
			soft and hard	✎	✎	✎	○	✎	✎

Table 2: Detailed information of the proposed RVSD datasets.

Dataset	Paired prompt textual description		
	Shadow-text pairs	Min word count	Max word count
RVSD	15,011 pairs	6 words	27 words

Example 1: "the soft shadow is located below"
Example 2: "the hard shadow of a person who is holding an umbrella and walking is in the upper left corner"

soft shadow's boundary might be fuzzy or indistinct. Soft shadow is more challenging to segment than hard shadow. The RVSD dataset includes 156 hard shadows and 42 soft shadows. Finally, the diversity of scenes is also an important factor, diverse scenes are closer to the real world, making the dataset more comprehensive. The RVSD dataset includes 74 daytime videos and 12 nighttime videos, as well as 24 indoor videos and 62 outdoor videos.

3.2.2 Language Description. One significant characteristic of the RVSD dataset is assigning language descriptions to each shadow on a frame-by-frame basis. Firstly, we incorporate language descriptions for shadows based on the ViSha dataset. The preceding referring object segmentation dataset [39] offered two kinds of language annotations: full-video expression and first-frame expression. The first-frame expression solely relies on the static attributes of the first frame image, whereas the full-video expression takes into account the whole video. However, neither of these expressions can perfectly align with the state of each target in the video (due to the movements of objects within the video), leading to instances where certain frames do not correspond with their language descriptions. Therefore, when adding language descriptions, we add corresponding language descriptions for each shadow target and its corresponding frames, making the content of the language descriptions match each frame. Meanwhile, we've established guidelines for the method of annotating language description, making the description more closely reflect the status of the target. In the language descriptions, we include the type of shadow, followed by the static attributes, the position information of the shadow, and

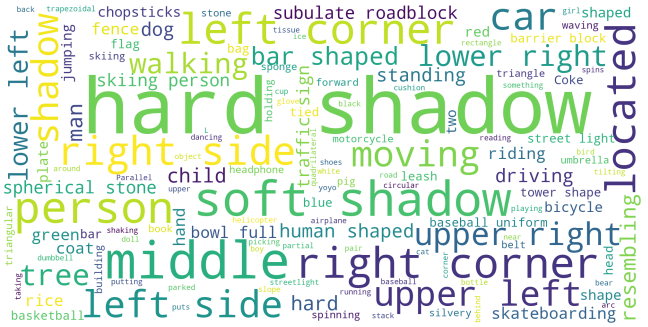

Figure 3: Word cloud of the RVSD dataset. The RVSD dataset encompasses a vast vocabulary that captures shadows from various perspectives, encompassing aspects like shadow type, location, shape, movement, and associated objects.

the action information of the object corresponding to the shadow. Table 2 presents the basic statistics of our language descriptions. Fig. 3 displays the word cloud for RVSD, featuring terms associated with shadow type (e.g., "hard shadow", "soft shadow"), actions of the shadow-casting subjects (e.g., "walking", "skateboarding"), position (e.g., "left", "upper right"), and the objects casting the shadows (e.g., "person" or "tree").

4 METHODOLOGY

4.1 Overview

Our approach is designed to generate a binary mask of the specified shadow of the given input video based on the corresponding natural language expression. Fig. 4 presents a schematic overview of our RSM-Net, specifically devised for the RVSD task. RSM-Net takes language descriptions and video frames as input, as depicted in the "Language and Image Integration" box to the right of Fig. 4. It reads the previously stored Twin-Track Synergistic Memory and processes the image through the MSA module to jointly achieve the accurate segmentation of the target shadow.

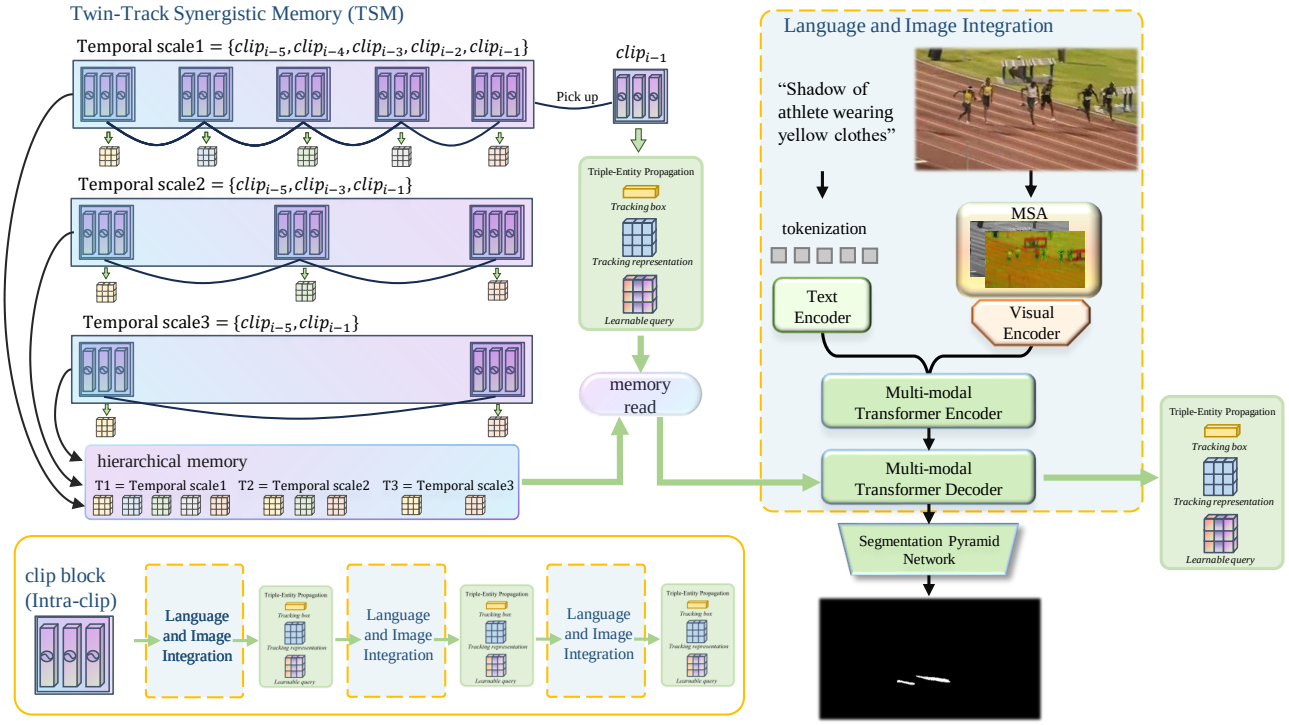

Figure 4: An overview of our approach. The TSM on the left side represents the construction phase of twin-track memory, which contains both inter-clip and intra-clip tracks of memory. The clip block (intra-clip) in the lower left corner signifies memory propagation between frames, with each clip in the figure containing three video frames (Language and Image Integration). Eventually, The hierarchical memory is strategically accessed for processing the current frame, ensuring comprehensive and context-aware shadow detection.

We first present how the twin-track synergistic memory is constructed and the content of our Triple-Entity memory propagation in Sec. 4.2. Subsequently, Sec. 4.3 is dedicated to a comprehensive exposition of our language and image integration, encompassing aspects of hierarchical memory reading and memory propagation. Lastly, in Sec. 4.4, we describe how our MSA effectively utilizes prior knowledge to facilitate the RVSD task.

4.2 Twin-Track Synergistic Memory (TSM) Construction

In response to the dynamic nature of shadows, we incorporate a twin-track memory structure into our framework. This encompasses both past hierarchical inter-clip memory and intra-clip memory, established before processing the current frame. The integration of these sequential memories in our network allows for a nuanced capture of evolving information across the video sequence, significantly contributing to the precise segmentation of the current frame. We emphasize that our hierarchical memory is dynamically updated. Specifically, when processing the current frame $clip_i$, we constructed memory using the preceding five clips $\{clip_{i-5}, ..., clip_{i-1}\}$. This dynamic updating and utilization in memory enhances our network's flexibility and efficiency while avoiding the disturbances from significant variations in very early video frames, allowing it to continuously adapt to the varying characteristics of shadows during the segmentation process.

First, our hierarchical inter-clip memory is generated from multiple clip blocks. The details of the clip block are in the yellow box in the lower left corner of Fig. 4. It mainly contains triple-entity memory propagation. In the triple-entity, the *tracking box* $t_{box} \in \mathbb{R}^{N_q \times 4}$ denotes the predicted shadow detection box, *tracking representation* $t_{rep} \in \mathbb{R}^{N_q \times d}$ refers to the encoded multi-modal features, and *learnable query* $t_{que} \in \mathbb{R}^{N_q \times d}$ signifies the shadow query. Our hierarchical inter-clip memory stores $t_{rep}s$ from three different time scales {T1,T2,T3}. Meanwhile, we pick up the triple-entity of the previous frame of the currently pending frame from $clip_{i-1}$ as the intra-clip memory. Next, we will elaborate on the processing of the current frame and the efficient retrieval of information from the twin-track memory.

4.3 Referring Video Frame Shadow Segmentation and Memory Read

In line with earlier researches [2, 49, 52], our learnable query-based referring segmentation largely follows a well-established paradigm, the Deformable DETR [52]. This involves processing a video frame, a textual expression, and a set of learnable queries. The output includes the target bounding box, segmentation mask, and associated output embeddings that match the input language expression.

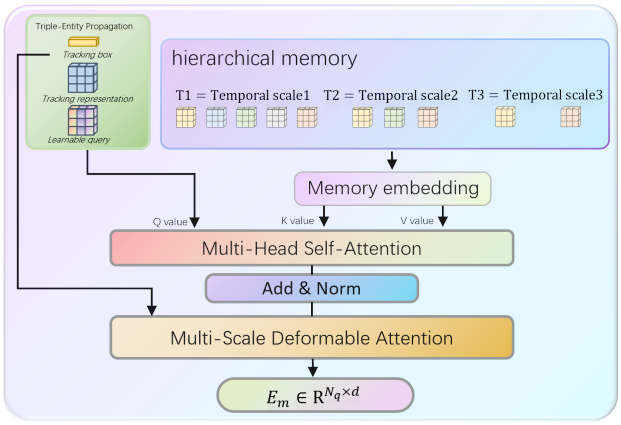

Figure 5: Details of hierarchical memory reading. First, perform memory embedding on it and then input it into the self-attention module together with the *L*earnable query.

To process a particular t_{th} frame image I_t, our network first performs attention enhancement of potential shadow regions via the MSA module, followed by extracting image feature $\mathcal{F}_i$ via the visual encoder. Meanwhile, the associated language description $\varepsilon = \{e_i\}_{i=1}^{N}$ with N words, where e_i is the i_{th} word is tokenized and fed into the text encoder, yielding $\mathcal{F}_\varepsilon$. Next, the visual and linguistic embeddings are linearly projected to a unified embedding space with the same dimension and concatenated to form a multi-modal embedding, denoted as $\mathcal{F}_m = \{\mathcal{F}_i, \mathcal{F}_\varepsilon\}$. This is then input into the transformer encoder, realizing cross-modal information fusion and interaction.

In the reading phase, it receives the triple-entity from the preceding frame along with the hierarchical memory, subsequently outputting the $E_m \in \mathbb{R}^{N_q \times d}$. Initially, we apply memory embedding to the hierarchical memory as shown in Fig. 5, thereby fusing features across various temporal scales as follows:

$$h_{\text{mem}} = \Phi_{\text{MLP}}^{(3)}\left(Concat\left(FC(\text{T1}), FC(\text{T2}), FC(\text{T3})\right)\right), \tag{1}$$

where $\Phi_{\text{MLP}}^{(3)}$ represents a three-layer MLP (Multi-Layer Perceptron). *Concat* and *FC* stand for concatenated operations and fully connected layers, respectively. The h_{mem} is fed into the Multi-Head Self-Attention module as both the Key (K) and Value (V), while the Query (Q) originates from the intra-clip *learnable query* t_{que}. During the propagation phase, the t_{rep} advances the output as $\Phi_{\text{MLP}}^{(3)}(E_m)$.

To generate the final mask, our segmentation pyramid network implements a cross-modal FPN (Feature Pyramid Network) [25] to enable multi-scale fusion between the linguistic features and visual feature maps (see Sec. 5 for model details). To summarize, the training objective of our network is to minimize the loss function as follows:

$$\underset{\alpha_{\text{box}},\beta_{\text{mask}}}{\arg\min}\ \mathcal{L}_{\text{refer}} = \alpha_{\text{box}}\mathcal{L}_{\text{box}} + \beta_{\text{mask}}\mathcal{L}_{\text{mask}}, \tag{2}$$

where $\mathcal{L}_{\text{box}}$ denotes the loss associated with bounding boxes, which is a composite of L1 loss and GIoU loss [37]. $\mathcal{L}_{\text{mask}}$ signifies the loss pertaining to masks, aggregating DICE loss [35] and the focal loss for binary masks. The coefficients for the losses are denoted as α_{box} for $\mathcal{L}_{\text{box}}$, and β_{mask} for $\mathcal{L}_{\text{mask}}$.

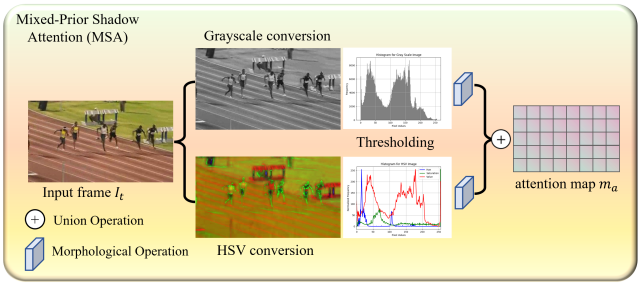

Figure 6: Illustration of MSA. The image is transformed into two distinct color spaces, enabling thresholding and morphological operations that result in the creation of a shadow attention map.

4.4 Mixed-Prior Shadow Attention (MSA)

The premise and foundation of the RVSD task is that the network can recognize and pay attention to the shadow region. Therefore, we design a Mixed-Prior Shadow Attention (MSA) module to utilize the physical prior knowledge [7, 11] to generate a weight map m_a to help the network focus more on the shadow region. Given the complexity and variability of the shadow background, our MSA identifies the shadow region from two different color spaces to improve robustness and generalization, as illustrated in Fig. 6.

Grayscale Thresholding. For a given image I_t, the grayscale representation is obtained by converting the RGB channels into a single intensity channel I_{gray}. Mathematically, the binary shadow mask M_{gray} is obtained as:

$$M_{\text{gray}} = \left\{(x,y) | T_{\text{min}} \le I_{\text{gray}}(x,y) \le T_{\text{max}}\right\}, \tag{3}$$

where the thresholds $[T_{\text{min}}, T_{\text{max}}]$ specify the acceptable range of grayscale values for identifying shadow regions.

HSV Thresholding. The input frame I_t is also converted to the HSV color space I_{HSV}. In the HSV color space, shadow regions typically exhibit lower values in the Saturation (S) and Value (V) channels due to color desaturation and reduced brightness caused by occlusion. Thus, two thresholds T_S and T_V are applied on the Saturation and Value channels respectively to discern the shadow regions. The binary shadow mask M_{HSV} is obtained as:

$$\begin{aligned} M_{HSV} = \{(x,y) | & T_{S_{\text{min}}} \le S(x,y) \le T_{S_{\text{max}}}, \\ & T_{V_{\text{min}}} \le V(x,y) \le T_{V_{\text{max}}}\}. \end{aligned} \tag{4}$$

The thresholds $T_{S_{\text{min}}}, T_{S_{\text{max}}}, T_{V_{\text{min}}}$, and $T_{V_{\text{max}}}$ specify the acceptable range for shadow regions in the S and V channels.

The combined shadow mask is then obtained by:

$$M_{\text{combined}}(x,y) = \Psi_{mor}(M_{\text{HSV}}(x,y)) \cup \Psi_{mor}(M_{\text{gray}}(x,y)). \tag{5}$$

The morphological operation (Ψ_{mor}) here specifically refers to the "opening" operation function with a kernel of size 5×5. This operation comprises an erosion followed by a dilation, which helps to remove noise and small interference while keeping the structure of the referring shadow. The final combined mask M_{combined} is then obtained by taking the union (logical OR operation denoted by $\cup$)

Table 3: Quantitative comparison of our network and other state-of-the-art methods on the RVSD Dataset.

Method	Pub.	Precision					IoU		mAP↑
		P@0.5↑	P@0.6↑	P@0.7↑	P@0.8↑	P@0.9↑	Overall↑	Mean↑	
URVOS [39]	ECCV 2020	51.2	44.8	37.5	28.6	16.1	58.5	45.9	33.5
CMPC-V [28]	TPAMI 2021	51.3	46.8	39.1	30.6	17.3	57.8	46.5	34.6
LBDT [9]	CVPR 2022	55.0	50.3	40.6	31.1	17.6	62.6	48.9	36.4
MTTR [2]	CVPR 2022	56.5	51.6	41.6	33.3	19.6	61.4	51.3	37.9
ReferFormer [49]	CVPR 2022	70.1	64.2	56.5	45.1	26.8	67.2	61.2	49.3
SgMg [34]	ICCV 2023	70.3	66.0	57.4	44.5	27.3	68.3	62.2	49.7
R2VOS [23]	ICCV 2023	71.2	66.9	58.3	45.0	27.0	69.8	62.4	50.2
OnlineRefer [48]	ICCV 2023	71.5	65.7	57.8	46.1	**27.9**	70.2	62.5	50.5
Our RSM-Net	-	**73.1**	**68.0**	**61.2**	**46.8**	27.1	**74.6**	**64.3**	**51.8**
Gain	-	↑1.6	↑1.1	↑2.9	↑0.7	↑-	↑4.4	↑1.8	↑1.3

Table 4: Ablation study of our methods on the RVSD Dataset. IntraC and InterC represent our network with only intra-clip memory features and only the inter-clip memory features, respectively. The IntraC-S and InC-H denote a single scale (Temporal scale 1) and a hierarchical scale (Temporal scale 1, 2, and 3), respectively for learning the inter-clip memory features.

Method	MSA	TSM			Precision					IoU		mAP↑
		IntraC	InerC-S	InterC-H	P@0.5↑	P@0.6↑	P@0.7↑	P@0.8↑	P@0.9↑	Overall↑	Mean↑	
Baseline	-	-	-	-	68.8	62.3	56.2	42.4	25.5	65.7	60.4	47.7
M1	✓	-	-	-	70.0	65.1	57.8	43.1	25.8	68.2	61.3	49.0
M2	✓	✓	-	-	71.9	66.6	59.5	45.6	27.0	71.8	62.8	50.7
M3	✓	✓	✓	-	72.6	67.1	59.5	45.7	**27.2**	72.0	63.3	51.0
Our RSM-Net	✓	✓	✓	✓	**73.1**	**68.0**	**61.2**	**46.8**	27.1	**74.6**	**64.3**	**51.8**

of these morphologically processed masks. The M_{combined} is then weighted as attention map m_a on the input frame I_t.

5 EXPERIMENTS

5.1 Evaluation Metrics and Data Setting.

Following referring segmentation works[2, 9, 44], we employ standard metrics for quantitative comparisons. These metrics are Precision@K, Overall IoU, Mean IoU, and mAP (mean Average Precision) across an IoU (Intersection over Union) range from 0.50 to 0.95 (a step of 0.05). IoU measures the overlap between predicted and ground truth regions, while precision@K evaluates the proportion of test instances surpassing the IoU threshold of K. The mAP metric computes the mean precision over varying IoU thresholds. To ensure a rich variety of scenarios in both the training and testing phases, the dataset is reasonably divided into training and testing subsets. The training subset encompasses 54 videos paired with 9,856 sentence-shadow combinations, while the testing subset contains 32 videos accompanied by 5,155 pairs.

5.2 Implementation and Training Details.

We utilize ResNet50 [13] as the visual backbones for extracting features. For the text encoding, we employ RoBERTa [30] and freeze its parameters throughout the training process. Following [48, 52], the final three-stage features from our visual backbone serve as inputs for both the Transformer encoder and the FPN [25]. The encoder and decoder of our Multi-modal Transformer framework have four layers, operating at a dimensionality of $d = 256$. The threshold range $[T_{S_{\min}}, T_{S_{\max}}]$, $[T_{V_{\min}}, T_{V_{\max}}]$ were empirically set as $[[0, 155], [6, 130]]$.

We perform all experiments using PyTorch on an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory for training. Our model optimization employs the AdamW optimizer [32] with an initial learning rate set at $1e-5$, while the visual backbone is adjusted at a lower rate of $5e-6$. The training spans 20 epochs, and the featuring learning rate reductions by a factor of 0.1 after the 3rd and 5th epochs. The initial frame's query number N_q is set to 5. Each video frame is resized to ensure a minimum dimension of 320 on the shorter side and a maximum of 576 on the longer side.

5.3 Comparison with State-of-the-art Methods

We compare our network against state-of-the-art methods on the RVSD dataset. These compared methods include URVOS [39], CMPC-V [28], LBDT [9], MTTR [2], ReferFormer [49], SgMg [34], R2VOS [23] and OnlineRefer [48]. In Table 3, we report the quantitative performance of our network and eight referring video segmentation methods on the RVSD datasets. According to Table 3, we can find that our RSM-Net clearly outperforms all compared methods, since our RSM-Net achieves the highest scores across nearly all evaluated metrics. Moreover, the last row ("Gain") in Table 3 highlights our method's improvements over competing approaches, and our network has an Overall IOU increase of 4.4%.

Moreover, Fig. 7 shows the visual comparisons of RVSD results produced by our network and other state-of-the-art methods. Apparently, with the same text descriptions and input video frames, our network can better identify the referred shadow regions than

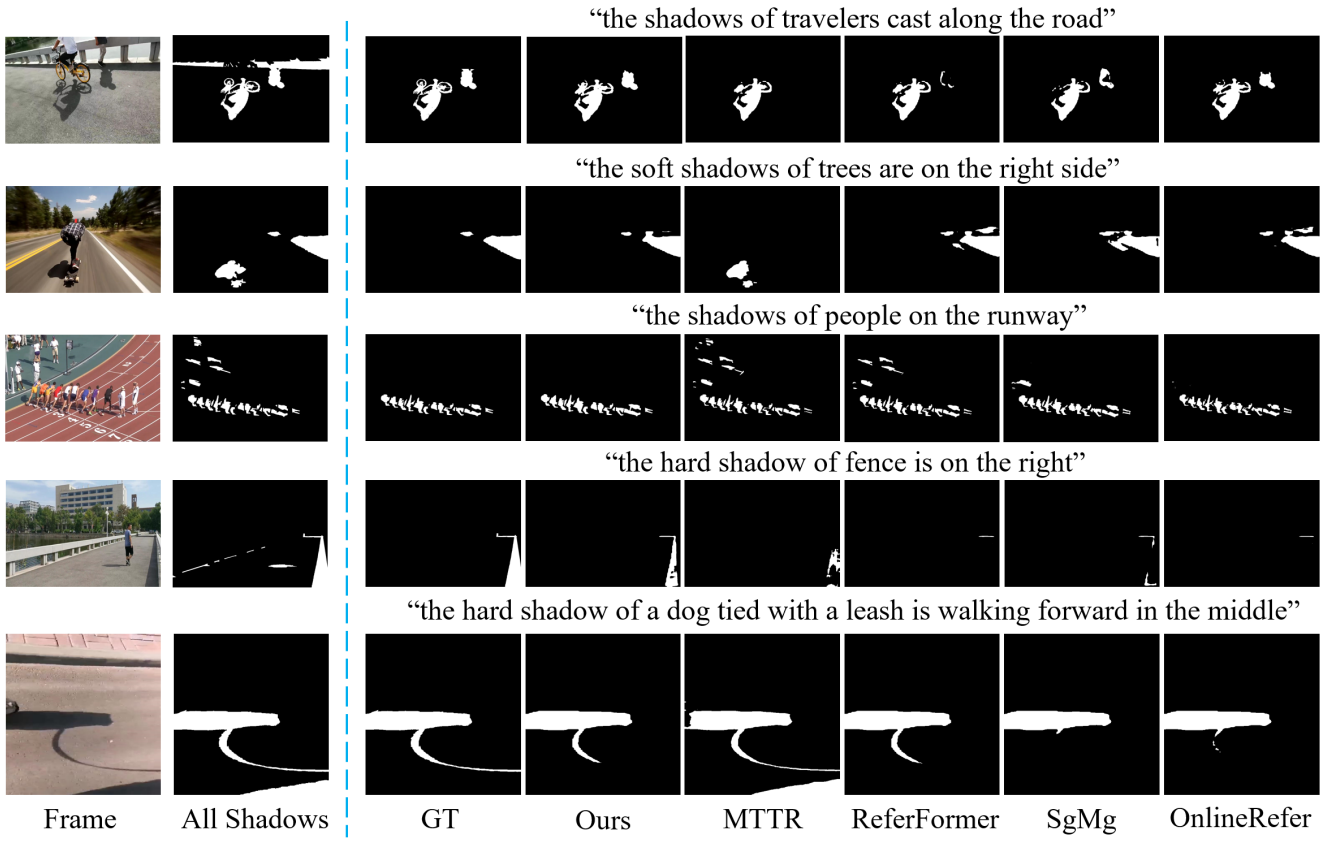

Figure 7: Visual comparisons of predicted video shadow detection results referring by the text descriptions. Apparently, our RSM-Net clearly outperforms compared methods and achieves more accurate shadow detection results based on the relevant text descriptions. Please zoom in for more details.

the compared methods, and our results are more consistent with the ground truths. For example, the first, second, and third rows of Fig. 7 show that our method can accurately segment the described region when there are multiple shadow regions present, whereas other methods (MTTR [2] in the second and third rows) may segment irrelevant shaded regions or may fail to recognize the corresponding shaded region (ReferFormer [49] and SgMg [34] in the first row). The fourth and fifth rows illustrate the segmentation for the shadow of interest through text descriptions when the object is not entirely visible or absent in the image, a feat not achievable with current instance shadow segmentation techniques. Our results can be observed to more closely align with the ground truths.

5.4 Ablation Studies

We further conduct ablation studies to validate the effectiveness of our MSA and TSM designs. To do so, we construct a baseline (denoted as "Baseline") by eliminating our MSA and TSM from our RSM-Net. Subsequently, we incrementally integrate MSA and TSM into the "Baseline" to formulate four networks, which are denoted as "M1", "M2", "M3", and our RSM-Net. Table 4 reports the quantitative results of these networks. Apparently, "M1" outperforms "Baseline", showcasing the effectiveness of the coarse shadow mask attention in our MSA for RVSD. The improvement in metrics from "M2" to "M1" demonstrates the contribution of the intra-clip memory features in our network. The progression from "M2" to "M3" with greater metric results further suggests that the incorporation of inter-clip features at a single temporal scale can bolster RVSD. while the larger metric results of "M3" than "M2", which further indicates that the inter-clip features at a single temporal scale can also enhance RVSD. In the end, our network has a superior performance over "M3", which demonstrates our hierarchical inter-clip features enable a better RVSD performance in our network.

6 CONCLUSION

In this study, we pioneer the RVSD task, which integrates linguistic prompts with video shadow detection, paving the way for new potential applications, such as interactive video editing. Our first contribution is the development and annotation of the dataset for RVSD, comprising 86 videos paired with 15,011 text descriptions and corresponding shadow masks. Furthermore, we devise a Twin-Track Synergistic Memory (TSM) module to learn intra-clip and hierarchical inter-clip memory features to boost segmentation performance and a Mixed-Prior Shadow Attention (MSA) module to learn a coarse shadow attention map for refining shadow areas for RVSD. Experimental results demonstrate that our method achieves better performance than other leading comparative methods.

REFERENCES

[1] A'Aeshah Alhakamy and Mihran Tuceryan. 2020. Real-time illumination and visual coherence for photorealistic augmented/mixed reality. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–34.

[2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. 2022. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4985–4995.

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

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

[5] Yung-Yu Chuang, Dan B Goldman, Brian Curless, David H Salesin, and Richard Szeliski. 2003. Shadow matting and compositing. In *ACM SIGGRAPH 2003 Papers*. 494–500.

[6] Runmin Cong, Yuchen Guan, Jinpeng Chen, Wei Zhang, Yao Zhao, and Sam Kwong. 2023. Sddnet: Style-guided dual-layer disentanglement network for shadow detection. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1202–1211.

[7] Rita Cucchiara, Costantino Grana, Massimo Piccardi, Andrea Prati, and Stefano Sirotti. 2001. Improving shadow suppression in moving object detection with HSV color information. In *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 01TH8585)*. IEEE, 334–339.

[8] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. 2023. MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions. *arXiv preprint arXiv:2308.08544* (2023).

[9] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. 2022. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4964–4973.

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

[11] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. 2005. On the removal of shadows from images. *IEEE transactions on pattern analysis and machine intelligence* 28, 1 (2005), 59–68.

[12] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. 2011. Single-image shadow detection and removal using paired regions. In *CVPR 2011*. IEEE, 2033–2040.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

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

[15] Yunzhong Hou and Liang Zheng. 2021. Multiview detection with shadow transformer (and view-coherent data augmentation). In *Proceedings of the 29th ACM International Conference on Multimedia*. 1673–1682.

[16] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 108–124.

[17] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. 2018. Direction-Aware Spatial Context Features for Shadow Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7454–7462. https://doi.org/10.1109/CVPR.2018.00778

[18] Xiang Huang, Gang Hua, Jack Tumblin, and Lance Williams. 2011. What characterizes a shadow boundary under the sun and sky?. In *2011 international conference on computer vision*. IEEE, 898–905.

[19] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 787–798.

[20] Salman Hameed Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. 2014. Automatic feature learning for robust shadow detection. In *2014 IEEE conference on computer vision and pattern recognition*. IEEE, 1939–1946.

[21] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. 2010. Detecting ground shadows in outdoor consumer photographs. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11*. Springer, 322–335.

[22] Hieu Le and Dimitris Samaras. 2020. From shadow segmentation to shadow removal. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 264–281.

[23] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. 2023. Robust Referring Video Object Segmentation with Cyclic Structural Consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22236–22245.

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

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

[26] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. 2017. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*. 1271–1280.

[27] Lihao Liu, Jean Prost, Lei Zhu, Nicolas Papadakis, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. 2023. SCOTCH and SODA: A Transformer Video Shadow Detection Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10449–10458.

[28] Si Liu, Tianrui Hui, et al. 2021. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 4761–4775.

[29] Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. 2023. Rotated Multi-Scale Interaction Network for Referring Remote Sensing Image Segmentation. *arXiv preprint arXiv:2312.12470* (2023).

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[31] Yanli Liu, Xingming Zou, Songhua Xu, Guanyu Xing, Housheng Wei, and Yanci Zhang. 2020. Real-Time Shadow Detection From Live Outdoor Videos for Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 7 (2020), 2748–2763.

[32] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[33] Yueming Lyu, Jing Dong, Bo Peng, Wei Wang, and Tieniu Tan. 2021. SOGAN: 3D-aware shadow and occlusion robust GAN for makeup transfer. In *Proceedings of the 29th ACM International conference on multimedia*. 3601–3609.

[34] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. 2023. Spectrum-guided Multi-granularity Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 920–930.

[35] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. Ieee, 565–571.

[36] Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras, and Nikos Paragios. 2011. Illumination estimation and cast shadow detection through a higher-order graphical model. In *CVPR 2011*. IEEE, 673–680.

[37] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 658–666.

[38] Elena Salvador, Andrea Cavallaro, and Touradj Ebrahimi. 2004. Cast shadow segmentation using invariant color features. *Computer vision and image understanding* 95, 2 (2004), 238–259.

[39] Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2020. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 208–223.

[40] Li Shen, Teck Wee Chua, and Karianto Leman. 2015. Shadow optimization from structured deep edge detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2067–2074. https://doi.org/10.1109/CVPR.2015.7298818

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

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

[43] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. 2016. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 816–832.

[44] Hongqiu Wang, Lei Zhu, Guang Yang, Yike Guo, Shichen Zhang, Bo Xu, and Yueming Jin. 2023. Video-Instrument Synergistic Network for Referring Video Instrument Segmentation in Robotic Surgery. *arXiv preprint arXiv:2308.09475* (2023).

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

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

[47] Scott Wehrwein, Kavita Bala, and Noah Snavely. 2015. Shadow detection and sun direction in photo collections. In *2015 International Conference on 3D Vision*. IEEE, 460–468.

[48] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. 2023. OnlineRefer: A Simple Online Baseline for Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2761–2770.

[49] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. 2022. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4974–4984.

[50] Zhenghao Xing, Tianyu Wang, Xiaowei Hu, Haoran Wu, Chi-Wing Fu, and Pheng-Ann Heng. 2022. Video Instance Shadow Detection. *arXiv preprint arXiv:2211.12827* (2022).

[51] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. 2010. Learning to recognize shadows in monochromatic natural images. In *2010 IEEE Computer Society conference on computer vision and pattern recognition*. IEEE, 223–230.

[52] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).

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