Wnet: Audio-Guided Video Object Segmentation via Wavelet-Based Cross-Modal Denoising Networks

Wenwen Pan[1,†], Haonan Shi[1,†], Zhou Zhao[1,*], Jieming Zhu[2], Xiuqiang He[2], Zhigeng Pan[3], Lianli Gao[4], Jun Yu[5], Fei Wu[1,6], Qi Tian[7],
[1]Zhejiang University, [2]Huawei Noah's Ark Lab, [3]Hangzhou Normal University,
[4]The University of Electronic Science and Technology of China, [5]Hangzhou Dianzi University,
[6]Shanghai Institute for Advanced Study of Zhejiang University, [7]Huawei Cloud & AI
{wenwenpan, shihn, zhaozhou}@zju.edu.cn, jiemingzhu@ieee.org, hexiuqiang1@huawei.com,
zgpan@hznu.edu.cn, lianli.gao@uestc.edu.cn, yujun@hdu.edu.cn, wufei@cs.zju.edu.cn,
tian.qi1@huawei.com

Abstract

Audio-Guided video object segmentation is a challenging problem in visual analysis and editing, which automatically separates foreground objects from the background in a video sequence according to the referring audio expressions. However, existing referring video object segmentation works mainly focus on the guidance of text-based referring expressions, due to the lack of modeling the semantic representations of audio-video interaction contents. In this paper, we consider the problem of audio-guided video semantic segmentation from the viewpoint of end-to-end denoising encoder-decoder network learning. We propose the wavelet-based encoder network to learn the cross-modal representations of the video contents with audio-form queries. Specifically, we adopt the multi-head cross-modal attention layers to explore the potential relations of video and query contents. A 2-dimension discrete wavelet transform is merged into the transformer encoder to decompose the audio-video features. Next, we maximize mutual information between the encoded features and multi-modal features after cross-modal attention layers to enhance the audio guidance. Then, a self attention-free decoder network is developed to generate the target masks with frequency-domain transforms. In addition, we construct the first large-scale audio-guided video semantic segmentation dataset. The extensive experiments show the effectiveness of our method[1].

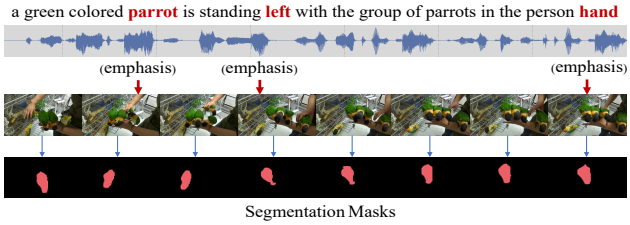

Figure 1. The audio-guided video object segmentation task.

1. Introduction

Referring video object segmentation aims to segment video objects referred by given language expressions, which has attracted wide attention due to its applicability to many practical problems including video analysis and video editing [33,35,49,50,61]. Currently, most referring video object segmentation approaches mainly focus on the guidance of text-guided referring expressions [18, 19, 31, 33, 35, 49, 61, 63], which can learn the multi-modal representation from the interaction network layer, and then generate the object masks to the given text references. The existing works have achieved promising performance in text-based video object segmentation, but they may still be ineffectively applied to the audio-guided video object segmentation due to the lack of modeling the semantic representation of audio-video interaction contents.

The audio-guided video analysis is a simulation of human cognition, comparing with the text-guided analysis [44]. Humankind use speech exclusively long before the invention of writing. People also learn and use language in the real world, as to collaborate, describe and relate their visual environment, talk about each other, and so on. Furthermore, in the natural scene, audio interaction is more convenient and common than text interaction. Although audio

[†]Equal contribution.
[*]Corresponding Author.
[1]Code is available at: https://github.com/asudahkzj/Wnet.git

inputs can be converted to text inputs through ASR models [3, 4, 46], the process will produce unavoidable losses. Since Harwath and Glass's collection of spoken captions for Flickr8k [14], more works address cognitive and linguistic questions [8, 10–12]. Other work addresses applied tasks, including multi-modal retrieval [22], cross-modality alignment [13, 27], retrieving speech in different languages using images as a pivot modality [1, 26, 39], and speech-to-speech retrieval [1, 39]. Our work focuses on the audio-guided video object segmentation tasks, shown as Fig. 1. The audio guidance often contains rich semantic information, such as the accent, emotion and speed. These extra factors can facilitate the object segmentation. The same object can correspond to different pronunciations, while the same pronunciation can point to different objects. Thus, the simple extension of the existing segmentation works based on text-based guidance is difficult for modeling the semantic representation of audio-video interaction contents. Inspired by MulT [51], we use multi-head cross-modal attention layers to fuse the video embeddings and audio embeddings. Different from the MulT model [51], we extend dimensions of inputs and apply it to large-scale natural language datasets. The cross-modal transformers referred to text embeddings are all removed.

One other bottleneck is the noise problem, derived from acquisition noise and fusing noise [7]. For the acquisition noises, we use a pre-trained MFCC model [5] to extract acoustic features, which is widely used in automatic speech and speaker recognition. In this paper, we focus on the processing of fusing noise. There is a large gap between video and audio representations. The joint representations reflect important information considering multi-modal alignment. Audio and video features have different redundant parts (i.e. irrelevant phonemes and pixels), likewise termed as noise. These noises are difficult to handle only by convolution operations and attention mechanisms in the time domain. As mentioned in [29], noises are likely to concentrate at high frequencies. Recently, Fnet [30] has been proposed to learn the frequency-domain-level representation with Fourier transforms for recognition tasks, while it only aims to speed up the encoder architectures but fails to obtain improvement in performances. Low-pass filtering on Fourier analysis cannot effectively distinguish the high-frequency parts of the required signal from the high-frequency interference caused by noise. If the low-pass filtering is too narrow, parts of the required signal are treated as noise and its morphological information is erased, which leads to the distortion of the original signal [45].

Motivated by this, we integrate the 2-dimension discrete wavelet (DWT) transform into the transformer encoder, which replaces self-attention layers with DWT layers. The DWT denoising has proved its effectiveness in image denoising [25, 45, 52], but has not been used in multi-modal representation yet to our knowledge. We are the first to devise the DWT-transformer for the audio-visual joint representation to filter the noise and outliers, as a priori. The layers of the whole transformer encoder are reduced, which obtains a sizable performance boost in terms of speed and model consumption. Inspired by the AMDIM [2], we maximize mutual information between the encoded features and multi-modal features after the cross-modal attention to enhance the audio guidance.

The main contributions of this paper are as follows: (i) Unlike the previous studies, we study the problem of audio-guided video object segmentation from the viewpoint of end-to-end denoising encoder-decoder network learning. (ii) We propose the wavelet-based encoder network to learn the cross-modal representations of the video contents with audio-form queries. (iii) We construct a large-scale dataset for audio-guided video object segmentation and validate the effectiveness of our proposed method through extensive experiments.

2. Related Work

2.1. Referring Expression object Segmentation

The referring expression segmentation task has attracted increasing research interest [18, 19, 31, 33, 35, 49, 61, 63] in recent years. Hu et al. [18] formulate this task as an image-region-wise classification problem. Li et al. [31] employ multi-scale image features from multiple convolutional layers. Qiu et al. [41] further enhance visual features and introduce an adversarial mechanism. Some works [33, 35, 49, 50, 61] make more interactions between the image and natural language query. Furthermore, the attention module [61, 63] is introduced to the segmentation task. To enhance the accuracy, further works successfully model the dependencies of cross-modal information [20], informative words of the expression [21] and localization information of the referent instances [24]. Moreover, Luo et al. [34] achieve a joint learning of referring expression comprehension and segmentation. [28] extends technologies to video data and incorporated temporal coherency. For the video data, existing methods commonly employ dynamic convolutions [9, 54] to adaptively generate convolutional filters, or leverage cross-modal attention [38, 55, 62] to compute the correlations among input visual and linguistic embeddings However, these works cannot handle the noise problem of audio-video joint representations.

2.2. Speech-Based Video Analysis

Comparing with the text-guided video analysis, audio-guided analysis is a more precise simulation of human cognition to the world [44]. Actually, people use speech exclusively long before the invention of writing. Harwath et al. [14] collect spoken captions for Flickr8k, and then much

Table 1. The statistics of the AVOS dataset.

	RVOS	A2D	J-HMDB	Total
Number of Audio	11,226	6,656	929	18,811

research [7, 12, 17, 47] begins to attach importance to this task. Some works emphasize the cognitive and linguistic questions, such as understanding how different learned layers correspond to visual stimuli [8, 10], learning linguistic units [11, 12] or how visually grounded representations can help understand lexical competition in phonemic processing [15]. Ramon Sanabria et al. [44] propose dual encoder models that can be used for efficient multimodal retrieval. However, these works take less consideration of video object segmentation.

3. Audio-Guided-VOS Dataset (AVOS)

There are previous works that constructed referring segmentation datasets for videos. Gavrilyuk et al. [9] extended the A2D [58] and J-HMDB [23] datasets with natural sentences. Seo et al. constructed the first large-scale referring video object segmentation dataset called RVOS [48].

To facilitate audio-based video object segmentation, we have constructed a large-scale audio-guided dataset, Audio-Guided-VOS (AVOS)[2], with referring audio expressions as Tab. 1. AVOS is the extension of RVOS [48], A2D [58] and J-HMDB [23]. We select the three datasets for their rich scene information. To obtain audio annotations, we employ 36 speakers to read the sentences totally. To ensure the recording quality, all the speakers are required to read proficiently, do not stammer, stuck and other situations. The sampling rate is 44,100K or above, the sampling number is 16 bits, and the speaking speed is 100-150 words per minute. Speaking speed should be normal speaking speed, or TV announcer speaking speed. The word accuracy of text files and audio files is not less than 99% under manual checking. The average length of each recording is 5 to 6 seconds, about 28 hours in total. Moreover, we have run two rounds of inspections. We not only correct the pronunciation in the recordings, but also correct grammar and spelling errors in the original texts. The ratio of the training set, the test set and the validation set is $75:15:10$.

4. Proposed Method

We present a video sequence as $\mathbf{v} = \{\mathbf{v}_i\}_{i=1}^{n}$, where $\mathbf{v}_i$ is the pre-extracted visual feature of the i-th frame and n is the frame number of the video. Each video is associated with an audio query, denoted by $\mathbf{q} = \{\mathbf{q}_i\}_{i=1}^{m}$ where $\mathbf{q}_i$ is the feature of i-th frame and m is the frame number of the audio. The goal of the audio-guided video object segmentation is to predict binary segmentation masks $\mathbf{S} = \{\mathbf{S}_i \in \{0,1\}^{W_o \times H_o}\}_{i=1}^{n}$.

[2] https://drive.google.com/drive/folders/Audio-Guide-Segmentation

4.1. Analysis on Wavelet Transform

As to the convolutional neural network, each convolutional layer is composed of several convolutional units, and the parameters of each convolutional unit are optimized by back propagation algorithm. Convolution operations aim to extract different features of inputs, represented as follows.

$$W(\tau) = \int_{-\infty}^{\infty} f(t)g(\tau - t)\mathrm{d}t. \tag{1}$$

The convolution kernel in the convolution layer is relatively fixed. Audio-video joint representations contain rich time-frequency characteristics, which are more suitable for window functions that vary in the time-frequency domain. The wavelet can be represented as follows.

$$W(a,\tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi(\frac{t-\tau}{a})\mathrm{d}t, \tag{2}$$

where scaling function $\psi_{a,\tau}(t) = a^{-\frac{1}{2}}\psi(\frac{t-\tau}{a})$ and a is the scale, which is inversely proportional to the frequency. The operation of the traditional convolution layer and wavelet have the commonality. The difference is $g(\tau - t)$ and $\psi(\frac{t-\tau}{a})$. Audio and video features have different redundant parts (i.e. irrelevant phoneme and pixel), termed as noises. The noises from the video and audio inputs are distributed among most features after the cross-modal attention. These noises are difficult to handle only by convolution operations in the time domain. As mentioned in [29], noises are likely to concentrate at high frequencies. Fnet [30] proposes to use Fourier sublayers to replace the self-attention layers. However, low-pass filtering on Fourier analysis cannot effectively distinguish the high-frequency part of the required signal from the high-frequency interference caused by noise. Wavelet can well retain the peak value and mutation part of the useful signal required in the original signal. It has good time-frequency localization characteristics and can be expressed linearly as:

$$W_x = W_f + W_e, \tag{3}$$

where W_e is the wavelet coefficients controlled by noise. We can use threshold quantization to reconstruct denoising joint representations. Furthermore, we can obtain improvements by replacing self-attention layers with the DWT layers in terms of the model consumption and speed.

4.2. Overview

As Fig. 2 illustrated, our model can be divided into five modules: visual encoder, audio encoder, transformer encoder, transformer decoder and segmentation module.

Visual Encoder. We employ ResNet-50 [16] as our backbone network to extract visual features from an input frame.

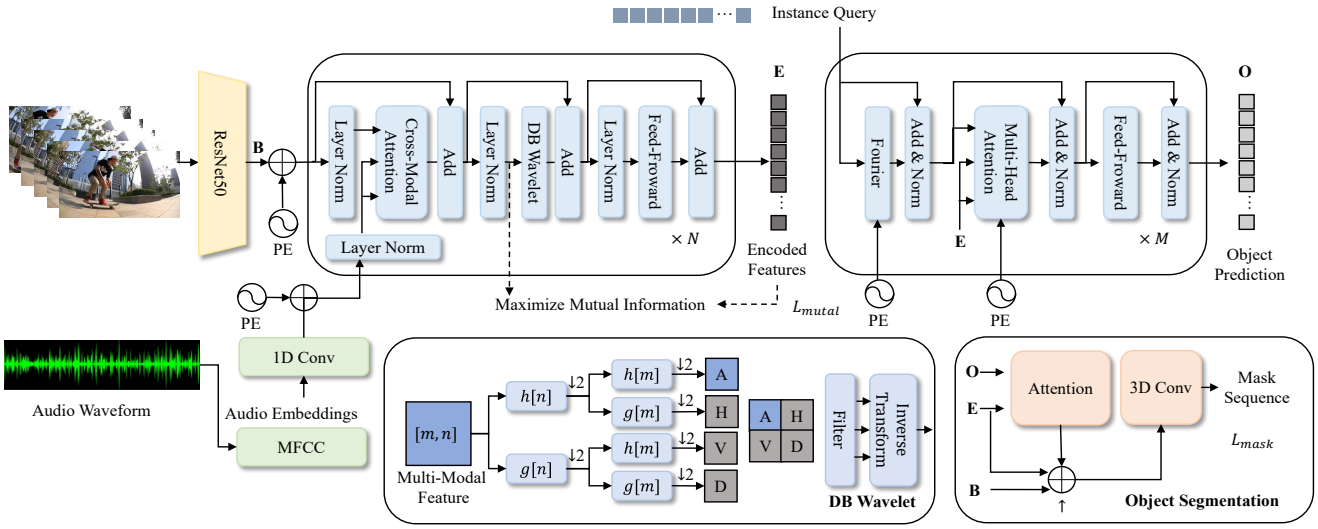

Figure 2. The whole framework for our segmentation model.

To include spatial information of the visual feature, we augment 3-dimensional spatial coordinates following [56], denoted as $\mathbf{v} = \{\mathbf{v}_i\}_{i=1}^{n}$. The output of the backbone is $\mathbf{B} \in \mathbf{R}^{n\times c\times H\times W}$. c represents the original video dimension. A 1×1 convolution is used to reduce the dimension to $\mathbf{R}^{n\times d\times H\times W}$. We then flatten the dimensions as $d \times (n \times H \times W)$.

Audio Encoder. MFCCs are the features widely used in automatic speech and speaker recognition. Following [37], we capture information about lower frequencies than higher frequencies by non-linear scaling and thus acts as a human ear. A set of MFCCs is encoded as a multi-hot vector and projected onto an embedding space using the 1D convolution, denoted as $\mathbf{q} = \{\mathbf{q}_i\}_{i=1}^{m}$.

Transformer Encoder and Decoder. We devise a transformer encoder-decoder framework for our audio-guided video object segmentation model. The model is in end-to-end manners. The transformer encoder is employed to learn the cross-modal representations of the video contents with audio-form queries. We first apply the layer normalization to the visual features and audio features, respectively. Next, we devise a wavelet-based cross-modal module to fuse the two modalities and achieve denoised joint representations. Each encoder layer consists of a multi-head attention module [53] and a fully connected feed-forward network. Then, we maximize the mutual information between the cross-modal representations and the encoded representations. During this stage, the temporal order is the same as the order of the initial input.

The transformer decoder aims to generate the top pixel features that can represent the target object of each frame. Motivated by Fnet [30], we also replace the self-attention sublayers with simple linear transformations. The self attention-free decoder can better handle audio-video encoding. Besides the Fourier layers, we follow the standard architecture of the transformer, using multi-headed encoder-decoder attention mechanisms. Following [56], the decoder then takes a small fixed number of learned positional embeddings (object queries) as inputs, and attends to the encoder output. The overall predictions follow the input frame order. We remove all self-attention layers in the transformer encoder-decoder framework to reduce model computation. Details of the transformer are in 4.3 and 4.4.

Object Sequence Segmentation. The module aims to predict the mask sequence for the target object. We capture the object predictions $\mathbf{O}$, backbone features $\mathbf{B}$ and encoded feature maps $\mathbf{E}$ from the previous layers, shown in Fig. 2. First, we employ an attention module to calculate the similarity map between $\mathbf{O}$ and $\mathbf{E}$. Following [56], we only compute the features of its corresponding frame. Next, we fuse the similarity map, $\mathbf{B}$ and $\mathbf{E}$ of the corresponding frames, following the DETR [6]. $\mathbf{B} \in \mathbf{R}^{n\times c\times H\times W}$, $\mathbf{E} \in \mathbf{R}^{d\times n\times (H\times W)}$, $\mathbf{O} \in \mathbf{R}^{n\times d}$, where n denotes the frame numbers, c and d denote the dimension. Then, we use a deformable convolution as the last layer of the fusion. Thus, the mask features for the target object of different frames are achieved. Finally, the 3D convolution, which three 3D convolutional layers and group normalization layers [57] with ReLU activation function, is employed to obtain the mask sequence.

4.3. DWT-Based Transformer Encoder

Comparing with the text-guided semantic segmentation, audio-based segmentation suffers from severe noise problems [44], derived from acquisition noise and fusing noise. For the acquisition noises, we use a pre-trained MFCC models [37] to extract acoustic features. For the fusing noise, we propose a DWT-based transformer encoder to realize multi-modal encoding and joint feature denoising.

We consider visual modality and audio modality, with

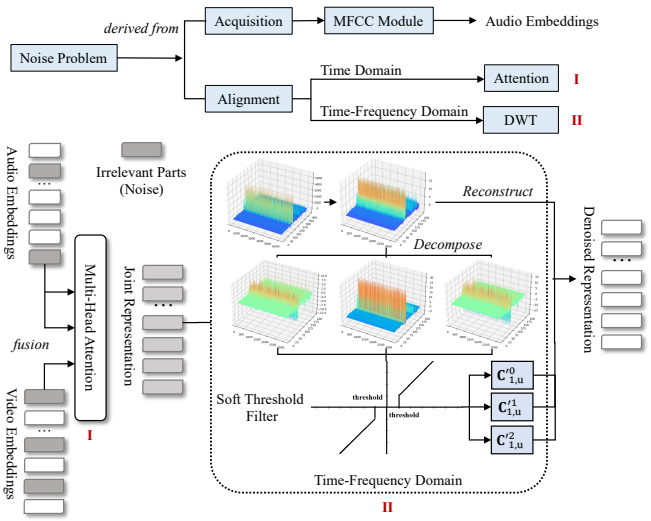

Figure 3. The process of the denoising operation.

two potentially non-aligned sequences from each of them, denoted as $\mathbf{v^P} \in \mathbf{R}^{\mathrm{T_v} \times \mathrm{d_v}}$ and $\mathbf{q^P} \in \mathbf{R}^{\mathrm{T_a} \times \mathrm{d_q}}$. $\mathrm{T}_{(\cdot)}$ represents the sequence length (audio or video), and $\mathrm{d}_{(\cdot)}$ represents the dimension, respectively. Inspired by the multi-modal transformer in MulT [51], we attend to interactions between multi-modal sequences across distinct time steps and latently adapt streams from audio modality to visual modality. We assume the input of the cross-modal attention is a sequence of queries $\mathbf{Q} = \mathbf{v}^p\mathbf{W}_Q$, keys $\mathbf{K} = \mathbf{q}^p\mathbf{W}_K$ and values $\mathbf{V} = \mathbf{q}^p\mathbf{W}_V$. The cross-modal attention is calculated by

$$\mathrm{Attention}_{a \to v}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}(\frac{\mathbf{Q}^\top \mathbf{K}}{\sqrt{\mathrm{d_k}}})\mathbf{V}^\top, \quad (4)$$

where $\mathrm{d_k}$ is the query dimensions and Softmax operation is performed on every row. We employ multi-head attention layers [53], which consists of H paralleled cross-modal attention layers. Finally, we obtain cross-modal representation $\mathbf{f} \in \mathbf{R}^{\mathrm{T_v} \times \mathrm{d_v}}$.

Noise problem can be derived from acquisition (pause, environment noise, etc.) and alignment. For the acquisition noises, we use a pre-trained MFCC model [5] to extract acoustic features, which is widely used in automatic speech and speaker recognition. There is a large gap between video and audio representations. The joint representations reflect important information considering multi-modal alignment. Attention is an importance estimate in the time domain, while the frequency domain can reflect another granularity of importance estimate. For multi-modal tasks, the importance of the frequency band requires consideration of interactive alignment information, and the high frequency part is more likely to be noisy information unrelated to alignment. For example, the noise can be parts of the audio irrelevant to the content of the video, or parts of the video irrelevant the audio features. To deal with the noise problem, we adopt 2d Discrete Wavelet Transform (DWT) for the joint representation. The reason for using DWT instead of DFT is that DFT is more likely to lose useful information at high frequencies, resulting in a decrease in actual performance.

The proposed algorithm is a hybrid approach that uses spatial and transform-domain information. Wavelet transform decomposes a signal into its sub-bands using a series of high-pass and low-pass filters. As noise is generally categorized as a high-frequency component, it is easier to separate it from the signal using wavelet transform. The decomposition of frequency content depends on the number of levels of DWT. The cross-modal representations $\mathbf{f} \in \mathbf{R}^{\mathrm{T_v} \times \mathrm{d_v}}$ serve as the input signal. The DWT separates filtering operations on rows and columns. $\mathrm{A_{j,u}}$ and $\mathrm{C^k_{j,u}}$ denote scaling and wavelet coefficients at scale j for the given signal $\mathbf{f}$ where $k = 1, 2, 3$. We'll be working with separable orthonormal filters so 2D filters can be expressed as a product between low pass filter h and high pass filter g. The coefficients at scale j can be obtained from coefficients at scale $j + 1$. We can obtain $\mathbf{A}_{j,u}$ and $\mathbf{C}^k_{j,u}$ as follows.

$$\begin{aligned}
\mathbf{A}_{j,u} &= \sqrt{2}\sum_u hh(l-2u)\mathbf{A}_{j+1,l}; \\
\mathbf{C}^1_{j,u} &= \sqrt{2}\sum_u hg(l-2u)\mathbf{A}_{j+1,l}; \\
\mathbf{C}^2_{j,u} &= \sqrt{2}\sum_u gh(l-2u)\mathbf{A}_{j+1,l}; \\
\mathbf{C}^3_{j,u} &= \sqrt{2}\sum_u gg(l-2u)\mathbf{A}_{j+1,l}.
\end{aligned} \quad (5)$$

To implement the filter bank, we use two-stage filter banks. In the first stage, rows of two-dimensional signal are convolved with h, g filters and then we downsample columns by 2. In the next stage, columns are convolved with the filters h, g and we keep only even indexed rows. A $\mathrm{n} \times \mathrm{d_v}$ cross-modal signal is transformed into four $\frac{\mathrm{n}}{2} \times \frac{\mathrm{d_v}}{2}$ signal after the two stages.

Next, we perform threshold quantization on the high-frequency coefficients $\mathrm{C^k_{j,u}}$ of wavelet decomposition. For the high-frequency coefficients (in three directions) of each layer from layer 1 to layer N, a threshold value is selected for threshold quantization. We adopt VisuShrink threshold α with a soft threshold function. For $\phi = \max|\mathrm{C'^k_{j,u}}|$, the filter operations can be represented as follows ($k \in [1, 2, 3]$).

$$\mathrm{C'}^k_{j,u}[x, y] = \mathrm{sgn}(\mathrm{C}^k_{j,u}[x, y])(\left|\mathrm{C}^k_{j,u}[x, y]\right| - \alpha\phi)_+ \quad (6)$$

Then, we perform the wavelet reconstruction of the signal. The wavelet is reconstructed according to the low frequency coefficients of the N-th layer of wavelet decomposition and the high-frequency coefficients from the 1st layer to the N-th layer after quantization.

$$\mathbf{f_n} = \mathrm{DWTInverse}(\mathbf{A}_{j,u}, \mathbf{C'}^k_{j,u}), \quad (7)$$

where $\mathbf{f_n}$ is the denoised joint representation. The $\mathbf{f_n}$ serves as the input of the following feed-forward layers and layer norm. Finally, we obtain the encoded representation $\mathbf{E}$.

4.4. Self Attention-Free Decoder

Due to the fusion between the audio and video signal, joint representations are more suitable for processing in the frequency domain. We adopt the decoder without the self-attention layers to achieve speedups. Inspired by [30], each layer consists of a Fourier mixing sublayer followed by a feed-forward sublayer. We replace the self-attention sublayer of each transformer decoder layer with a Fourier sublayer. A 2D DFT is applied to its embedding input. One 1D DFT is along the sequence dimension, $\mathbb{F}_{seq}$, and one 1D DFT is along the hidden dimension $\mathbb{F}_{hidden}$.

$$y = \mathbb{R}(\mathbb{F}_{seq}(\mathbb{R}(\mathbb{F}_{hidden}(x)))), \tag{8}$$

We only keep the real part of the result. After the Fourier layers, we employ the multi-head attention layers. The object prediction $\mathbf{O}$ is obtained.

4.5. Training of Wnet

Among the whole model, the loss function includes the mask loss, the box loss and the mutual loss.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{mask} + \lambda_2 \mathcal{L}_{box} + \mathcal{L}_{mutual}, \tag{9}$$

where λ_1, λ_2 aim to adjust the three losses. The mask loss for supervising the predictions is defined as a combination of the Dice [36] and Focal [32] loss:

$$\mathcal{L}_{mask}(m_i, m_{\sigma(i)}) = \frac{1}{T}\sum_{t=0}^{T}[\mathcal{L}_{Dice}(m_{i,t}, m_{\sigma(i),t}) + \mathcal{L}_{Focal}(m_{i,t}, m_{\sigma(i),t})], \tag{10}$$

where m is the predicted mask, m_σ is the target mask and T is the number of frames in the video. $\mathcal{L}_{box}$ scores the bounding boxes. We use a linear combination of the sequence level L1 loss and the generalized IOU [43] loss.

$$\mathcal{L}_{box}(b_i, b_{\sigma(i)}) = \frac{1}{T}\sum_{t=0}^{T}[\mathcal{L}_{iou}(b_{i,t}, b_{\sigma(i),t}) + ||b_{i,t} - b_{\sigma(i),t}||_1], \tag{11}$$

We use KL divergence [60] to maximize the mutual information between the cross-modal representation $\mathbf{f}$ and the encoded representation $\mathbf{E}$.

$$\mathcal{L}_{mutual}(\mathbf{E}(i,j)||\mathbf{f}(i,j)) = \sum \mathbf{E}(i,j)(\log \frac{\mathbf{E}(i,j)}{\mathbf{f}(i,j)}), \tag{12}$$

where i stands for the sequence and j stands for the dimension. $\mathbf{E}$ and $\mathbf{f}$ are sent to the softmax function before computing KL divergence. The KL divergence is used to pull in the distance between the cross-modal representation and the encoded representation. Thus, the audio guidance is strengthened, which avoids the DWT operation filtering too many audio factors.

Table 2. The comparison of different method for audio-guided semantic segmentation on AVOS.

Model	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
URVOS+ [48]	37.1%	39.2%	38.2%
PAM+ [38]	38.6%	38.9%	38.8%
VisTR+ [56]	38.0%	39.5%	38.8%
Wnet (Ours)	**43.0%**	**45.0%**	**44.0%**

Table 3. The results of different datasets in the AVOS dataset. In this table, we use the same dataset for training and testing.

Dataset	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
RVOS	43.0%	44.1%	43.6%
A2D	49.8%	55.1%	52.5%
J-HMDB	65.6%	56.7%	61.2%

Note: JHMDB-Sentences is only for evaluation, not training, so it is directly evaluated using the checkpoint trained on A2D-Sentences.

5. Experiments

5.1. Performance Criteria

We evaluate the performance of our Wnet method based on two widely-used evaluation criteria for audio-guided video semantic segmentation following [40]. Given the testing video sequence $\mathbf{v}$ and audio query $\mathbf{q}$ with the ground-truth masks $\mathbf{G}$, we denote the generated masks from our Wnet method by $\mathbf{S}$. We employ the Jaccard index $\mathcal{J}$ defined as the intersection-over-union of the generated segmentation and the ground-truth mask ($\mathcal{J} = |\frac{S \cap G}{S \cup G}|$). From a contour-based perspective, one can interpret S as a set of closed contours $c(S)$ delimiting the spatial extent of the mask. Therefore, one can compute the contour-based precision and recall P_c and R_c between the contour points of $c(S)$ and $c(G)$. We adopt F-measure as a trade-off between the two ($\mathcal{F} = \frac{2P_c R_c}{P_c + R_c}$).

5.2. Implementation Details

Visual Feature Extraction. We use a ResNet-50 backbone to extract visual features, which has the same settings as DETR [6]. And it is then fed into a 2D convolution with kernel size 1 to map the model dimension and each frame is concatenated to form the clip level feature.

Acoustic Feature Extraction. We use a 39-dimensional MFCC to represent its acoustic feature. Then, we use 1D convolution to further extract features and map them to the corresponding dimensions of the model following the implementation by Tsai et al. [51]. The kernel size of 1D convolution is 1.

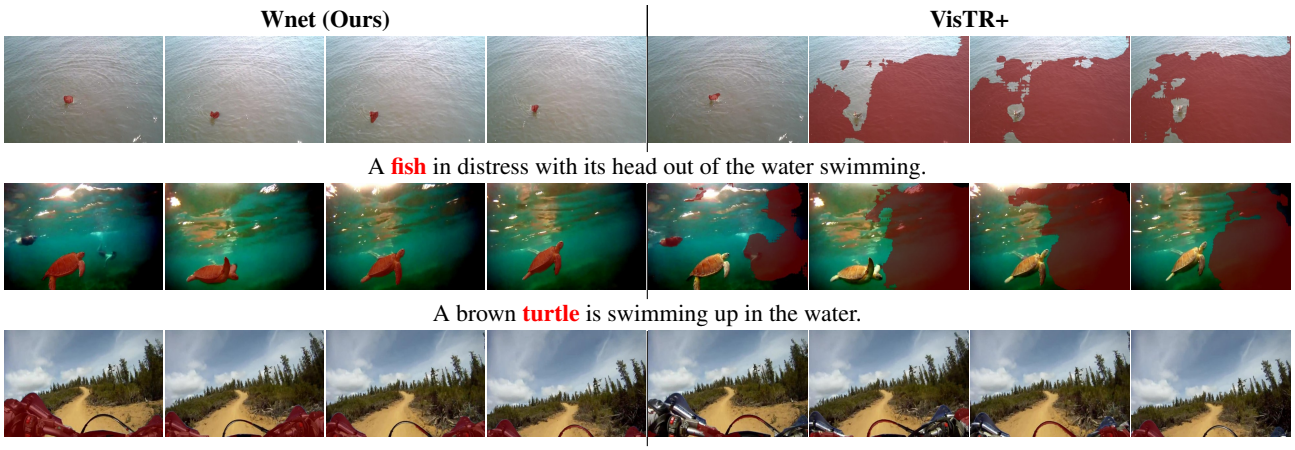

Figure 4. Visualization of Wnet and VisTR+ on the AVOS.

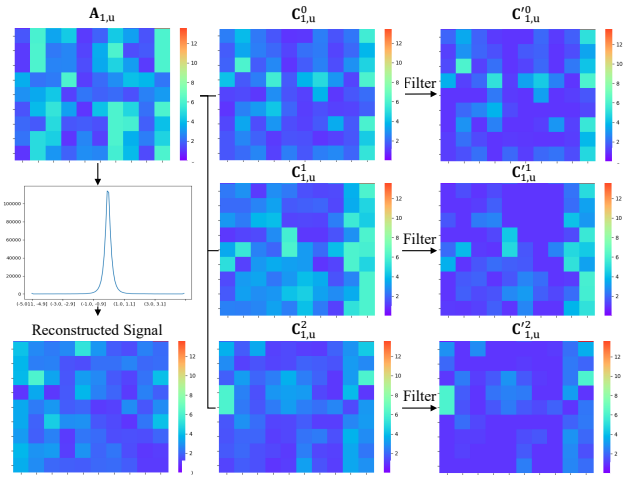

Figure 5. Visualization of DWT-based denoised features.

Dataset Processing. Our dataset (AVOS) contains three parts (RVOS, A2D and J-HMDB). For the RVOS part, we use the same videos in Youtube-VIS [59], as VisTR [56]. The mask annotations in validation and test set are unavailable, so we divide the training set into the training, validation and test set in our experiments. However, we also offer the audio queries for original validation and test set.

Model Setting. We adopt a 2-layer, 8-head multi-head cross-attention [51] module with the width of 3 to fuse visual and audio features. Between the attention layer and the feed-forward layer, a wavelet transform filter layer is used to remove noise from joint representations.For the transformer decoder, we use Fourier transform [30] instead of the self-attention layer. After obtaining the prediction of the decoder and the encoder, for each corresponding frame, we send them to an attention module to obtain the attention map, which is not multiplied by the value. Then it will be fused with the backbone features and the memory to get the mask features for each instance of each frame, following the same practice with VisTR [56]. We expand the number of frames per video to 36 for end-to-end training, and applied 36 query slots for 36 objects throughout the video. Finally, we use three Conv3d layers and GroupNorm layers [57] with ReLU activation. The Conv3d layers have the kernel size of 3, padding of 2 and dilation of 2. And we use a last Conv3d layer with the kernel size of 1 to obtain the mask. More details are in supplementary material.

5.3. Performance Comparisons

We compare our proposed method with other existing methods for the problem as follows:

VisTR+ is the extension of the transformer-based video instance segmentation algorithm [56], where the cross-modal attention layer is added to fuse the two modalities. For the VisTR+, we use the Hungarian loss as [56]. For our Wnet, we use the box and mask loss.

URVOS+ is the extension of the unified referring video segmentation network [48], where the MFCC layer [5] is added to encode the audio inputs.

PAM+ is the extension of the polar relative positional encoding mechanism [38], where the MFCC layer [5] is added to encode the audio inputs.

Tab. 2 and Tab. 3 presents the performance on AVOS. We exceed VisTR+, URVOS+ and PAM+ with 5.0%, 5.9% and 4.4% in the region similarly. Wnet has an absolute improvement of 5.5%, 5.8% and 6.1% for the contour accuracy. These comparisons mean that the audio-guided video object segmentation is quite different from the text-guided task. There is also a vast difference between the recordings collected from the natural environment and those generated from text-to-speech model. Therefore, it is not suitable to treat audio-guided segmentation as the combination of automatic speech recognition and text-based segmentation.

The visualization of Wnet on the AVOS test dataset is shown in Fig. 4, with each row containing images sampled from the same video. The comparison between Wnet and VisTR+ shows our efficiency in the audio-guided models.

Table 4. The results for different components.

Model	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Base	39.3%	41.6%	40.4%
Base + AFD	41.2%	42.5%	41.8%
Base + AFD + MIM	41.8%	43.0%	42.4%
Base + AFD + MIM + DWT	**42.9%**	**44.0%**	**43.5%**

Note: We use self-attention layers to replace DWT layers in Base, Base+AFD and Base+AFD+MIM.

Table 5. The results for different wavelet basis, mentioned in [42].

Wavelet Basis	Daubechies	Symlets	Coiflets	Meyer
$\mathcal{J}$	**42.9%**	41.9%	41.7%	40.7%
$\mathcal{F}$	**44.0%**	42.6%	42.9%	42.7%

Table 6. The results for DWT. $[a, b]$ means the retained coefficients (value is in interval $[a \cdot max\_value, b \cdot max\_value]$) after filters. For the low pass and high-low pass, we use the hard threshold function. For the high pass, we use the soft threshold function.

Model		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Low Pass	$[0, 0.9]$	41.8%	43.3%	42.5%
	$[0, 0.8]$	42.1%	43.3%	42.7%
	$[0, 0.7]$	40.3%	41.3%	40.6%
High-Low Pass	$[0.008, 0.9]$	42.1%	43.7%	42.9%
High Pass	$[0.01, 1]$	42.8%	43.2%	43.1%
	$[0.008, 1]$	**42.9%**	**44.0%**	**43.5%**
	$[0.006, 1]$	42.3%	43.8%	43.1%

Table 7. The results for threshold function selection. Take high pass $[0.008, 1]$ for example.

Function Selction for High Pass	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Hard Function	42.1%	41.8%	42.0%
Soft Function	**42.9%**	**44.0%**	**43.5%**

Table 8. The results for J selection. For different J, the number of the high frequency coefficient matrix is 3J.

J	Number of Coefficient Matrix	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
1	3 ($\mathbf{C^k_{1,u}}$)	**42.9%**	**44.0%**	**43.5%**
2	6 ($\mathbf{C^k_{1,u}}$; $\mathbf{C^k_{2,u}}$)	41.0%	41.7%	41.4%

Table 9. The comparison of average inference latency among Wnet and audio-text-segmentation model. The evaluation is conducted on a server with 1 NVIDIA 3090Ti GPU, 12 Intel Xeon CPU. The batch size is set to 1.

Method	$\mathcal{J}$	Latency	Speedup
Wnet	42.9%	0.0014s	2.10×
ASR+RVOS	38.4%	0.0032s	1.00×

Wnet can segment small objects from the nature environment, while the VisTR+ performs poorly under this circumstance. Furthermore, the visualization of the DWT process in Fig. 5 shows the denoising performance.

5.4. Ablation Study

In the ablation study, we fine-tune parameters on the validation set. We take the audio-guided RVOS dataset (parts of AVOS dataset) as the example.

About the model components. As shown in Tab. 4, we conduct the experiments to verify the effectiveness of our model design, including the DWT-Based denoising (DWT), mutual information maximum (MIM) and self attention-free decoder (AFD). We use the self-attention layers in the models without the DWT layers. The full model achieves better results than the model (w/o. DWT). It suggests that the DWT layers can filter the noise generated in the audio-video fusion and improve subsequent segmentation results.

About the wavelet basis. Tab. 5 shows the comparison of different wavelet bases. Results verify that the Daubechies wavelet basis is proper for the discrete joint representations.

About the threshold parameter selection. We conduct the experiments under high-pass filters and low-pass filters with different threshold parameter selections. Results in Tab. 6 shows that we can get the best performance under the low-pass filters (0.008) with the soft function.

About the threshold function selection. Two common threshold functions are the hard and the soft function. The hard threshold method can preserve the local features such as the edge of the signal well, while the soft threshold method is relatively smooth. As Tab. 7 shown, we choose the soft function for our model.

About the J selection. Tab. 8 shows the results for the J selection. The performance will be worse when the order is increasing. We select $J = 1$ for our model, with 1 low frequency and 3 high-frequency coefficients.

About the audio-text-segment model. The audio-text-segment model means that we use an ASR model first and then employ the latter referring segmentation model. Tab. 9 shows the results for comparison of Wnet and audio-text-segment model, in terms of speed and quality. We achieve the better performance in these two factors.

6. Conclusion

In this paper, we present the problem of open-ended audio-guided video semantic segmentation, which can be applied in video analysis, video editing, virtual human and so on, from the viewpoint of end-to-end denoising encoder-decoder network learning. We propose the wavelet-based encoder network to learn the cross-modal representations of the video contents with audio-form queries. Then, a self attention-free decoder network is developed to generate the target masks with frequency-domain transforms. In addition, we construct the first large-scale audio-guided video semantic segmentation dataset. The extensive experiments show the effectiveness of our method.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No.61836002, No.62020106007, No.62072150, No.62072397) and Zhejiang Natural Science Foundation (LR19F020006).

References

[1] Emmanuel Azuh, David Harwath, and James R. Glass. Towards bilingual lexicon discovery from visually grounded speech audio. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 276–280. ISCA, 2019. 2

[2] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 15509–15519, 2019. 2

[3] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[5] Lallouani Bouchakour and Mohamed Debyeche. Mfccs and gabor features for improving continuous arabic speech recognition in mobile communication modified. In *Proceedings of the 3rd International Conference on Advanced Aspects of Software Engineering, ICAASE 2018, Constantine, Algeria, December 1-2, 2018*, volume 2326 of *CEUR Workshop Proceedings*, pages 115–121. CEUR-WS.org, 2018. 2, 5, 7

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. 4, 6

[7] Grzegorz Chrupala. Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *CoRR*, abs/2104.13225, 2021. 2, 3

[8] Grzegorz Chrupala, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 613–622. Association for Computational Linguistics, 2017. 2, 3

[9] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees G. M. Snoek. Actor and action video segmentation from a sentence. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5958–5966. IEEE Computer Society, 2018. 2, 3

[10] Lieke Gelderloos and Grzegorz Chrupala. From phonemes to images: levels of representation in a recurrent neural model of visually-grounded language learning. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1309–1319. ACL, 2016. 2, 3

[11] David Harwath and James R. Glass. Towards visually grounded sub-word speech unit discovery. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 3017–3021. IEEE, 2019. 2, 3

[12] David Harwath, Wei-Ning Hsu, and James R. Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2, 3

[13] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James R. Glass. Jointly discovering visual objects and spoken words from raw sensory input. *Int. J. Comput. Vis.*, 128(3):620–641, 2020. 2

[14] David F. Harwath and James R. Glass. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015*, pages 237–244. IEEE, 2015. 2

[15] William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. Word recognition, competition, and activation in a model of visually grounded speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 339–348. Association for Computational Linguistics, 2019. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3

[17] Bertrand Higy, Lieke Gelderloos, Afra Alishahi, and Grzegorz Chrupala. Discrete representations in neural models of spoken language. *CoRR*, abs/2105.05582, 2021. 3

[18] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 108–124. Springer, 2016. 1, 2

[19] Ronghang Hu, Marcus Rohrbach, Subhashini Venugopalan, and Trevor Darrell. Utilizing large scale vision and text datasets for image segmentation from referring expressions. *CoRR*, abs/1608.08305, 2016. 1, 2

[20] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4423–4432. IEEE, 2020. 2

[21] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image

segmentation via cross-modal progressive comprehension. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10485–10494. IEEE, 2020. 2

[22] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 55–65. Association for Computational Linguistics, 2019. 2

[23] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 3192–3199. IEEE Computer Society, 2013. 3

[24] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. *CoRR*, abs/2103.16284, 2021. 2

[25] Kevin Kaergaard, Søren Hjøllund Jensen, and Sadasivan Puthusserypady. A comprehensive performance analysis of EEMD-BLMS and DWT-NN hybrid algorithms for ECG denoising. *Biomed. Signal Process. Control.*, 25:178–187, 2016. 2

[26] Herman Kamper and Michael Roth. Visually grounded cross-lingual keyword spotting in speech. In Shyam S. Agrawal, editor, *6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages, SLTU 2018, 29-31 August 2018, Gurugram, India*, pages 253–257. ISCA, 2018. 2

[27] Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. Visually grounded learning of keyword prediction from untranscribed speech. In Francisco Lacerda, editor, *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 3677–3681. ISCA, 2017. 2

[28] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part IV*, volume 11364 of *Lecture Notes in Computer Science*, pages 123–141. Springer, 2018. 2

[29] Salim Lahmiri and Mounir Boukadoum. Physiological signal denoising with variational mode decomposition and weighted reconstruction after DWT thresholding. In *2015 IEEE International Symposium on Circuits and Systems, ISCAS 2015, Lisbon, Portugal, May 24-27, 2015*, pages 806–809. IEEE, 2015. 2, 3

[30] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. Fnet: Mixing tokens with fourier transforms. *CoRR*, abs/2105.03824, 2021. 2, 3, 4, 6, 7

[31] Ruiyu Li, Kai-Can Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5745–5753. IEEE Computer Society, 2018. 1, 2

[32] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327, 2020. 6

[33] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan L. Yuille. Recurrent multimodal interaction for referring image segmentation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1280–1289. IEEE Computer Society, 2017. 1, 2

[34] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10031–10040. IEEE, 2020. 2

[35] Edgar Margffoy-Tuay, Juan C. Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pages 656–672. Springer, 2018. 1, 2

[36] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 565–571. IEEE Computer Society, 2016. 6

[37] P. P. Mini, Tessamma Thomas, and R. Gopikakumari. EEG based direct speech BCI system using a fusion of SMRT and MFCC/LPCC features with ANN classifier. *Biomed. Signal Process. Control.*, 68:102625, 2021. 4

[38] Ke Ning, Lingxi Xie, Fei Wu, and Qi Tian. Polar relative positional encoding for video-language segmentation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 948–954. ijcai.org, 2020. 2, 6, 7

[39] Yasunori Ohishi, Akisato Kimura, Takahito Kawanishi, Kunio Kashino, David Harwath, and James R. Glass. Trilingual semantic embeddings of visually grounded speech with self-attention mechanisms. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 4352–4356. IEEE, 2020. 2

[40] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 724–732. IEEE Computer Society, 2016. 6

[41] Shuang Qiu, Yao Zhao, Jianbo Jiao, Yunchao Wei, and Shikui Wei. Referring image segmentation by generative adversarial learning. *IEEE Trans. Multim.*, 22(5):1333–1344, 2020. 2

[42] V. J. Rehna and M. K. Jeya Kumar. Wavelet based image coding schemes : A recent survey. *CoRR*, abs/1209.2515, 2012. 8

[43] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 658–666. Computer Vision Foundation / IEEE, 2019. 6

[44] Ramon Sanabria, Austin Waters, and Jason Baldridge. Talk, don't write: A study of direct speech-based image retrieval. *CoRR*, abs/2104.01894, 2021. 1, 2, 3, 4

[45] Gayatri Saripalli, Priyank H. Prajapati, and Anand D. Darji. CSD optimized DWT filter for ECG denoising. In *2020 24th International Symposium on VLSI Design and Test (VDAT), Bhubaneswar, India, July 23-25, 2020*, pages 1–6. IEEE, 2020. 2

[46] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 3465–3469. ISCA, 2019. 2

[47] Sebastiaan Scholten, Danny Merkx, and Odette Scharenborg. Learning to recognise words using visually grounded speech. In *IEEE International Symposium on Circuits and Systems, ISCAS 2021, Daegu, South Korea, May 22-28, 2021*, pages 1–5. IEEE, 2021. 3

[48] Seonguk Seo, Joon-Young Lee, and Bohyung Han. 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*, volume 12360 of *Lecture Notes in Computer Science*, pages 208–223. Springer, 2020. 3, 6, 7

[49] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 38–54. Springer, 2018. 1, 2

[50] Hengcan Shi, Hongliang Li, Qingbo Wu, and King Ngi Ngan. Query reconstruction network for referring expression image segmentation. *IEEE Trans. Multim.*, 23:995–1007, 2021. 1, 2

[51] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6558–6569. Association for Computational Linguistics, 2019. 2, 5, 6, 7

[52] Naveed ur Rehman, Ubaid ur Rehman, Syed Zain Abbas, Anum Asif, and Anum Javed. Translation invariant DWT based denoising using goodness of fit test. In *IEEE Statistical Signal Processing Workshop, SSP 2016, Palma de Mallorca, Spain, June 26-29, 2016*, pages 1–5. IEEE, 2016. 2

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 4, 5

[54] Hao Wang, Cheng Deng, Fan Ma, and Yi Yang. Context modulated dynamic networks for actor and action video segmentation with language queries. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12152–12159. AAAI Press, 2020. 2

[55] Hao Wang, Cheng Deng, Junchi Yan, and Dacheng Tao. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3938–3947. IEEE, 2019. 2

[56] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8741–8750. Computer Vision Foundation / IEEE, 2021. 4, 6, 7

[57] Yuxin Wu and Kaiming He. Group normalization. *Int. J. Comput. Vis.*, 128(3):742–755, 2020. 4, 7

[58] Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J. Corso. Can humans fly? action understanding with multiple classes of actors. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2264–2273. IEEE Computer Society, 2015. 3

[59] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5187–5196. IEEE, 2019. 7

[60] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *CoRR*, abs/2106.01883, 2021. 6

[61] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10502–10511. Computer Vision Foundation / IEEE, 2019. 1, 2

[62] Linwei Ye, Mrigank Rochan, Zhi Liu, Xiaoqin Zhang, and Yang Wang. Referring segmentation in images and videos with cross-modal self-attention network. *CoRR*, abs/2102.04762, 2021. 2

[63] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension.

In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1307–1315. IEEE Computer Society, 2018. 1, 2