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

Wenwen Pan1,†, Haonan Shi1,†, Zhou Zhao1,*, Jieming Zhu2, Xiuqiang He2, Zhigeng Pan3, Lianli Gao4, Jun Yu5, Fei Wu1,6, Qi Tian7,
1Zhejiang University, 2Huawei Noah’s Ark Lab, 3Hangzhou Normal University, 4The University of Electronic Science and Technology of China, 5Hangzhou Dianzi University, 6Shanghai Institute for Advanced Study of Zhejiang University, 7Huawei 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 method1.

1Equal contribution.
2Corresponding Author.
3Code is available at: https://github.com/asudahkzj/Wnet.git

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
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
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 \(\mathbf{v} = \{v_i\}_{i=1}^n\), where \(v_i\) 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 \(\mathbf{q} = \{q_j\}_{j=1}^m\) where \(q_j\) 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 \(\mathbf{S} = \{S_i \in \{0, 1\}^{W_v \times H_v}\}_{i=1}^n\).

### 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.

\[
W(\tau) = \int_{-\infty}^{\infty} f(t)g(t - \tau)dt.  \quad (1)
\]

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.

\[
W(a, \tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi(t - \tau \frac{t}{a})dt,  \quad (2)
\]

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 \(\psi(t\frac{t - \tau}{a})\). 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:

\[
W_z = W_f + W_e,  \quad (3)
\]

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.
To include spatial information of the visual feature, we augment 3-dimensional spatial coordinates following [56], denoted as $v = \{v_i\}_{i=1}^n$. The output of the backbone is $B \in \mathbb{R}^{n \times c \times H \times W}$, $c$ represents the original video dimension. A 1x1 convolution is used to reduce the dimension to $\mathbb{R}^{n \times d \times H \times W}$. 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 $q = \{q_i\}_{i=1}^m$.

**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 $O$, backbone features $B$ and encoded feature maps $E$ from the previous layers, shown in Fig. 2. First, we employ an attention module to calculate the similarity map between $O$ and $E$. Following [56], we only compute the features of its corresponding frame. Next, we fuse the similarity map, $B$ and $E$ of the corresponding frames, following the DETR [6]. $B \in \mathbb{R}^{n \times c \times H \times W}$, $E \in \mathbb{R}^{4 \times n \times (H \times W)^3}$, $O \in \mathbb{R}^{n \times d}$, 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
two potentially non-aligned sequences from each of them, denoted as $v^p \in \mathbb{R}^{T_p \times d_v}$ and $q^p \in \mathbb{R}^{T_q \times d_q}$. $T_i$ represents the sequence length (audio or video), and $d_i$ represents the dimension, respectively. Inspired by the multi-modal transformer in MuLiT [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 $Q = v^p W_Q$, keys $K = q^p W_K$ and values $V = q^v W_V$. The cross-modal attention is calculated by

\[
\text{Attention}_{q \rightarrow v}(Q, K, V) = \text{Softmax}\left(\frac{Q^\top K}{\sqrt{d_q}}\right)V^V,
\]

where $d_q$ 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 $f \in \mathbb{R}^{T \times d_v}$.

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 $f \in \mathbb{R}^{T \times d_v}$ serve as the input signal. The DWT separates filtering operations on rows and columns. $A_{j,u}$ and $C_{j,u}^k$ denote scaling and wavelet coefficients at scale $j$ for the given signal $f$ 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 $A_{j,u}$ and $C_{j,u}^k$ as follows.

\[
\begin{align*}
A_{j,u} &= \sqrt{2} \sum_u h(l - 2u)A_{j+1,u}; \\
C_{j,u}^1 &= \sqrt{2} \sum_u g(l - 2u)A_{j+1,u}; \\
C_{j,u}^2 &= \sqrt{2} \sum_u gh(l - 2u)A_{j+1,u}; \\
C_{j,u}^3 &= \sqrt{2} \sum_u gg(l - 2u)A_{j+1,u}.
\end{align*}
\]

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 $n \times d_v$ cross-modal signal is transformed into four $\frac{n}{2} \times \frac{d_v}{2}$ signal after the two stages.

Next, we perform threshold quantization on the high-frequency coefficients $C_{j,u}^k$ 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 $\alpha$ with a soft threshold function. For $\hat{\alpha} = \max \{C_{j,u}^k\}$, the filter operations can be represented as follows ($k \in [1, 2, 3]$).

\[
C_{j,u}^k[x, y] = \text{sgn}(C_{j,u}^k[x, y])(|C_{j,u}^k[x, y]| - \alpha \hat{\alpha})_+
\]

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.

\[
f_n = \text{DWTInverse}(A_{j,u}, C_{j,u}^k),
\]
where \( f_n \) is the denoised joint representation. The \( f_n \) serves as the input of the following feed-forward layers and layer norm. Finally, we obtain the encoded representation \( E \).

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, \( F_{\text{seq}} \), and one 1D DFT is along the hidden dimension \( F_{\text{hidden}} \).

\[
y = \mathbb{R}(F_{\text{seq}}(\mathbb{R}(F_{\text{hidden}}(x))))
\]

We only keep the real part of the result. After the Fourier layers, we employ the multi-head attention layers. The object prediction \( O \) 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.

\[
\mathcal{L} = \lambda_1 \mathcal{L}_{\text{mask}} + \lambda_2 \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mutual}},
\]

where \( \lambda_1, \lambda_2 \) 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:

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

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

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

We use KL divergence [60] to maximize the mutual information between the cross-modal representation \( f \) and the encoded representation \( E \).

\[
\mathcal{L}_{\text{mutual}}(E(i,j)||f(i,j)) = \sum \mathbb{E}(i,j)(\log \frac{E(i,j)}{f(i,j)}),
\]

where \( i \) stands for the sequence and \( j \) stands for the dimension. \( E \) and \( f \) 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.

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 \( v \) and audio query \( q \) with the ground-truth masks \( G \), we denote the generated masks from our Wnet method by \( S \). We employ the Jaccard index \( J \) defined as the intersection-over-union of the generated segmentation and the ground-truth mask (\( J = \frac{|S \cap G|}{|S \cup G|} \)). 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 (\( F = \frac{2P_cR_c}{P_c+R_c} \)).

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.
A fish in distress with its head out of the water swimming.

A brown turtle is swimming up in the water.

An atv being ridden down a path.

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.
For the high pass, we use the soft threshold function. For the low pass and high-low pass, we use the hard threshold function. 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


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


