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

In the supplementary material, we provide here more details of our model architecture, more experiment results of the ablation study, and show more visual results of our segmentation model. As to the details of our network architecture, we have opened source the code and dataset (AVOS). Our code is available at: https://github.com/asudahkzj/Wnet.git. Our dataset is available at: https://drive.google.com/drive/folders/Audio-Guide-Segmentation. The allocations of training sets, test sets, and verification sets is also provided (.json).

1. About the Wavelet Basis

Wavelet transform is different from Fourier transform, and the results of wavelet transform are not the same according to the different wavelet generating function. Tab. 1 shows the comparison of different wavelet bases. Results verify that the Daubechies wavelet basis is proper for the discrete joint representations. Because of the page limit in the main body, we list other wavelet bases in this part.

Table 1. The results for different wavelet basis, mentioned in [2].

Wavelet Basis	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Daubechies	42.9%	44.0%	43.5%
Symlets	41.9%	42.6%	42.3%
Meyer	40.7%	42.7%	41.7%
Coiflets	41.7%	42.9%	42.3%
Biorthogonals	42.1%	43.1%	42.6%
Haar	41.8%	42.6%	42.2%

2. About the Loss Weight

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

$$\mathcal{L} = \lambda_{mutual}\mathcal{L}_{mutual} + \lambda_{box}\mathcal{L}_{box} + \mathcal{L}_{mask}, \quad (1)$$

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 and Focal loss. $\mathcal{L}_{box}$ scores the bounding boxes. We use a linear combination of the sequence level L1 loss and the generalized IOU loss. We use KL divergence to maximize the mutual information between the cross-modal representation $\mathbf{f}$ and the encoded representation $\mathbf{E}$. We can obtain the best performance when $\lambda_{mutual} = 500$, $\lambda_{box} = 7$.

Table 2. The results for different λ_{mutual}.

λ_{mutual}	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
50	41.7%	41.0%	41.4%
250	42.9%	44.0%	43.5%
500	39.1%	40.2%	39.7%

Table 3. The results for different λ_{box}.

λ_{box}	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
1.75	41.3%	42.6%	42.0%
3.5	42.9%	44.0%	43.5%
5.25	42.7 %	43.6%	43.2%
7	41.9%	43.5%	42.7%

3. About the Vanishing Moment

We conduct a series of experiments on the vanishing moment of the selected DWT. The order of vanishing moment is a concept often used in wavelet transform. In practice, the basic wavelet is not only required to satisfy the admissible condition, but also imposed on the so-called vanishing moment condition, so that as many wavelet coefficients as possible are zero or as few non-zero wavelet coefficients are generated, which is beneficial to data compression and noise elimination.

The larger the vanishing moment is, the more wavelet coefficients with a value of 0 will be generated during the wavelet decomposition, which makes the signal decomposition more sparse. But at the same time, a larger vanishing moment will also produce a larger support interval, which is a trade-off relationship.

Table 4. The results for different vanishing moment.

Vanishing Moment	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
db1	41.5%	42.9%	42.2%
db2	42.9%	44.0%	43.5%
db3	41.0%	41.7%	41.4%

4. About the Other Evaluation Standard

We also use other evaluation standard to test our model. We measure prec@X, the percentage of correctly segmented frames in the whole dataset, given a predefined threshold X sampled from the range [0.5, 0.9]. Note that segmentation in a frame is regarded as successful if its $\mathcal{J}$ score is higher than a threshold.

Table 5. The results of prec@X. We use self-attention layers to replace DWT layers in Wnet (w/o. DWT).

Models	p0.5	p0.6	p0.7	p0.8	p0.9
Wnet (w/o. DWT)	42.6%	34.9%	27.4%	17.6%	7.1%
Wnet	43.4%	36.2%	28.0%	18.9%	7.4%

For the DWT selection, we conduct abundant experiments to figure out what the appropriate threshold is. Also, we test the AP to verify the performance of our method.

Table 6. The results of AP. $[a, b]$ 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		AP_{25}	AP_{50}	AP_{75}
Low Pass	$[0, 0.9]$	64.8%	35.5%	15.2%
	$[0, 0.8]$	64.6%	35.8%	18.0%
	$[0, 0.7]$	61.7%	32.6%	11.7%
High-Low Pass	$[0.008, 0.09]$	62.6%	34.0%	13.6%
High Pass	$[0.006, 1]$	63.4%	35.5%	14.1%
	$[0.004, 1]$	61.8%	32.6%	12.5%
	$[0.002, 1]$	61.2%	32.1%	12.1%

5. About the Positional Encoding

Position information is important for the dense prediction problem. As the original feature sequence contains no positional information, we supplement with the spatial and temporal positional encodings, which indicate the relative positions in the video sequence. The ordered format of the sequence supervision and the correspondence between the input and output order of the transformer provide some relative positional information implicitly. Experiments of models with and without positional encoding verify the necessity of explicit positional encoding.

Table 7. The results for model with or without PE.

Models	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Wnet (w/o. PE)	41.8%	43.2%	42.5%
Wnet	42.9%	44.0%	43.5%

6. About the Length

Tab. 8 shows that Wnet can obtain good performance when the video sequence is long. The reason of the performance is that the number of the video (length 20-30) is rather small in dataset.

Table 8. The performance for different video sequence length.

Length	$\mathcal{J}$	prec@0.5	prec@0.7	prec@0.9
10	28.1%	29.5%	14.6%	3.6%
20	41.9%	43.4%	28.3%	8.3%
30	40.5%	41.4%	23.2%	5.0%
40	44.9%	45.2%	31.1%	8.4%

7. About the DCT

The difference between DWT and DCT lies in that the wavelet domain of the image is divided into four sub-bands after DWT transformation, and each sub-band includes not only the frequency domain component of the image but also its spatial component. And its upper left subband (LL subband) containing the main information of the image can continue DWT transformation again, so as to continuously decompose it into many signal components of different resolutions. We test the DCT-based encoder under the same condition of Wnet, and the J-score is 41.6% and 42.0% for soft and hard threshold, which also verifies the effectiveness of our method.

8. About the KL divergence

We conduct experiments under the L1, L2 and cosine distance. The J-score is 41.2%, 40.7% and 41.7%, respectively.

9. Model Details

There are details for the whole model and training agent. We list some parameters in the Tab. 9.

We adopt a 2-layer, 8-head multi-head cross-attention [3] module with the width of 3 to fuse visual and audio features. For the transformer, we use 4 encoder, 4 decoder layers of 384 width with 8 attention heads. Between the attention layer and the feed-forward layer, a wavelet transform filter layer is used to remove noise from joint representations. We employ db2 wavelet basis and 1-level decomposition. The threshold is set to 0.008 for high-pass filters with the soft threshold function. For the transformer decoder, we

Table 9. Model Details.

Paramter	Value
Encoder Layers	4
Decoder Layers	4
Hidden Dim	384
Feed-Forward Dim	2048
Frame Number	36
Query Number	36
Learning Rate (Backbone)	1e-5
Learning Rate	1e-4
Optimizer	AdamW
Weight Decay	1e-4
Clip Max Norm	0.1
α	0.008
batch size	1
λ_{mutual}	250
λ_{box}	3.5

use Fourier transform [1] instead of the self-attention layer. After obtaining the prediction of the decoder and the encoder, for each corresponding frame, we send them to a self-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 [4]. 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 [5] 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. The batch size is set to 1. The model is trained using Adam optimizer. The backbone has a learning rate of 1e-5. The number of parameters is 42,974,583 (Wnet) and 47,702,391 (the model with self-attention).

10. Visualization Results

Fig. 1 and Fig. 2 show some results of our Wnet. Therefore, there is some challenging instances in Fig. 2. The video and the sentence are more complex, which contain more than one object. The red is the target object, and the blue is the distraction.

References

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

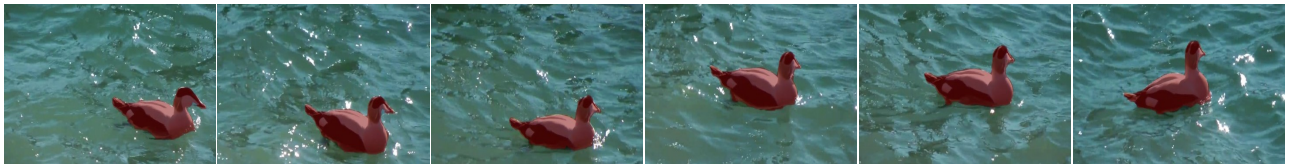
[2] V. J. Rehna and M. K. Jeya Kumar. Wavelet based image coding schemes : A recent survey. *CoRR*, abs/1209.2515, 2012. 1

[3] 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

[4] 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. 3

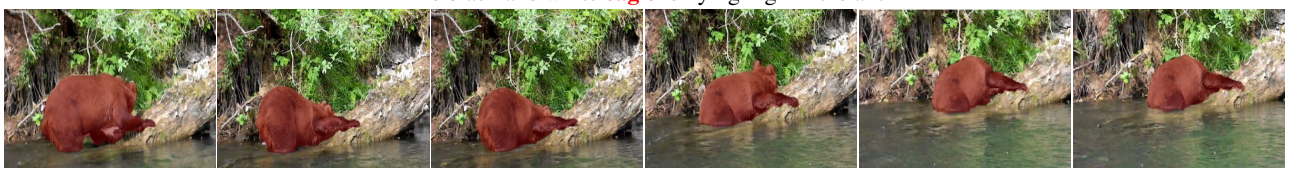
[5] Yuxin Wu and Kaiming He. Group normalization. *Int. J. Comput. Vis.*, 128(3):742–755, 2020. 3

Wnet (Ours)



A **duck** is swimming in the water.

The black and white **eagle** is flying high in the air.



A brown **bear** partially submerged in water attacking a small animal.

A **person** knee boarding behind a boat.

A sliver ram **car** in a parking lot with a sale price on it.

A small **monkey** is standing on the green grass looking to the left.

A **lizard** on a white towel.

A dark orange puffy **fish** is floating in a bowl of water.

Figure 1. Visualization of Wnet on the AVOS.

Wnet (Ours)

A **person** skateboarding on a **ramp**.

A **skateboard** ridden on **benches**.

A **person** has landed on the ground after **parachuting**.

A **person** grabbing a **crocodile**.

A white **cat** sitting in front of a flatscreen **tv**.

The **person** controlling the **ducks** with a **stick** in his right hand.

A **giraffe** eating **leaves** off of a tree.

A green **parrot** standing on a tree stump with its face overlooking a **stump**.

Figure 2. Visualization of Wnet for challenging prediction on the AVOS.