Supplementary of "FastForensics: Efficient Two-Stream Design for Real-Time Image Manipulation Detection"

BMVC 2024 Submission # 339

1 Analysis on Loss Term Weights of λ_1, λ_2 and λ_3

We analyze these loss term weights on the NIST16 dataset using different combinations. The results are shown in Table 1, exhibiting that our method performs stably along with the variations of weights. This demonstrates our method is not sensitive to the loss term weights.

Table 1: Effect of $\lambda_1, \lambda_2, \lambda_3$.

$\lambda_1, \lambda_2, \lambda_3$	AUC↑	F_1 ↑
1,0.5,5	98.9	85.8
1,0.5,10	98.8	85.2
1,1,1	98.6	85.3
1,1,5	98.7	86.0
1,2,5	98.9	86.4
1,2,10	98.8	85.1
1,1,10 (Ours)	98.9	86.5

2 More Explanation of Why EWTB Is Efficient and Effective?

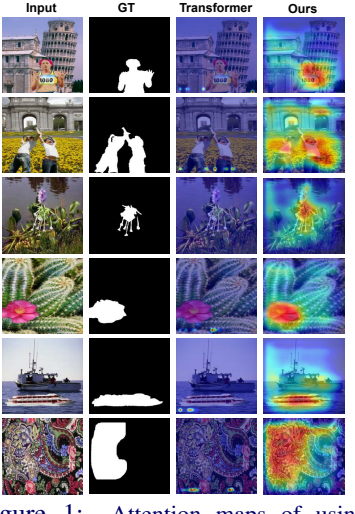

Figure 1: Attention maps of using vanilla Transformer block and ours using GradCAM.

In vanilla Transformer blocks, Self-Attention is the most time-consuming operation due to the multiplication of query, key, and value features. To improve the computational efficiency, we follow [1] to separate the input features of the Transformer block equally into several pieces and perform self-attention inside each piece. Then we reduce the dimension of query, key, and value features to further save the cost. Moreover, in the cognitive branch, we only employ four blocks for a good balance between performance and efficiency. For a fair comparison, we adapt vanilla Transformer blocks into our architecture by only substituting our EWTB in the cognitive branch, and maintaining other settings as the same. Due to the nature of the vanilla Transformer block, its output channel dimension is 768, different from $128, 256, 384$ in ours. The Flops and Parameters comparison is shown in (*the body of the paper Table 4*). It can be seen that our architecture is three times less than using conventional Transformer blocks.

The comparison in performance on CASIA, NIST16, and COVERAGE is also shown in (*the body of the paper Table 4*). We can observe that our method surprisingly outperforms

the vanilla version by a large margin in F1 score, 61.9% on average. This may be due to the following reasons. First, without a large number of training samples (*e.g.*, ImageNet), the vanilla Transformer can hardly be well-trained compared to a compact version such as EWTB in our method. Second, the inspective branch in our method is also a lightweight architecture. Thus we design EWTB to cooperate with this branch for nearly equal knowledge sharing. However, the large number of parameters in the vanilla Transformer may enable the model to over-focus on cognitive perspective, overlooking the importance of capturing the texture-aware traces. For further verification, we visualize the attention maps using the last layer of the detection head. As shown in Fig. 1, the attention of vanilla Transformer blocks disperses over the entire image, having no salient highlights on manipulated regions.

3 Why Using DWT in Self-attention?

Discrete Wavelet Transform (DWT) can decompose the feature map into four wavelet subbands, which are the combination of the low-level component of the basic image structure, and the high-frequency component with texture details and the size of these four wavelet subbands are $1/2$ of feature map. By combining these four subbands, the feature map can be recovered with information loss.

This process has two advantages:

1. Downsampling of QKV is a general way to reduce the computional complexity in Self-attention. However, downsampling may lead to information loss. Benefits from the nature of DWT, we can achieve the downsampling and recover the feature map without any loss. In contrast, other general frequency transformation operations such as DCT do not downsample the feature map.
2. DWT can capture traces from different perspectives, which extract not only the global frequency information, but also the fine-grained levels of frequency information. In contrast, other general frequency transformation operations such as DCT only target for global frequency information.

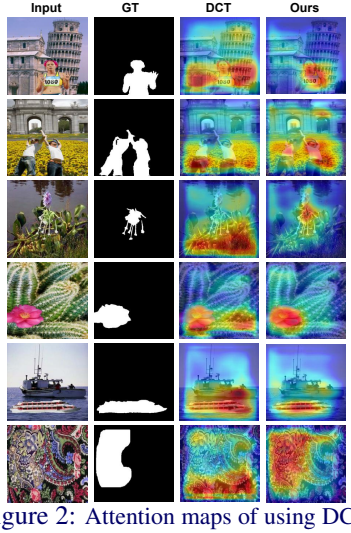

Figure 2: Attention maps of using DCT and DWT (ours) using GradCAM.

For better demonstration, we visualize the attention maps using DCT and DWT in Fig 2. It can be seen that using DWT in our method can more precisely locate the manipulated regions.

4 More Study of Shared Global Q and Support V

Effect of Shared Global Q. For further analysis, we show the attention maps without using shared global Q (SQ) and ours using GradCAM in Fig 3. It can be seen that using shared global Q can provide more global information, whereas the attention is scattered without using shared global Q.

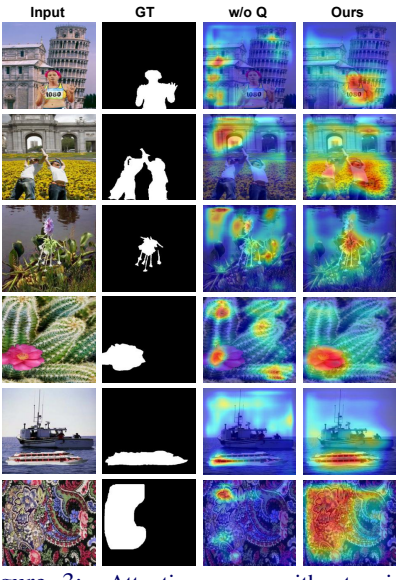

Figure 3: Attention maps without using shared global Q (SQ) and ours using GradCAM.

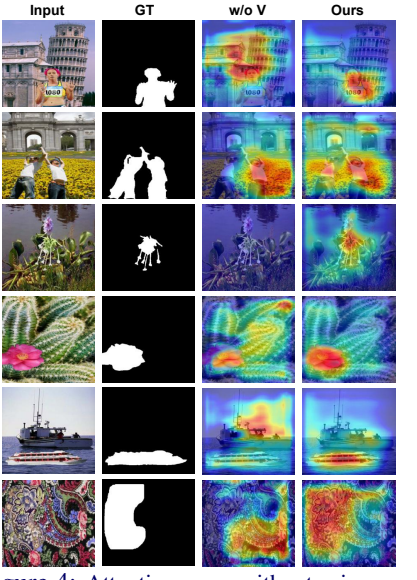

Figure 4: Attention maps without using support V and ours using GradCAM.

Effect of Support V. Fig 4 shows the attention maps without using support V and ours using GradCAM. It can also be seen that the lack of support V leads to the attention dispersion to incorrect regions, representing the ineffectiveness of spotting manipulated regions.

5 Various Model Architectures

To explore the effect of model size on the final results, we adjusted the parameter settings for each stage and observed the effect of different module parameter sizes on the final results.See Table 2 for detailed experiments In the slim and large models, the number of conv in the basic residual network in the stem stage and in the inspective branch is 1 and 2, respectively. in the cognitive branch of slim, the dimension of the EWTB module is reduced to [64, 128 , 224], while in the large cognitive branch, the number of EWTBs is raised to [1, 2, 3], and the dimensionality setting is kept the same as that of base. the specific settings of the base model can be found in (*the body of the paper Sec 3*). From the experimental results, it can be seen that when the model pays too much attention to the cognitive branch, it may lead to losing part of the fine-grained information, which negatively affects the final result of the model. Similarly, if the cognitive branch is too lightweight, the model may lack the necessary guidance information for tamper detection.

Table 2: Configuration of different architectures and their performance.

Architectures	Stem	Stage 1	Stage 2	Stage 3	$F_1 \uparrow$	AUC↑	Params	FLOPs
Ours-Slim	#Conv: 1 #Channel: 128	#Block: 1 #Channel: 64	#Block: 1 #Channel: 128	#Block: 1 #Channel: 224	80.4	98.8	2.92M	1.12G
Ours-Base	#Conv: 2 #Channel: 128	#Block: 1 #Channel: 128	#Block: 1 #Channel: 256	#Block: 1 #Channel: 384	**86.5**	**98.9**	8.37M	2.16G
Ours-Large	#Conv: 2 #Channel: 128	#Block: 1 #Channel: 128	#Block: 2 #Channel: 256	#Block: 3 #Channel: 384	83.1	98.7	14.44M	4.63G

6 Various Feature Fusing Strategies

In delving into the impact of feature fusion strategies, we design a series of experiments. Specifically, "Attn" stands for fusing the feature from the cognitive branch to the inspective branch using an attention manner. "Mult" denotes performing the multiplication instead of adding operations. "Concat" denotes concatenating these features. Experimental results show that the adding strategy is the most effective in our method in F1 score. We conjecture that the element-by-element summation enables the model to more fully utilize the feature information of both branches.See Table 3 for detailed results.

Table 3: Effect of various feature fusing strategies.

Setting	$F_1 \uparrow$	AUC↑
Attn	79.6	98.5
Mult	82.6	**99.1**
Concat	81.5	98.7
Ours	**86.5**	98.9

7 Positions of using WaveLet

Note that we employ four heads in each EWTB. Thus We analyze which head should use Wavelet-guided Self-Attention. From Table 4, it can be observed that the best results in the F1 score are achieved by applying Wavelet-guided Self-Attention at the positions of head 2 and head 4. This may be because a proper combination of frequency information with color information enables the model to effectively capture the manipulation traces.

Table 4: Positions of using Wavelet in EWTB.

Pos	$F_1 \uparrow$	AUC↑
(1,3)	81.3	98.4
(2,4)	**86.5**	98.9
(1,2,3,4)	81.5	**99.1**

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.