Dynamic Difference Learning With Spatio–Temporal Correlation for Deepfake Video Detection

23年 TIFS（开源）

1. 目标：视频，伪造检测

2. 核心思想：动态定位由视频伪造方法引起的帧间不一致性，并建模连续帧中的时空不一致性，而不受面部运动的干扰。

3. 方法：Xception骨干网络+两个模块（即插即用）：

   ①动态细粒度差异捕获模块（DFDC）

   用于精确挖掘假视频中的空间信息，包括局部运动不一致和人脸纹理不一致。

   ②多尺度时空聚合模块（MSA）

   多尺度池化操作融合了长距离和短距离时间信息，以增强时空不一致性。

4. 网络分为两个阶段：

   ①下采样阶段； ②特征提取阶段。

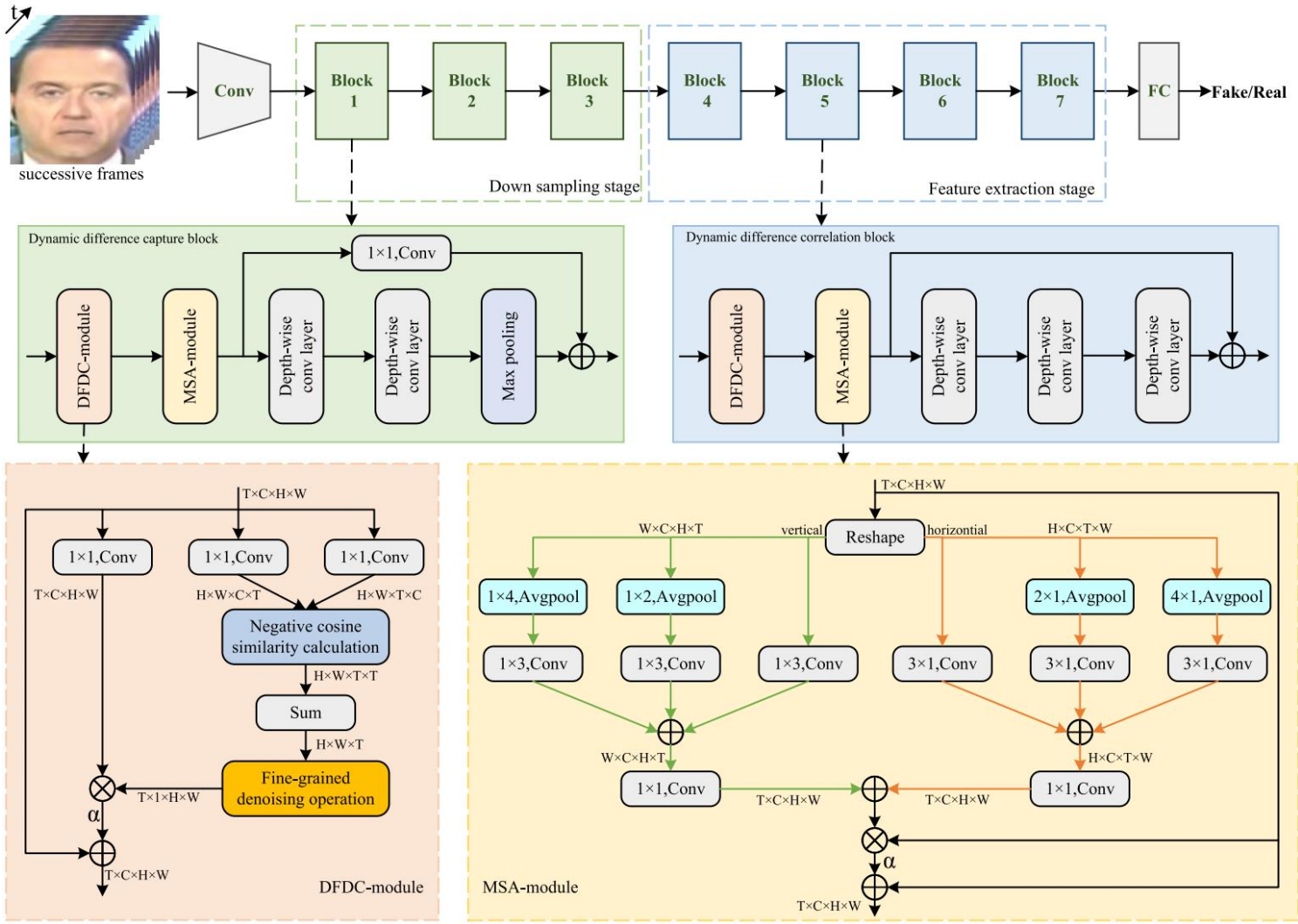
5. DFDC 模块和 MSA 模块分别完成：

   ①在空间域中跨不同帧捕获帧间差异；

   ②在时间域中跨不同帧聚合帧间差异的任务。

   这两个模块以串联方式添加到每个网络阶段。

6. 密集采样策略，视频分为n个片段，以T帧为单位



Dynamic Difference Learning With Spatio–Temporal Correlation for Deepfake Video Detection

23年 TIFS（开源）

7. 动态细粒度差异捕获模块（DFDC）：

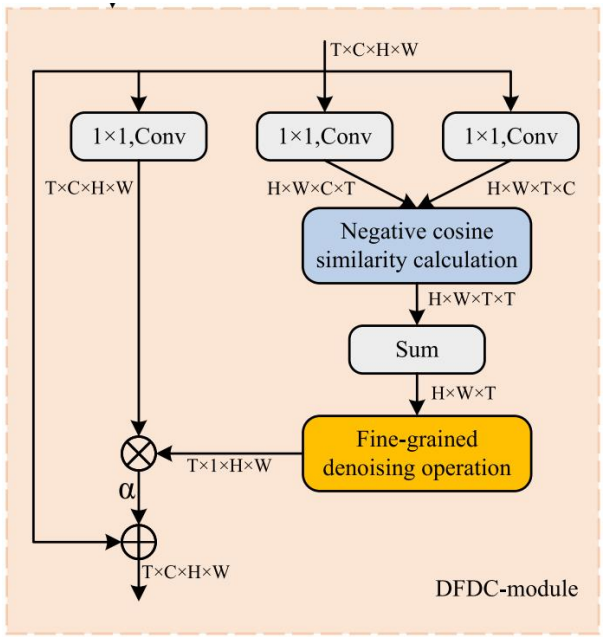
①HxWxCxT，HxW看作空间信息，CxT即以帧T为单位去观察通道C；

②负余弦相似度的计算 $\operatorname{sim}\left(x_{h,w,i}, x_{h,w,j}\right) = -\frac{x_{h,w,i}}{\left|x_{h,w,i}\right|_2} \cdot \frac{x_{h,w,j}}{\left|x_{h,w,j}\right|_2}$

③对每个块对应的T个负余弦值求和，生成全局不一致性图$M^{\wedge}$

④面部运动引起大量帧间差异信息，提出一种**细粒度去噪操作**：

- 核心思想：伪造差异的强度波动大，而运动差异的强度波动小
- 帧间差异强度波动图$M^{\sim}$通过计算中每个位置的T个分数的方差来获得
- 对方差排序，使k=H作为阈值，小于阈值的$M^{\wedge}$相应位置更新为-T



Algorithm 1 DFDC-Module

Input: input feature $\mathbf{F}$
Output: enhanced feature $\widehat{\mathbf{F}}$
$\mathbf{F}_q = \sigma_q(\mathbf{F})$;
$\mathbf{F}_k = \sigma_k(\mathbf{F})$;
$\mathbf{F}_v = \sigma_v(\mathbf{F})$;
for $x_{h,w,i} \in \mathbf{F}_q$; $x_{h,w,j} \in \mathbf{F}_k$ **do**
  Calculate $\mathbf{M}$ by Eq. (3)
end
for $h = 1 : H$; $w = 1 : W$ **do**
  for $i = 1 : T$ **do**
    for $j = 1 : T$ **do**
      $\widehat{\mathbf{M}}_{h,w,i} \leftarrow$ sum all score $\mathbf{M}_{h,w,i,j}$
    end
    $\widetilde{\mathbf{M}}_{h,w} \leftarrow$ var all score $\widehat{\mathbf{M}}_{h,w,i}$
  end
end
Sort $\widetilde{\mathbf{M}}$ and pick k-th value as threshold $\mathbb{T}$;
for $h = 1 : H$; $w = 1 : W$ **do**
  if $\widetilde{\mathbf{M}}_{h,w} \leq \mathbb{T}$ **then**
    for $i = 1 : T$ **do**
      $\widehat{\mathbf{M}}_{h,w,i} = -T$
    end
  end
end
Return:$\widehat{\mathbf{F}} \leftarrow \mathbf{F} + \alpha \cdot \mathbf{F}_v \cdot \widehat{\mathbf{M}}$

8. 多尺度时空聚合模块（MSA）：

①以对连续帧进行双向切片，应用条带池化操作和卷积来压缩时间维度

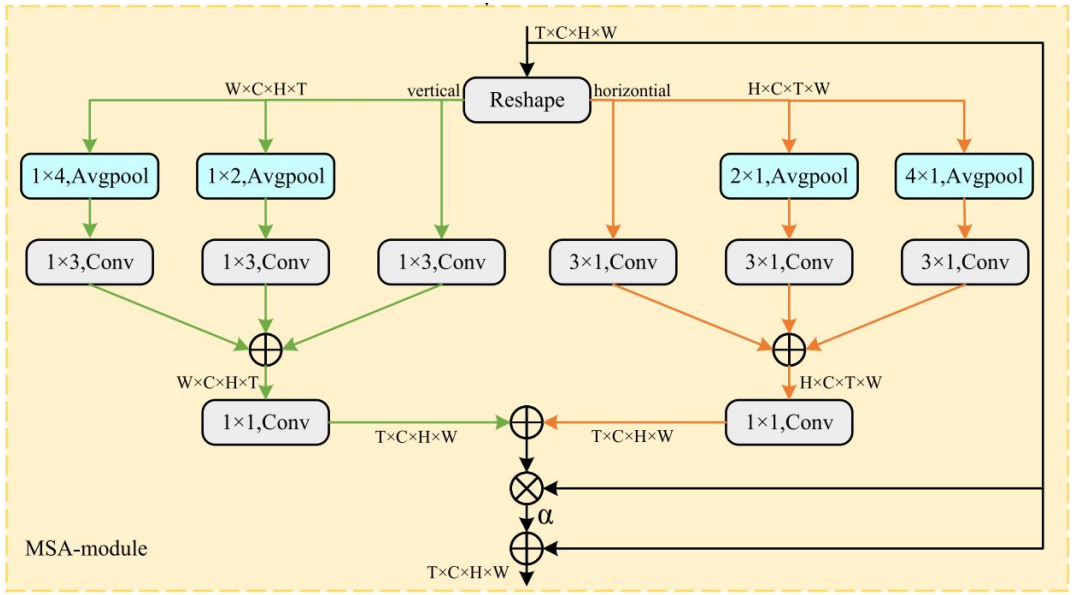
②过程可定义为：

$$\mathbf{D}^V = \mathrm{Conv}_{1\times1}\Big(\frac{1}{3}\cdot(\mathrm{Sconv}_{1\times3}(\widehat{\mathbf{F}}^V)) + \frac{1}{3}\cdot(\mathrm{Sconv}_{1\times3}(\mathrm{Sap}_{1\times2}(\widehat{\mathbf{F}}^V))) + \frac{1}{3}\cdot(\mathrm{Sconv}_{1\times3}(\mathrm{Sap}_{1\times4}(\widehat{\mathbf{F}}^V)))\Big),$$

$$\mathbf{D}^H = \mathrm{Conv}_{1\times1}\Big(\frac{1}{3}\cdot(\mathrm{Sconv}_{3\times1}(\widehat{\mathbf{F}}^H)) + \frac{1}{3}\cdot(\mathrm{Sconv}_{3\times1}(\mathrm{Sap}_{2\times1}(\widehat{\mathbf{F}}^H))) + \frac{1}{3}\cdot(\mathrm{Sconv}_{3\times1}(\mathrm{Sap}_{4\times1}(\widehat{\mathbf{F}}^H)))\Big),$$

③统一的时空不一致性可以建模为：

$$\widetilde{\mathbf{F}} = \widehat{\mathbf{F}} + \alpha\cdot\widehat{\mathbf{F}}\cdot\Big(\frac{1}{2}\cdot(\mathrm{sigmoid}(\mathbf{D}^V) - \frac{1}{2}) + \frac{1}{2}\cdot(\mathrm{sigmoid}(\mathbf{D}^H) - \frac{1}{2})\Big),$$



Dynamic Difference Learning With Spatio–Temporal Correlation for Deepfake Video Detection

23年 TIFS（开源）

Methods	Frame-level						Video-level					
	Celeb-DF			DFDC			Celeb-DF			DFDC		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
ShallowNet [22]	0.8303	0.9056	0.8792	0.7931	0.8577	0.8403	0.8552	0.8281	0.8987	0.8047	0.8042	0.8493
Mesonet [62]	0.9464	0.9852	0.9595	0.8621	0.9261	0.8946	0.9768	0.9774	0.9826	0.8750	0.8749	0.9047
Xception [25]	0.9780	0.9981	0.9833	0.8891	0.9601	0.9144	0.9923	0.9901	0.9941	0.8984	0.8978	0.9226
F3Net [27]	0.9876	0.9939	0.9904	**0.9245**	0.9778	**0.9421**	0.9942	0.9916	0.9956	0.9375	0.9245	0.9526
Multi-att [29]	0.9863	0.9952	0.9894	0.9102	0.9753	0.9394	0.9903	0.9887	0.9927	0.9219	0.9396	0.9535
GFF [32]	0.9859	0.9956	0.9874	0.9233	0.9651	0.9407	0.9865	0.9901	0.9912	0.9219	0.9223	0.9408
CNN+LSTM [13]	0.8620	0.9337	0.8953	0.8509	0.9122	0.8827	0.8994	0.9546	0.9248	0.8789	0.9129	0.9057
DBiRNN [14]	0.9629	0.9948	0.9718	0.8780	0.9470	0.9007	0.9788	0.9942	0.984	0.8828	0.9483	0.9056
CLRNet [39]	0.8951	0.9536	0.9220	0.8208	0.9094	0.8498	0.9402	0.9671	0.9564	0.8281	0.9182	0.8534
ConvLSTM [49]	0.9610	0.9902	0.9701	0.8785	0.9474	0.9010	0.9729	0.9892	0.9796	0.8633	0.9543	0.8895
DIL [36]	0.7705	0.8279	0.8296	0.7319	0.8006	0.7861	0.8085	0.8349	0.8611	0.7070	0.8011	0.7648
Ours	**0.9907**	**0.9998**	**0.9933**	0.9217	**0.9797**	0.9407	**0.9961**	**0.9970**	**0.9985**	**0.9414**	**0.9586**	**0.9589**

Methods	Video-level					
	FaceForensics++		Celeb-DF		DFDC	
	ACC	AUC	ACC	AUC	ACC	AUC
ResNet18	0.9536	0.9594	0.9884	0.9872	0.8594	0.8575
ResNet18+DFDC	0.9679	0.9873	0.9923	0.9967	0.8906	0.9248
ResNet18+MSA	0.9643	0.9788	0.9891	0.9955	0.8783	0.9457
ResNet18+DFDC+MSA	0.9714	0.9893	**0.9981**	**0.9982**	0.8945	0.9387
Xception	0.9679	0.9764	0.9923	0.9901	0.8984	0.8978
Xception+DFDC	0.9821	0.9927	0.9942	0.9962	0.9336	0.9480
Xception+MSA	0.9786	0.9815	0.9942	0.9952	0.9142	0.9223
Xception+DFDC+MSA	**0.9893**	**0.9929**	0.9961	0.9970	**0.9414**	**0.9586**

9. 在 Celeb - DF 和 DFDC 上的比较结果（上表）

10. 消融实验（下左表+下右表）

DFDC	MSA	DF	F2F	FS	NT
		0.9458	0.8935	0.9242	0.7435
✓		0.9668	0.9193	0.9506	0.7828
	✓	0.9673	0.9136	0.9453	0.7724
✓	✓	**0.9743**	**0.9228**	**0.9601**	**0.7935**