t-SNE for temporal feature

2022.06.27

최종욱

1. Problems

Problems

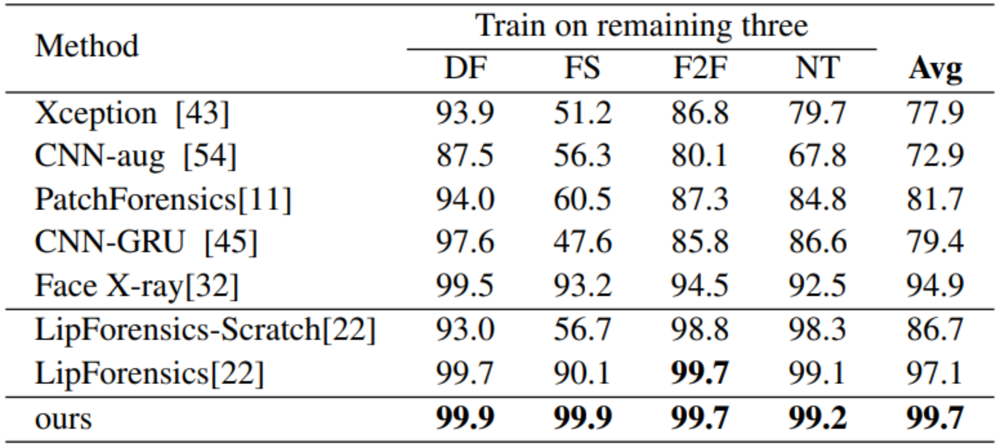
- Deepfake Detection (Input)
  - Video → temporal inconsistency
  - Frame → spatial inconsistency
- Model Generalization
  - Cross-manipulation
  - Cross-dataset

Model Generalization

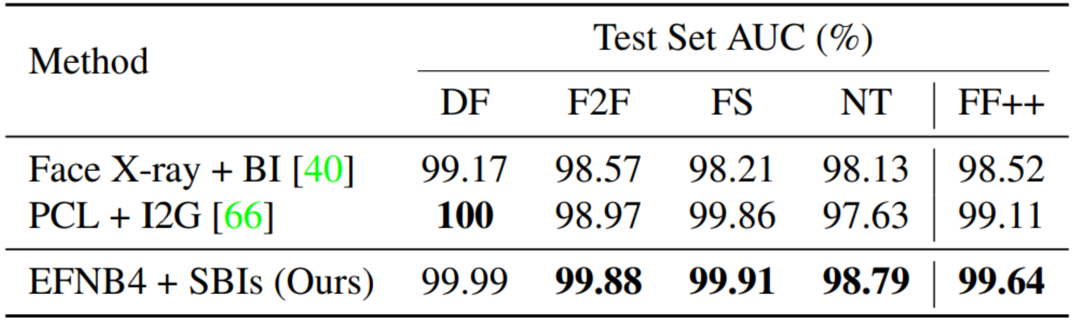
- Cross Manipulation

FTCN (Video based)[1]

SBI (Frame based)[2]



Method	Train on remaining three				
	DF	FS	F2F	NT	**Avg**
Xception [43]	93.9	51.2	86.8	79.7	77.9
CNN-aug [54]	87.5	56.3	80.1	67.8	72.9
PatchForensics[11]	94.0	60.5	87.3	84.8	81.7
CNN-GRU [45]	97.6	47.6	85.8	86.6	79.4
Face X-ray[32]	99.5	93.2	94.5	92.5	94.9
LipForensics-Scratch[22]	93.0	56.7	98.8	98.3	86.7
LipForensics[22]	99.7	90.1	**99.7**	99.1	97.1
ours	**99.9**	**99.9**	**99.7**	**99.2**	**99.7**

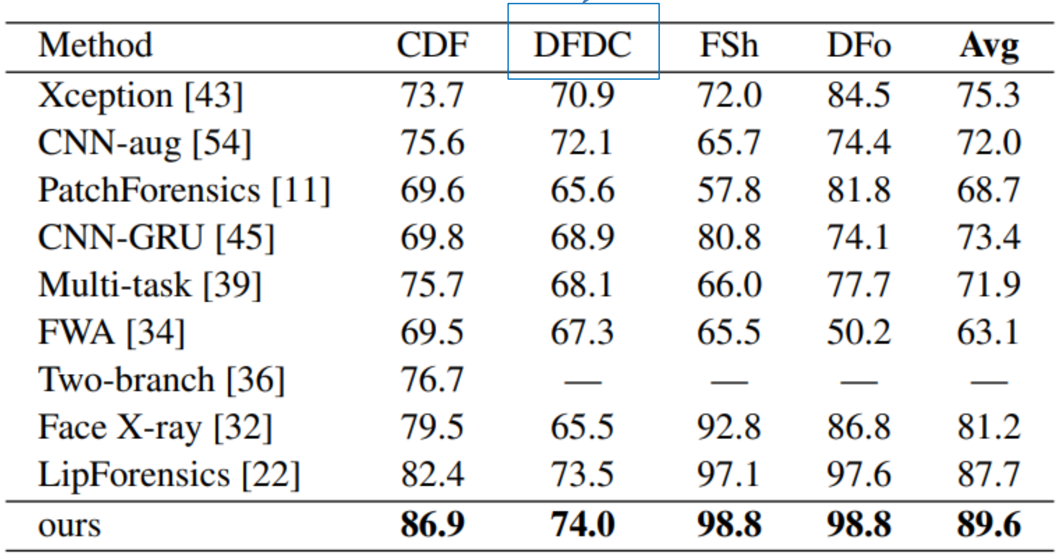


Method	Test Set AUC (%)				
	DF	F2F	FS	NT	FF++
Face X-ray + BI [40]	99.17	98.57	98.21	98.13	98.52
PCL + I2G [66]	**100**	98.97	99.86	97.63	99.11
EFNB4 + SBIs (Ours)	99.99	**99.88**	**99.91**	**98.79**	**99.64**

Model Generalization

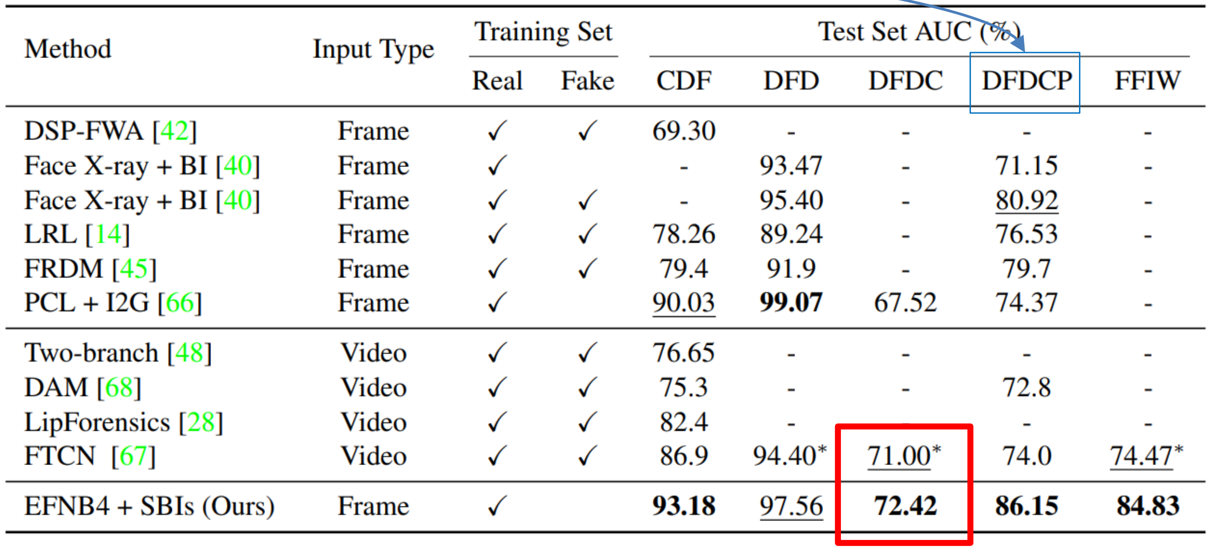
- Cross Dataset

FTCN (Video based)[1]



Method	CDF	DFDC	FSh	DFo	**Avg**
Xception [43]	73.7	70.9	72.0	84.5	75.3
CNN-aug [54]	75.6	72.1	65.7	74.4	72.0
PatchForensics [11]	69.6	65.6	57.8	81.8	68.7
CNN-GRU [45]	69.8	68.9	80.8	74.1	73.4
Multi-task [39]	75.7	68.1	66.0	77.7	71.9
FWA [34]	69.5	67.3	65.5	50.2	63.1
Two-branch [36]	76.7	—	—	—	—
Face X-ray [32]	79.5	65.5	92.8	86.8	81.2
LipForensics [22]	82.4	73.5	97.1	97.6	87.7
ours	**86.9**	**74.0**	**98.8**	**98.8**	**89.6**

SBI(Frame based)[2]

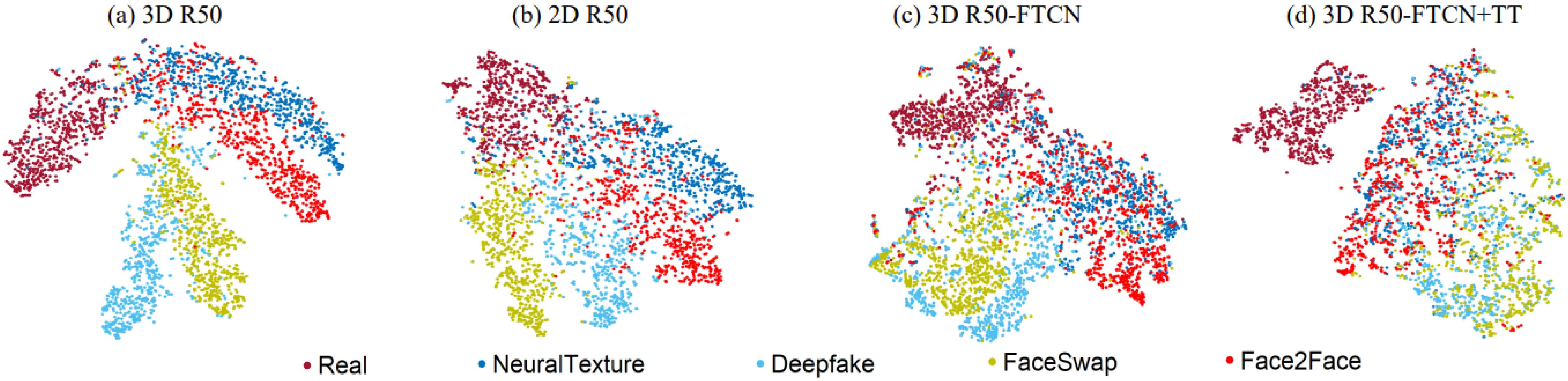


Method	Input Type	Training Set		Test Set AUC (%)				
		Real	Fake	CDF	DFD	DFDC	DFDCP	FFIW
DSP-FWA [42]	Frame	✓	✓	69.30	-	-	-	-
Face X-ray + BI [40]	Frame	✓		-	93.47	-	71.15	-
Face X-ray + BI [40]	Frame	✓	✓	-	95.40	-	80.92	-
LRL [14]	Frame	✓	✓	78.26	89.24	-	76.53	-
FRDM [45]	Frame	✓	✓	79.4	91.9	-	79.7	-
PCL + I2G [66]	Frame	✓		90.03	**99.07**	67.52	74.37	-
Two-branch [48]	Video	✓	✓	76.65	-	-	-	-
DAM [68]	Video	✓	✓	75.3	-	-	72.8	-
LipForensics [28]	Video	✓	✓	82.4	-	-	-	-
FTCN [67]	Video	✓	✓	86.9	94.40*	71.00*	74.0	74.47*
EFNB4 + SBIs (Ours)	Frame	✓		**93.18**	97.56	**72.42**	**86.15**	**84.83**

2. t-SNE

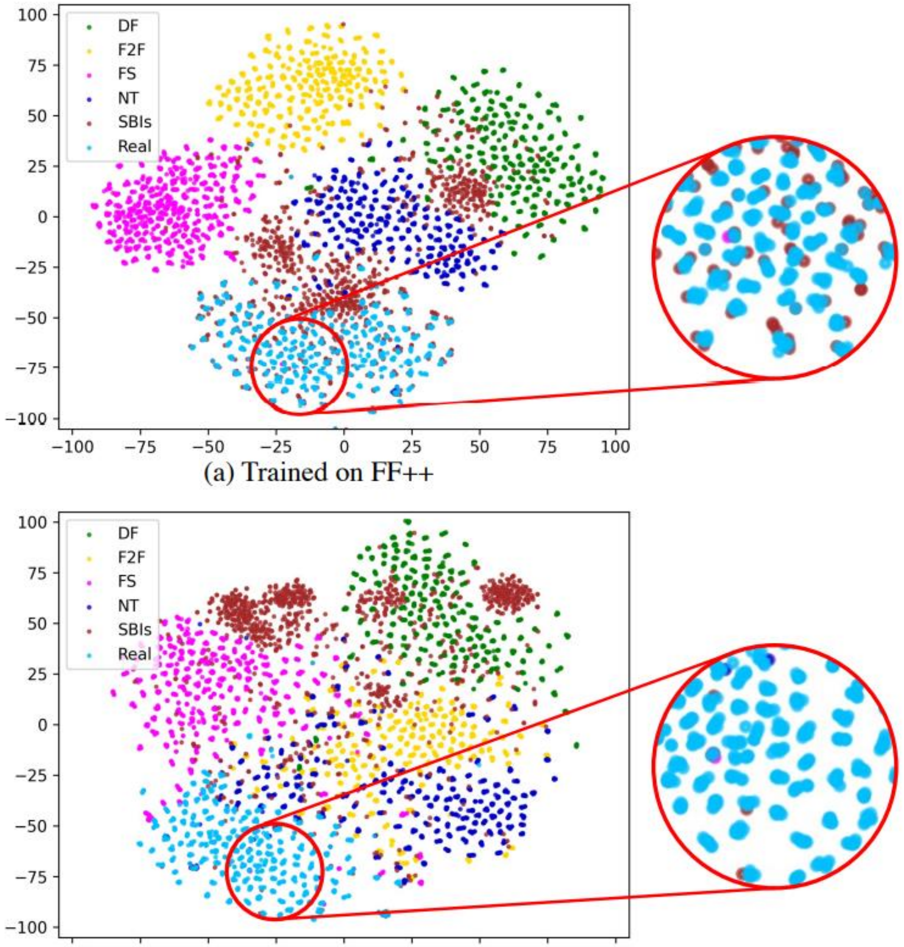
T-SNE

- FTCN[1]


(a) 3D R50
(b) 2D R50
(c) 3D R50-FTCN
(d) 3D R50-FTCN+TT
● Real
● NeuralTexture
● Deepfake
● FaceSwap
● Face2Face

T-SNE

- SBI[2]



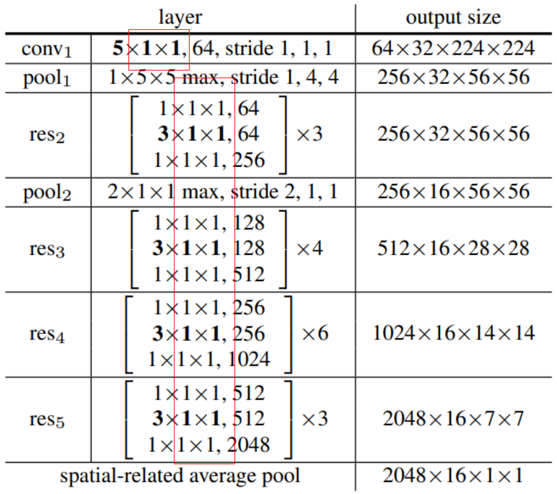
(a) Trained on FF++

(b) Trained on SBIs

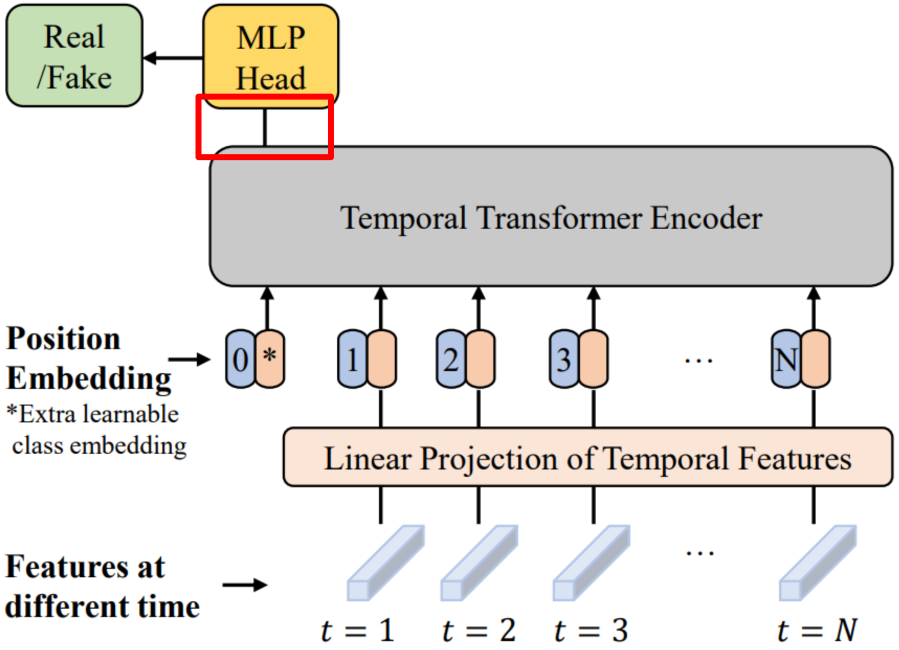
3. Experiment

Experiment

- FTCN
  - Video feature



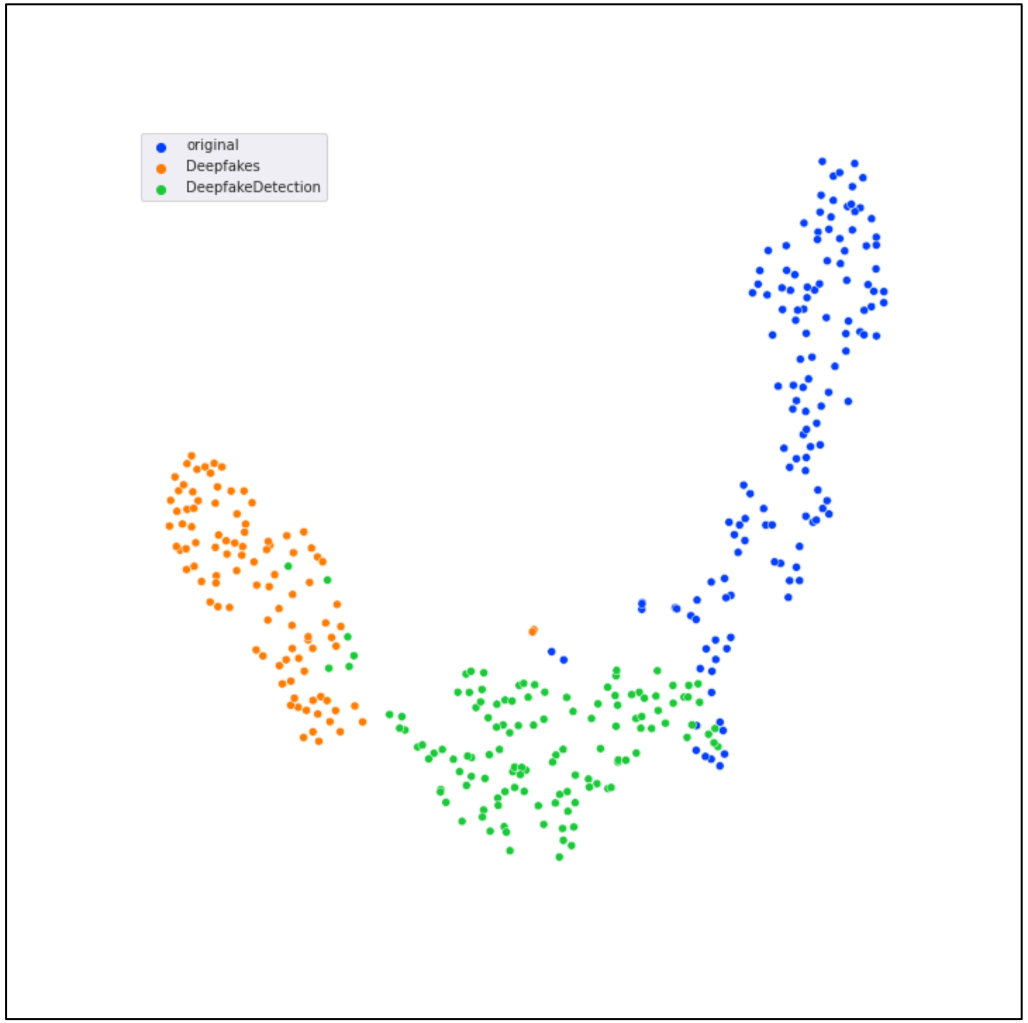
layer		output size
conv_1	$\mathbf{5}\times\mathbf{1}\times\mathbf{1}$, 64, stride 1, 1, 1	$64\times32\times224\times224$
pool_1	$1\times5\times5$ max, stride 1, 4, 4	$256\times32\times56\times56$
res_2	$\left[\begin{array}{c} 1\times1\times1, 64 \\ \mathbf{3}\times\mathbf{1}\times\mathbf{1}, 64 \\ 1\times1\times1, 256 \end{array}\right]\times3$	$256\times32\times56\times56$
pool_2	$2\times1\times1$ max, stride 2, 1, 1	$256\times16\times56\times56$
res_3	$\left[\begin{array}{c} 1\times1\times1, 128 \\ \mathbf{3}\times\mathbf{1}\times\mathbf{1}, 128 \\ 1\times1\times1, 512 \end{array}\right]\times4$	$512\times16\times28\times28$
res_4	$\left[\begin{array}{c} 1\times1\times1, 256 \\ \mathbf{3}\times\mathbf{1}\times\mathbf{1}, 256 \\ 1\times1\times1, 1024 \end{array}\right]\times6$	$1024\times16\times14\times14$
res_5	$\left[\begin{array}{c} 1\times1\times1, 512 \\ \mathbf{3}\times\mathbf{1}\times\mathbf{1}, 512 \\ 1\times1\times1, 2048 \end{array}\right]\times3$	$2048\times16\times7\times7$
spatial-related average pool		$2048\times16\times1\times1$


Real /Fake
MLP Head
Temporal Transformer Encoder
Position Embedding
*Extra learnable class embedding
0
*
1
2
3
...
N
Linear Projection of Temporal Features
Features at different time
t = 1
t = 2
t = 3
t = N

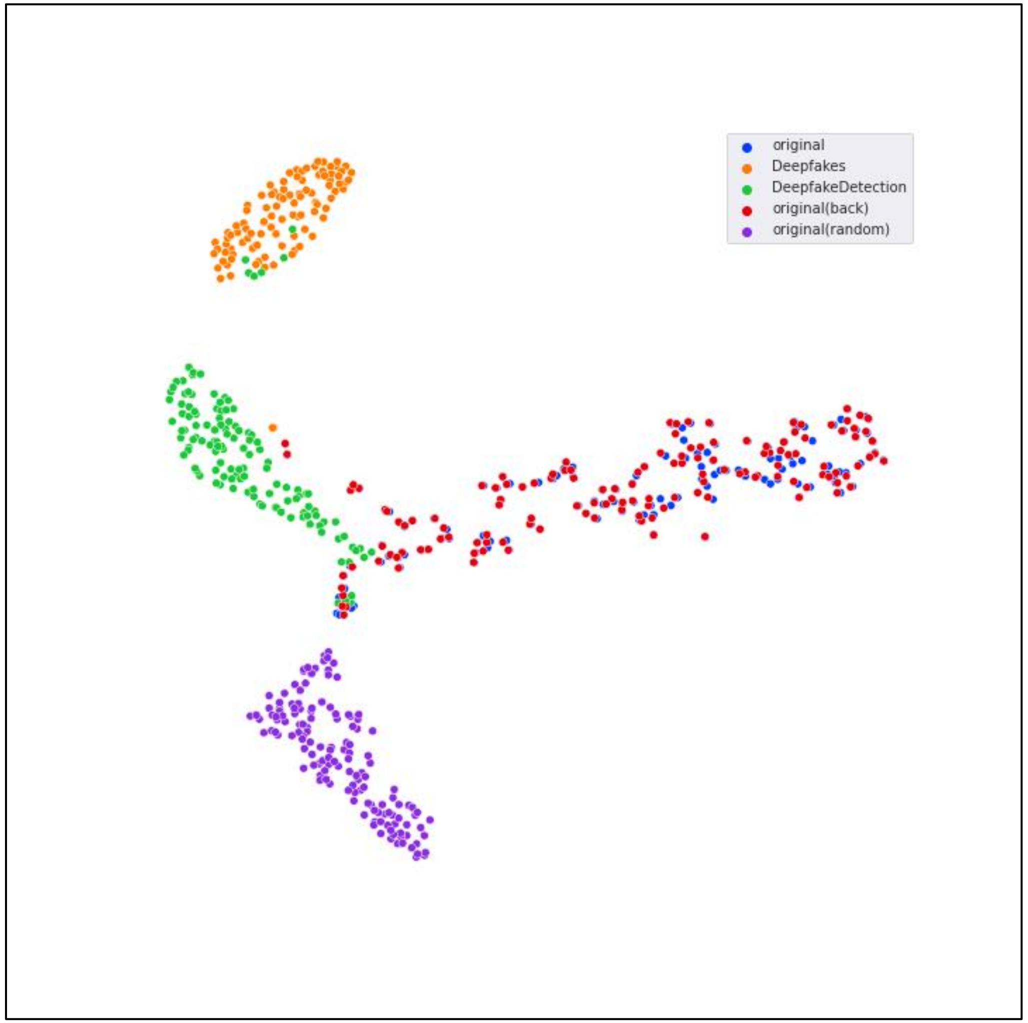
- Check
  - Dataset (1st , 2nd)
  - Temporal Consistency (back, random)

Experiment

Dataset

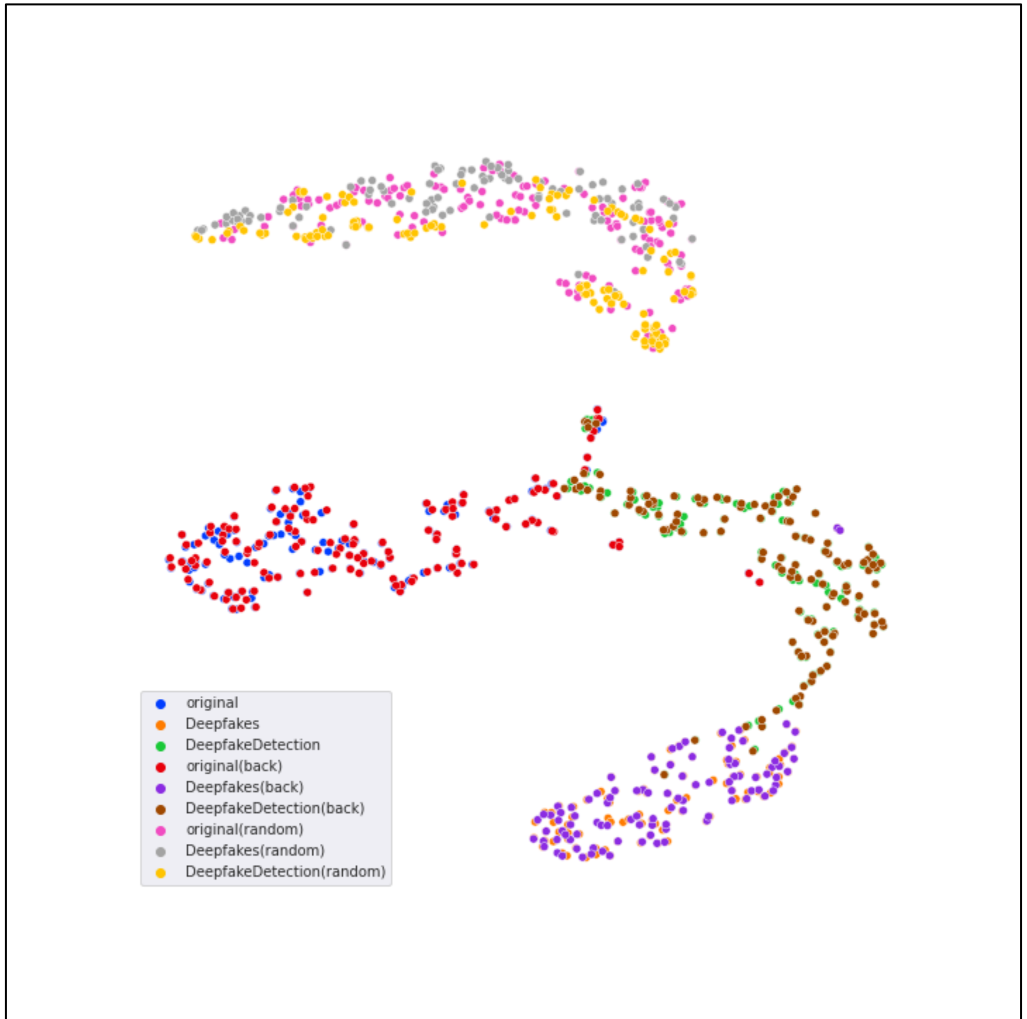


Temporal Consistency



Experiment

Dataset + Temporal Consistency



4. Conclusion

Conclusion

- FTCN lacks generalization performance for 2nd generation Dataset

- Frame back

- Frame random

Reference

[1] Exploring Temporal Coherence for More General Video Face Forgery Detection (ICCV 2021)

[2] Detecting Deepfakes with Self-Blending Images (CVPR 2022 oral)

감사합니다