Supplementary Material: Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection

Baseline Results Using Hand-crafted features v/s Deep Learning for DeepFakes Detection

Model	Train set	Test set	Accuracy	Precision	Recall	F1 score	F2 score
SIFT + SVC	DF	DF	61.48	78.41	29.74	43.12	33.95
		F2F	50.20	56.12	8.32	14.49	10.03
		FS	54.56	44.90	7.29	12.55	8.76
		NT	55.84	55.48	10.77	18.04	12.84
	F2F	DF	51.30	50.66	31.16	38.59	33.76
		F2F	60.16	63.89	49.33	55.67	51.69
		FS	53.06	46.17	30.25	36.55	32.49
		NT	52.22	45.32	28.55	35.03	30.83
	FS	DF	52.85	56.88	17.68	26.97	20.50
		F2F	50.59	41.28	10.11	16.25	11.91
		FS	60.57	64.42	49.90	56.24	52.26
		NT	52.11	51.78	32.13	39.65	34.77
	FS	DF	53.00	56.88	17.68	26.97	20.50
		F2F	51.26	56.98	15.93	24.90	18.61
		FS	52.64	38.15	9.57	15.31	11.26
		NT	59.44	63.23	24.16	34.96	27.56
HoG + SVC	DF	DF	74.26	65.74	78.35	71.50	75.46
		F2F	51.78	22.36	56.19	31.99	43.14
		FS	51.44	14.05	38.25	20.55	28.45
		NT	56.54	25.75	53.85	34.84	44.20
	F2F	DF	49.82	27.21	48.06	34.75	41.67
		F2F	71.80	71.10	72.70	71.89	72.37
		FS	56.00	34.77	51.15	41.40	46.75
		NT	56.28	36.26	52.23	42.80	48.00
	FS	DF	46.84	14.91	39.14	21.59	29.54
		F2F	51.66	27.13	54.73	36.28	45.48
		FS	73.60	68.05	71.51	69.74	70.79
		NT	52.36	21.81	44.32	29.23	36.74
	FS	DF	58.14	45.78	59.60	51.79	56.21
		F2F	56.96	44.05	60.38	50.93	56.21
		FS	50.94	28.05	42.60	33.83	38.59
		NT	64.80	59.49	61.33	60.40	60.96

Table 4: Classification results on different test sets using SIFT+SVC and HoG+SVC

Model	Train set	Test set	Accuracy	Precision	Recall	F1 score	F2 score
Xception	DF	DF	98.77	99.28	98.23	98.75	99.06
		F2F	51.10	83.72	3.57	6.85	15.25
		FS	55.05	8.33	0.08	0.16	0.38
		NT	57.34	85.69	5.23	9.86	21.02
	F2F	DF	52.55	4.68	91.77	8.91	19.43
		F2F	99.36	99.14	99.59	99.37	99.50
		FS	55.72	1.22	70.43	2.40	5.71
		NT	55.65	1.09	67.98	2.14	5.11
	FS	DF	50.24	0.08	13.31	0.15	0.37
		F2F	50.13	1.42	74.97	2.79	6.60
		FS	99.43	99.33	99.40	99.37	99.39
		NT	55.17	0.09	13.31	0.18	0.45
	NT	DF	74.80	93.88	52.59	67.41	57.66
		F2F	53.45	76.55	10.86	19.02	13.11
		FS	54.03	21.12	1.12	2.13	1.38
		NT	93.61	95.55	89.86	92.62	90.94
CNN+LSTM	DF	DF	88.53	89.88	86.60	88.21	87.24
		F2F	61.24	77.45	32.45	45.74	36.72
		FS	51.79	24.26	3.81	6.59	4.58
		NT	68.42	77.55	41.10	53.73	45.37
	F2F	DF	63.42	75.53	38.75	51.22	42.93
		F2F	85.47	87.26	83.30	85.24	84.07
		FS	57.43	56.46	19.87	29.39	22.82
		NT	65.52	71.28	38.02	49.59	41.93
	FS	DF	48.05	38.64	8.20	13.52	9.73
		F2F	55.05	64.91	23.33	34.33	26.76
		FS	86.34	84.30	85.25	84.77	85.06
		NT	52.84	39.01	10.16	16.12	11.92
	NT	DF	69.21	80.23	50.26	61.81	54.32
		F2F	64.44	77.52	41.37	53.95	45.63
		FS	52.58	36.80	8.80	14.20	10.38
		NT	84.12	84.03	79.52	81.71	80.38
Meso Inception-4	DF	DF	88.96	99.89	77.80	87.47	81.40
		F2F	50.12	92.48	0.99	1.97	1.24
		FS	55.36	3.85	0.00	0.01	0.01
		NT	56.05	93.93	1.58	3.10	1.96
	F2F	DF	50.44	29.31	0.06	0.11	0.07
		F2F	76.98	99.90	54.34	70.39	59.79
		FS	55.46	74.24	0.20	0.40	0.25
		NT	55.41	59.52	0.10	0.20	0.13
	FS	DF	49.76	2.04	0.03	0.05	0.03
		F2F	50.73	70.61	3.65	6.95	4.51
		FS	96.55	98.01	94.18	96.05	94.92
		NT	55.41	59.52	0.10	0.20	0.13
	NT	DF	54.72	99.01	8.73	16.04	10.67
		F2F	49.89	86.50	0.54	1.08	0.68
		FS	55.35	2.78	0.00	0.01	0.00
		NT	69.23	99.66	31.12	47.43	36.09

Table 5: Classification results on different test sets using Xception, CNN+LSTM and MesoInception-4 models

Results in terms of AUC Score for proposed SupCon Model, Xception Model and proposed fusion model

Area Under the Curve (AUC) score is the measure of the capability of a classifier to distinguish among classes, and it is used as a review of the ROC curve. AUC ranges in value from 0 to 1. The higher the AUC is, the better the model is. AUC score is scale-invariant and classification-threshold-invariant. Table 6 shows the AUC score results on all three models. We can observe the similar regular pattern in Section 6.3. The performance of the models on unknown attacks attends to be not satisfying. In the meantime, fusion model always outperforms a single model.

Table 6: Test AUC score of SupCon model, Xception and fusion model. Highlighted rows indicate unknown attack detection result.

Training set	Test set	SupCon	Xception	Fusion
F2F + FS + NT	DF	0.8792	0.9156	0.9279
	F2F	0.9817	0.9939	0.9954
	FS	0.9833	0.9956	0.9962
	NT	0.9372	0.9819	0.9820
DF + FS + NT	DF	0.9776	0.9944	0.9956
	F2F	0.6953	0.7920	0.7920
	FS	0.9780	0.9958	0.9970
	NT	0.9509	0.9642	0.9722
DF + F2F + NT	DF	0.9883	0.9969	0.9977
	F2F	0.9843	0.9907	0.9927
	FS	0.3839	0.4925	0.4925
	NT	0.9513	0.9698	0.9750
DF + F2F + NT	DF	0.9889	0.9982	0.9984
	F2F	0.9935	0.9986	0.9985
	FS	0.9928	0.9993	0.9991
	NT	0.6386	0.6868	0.7001

Role of weights in proposed approach

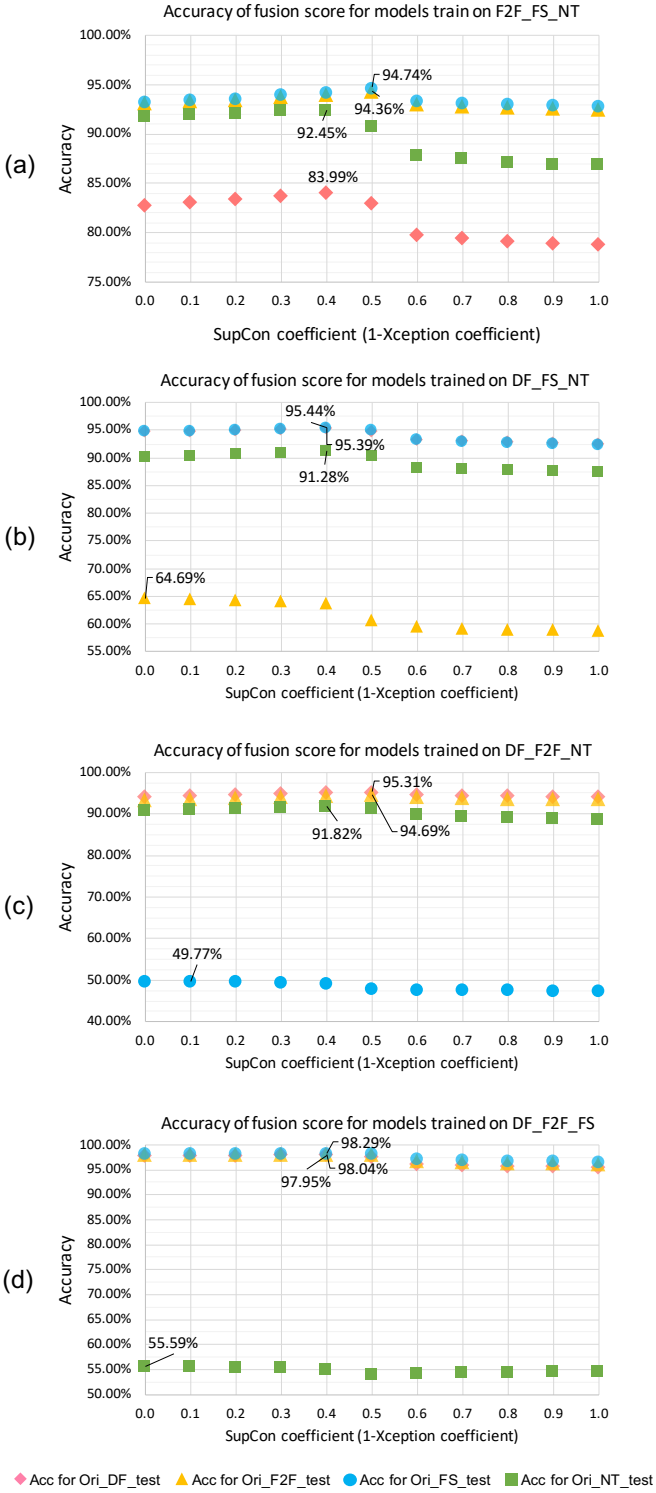


Figure 10: Fusion accuracy of SupCon and Xception final scores

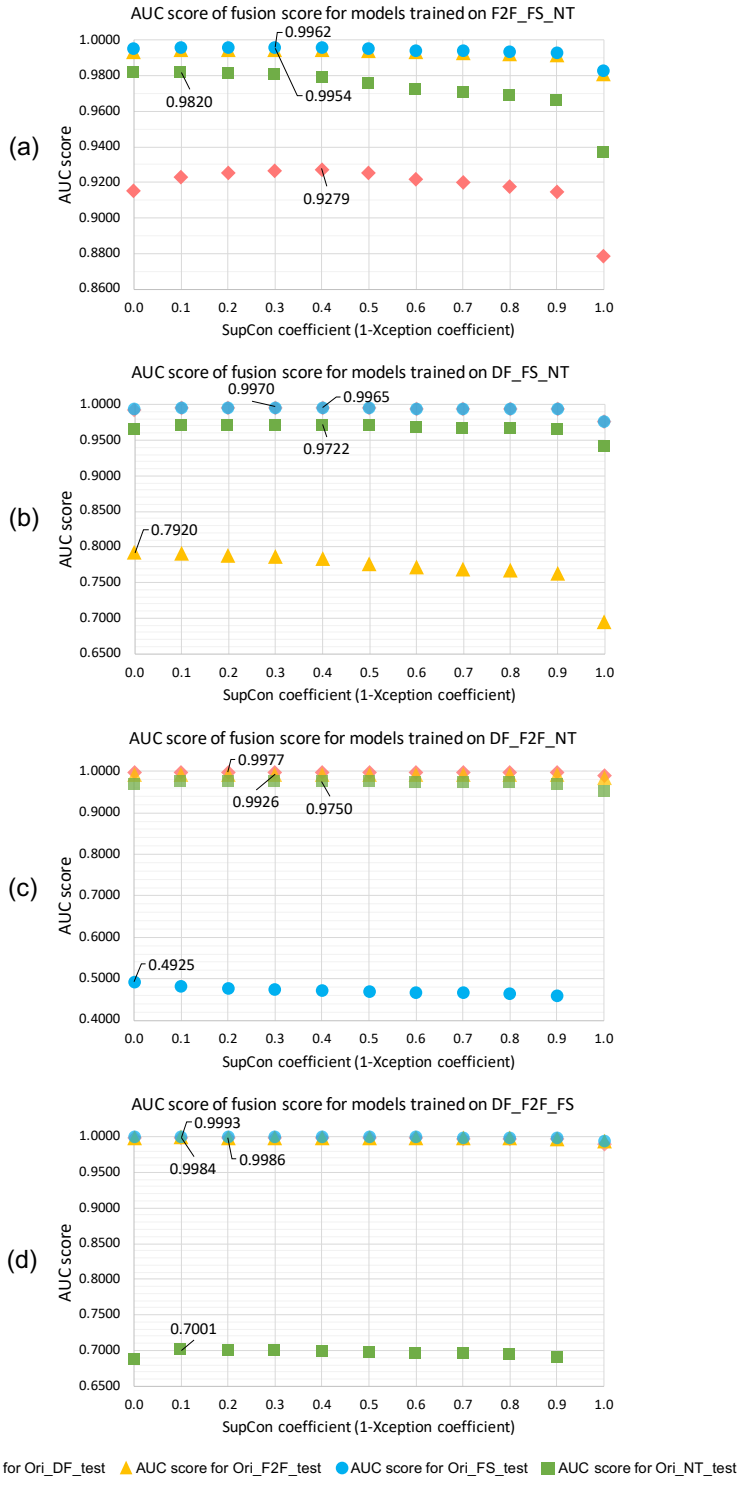


Figure 11: The fusion AUC score of SupCon and Xception final scores

Xception trained by two attacks

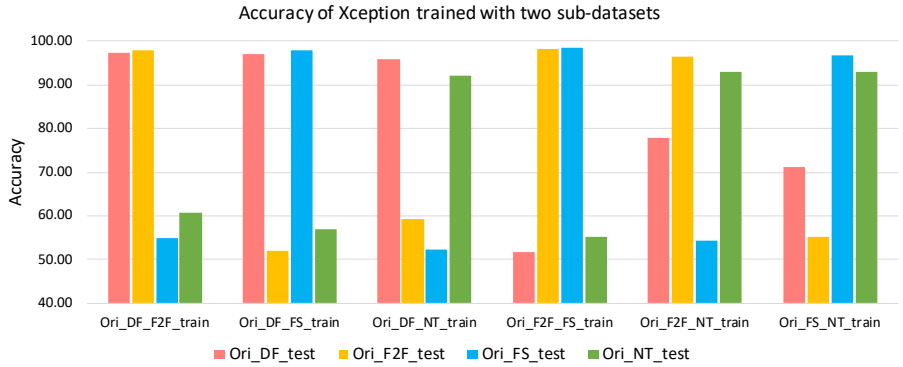
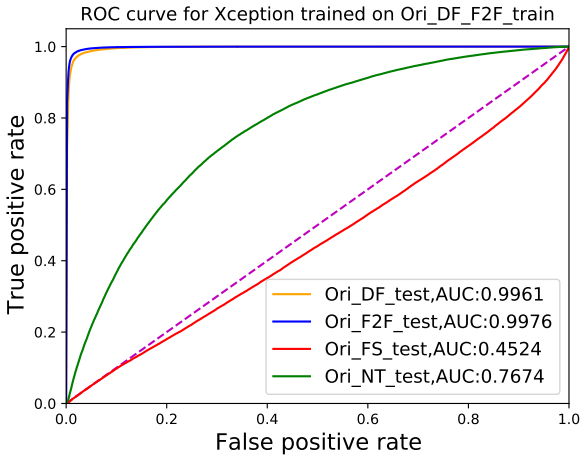
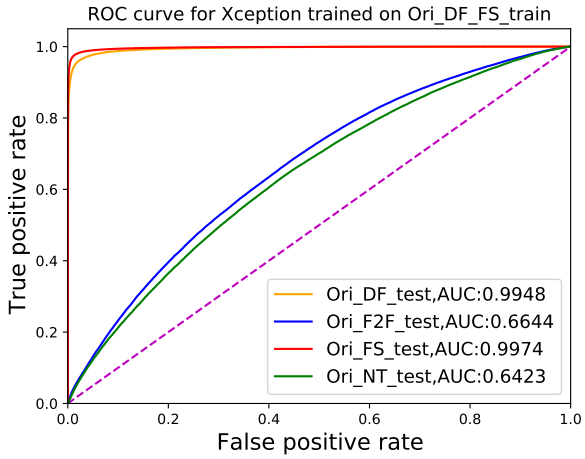


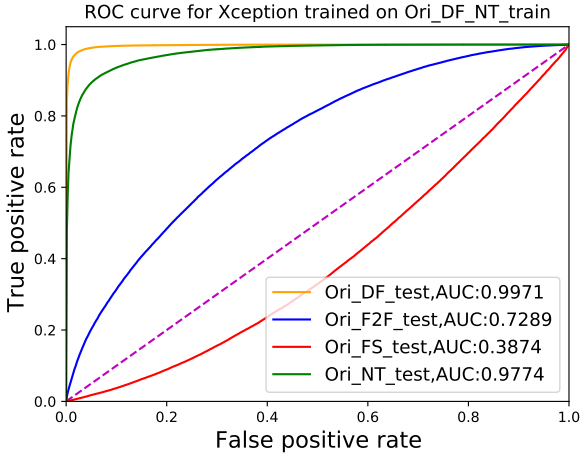
Figure 12: Accuracy of Xception trained with two known attacks



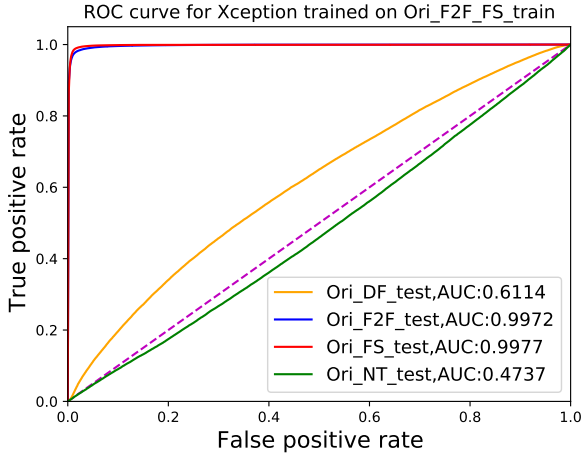
(a) ROC curve for Xception trained on *DF + F2F*



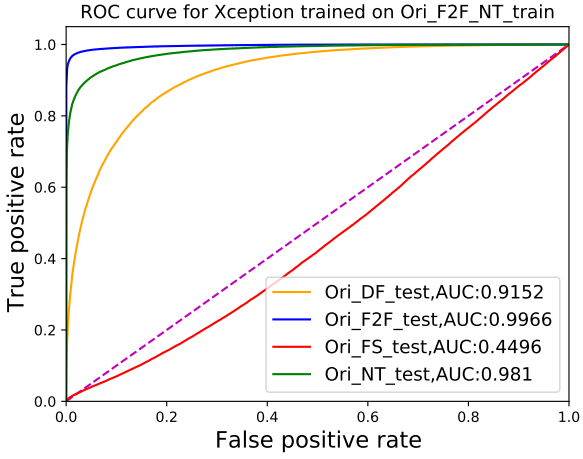
(b) ROC curve for Xception trained on *DF + FS*



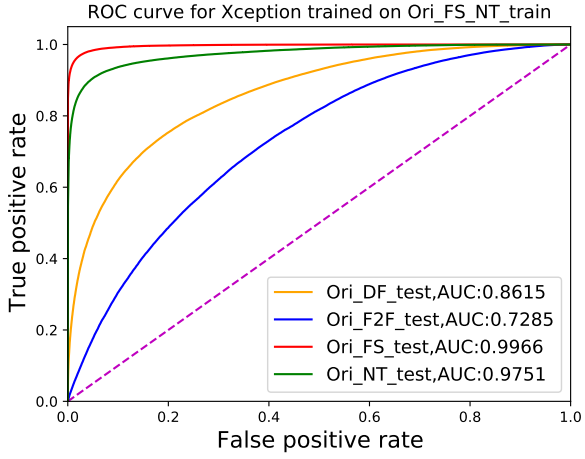
(c) ROC curve for Xception trained on *DF + NT*



(d) ROC curve for Xception trained on *F2F + FS*



(e) ROC curve for Xception trained on *F2F + NT*



(f) ROC curve for Xception trained on *FS + NT*

Figure 13: The ROC curves tested on four test sets for the Xception Network trained on two known attacks. (a) Trained on *DF + F2F*. (b) Trained on DF + FS. (c) Trained on *DF + NT*. (d) Trained on *F2F + FS*. (e) Trained on *F2F + NT*. (f) Trained on *FS + NT*.