MixNet for Generalized Face Presentation Attack Detection

Nilay Sanghvi[1], Sushant Kumar Singh[1], Akshay Agarwal[1,2], Mayank Vatsa[3], and Richa Singh[3]
[1]IIIT Delhi, India, [2]Texas A&M University, Kingsville, USA; [3]IIT Jodhpur, India

http://iab-rubric.org/index.html

Introduction

- Facial recognition systems are highly deployed, but are still vulnerable to variety of presentation attack mediums, broadly:
  - Print (2D)
  - Replay (2D)
  - Mask (3D)

- **Generalizability** against multiple attack mediums is a major problem with existing face presentation attack detection (PAD) algorithms.

- We propose a novel Face PAD algorithm called **MixNet**, which utilizes state-of-the-art convolutional neural networks (CNNs) and learns the feature mapping for each attack category.

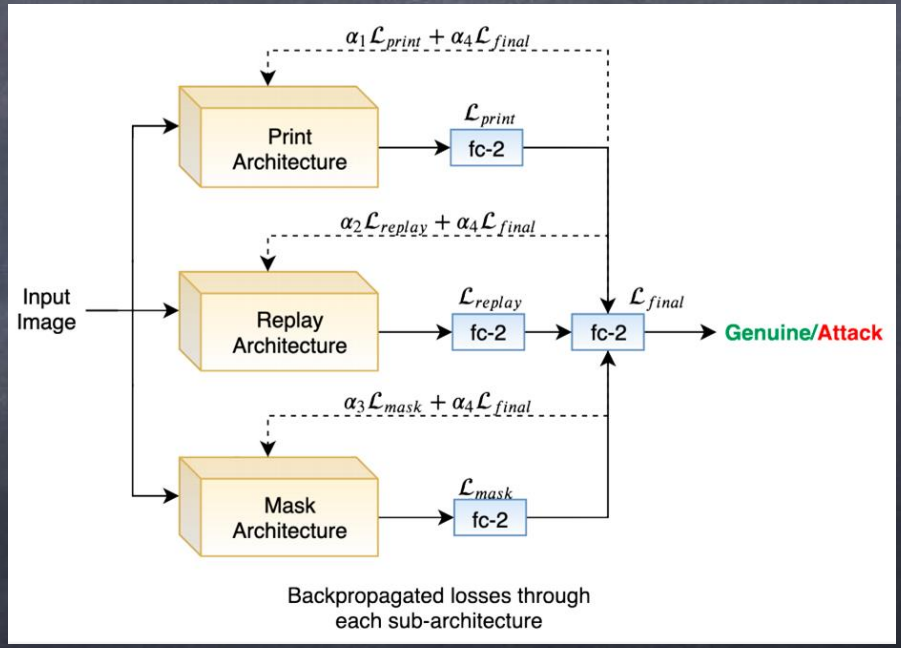
- MixNet, can further identifies the attack type without an extra computational overhead.

Proposed Algorithm: MixNet (Training)

- Most algorithms pose face PAD as a **binary classification problem**, i.e., classifying genuine vs attack, leading to poor generalizability against unseen attacks.

- In MixNet, we add an intermediate step of detecting the three broad attacks using dedicated sub-architectures.

- On passing an attack sample, only the sub-architecture responsible for detecting that attack should output a score close to 1. Other two sub-architectures should output a score close to 0.



Proposed Algorithm: MixNet (Data Labeling)

- We label each data sample as a quadruple.
- The first three entries correspond to the desired output from the three sub-architectures.
- The last entry corresponds to the final classification output of MixNet.

Type of Sample	Print Label	Replay Label	Mask Label	Final Label
Genuine	0	0	0	0
Print Attack	1	0	0	1
Replay Attack	0	1	0	1
Mask Attack	0	0	1	1

Labeling of the input data for training proposed MixNet

Proposed Algorithm: MixNet (Loss Function)

- Each sub-architecture has a loss associated with it - print, replay and mask loss.
- Further, there is final classification loss for the final output layer.
- During training, MixNet tries to minimize the total loss.

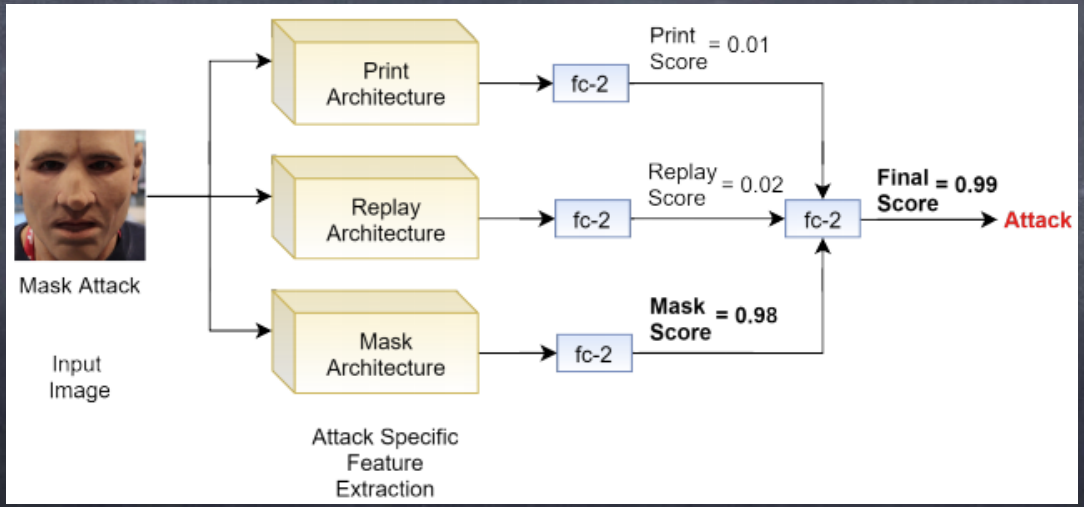
$$\mathcal{L}_{total} = \alpha_1 \mathcal{L}_{print} + \alpha_2 \mathcal{L}_{replay} + \alpha_3 \mathcal{L}_{mask} + \alpha_4 \mathcal{L}_{final} \quad (1)$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are the regularization coefficients for the four losses. Each of these losses is categorical cross-entropy loss represented as:

$$\mathcal{L}_{cross-entropy} = -\sum_i y_i \log(p_i) \quad (2)$$

Proposed Algorithm: MixNet (Testing)

- Once MixNet is trained, each sub-architecture outputs a score between 0 and 1, indicating the confidence that the corresponding attack is present in input image.
- For the final classification, MixNet combines the scores from the three sub-architectures to give a final score between 0 and 1 denoting the confidence that the input image is an attack image.



Experimental Settings

- We have merged two challenging databases:
  - Silicone Mask Attack Database (**SMAD**) [1]
  - Spoof In the Wild with Multiple Attack Types (**SiW-M**) [2]

- We divide the merged database into 2 non-overlapping parts.
  - Intra-Database
  - Cross-Database and Unseen Attack

Intra-Database Protocol

- Videos from each class (genuine, print, replay, and mask) of this part are equally divided into 3 non-overlapping folds.

- In each iteration of 3 fold cross-validation, the model is trained on two folds and tested on the third.

Video Type	Database	Number of Videos
Genuine	SMAD	65
Genuine (from train split)	SiW-M	217
Print Attack	SiW-M	104
Replay Attack	SiW-M	99
Mask Attack	SMAD	65

Video Type	Number of Videos	Scenario
Genuine (from test split)	131	Seen
Silicone Mask	27	Cross
Paper Mask	17	Unseen
Half Mask	72	Unseen
Transparent Mask	88	Unseen
Mannequin	40	Unseen

Cross-Database: The three trained models, each from the three iterations of cross-validation performed in the intra-database protocol, are evaluated on this part.

[1] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," IEEE TIFS, vol. 12, no. 7, pp. 1713–1723, July 2017.
[2] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in IEEE CVPR, 2019, pp. 4680–4689.

Experimental Results (Intra-database)

- The algorithms based on CNN models outperform the hand-crafted features based algorithms.
- Among the two CNN models used, the deeper model with 121 layers yields the lowest error rates.
- The MixNet corresponding to ResNet model pre-trained on face images (VGG-Face2) shows a slightly higher error rate than object images (ImageNet) based counterpart

Architecture	ACER	APCER	BPCER
LBP+HOG	14.98 ± 2.90	14.99 ± 6.15	14.96 ± 4.81
Multi-scale LBP [3]	16.01 ± 1.64	12.60 ± 0.65	19.43 ± 3.06
ResNet50 [4]	10.05 ± 2.82	10.91 ± 5.21	9.18 ± 5.43
MixNet-ResNet50*	6.41 ± 0.69	**2.34 ± 1.34**	10.49 ± 2.06
MixNet-ResNet50-VF2*	6.85 ± 2.89	7.24 ± 4.46	**6.47 ± 1.86**
DenseNet121	**6.02 ± 0.63**	7.16 ± 2.61	**4.88 ± 3.87**
MixNet-DenseNet121*	**4.52 ± 0.90**	**1.76 ± 0.31**	7.28 ± 1.61

Results in terms of μ ± σ for **intra-db** protocol. Top-2 results are in green.

***MixNet-ResNet50** refers to MixNet using ResNet50 as the 3 sub-architectures. Similarly, MixNet-ResNet50-VF2 and MixNet-DenseNet121 refer to MixNet using ResNet50-VGG-Face2 [5] and DenseNet121 respectively.

[3] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in IEEE CVPR, 2019, pp. 4680–4689.
[4] J. Ma¨att ¨a, A. Hadid, and M. Pietik ¨ ainen, "Face spoofing detection from ¨ single images using micro-texture analysis," in IJCB, 2011, pp. 1–7.
[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in IEEE FG, 2018, pp. 67–74

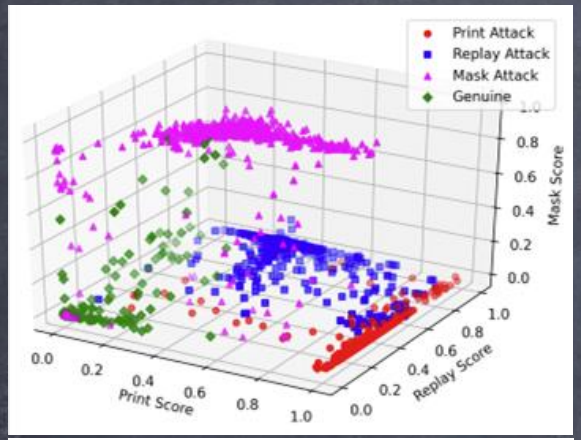
Experimental Results (Cross-db and unseen attack)

- The hand-crafted algorithms lack generalizability and perform poorly.
- Paper mask is found to be the easiest attack to be detected, which might be because it lacks the smooth texture and suffers from edge artifacts.
- The effectiveness of the proposed algorithm on attacks such as mannequin, which is not explored previously in the literature, shows that it is generalizable to handle the real-world scenarios.

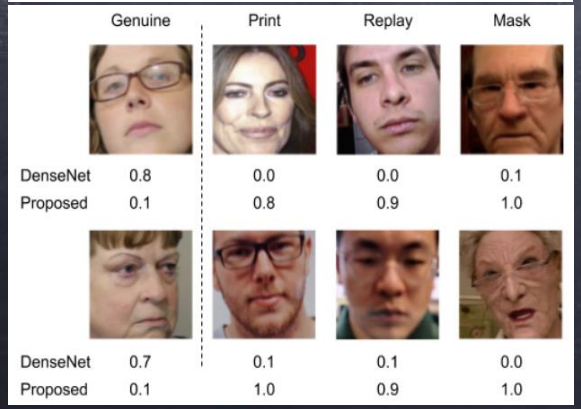
Architecture	Silicone Mask	Paper Mask	Half Mask	Transparent Mask	Mannequin
LBP+HOG	53.33	21.44	54.06	83.85	26.98
Multi-scale LBP	45.68	4.84	42.91	84.74	21.03
ResNet50	**12.84**	97.50	44.28	98.18	31.32
MixNet-ResNet50	16.22	**1.00**	**26.41**	**71.54**	**4.78**
MixNet-ResNet50-VF2	17.74	10.20	61.73	82.46	23.67
DenseNet121	23.12	26.12	39.92	92.94	9.34
MixNet-DenseNet121	**11.54**	**4.54**	**21.56**	**81.56**	**3.54**

APCER % attack-wise for **cross-db and unseen attack protocol**. Top-2 results are in green.

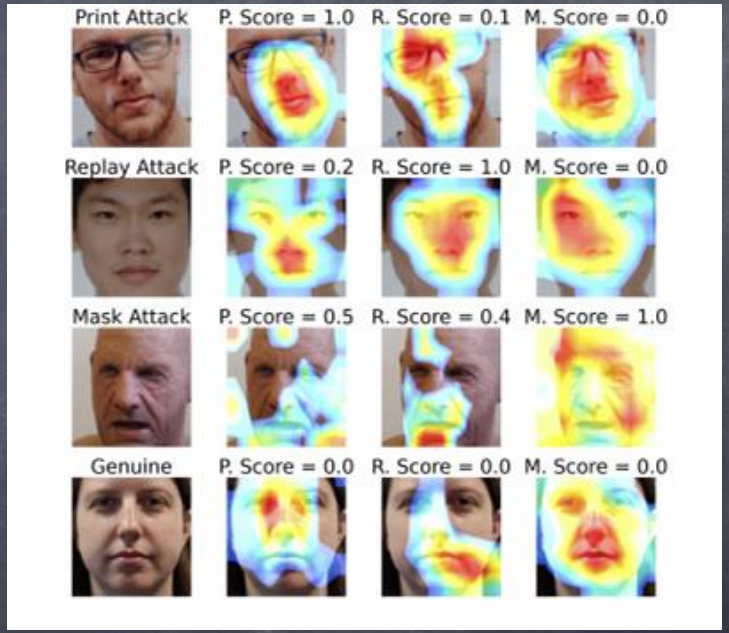
Visualizations and Analysis



Visualisation of output scores from the three sub-architectures of MixNet-DenseNet121 for a subset of test samples



Images of genuine and different types of attack classes. The classification scores computed using DenseNet121 and the proposed algorithm are also written.



The Class Activation Maps (CAMs) [6] of four kinds of images obtained from MixNet-DenseNet121. From left to right: original image, CAM of print, replay, and mask architecture, respectively. P. Score, R. Score, M. Score represent the scores of Print, Replay, and Mask class, respectively

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in IEEE CVPR, 2016, pp. 2921–929.

Ablation Study

Results on Existing Databases - We performed experiments on **Replay-Attack [6]** and **MSU-MFSD [7]** for an extensive comparison of MixNet with other face PAD algorithms.

Simultaneous vs. Independent Sub-architectures Training - We compared proposed MixNet to a method where we separately train models, each dedicated to classifying a specific attack type.
Then, take the **maximum** or **average** of these model's output for **final output** score. For each attack sub-architecture, we selected the best model among Xception, DenseNet121, and ResNet50 based on **HTER** on the validation set.

Method	Replay Attack	MSU-MFSD
Haralick Features [8]	-	5.0
Deep Learning [9]	2.1	5.8
DR-UDA (SE-ResNet18) [10]	1.3	6.3
Multi-Regional CNN [11]	1.6	-
CCoLBP + Ensemble Learning [12]	4.0	5.0
Independently Optimized Sub-Nets	1.3	2.4
Ours (MixNet-DenseNet)	0.6	0.4

Intra database evaluation on Replay Attack (HTER%) and MSU-MSFD (EER%)

[6] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in BIOSIG, 2012, pp. 1–7,
[7] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," IEEE TIFS, vol. 10, no. 4, pp. 746–761, 2015.
[8] A. Agarwal, R. Singh, and M. Vatsa, "Face anti-spoofing using haralick features," in IEEE BTAS, 2016, pp. 1–6.
[9] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," IEEE TIFS, vol. 13, no. 7, pp. 1794–1809, 2018.
[10] G. Wang, H. Han, S. Shan, and X. Chen, "Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection," IEEE TIFS, vol. 13, no. 7, pp. 1794–1809, 2020.
[11] Y. Ma, L. Wu, Z. Li et al., "A novel face presentation attack detection scheme based on multi-regional convolutional neural networks," Pattern Recognition Letters, vol. 131, pp. 261–267, 2020.
[12] F. Peng, L. Qin, and M. Long, "Face presentation attack detection based on chromatic co-occurrence of local binary pattern and ensemble learning," JVCIR, vol. 66, p. 102746, 2020.

Conclusion

- Facial recognition algorithms are vulnerable to presentation attacks which limits their usability for security purposes. Thus, it is essential to develop more reliable and robust algorithms to detect such attacks.

- Instead of learning generalized features for all attack types, learning attack specific features tackles the problem more effectively. Extensive experimental comparisons of MixNet with such generalized networks establish the efficacy of the proposed idea.

Future Work

- Currently, the same network has been used for each sub-architecture in MixNet. We may select different networks for each sub-architecture such that they are state of the art for detecting the corresponding attack.

Thank You!

Paper: https://arxiv.org/abs/2010.13246