EUROPEAN CONFERENCE ON COMPUTER VISION
TEL AVIV 2022

RepMix: Representation Mixing for Robust Attribution of Synthesized Images

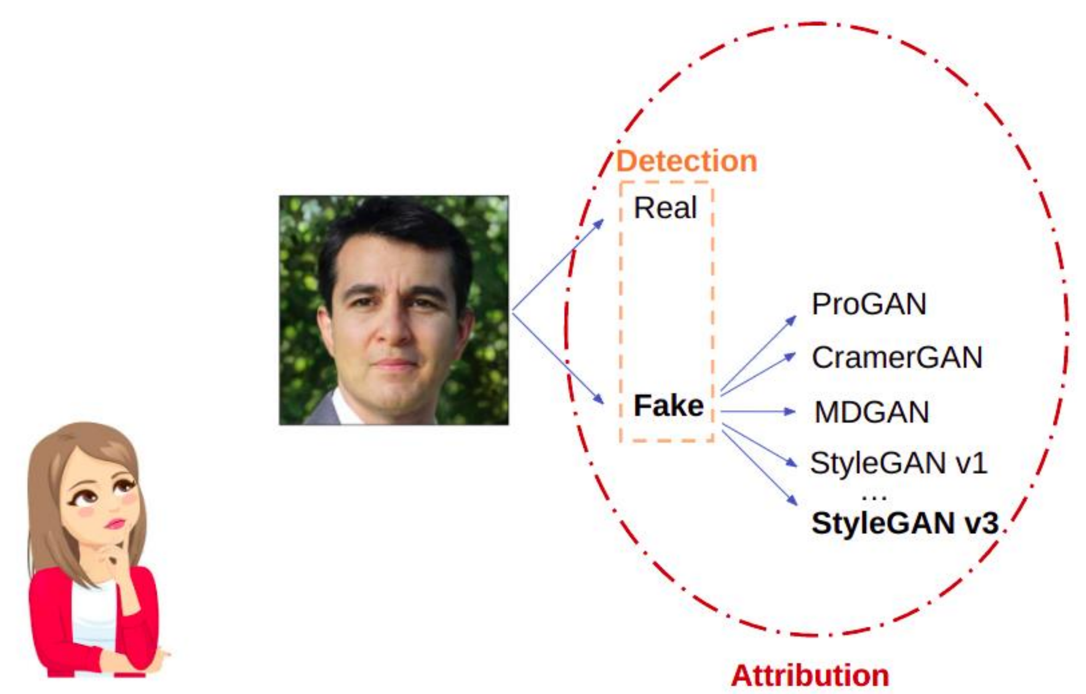
Tu Bui*[1], Ning Yu[2] and John Collomosse[1,3]

[1]University of Surrey, [2]Salesforce, [3]Adobe Research, *t.v.bui@surrey.ac.uk

University of Surrey · Adobe · salesforce

1. Problems

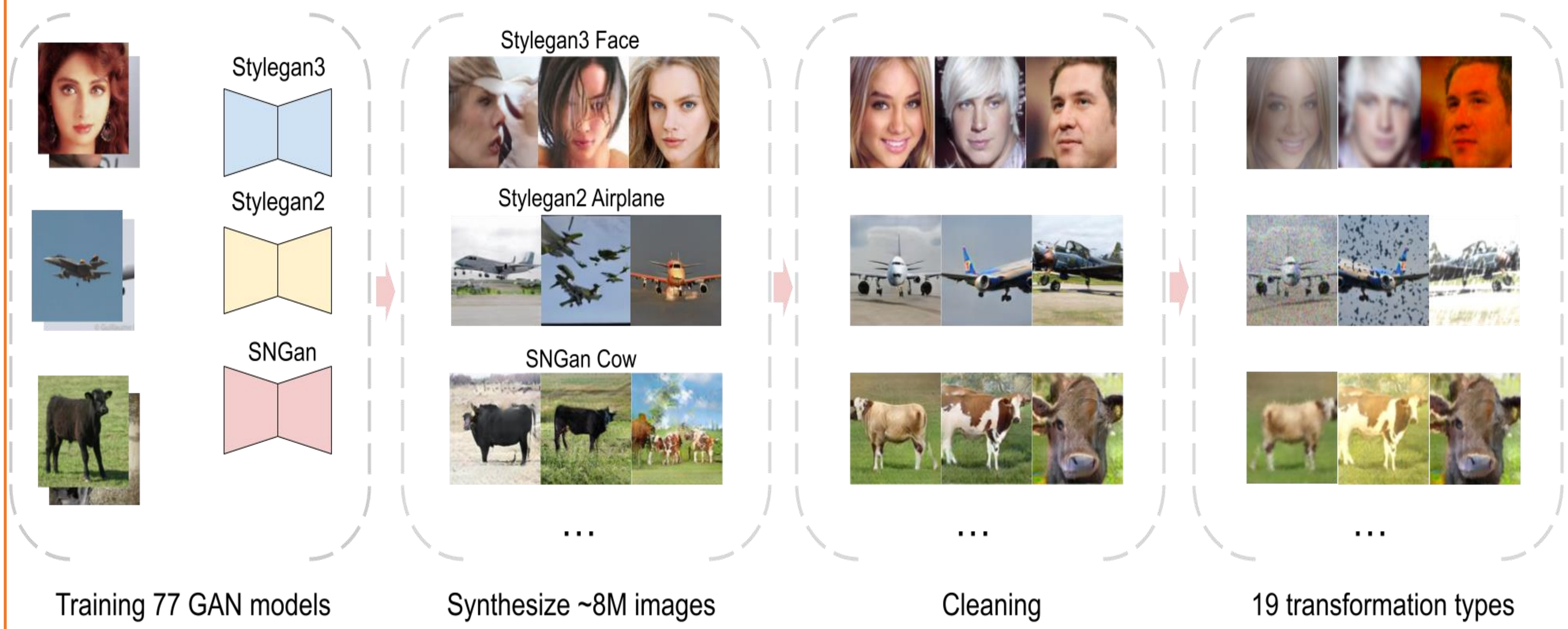
Image Attribution is the process of detecting if an image is synthetic, and if yes, which **synthesizer** generates that image.



Challenges:

- Attribution to model **architecture** is more challenging than to model **instance** → require generalization to unseen training categories.
- Robustness to **noises**: perturbations during online redistribution typically distort high frequency details which would be useful for image attribution.

2. The Attribution88 Dataset



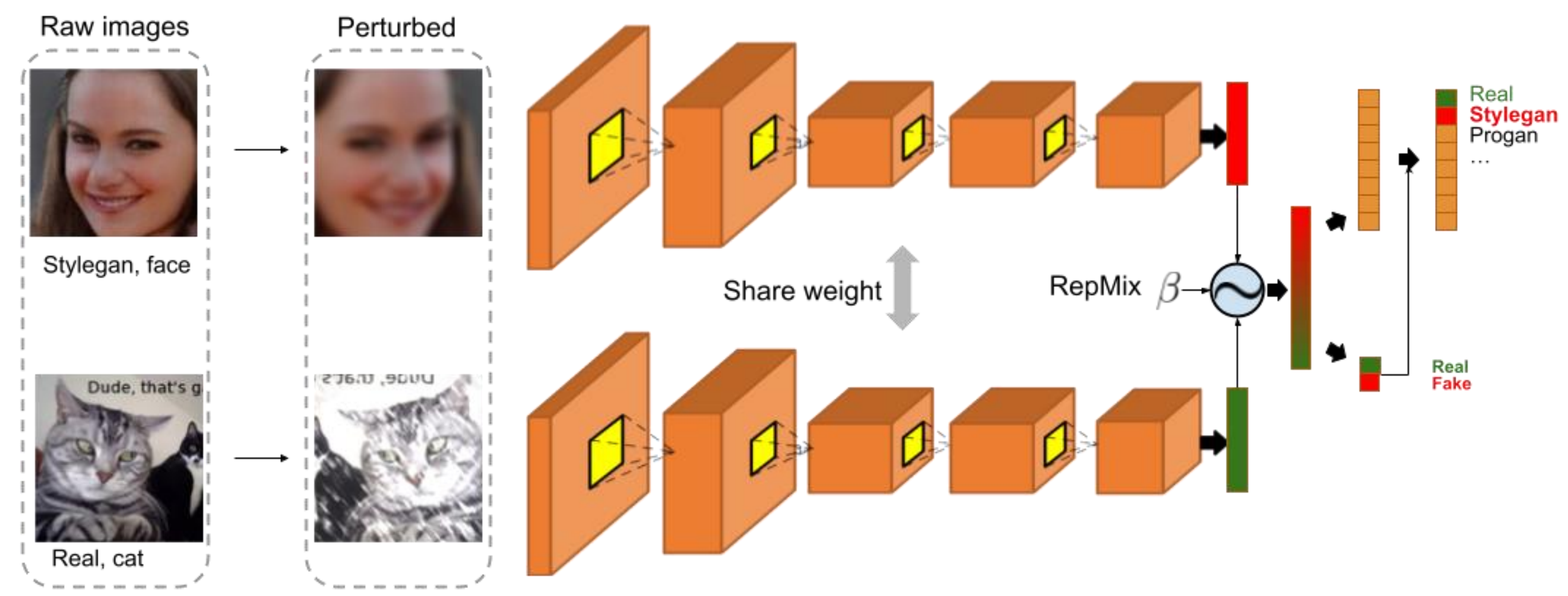
Motivation:

- Existing attribution datasets [5] are **non-deterministic**, lack of semantic **diversity**, source model's **quality** and levels of perturbations.
- The needs of a new benchmark for model architecture attribution instead of model instance attribution.

The **Attribution88** dataset:

- **1,056,000** RGB images of size 128x128,
- **8** source classes (real + 7 GAN architectures) x **11** semantics,
- **77** GAN models have been collected or trained using LSUN real images.
- Data cleaning: removing artifacts and balancing diversity.
- Noises: **19** ImageNet-C random perturbation sources.
- Deterministic test set: semantics (**6** seen + **5** unseen), perturbations (**15** seen + **4** unseen).

3. RepMix Design



RepMix Architecture:

$$\boldsymbol{f}(x_i, x_j, \alpha) = \boldsymbol{f}_t(\alpha \boldsymbol{f}_b(x_i) + (1-\alpha)\boldsymbol{f}_b(x_j))$$

where f_t and f_b are top and bottom layers; α is drawn randomly from Beta distribution; x_i and x_j are 2 input images.

- Mix random image features of different semantics and sources (GANs or real).
- Predict the source mixture ratio.
- Mixing at top fully connected layer gives the best performance.

Compound loss:

- Account for the hierarchical nature of attribution: image → real or fake (synthesised) → generator sources,
- Consist of real/fake detection loss and attribution loss,
- Detection (real/fake) scores factor in attribution scores,
- Weighted cross-entropy losses.

Contribution of different design components

	Detection Acc. ⇑	Attribution Acc. ⇑	Attribution NMI ⇑
All (ResNet backbone)	**0.9426**	**0.7400**	**0.5546**
w/o compound loss	0.9364	0.7204	0.5280
w/o RepMix	0.9296	0.7188	0.5205
w/o RepMix+Compound loss	0.9283	0.7129	0.5167
w/o Augmentation	0.7044	0.2762	0.0856
VGG16	0.9493	0.7150	0.5315
AlexNet	0.8818	0.5280	0.2817

5. References

[1] Asnani, Vishal, et al. "Reverse engineering of generative models: Inferring model hyperparameters from generated images." arXiv preprint arXiv:2106.07873 (2021).

[2] Frank, Joel, et al. "Leveraging frequency analysis for deep fake image recognition" Proc. ICLR, 2020.

[3] Marra, Francesco, et al. "Do gans leave artificial fingerprints?." Proc. IEEE MIPR, 2019.

[4] Sirovich, Lawrence, and Michael Kirby. "Low-dimensional procedure for the characterization of human faces." Josa a 4.3 (1987): 519-524.

[5] Yu, Ning, Larry S. Davis, and Mario Fritz. "Attributing fake images to gans: Learning and analyzing gan fingerprints." Proc. ICCV, 2019.

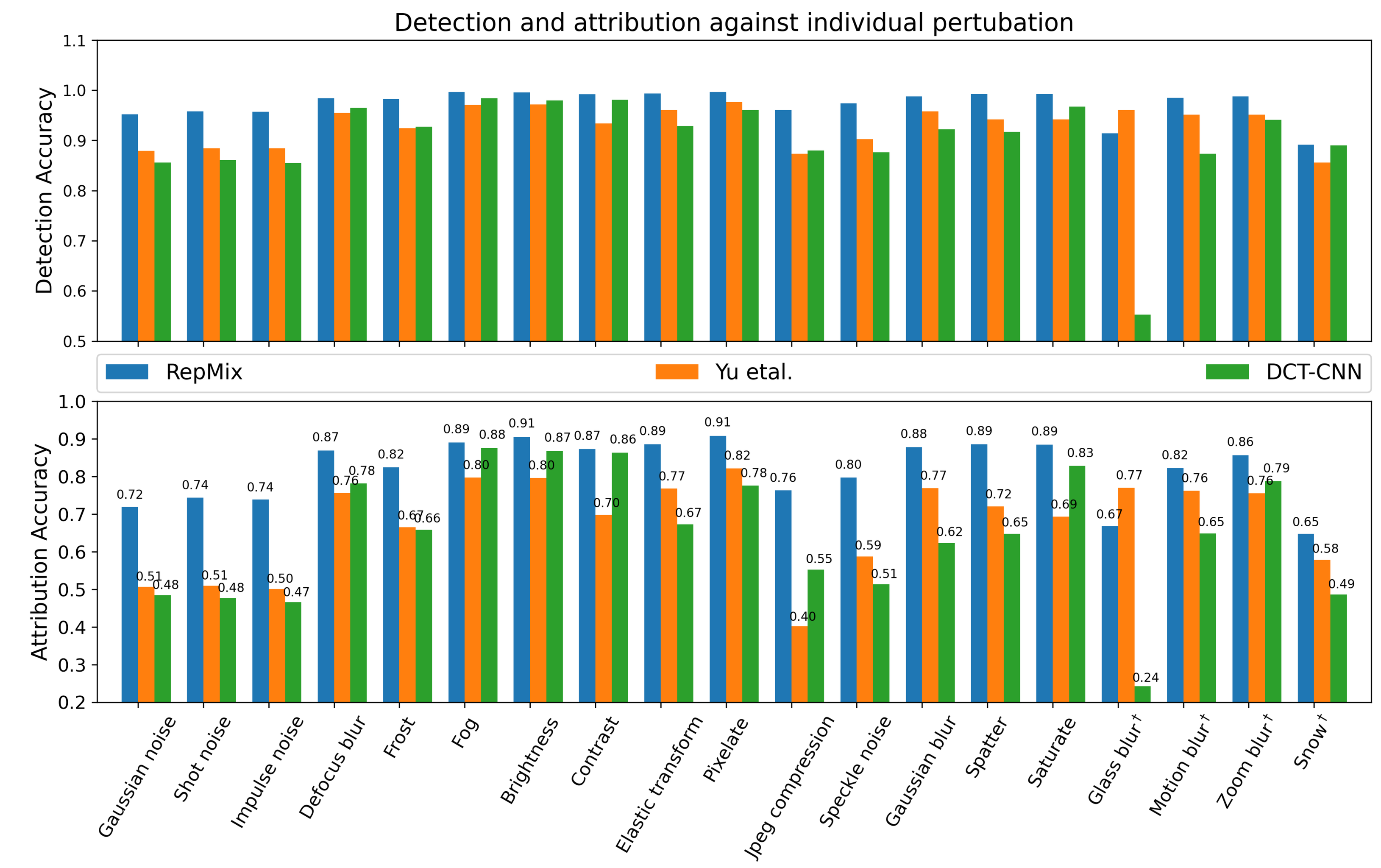
4. Experiments

Baseline Comparison

Table 1: Performance of RepMix and other baselines on a control set and Attribution88 test set

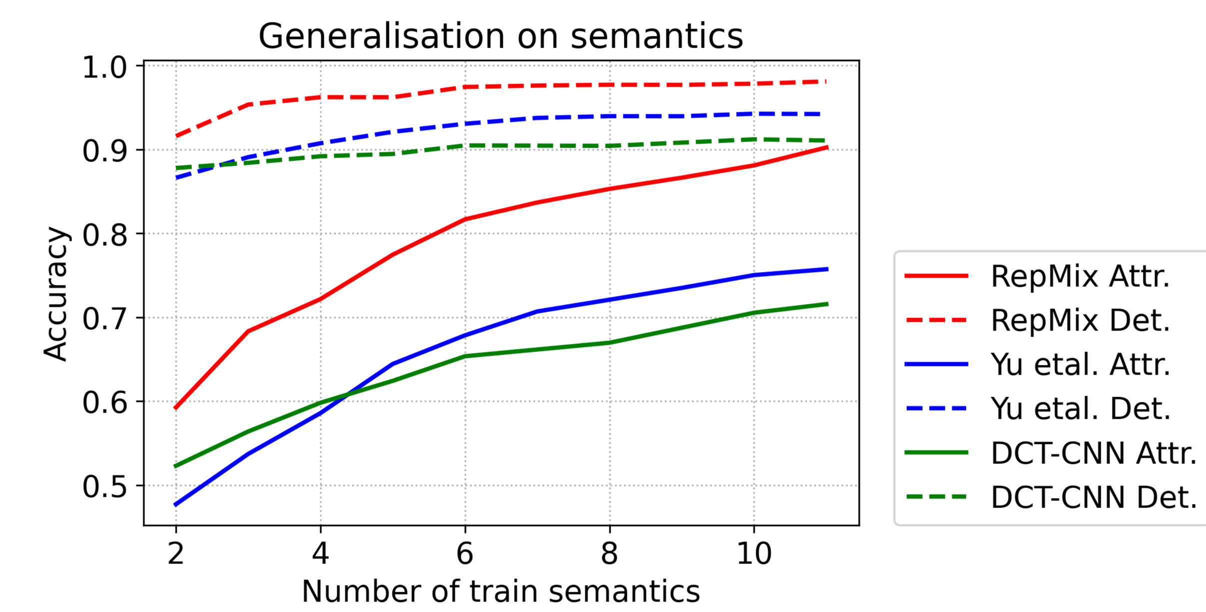
	1 Sem., Clean			Attribution88		
	Det. Acc. ⇑	Attr. Acc. ⇑	Attr. NMI ⇑	Det. Acc. ⇑	Attr. Acc. ⇑	Attr. NMI ⇑
RepMix	**1.0000**	**0.9994**	**0.9975**	**0.9745**	**0.8207**	**0.6679**
Yu *et al.* (reimp.)	0.9910	0.9838	0.9458	0.9306	0.6784	0.4666
Yu *et al.* [5]	0.9888	0.9844	0.9455	0.9190	0.6322	0.4028
DCT-CNN [2]	0.9922	0.9838	0.9526	0.9001	0.6447	0.4061
Reverse Eng. [1]	0.9976	0.9960	0.9834	0.8665	0.5637	0.3653
EigenFace [4]	0.8262	0.6538	0.4515	0.7829	0.1515	0.0034
PRNU [3]	0.8544	0.8482	0.7389	0.7845	0.1252	0.0003

Robustness to Perturbation



† indicates unseen transformations.

Robustness to Semantics



Number of semantics in the test set is fixed at 11s.

Robustness to Adversarial Attacks

Table 2: Adversarial attacks at different levels of max perturbation ϵ

Error ⇓	$\epsilon = 2/255$	$\epsilon = 4/255$	$\epsilon = 8/255$	$\epsilon = 16/255$	$\epsilon = 24/255$	$\epsilon = 32/255$
RepMix	**0.1509**	**0.1952**	**0.2454**	**0.3008**	**0.3333**	**0.3572**
Yu *et al.* [5]	0.2113	0.2709	0.3328	0.3945	0.4303	0.4534
DCT-CNN [2]	0.1545	0.2190	0.2831	0.3375	0.3642	0.3812