Open Set Synthetic Image Attribution

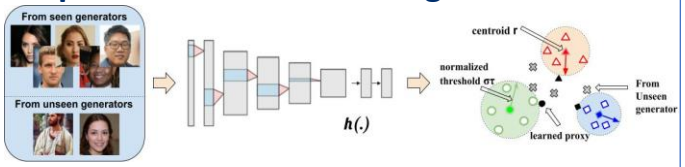
Shengbang Fang, Tai D. Nguyen, Matthew C. Stamm

Motivation: Attribute a synthetic image's source generator, in an open-set scenario.

Taming Transformer

? Unknown Generator

Proposed: A Metric-learning based method

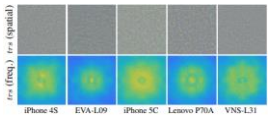


- Identify an existing image generator
- Detect the image from an unseen generator

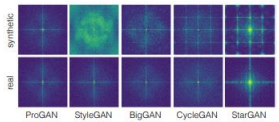
Training:

1. Transferable Embedding Initialization

- Camera traces transfer to other forensic tasks



Camera model traces (frequency) [1]



Synthetic image traces (frequency) [2]

[1] Chang Chen, Zhiwei Xiong, Xiaoming Liu, Feng Wu. "Camera Trace Erasing". In CVPR 2020.
[2] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, Alexei A. Efros. "CNN-generated images are surprisingly easy to spot...for now". In CVPR, 2020.

- Pre-train on Camera model classification

$$L_{init} = -\sum_k y_k log(\hat{y}_k)$$

- Remove last soft-max layer

2. Embedding Learning

- Loss Function: ProxyNCA++

$$P_i = \frac{exp(-d(h(x_i), p(y_i)))}{\sum_{p(a)\in p(A)} exp(-d(h(x_i), p(a)))}$$

$$L_{proxyNCA++} = -log(Pi)$$

Inference:

1. Finding Class References

- Compute reference point

$$r_i = \frac{\sum_{x_i \in T_i} h(xi)}{|T_i|}$$

- Find nearest point

$$\underset{k}{\operatorname{argmax}}\, d(h(x), rk)$$

2. Reject Criteria

- Compute normalized score

$$\sigma i = \sqrt{E(d(h(xi), ri)2)} = \sqrt{E(||h(xi) - ri||2)}$$

$$= \sqrt{\frac{\sum_{xi \in Ti} ||h(xi) - ri||2}{|Ti| - 1}}$$

$$s(x, ri) = \frac{d(h(x), ri)}{\sigma_i}$$

- Accept/Reject Criterion

$$R(x) = \begin{cases} g_i \in G\ (seen),\ if\ s(x,ri) < \tau \\ g_i \notin G\ (unseen),\ if\ s(x,ri) \geq \tau \end{cases}$$

Experimental Results

1. Evaluation Metrics

- Seen Generators

$$aF_1 = \frac{\sum_{j=1}^{N} F_{1j}}{N}, \qquad where\ F_{1j} = \frac{2\ TPj}{2TPj + FPj + FNj}$$

- Unseen Generators

$$CRR = \frac{|\{mini\ s(x, ri) > \tau, \forall x \in \{gu \notin \{G\}\}|}{|\{\forall x \in \{g_u \notin \{G\}\}|}$$

2. Results

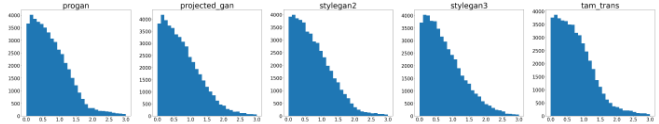


Figure 2: Distribution of training set's embeddings' normalized distance from reference point

Embedding Arch.	Xception	ResNet50	CamID-CNN	Stega-CNN	MISLNet
Train From Scratch	0.827	0.671	0.519	0.787	0.761
With Pre-Training	**0.868**	0.714	0.574	0.808	**0.868**

Table 1: AUC of aF_1-CRR response curve of different CNN models. Pre-trainining on camera model significantly improved the models generatlizability and performance on open-set scenario.

Method	ProGAN	Proj.-GAN	StyleGAN2	StyleGAN3	Taming Trans.	aF_1	StyleGAN Stable	Diffusion	CRR
RepMix	0.669	0.827	0.762	0.839	0.860	0.791	0	0	0
DCT-CNN	0.673	0.929	0.687	0.609	0.851	0.750	0	0	0
ResNet-50	0.572	0.995	0.995	0.797	0.976	0.867	0	0	0
Proposed	0.744	0.974	0.875	0.969	0.940	**0.900**	0.484	0.806	0.645
FSM	0.000	0.032	0.000	0.385	0.585	0.200	0.910	0.363	0.637
EXIF-Net	0.374	0.245	0.124	0.187	0.163	0.219	0.525	0.741	0.633

Table 2: Comparison with other existing approaches (closed-set synthetic image attribution, open-set image source identification) on open-set synthetic image attribution. Our method achieve outperformed both of them.