Discovering Transferable Forensic Features for CNN-generated Images Detection

Keshigeyan Chandrasegaran [1], Ngoc-Trung Tran [1], Alexander Binder [2,3], Ngai-Man Cheung [1]

{ keshigeyan, ngoctrung_tran, ngaiman_cheung }@sutd.edu.sg, alexabin@uio.no

[1] Singapore University of Technology and Design (SUTD) [2] Singapore Institute of Technology (SIT) [3] University of Oslo (UIO)

ECCV TEL AVIV 2022

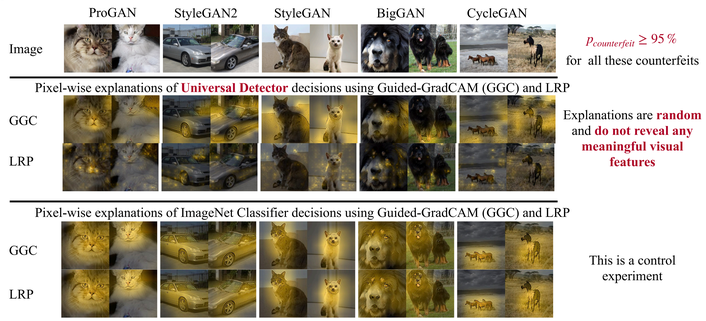
Introduction

Background: Visual counterfeits are increasingly causing an existential conundrum in mainstream media. With rapid improvements in generative modelling, detecting such counterfeits is increasingly becoming challenging and critical. However, a recent class of forensic detectors known as *universal detectors* (Wang et al., 2020) can surprisingly spot counterfeits regardless of generator architectures, loss functions, datasets or resolutions. This intriguing cross-model forensic transfer suggests the existence of **Transferable Forensic Features (*T-FF*)** in universal detectors.

Reserach question: What **Transferable Forensic Features *(T-FF)*** are used by universal detectors for counterfeit detection?

Our main contribution: Our work conducts the **first** analytical study to **discover & understand Transferable Forensic Features *(T-FF)*** in universal detectors for counterfeit detection.

Existing Pixel-wise explanation methods are not informative to discover *T-FF*



Forensic Feature Relevance Statistic (FF-RS) (ω)

Which **feature maps** in universal detectors are responsible for **cross-model forensic transfer?**

FF-RS ($\boldsymbol{\omega}$) is a scalar ([0,1]) that **quantifies** the forensic relevance of every feature map.

$\boldsymbol{\omega}$ for a feature map quantifies

$$\frac{\text{positive forensic relevance of the feature map}}{\text{total unsigned forensic relevance of the entire layer}}$$

top-k : **Set of *T-FF* (large-ω)**
random-k : Random feature maps
low-k : Low-ω feature maps

Algorithm 1: Calculate FF-RS (ω) (Non-vectorized)

Input:
forensics detector M,
data $D = \{x\}_{i=1}^{n}$, D is a large counterfeit dataset where x_i indicates the i^{th} counterfeit image.
Output:
$\omega(l_c)$ where l, c indicates the layer and channel index of forensic feature maps. Every forensic feature map can be characterized by a unique set of l, c.

$R \leftarrow [\,]$; /*List to store feature map relevances*/
Set M to evaluation mode
for i *in* $\{0, 1, ..., n\}$ **do**
  $f(x_i) \leftarrow M(x_i)$; /*logit output*/
  $r_i \leftarrow LRP(M, x_i, f(x_i))$; /*calculate LRP scores for counterfeits*/
  for l' *in* $r_i.size(0)$ **do**
    for c' *in* $r_i.size(1)$ **do**
      $r_i(l', c', h, w) \leftarrow \frac{max(0, r_i(l', c', h, w))}{\sum_{c,h,w} ||r_i(l', c, h, w)||}$
      $R.append(r_i)$; /*r_i.size():(layer, channel, height, width)*/
    end
  end
end
$\omega(l_c) \leftarrow \sum_{h,w} \frac{1}{k} \sum_i^n R_i(l, c, h, w)$; /*forensic feature relevance*/
return $\omega(l_c)$

Validation of our proposed *FF-RS (ω)* for discovering *T-FF*

ResNet-50 Detector feature map dropout results (Sensitivity assessments)

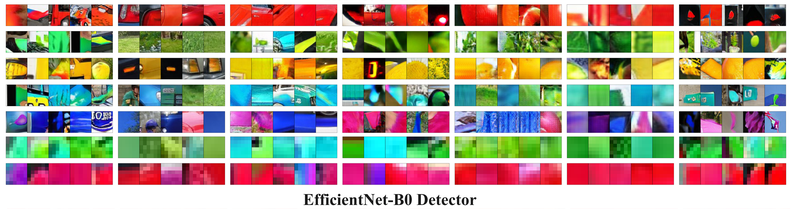
AP / Acc	ProGAN			StyleGAN2			StyleGAN			BigGAN			CycleGAN			StarGAN			GauGAN		
k = 114	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN
baseline	100.0	100.0	100.0	99.1	95.5	95.0	99.3	96.0	95.6	90.4	83.9	85.1	97.9	93.4	92.6	97.5	94.0	89.3	98.8	93.9	96.4
top-k	**69.8**	**99.4**	**3.2**	**55.3**	**89.4**	**11.3**	**56.6**	**90.6**	**13.7**	**55.4**	**86.4**	**18.3**	**61.2**	**91.4**	**17.4**	**72.6**	**89.4**	**35.9**	**71.0**	**95.0**	**18.8**
random-k	100.0	99.9	96.1	98.6	89.4	96.9	98.7	91.4	96.1	88.0	79.4	85.1	96.6	81.0	96.2	97.0	88.0	91.7	98.7	91.9	97.1
low-k	100.0	100.0	100.0	99.1	95.6	95.0	99.3	96.0	95.6	90.4	83.9	85.1	97.9	93.4	92.6	97.5	94.0	89.3	98.8	93.9	96.4

EfficientNet-B0 Detector feature map dropout results (Sensitivity assessments)

AP / Acc	ProGAN			StyleGAN2			StyleGAN			BigGAN			CycleGAN			StarGAN			GauGAN		
Sk = 27S	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN
baseline	100.0	100.0	100.0	95.9	95.2	85.4	99.0	96.1	94.3	84.4	79.7	75.9	97.3	89.6	93.0	96.0	92.8	85.5	98.3	94.1	94.4
top-k	**50.0**	**100.0**	**0.0**	**54.5**	**94.3**	**7.0**	**52.1**	**97.3**	**2.6**	**53.5**	**97.4**	**3.8**	**47.5**	**100.0**	**0.0**	**50.0**	**100.0**	**0.0**	**46.2**	**100.0**	**0.0**
random-k	100.0	99.9	100.0	96.5	91.9	89.8	99.2	91.2	97.5	84.5	59.4	89.1	96.9	82.6	95.8	96.7	82.5	93.3	98.1	87.8	96.2
Low-k	100.0	100.0	100.0	95.3	88.7	88.3	98.9	90.8	96.1	83.5	70.8	80.8	96.6	85.2	94.1	95.4	91.0	85.4	98.1	91.2	96.4

FF-RS (ω) successfully quantifies and discovers *T-FF*

Color is a critical *T-FF* in universal detectors for counterfeit detection (LRP-max)



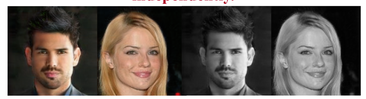
Algorithm 2: Obtain LRP-max pixel-wise explanations (For a single feature map, for a single sample)

Input:
forensics detector M,
counterfeit image x where $x.size() = (3, x_{height}, x_{width})$,
forensic feature map l, c where l, c indicate layer and channel index respectively.
Output:
$\tilde{E}_{l_c}(x)$ where E indicates the LRP-max pixel-wise explanations for sample x corresponding to forensic feature map at layer index l and channel index c. Do note that $\tilde{E}_{l_c}(x).size()$ is (x_{height}, x_{width}).
Every forensic feature map can be characterized by a unique set of l, c.

$z_{l_c}(x) \leftarrow LRP - FORWARD(M_{l_c}(x_i))$; /*(h, w) relevance scores*/
$h^*, w^* \leftarrow argmax(z_{l_c}(x))$; /*find index of max relevance*/
$z_{l_c}^{max}(x) \leftarrow z_{l_c}(x)[h^*, w^*]$; /*LRP-max response neuron*/
$\tilde{E}_{l_c}(x) \leftarrow LRP - BACKWARD(z_{l_c}^{max}(x))$; /*explain LRP-max neuron*/
$\tilde{E}_{l_c}(x) \leftarrow \sum_{k=0}^{3}(E_{l_c}(x)(k, x_{height}, x_{width})$; /*spatial LRP-max*/
return $\tilde{E}_{l_c}(x)$

We introduce a novel pixel-wise visualization method **LRP-max** for visualizing which pixels in the input space correspond to **maximum spatial relevance scores for each *T-FF*.**

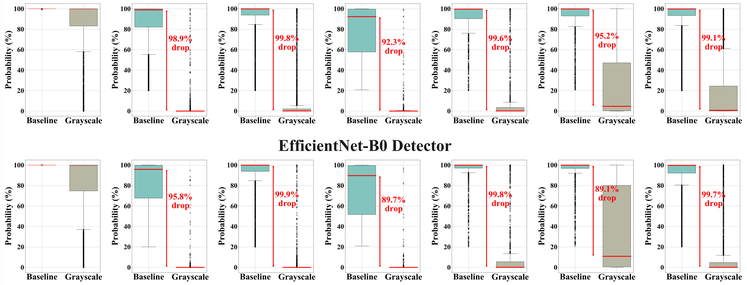
Principal idea : Instead of back-propagating using detector logits, **back-propagate from the maximum spatial relevance neuron** for **each *T-FF* independently.**



Color is a critical *T-FF* in universal detectors for counterfeit detection (Quantitative Studies)

Color Ablation Studies

Median Counterfeit probability analysis results based on Color Ablation (Grayscale)



Color ablation causes the median counterfeit probability to **drop by > 89% across all unseen GANs**

Statistical Test Results

Standard Detectors

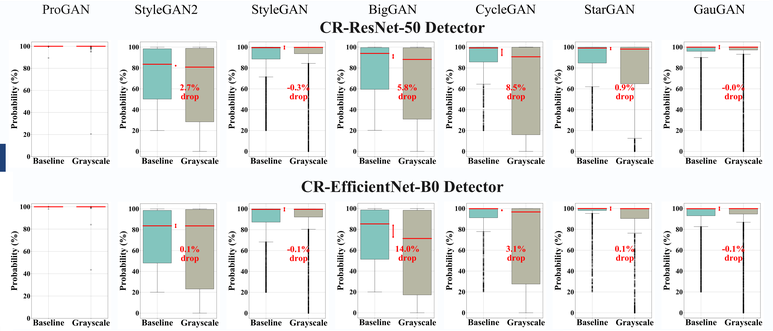
Color-Conditional (%)	ResNet-50	EfficientNet-B0
ProGAN	85.1	51.9
StyleGAN2	74.6	48.1
StyleGAN	73.7	40.7
BigGAN	68.4	40.7
CycleGAN	86.8	44.4
StarGAN	71.1	44.4
GauGAN	70.2	37.0

Color-Robust (CR) Detectors

Color-Conditional (%)	CR-ResNet-50	CR-EfficientNet-B0
ProGAN	**55.3**	**20.0**
StyleGAN2	**33.3**	**30.0**
StyleGAN	**48.2**	**20.0**
BigGAN	**31.6**	**10.0**
CycleGAN	**56.1**	**20.0**
StarGAN	**48.2**	**20.0**
GauGAN	**39.5**	**10.0**

Color Robust (CR) Universal Detectors

Idea : **Randomly remove color information** from samples during training (both real and counterfeits) to manoeuvre detectors to **learn *T-FF* that do not substantially rely on color information (Random Grayscaling with $p = 50\%$).**



References

Wang, Sheng-Yu, et al. "CNN-generated images are surprisingly easy to spot... for now." *CVPR*. 2020.

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *CVPR*. 2017.

Bach S, Binder A, Montavon G, Klauschen F, Müller KR, et al. (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE

Karras, Tero, et al. "Progressive Growing of GANs for Improved Quality, Stability, and Variation." *ICLR*. 2018.

Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *CVPR*. 2019.

Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." CVPR. 2020.

Brock, Andrew, Jeff Donahue, and Karen Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis." *ICLR*. 2018.

Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *CVPR*. 2017.

Choi, Yunjey, et al. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." *CVPR* 2018.

Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." *CVPR*. 2019.

Key Takeaways

Code & Pre-trained Models

- We propose a novel **Forensic Feature Relevance Statistic (FF-RS)** to **quantify & discover Transferable Forensic Features *(T-FF)*** in universal detectors for counterfeit detection.
- We qualitatively and quantitatively show that **color is a critical Transferable Forensic Feature *(T-FF)*** in universal detectors for counterfeit detection
- Based on our findings, we propose a **simple data augmentation** scheme to train **Color-Robust (CR)** universal detectors