

Discovering Transferable Forensic Features for CNN-generated Images Detection

Keshigeyan Chandrasegaran¹, Ngoc-Trung Tran¹, Alexander Binder^{2,3}, and Ngai-Man Cheung¹

¹ Singapore University of Technology and Design (SUTD)
 {keshigeyan, ngoctrung.tran, ngaiman.cheung}@sutd.edu.sg

² Singapore Institute of Technology (SIT)

³ University of Oslo (UIO)
 alexander.binder@singaporetech.edu.sg alexabin@uio.no

Abstract. Visual counterfeits ⁴ are increasingly causing an existential conundrum in mainstream media with rapid evolution in neural image synthesis methods. Though detection of such counterfeits has been a taxing problem in the image forensics community, a recent class of forensic detectors – *universal detectors* – are able to surprisingly spot counterfeit images regardless of generator architectures, loss functions, training datasets, and resolutions [87]. This intriguing property suggests the possible existence of *transferable forensic features (T-FF)* in *universal detectors*. In this work, we conduct the first analytical study to discover and understand *T-FF* in *universal detectors*. Our contributions are 2-fold: 1) We propose a novel *forensic feature relevance statistic (FF-RS)* to quantify and discover *T-FF* in *universal detectors* and, 2) Our qualitative and quantitative investigations uncover an unexpected finding: *color* is a critical *T-FF* in *universal detectors*. Code and models are available at <https://keshik6.github.io/transferable-forensic-features/>

1 Introduction

Visual counterfeits are increasingly causing an existential conundrum in mainstream media [21,70,1,26,53,61,33,32,74]. With rapid improvements in CNN-based generative modelling [30,39,38,94,66,19,10,62,97,45,3,81,96,82,83,48,86,44], detection of such counterfeits is increasingly becoming challenging and critical. Nevertheless, a recent class of forensic detectors known as *universal detectors* are able to surprisingly spot counterfeits regardless of generator architectures, loss functions, datasets and resolutions without any extensive adaptation [87]. i.e.: Publicly released ResNet-50 [35] universal detector by Wang *et al.* [87] trained only on ProGAN [37] counterfeits, surprisingly generalizes well to detect counterfeits from unseen GANs including StyleGAN2 [41], StyleGAN [40], BigGAN [10], CycleGAN [97], StarGAN [18] and GauGAN [62]. This intriguing cross-model forensic transfer property suggests the existence of *transferable forensic features (T-FF)* in *universal detectors*.

⁴ We refer to CNN-generated images as counterfeits throughout this paper

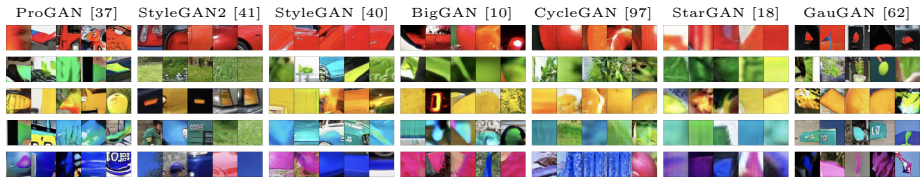


Fig. 1. Color is a critical *transferable forensic feature (T-FF)* in universal detectors: Large-scale study on visual interpretability of *T-FF* discovered through our proposed *forensic feature relevance statistic (FF-RS)*, reveal that color information is critical for cross-model forensic transfer. Each row represents a color-conditional *T-FF* and we show the LRP-max response regions for different GAN counterfeits for the publicly released ResNet-50 universal detector by Wang *et al.* [87]. This detector is trained with ProGAN [37] counterfeits [87] and cross-model forensic transfer is evaluated on unseen GANs. All counterfeits are obtained from the ForenSynths dataset [87]. The consistent color-conditional LRP-max response across all GANs for these *T-FF* clearly indicate that *color* is critical for cross-model forensic transfer in universal detectors. We further observe similar results using an EfficientNet-B0-based [78] universal detector following the exact training / test strategy proposed by Wang *et al.* [87] in Fig. 3. More visualizations are included in Supplementary G.

1.1 Transferable Forensic Features (T-FF) in Universal Detectors

This work is motivated by a profound and challenging thesis statement: *What transferable forensic features (T-FF) are used by universal detectors for counterfeit detection?* A more elemental representation of this thesis statement would be: given an image of a real car and a high fidelity synthetic car generated by an unseen GAN (i.e.: StyleGAN2 [41]), what *T-FF* are used by the universal detector, such that it detects the synthetic car as counterfeit accurately? Though Wang *et al.* [87] hypothesize that universal detectors may learn low-level CNN artifacts for detection, no qualitative / quantitative evidence is available in contemporary literature to understand *T-FF* in universal detectors. *Our work takes the first step towards discovering and understanding T-FF in universal detectors for counterfeit detection.* A foundational understanding on *T-FF* and their properties are of paramount importance to both image forensics research and image synthesis research. Understanding *T-FF* will allow to build robust forensic detectors and to devise techniques to improve image synthesis methods to avoid generation of forensic footprints.

1.2 Our contributions

Our work conducts the *first analytical study to discover and understand T-FF in universal detectors for counterfeit detection.* We begin our study by comprehensively demonstrating that input-space attribution – using 2 popular algorithms namely Guided-GradCAM [72] and LRP [5] – of universal detector decisions are not informative to discover *T-FF*. Next, we study the forensic feature space of universal detectors to discover *T-FF*. But investigating the feature space is an

extremely daunting task due to the sheer amount of feature maps present. i.e.: ResNet-50 [35] architecture contains approximately 27K feature maps. To tackle this challenging task, *we propose a novel forensic feature relevance statistic (FF-RS), to quantify and discover T-FF in universal detectors*. Our proposed FF-RS (ω) is a scalar which quantifies the ratio between positive forensic relevance of the feature map and the total unsigned relevance of the entire layer that contains the particular feature map. Using our proposed FF-RS (ω), we successfully discover *T-FF* in the publicly released ResNet-50 universal detector [87].

Next, to understand the discovered *T-FF*, *we introduce a novel pixel-wise explanation method based on maximum spatial Layer-wise Relevance Propagation response (LRP-max)*. Particularly we visualize the pixel-wise explanations of each discovered *T-FF* in universal detectors independently using LRP-max visualization method. Large-scale study on visual interpretability of *T-FF* reveal that color information is critical for cross-model forensic transfer. Further large-scale quantitative investigations using median counterfeits probability analysis and statistical tests on maximum spatial activation distributions based on color ablation show that *color* is a critical *T-FF* in universal detectors. Our findings are intriguing and new to the research community, as many contemporary image forensics works focus on frequency discrepancies between real and counterfeit images [24,25,92,12,71,42]. In summary, our contributions are as follows:

- We propose a novel *forensic feature relevance statistic (FF-RS)* to quantify and discover *transferable forensic features (T-FF)* in universal detectors for counterfeit detection.
- We qualitatively – using our proposed LRP-max visualization for feature map activations – and quantitatively – using median counterfeits probability analysis and statistical tests on maximum spatial activation distributions based on color ablation – show that *color* is a critical *transferable forensic feature (T-FF)* in universal detectors for counterfeit detection.

2 Related Work

Counterfeit detection. Recent works have studied counterfeit detection both in the RGB domain [67,54,20,92,60,85,87] and frequency domain [25,24,12,27,51]. Particularly, notable number of works have proposed to use hand-crafted features for counterfeit detection [25,24,12,60]. Using simple experiments, McCloskey *et al.* [56] showed that detection based on the frequency of over-exposed pixels can provide good discrimination between real images and counterfeits. Li *et al.* observed disparities between GAN images and real images in the residual domain of the chrominance color components [46]. Some recent works have also proposed methods to detect and attribute counterfeits to the generating architectures [91,55]. Anomaly detection techniques leveraging on pre-trained face recognition models have also been proposed [85].

Cross-model forensic transfer. Most counterfeit detection works do not focus on cross-model forensic transfer. Among the works that study forensic transfer,

Cozzolino *et al.* [20] and Zhang *et al.* [92] observed that counterfeit detectors generalized poorly during cross-model forensic transfer. In order to solve poor forensic transfer performance, Cozzolino *et al.* [20] proposed an autoencoder based adaptation framework to improve cross-model forensic transfer. The work by Wang *et al.* [87] was the first work to show that counterfeit detectors – universal detectors – can generalize well during cross-model forensic transfer without any re-training / fine-tuning / adaptation on the target samples suggesting the possible existence of *transferable forensic features*. Furthermore, Chai *et al.* [11] showed that patch-based detectors with limited receptive fields often perform better at detecting unseen counterfeits compared to full-image based detectors.

Interpretability methods. A number of interpretability methods in machine learning aim to summarize the relations which a model has learnt as a whole, such as PCA and t-SNE [63,52], or to explain single decisions of a neural network. The latter may follow different lines of questioning, such as identifying similar training samples in k-NN and prototype CNNs [49,14], finding modified samples such as pertinent negatives [23], or model-based uncertainty estimates [29]. One class of algorithms aims at computing input space attributions. This includes Shapley values [77,50,15] suitable for tabular data types, and methods for data types for which dropping a feature is not well defined, relying on modified gradients such as Guided Backprop [75], Layer-wise Relevance Propagation (LRP) [5], Guided-GradCAM [72], Full-Grad [76], and class-attention-mapping inspired research [22,84,36,28,59]. Bau *et al.* proposed frameworks for interpreting representations at the feature map level for classifiers [7] and GANs [8].

3 Dataset / Metrics

We use the ForenSynths dataset proposed by Wang *et al.* [87]. ForenSynths is the largest counterfeit benchmark dataset containing CNN-generated images from multiple generator architectures, datasets, loss functions and resolutions. In addition to ProGAN [37], we select 6 candidate GANs to comprehensively study cross-model forensic transfer in this work namely, StyleGAN2 [41], StyleGAN [40], BigGAN [10], CycleGAN [97], StarGAN [18] and GauGAN [62]. Following Wang *et al.* [87], we use AP (Average Precision) to measure cross-model forensic transfer of universal detectors. Particularly, we also show the accuracies for real and counterfeit images as we intend to understand counterfeit detection. For detector calibration, we follow [87] and use the oracle threshold obtained using geometric mean of sensitivity and specificity.

4 Discovering Transferable Forensic Features (T-FF)

4.1 Input-space attribution methods

Interpretable machine learning algorithms are useful exploratory tools to visualize neural networks’ decisions by input-space attribution [9,72,76,22,84,36,28,59]. We start from the following question: *Are interpretability methods suitable to discover T-FF in universal detectors?*

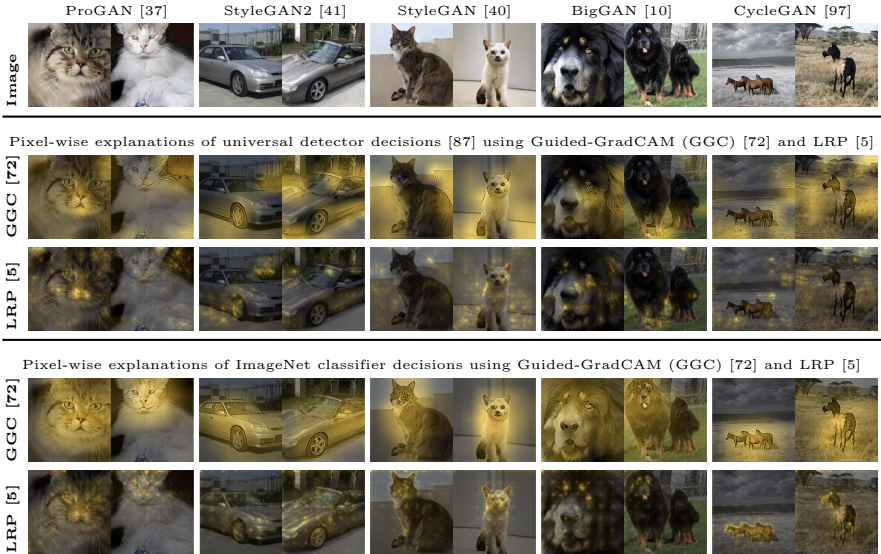


Fig. 2. Pixel-wise explanations of universal detector decisions are not informative to discover T -FF: We show pixel-wise explanations using Guided-GradCAM (GGC) (row 2) [72] and LRP (row 3) [5] for the ResNet-50 universal detector [87] for ProGAN [37], CycleGAN [97], StarGAN [18], BigGAN [10] and StyleGAN2 [41]. The universal detector predicts probability $p \geq 95\%$ for all counterfeit images shown above. All these counterfeits are obtained from ForenSynths dataset [87]. For LRP [5], we only show positive relevances. We also show pixel-wise explanations of ImageNet classifier decisions for the exact counterfeits using GGC (row 4) and LRP (row 5). This is shown as a control experiment to emphasize the significance of our observations. As one can clearly observe, pixel-wise explanations of universal detector decisions are not informative to discover T -FF (rows 2, 3) as the explanations appear to be random and not reveal any meaningful visual features used for counterfeit detection. Particularly, it remains unknown as to why the universal detector outputs high detection probability ($p \geq 95\%$) for these counterfeits. On the other hand, pixel-wise explanations of ImageNet classifier decisions produce meaningful results. i.e.: GGC (row 4) and LRP (row 5) explanation results for cat samples (columns 1, 2, 5, 6) show that ImageNet uses features such as eyes and whiskers to classify cats. This shows that interpretability techniques such as GGC and LRP are not informative to discover T -FF in universal detectors. In other words, we are unable to discover any forensic footprints based on pixel-wise explanations of universal detectors. More examples in Supplementary I.

We use 2 popular interpretability methods namely Guided-GradCAM [72] and LRP [5] to analyse the pixel-wise explanations of universal detector decisions. These methods were chosen due to their relatively low amount of gradient shattering noise [6]. We show the pixel-wise explanation results of ResNet-50 universal detector [87] decisions for ProGAN [37] and 4 GANs not used for training – CycleGAN [97], StarGAN [18], BigGAN [10] and StyleGAN2 [41]– in Fig. 2. As one can observe in Fig. 2, pixel-wise explanations of universal detector decisions are not informative to discover T -FF due to their focus on spatial localization. Particularly, we are unable to discover any forensic footprints based

Table 1. Sensitivity assessments using feature map dropout showing that our proposed *FF-RS* (ω) successfully quantifies and discovers *T-FF*: We show the results for the publicly released ResNet-50 universal detector [87] (top) and our own version of EfficientNet-B0 [78] universal detector (bottom) following the exact training and test strategy proposed in [87]. We show the AP, real and GAN image detection accuracies for baseline [87], top- k , random- k and low- k forensic feature dropout. The random- k experiments are repeated 5 times and average results are reported. Feature map dropout is performed by suppressing (zeroing out) the resulting activations of target feature maps (i.e.: top- k). We can clearly observe that feature map dropout of top- k corresponding to *T-FF* results in substantial drop in AP and GAN detection accuracies across ProGAN and all 6 unseen GANs [41,40,10,97,18,62] compared to baseline, random- k and low- k results. This is consistently seen in both ResNet-50 and EfficientNet-B0 universal detectors. This shows that our proposed *FF-RS* (ω) can successfully quantify and discover the *T-FF* in universal detectors. $k \approx 0.5\%$ of total feature maps. More details included in Supplementary D.

ResNet-50																						
		ProGAN [37]			StyleGAN2 [41]			StyleGAN [40]			BigGAN [10]			CycleGAN [97]			StarGAN [18]			GauGAN [62]		
$k = 114$		AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN
baseline [87]		100.	100.0	100.	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	1.3	56.6	90.6	13.7	55.4	86.3	18.3	61.2	91.4	17.4	72.6	89.4	35.9	71.0	95.0	18.8
random-k		100.	99.9	96.1	98.6	89.4	96.9	98.7	91.4	96.1	88.0	79.4	85.0	96.6	81.0	96.2	97.0	88.0	91.7	98.7	91.9	97.1
low-k		100.	100.	100.	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																						
		ProGAN [37]			StyleGAN2 [41]			StyleGAN [40]			BigGAN [10]			CycleGAN [97]			StarGAN [18]			GauGAN [62]		
$k = 27$		AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN
baseline [87]		100.	100.	100.	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	54.5	94.3	7.0	52.1	97.3	2.6	53.5	97.4	3.8	47.5	100.	0.0	50.0	100.	0.0	46.2	100.	0.0
random-k		100.	99.9	100.	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.	100.	100.	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

on pixel-wise explanations of universal detector decisions. This is consistently seen across both Guided-GradCAM [72] and LRP [5] methods. We remark that these observations do not indicate failure modes of Guided-GradCAM [72] or LRP [5] methods, but rather suggest that universal detectors are learning more complex *T-FF* that are not easily human-parsable.

4.2 Forensic Feature Space

Given that input-space attribution methods are not informative to discover *T-FF*, we study the feature space to discover *T-FF* in universal detectors for counterfeit detection. Particularly, we ask the question: which feature maps in universal detectors are responsible for cross-model forensic transfer? This is a challenging problem as it requires quantifying the importance of every feature map in universal detectors for counterfeit detection. The ResNet-50 universal detector [87] consists of approximately 27K intermediate feature maps.

Forensic feature relevance statistic (FF-RS). We propose a novel *FF-RS* (ω) to quantify the relevance of every feature map in universal detectors for counterfeit detection. Specifically, for feature map at layer l and channel c , $\omega(l_c)$ computes the forensic relevance of this feature map for counterfeit detection. We

describe the important design considerations and intuitions behind our proposed *FF-RS* (ω) below and include the pseudocode in Algorithm 1:

- We postulate the existence of a set of feature maps in universal detectors that are responsible for cross-model forensic transfer. In particular, we hypothesize that there is a set of *common transferable forensic feature maps* that mostly gets activated when passing counterfeits from ProGAN [37] and unseen GANs.
- Our proposed *FF-RS* (ω) is a scalar that quantifies the forensic relevance of every feature map. In particular, ω for a feature map quantifies the ratio between positive forensic relevance of the feature map and the total unsigned forensic relevance of the entire layer that contains the particular feature map. This is shown in Line 8 in Algorithm 1. For the numerator we are only interested in positive relevance, therefore use a max operation to select only positive relevance (identical to a ReLU operation).
- The relevance scores are calculated using LRP [5] (More details on LRP [5] in Supplementary A). This is shown in Line 5 in Algorithm 1 where $r_i(l, c, h, w)$ is the estimated relevance of the feature map at layer l , channel c at the spatial location h, w
- ω is calculated over large number of counterfeit images and is bounded between $[0, 1]$. i.e.: $\omega = 1$ indicates that the particular feature map is the most relevant forensic feature and $\omega = 0$ indicates vice versa.
- Finally we use ω to rank all the feature maps and identify the set of *T-FF*. We refer to this set as top-k in our experiments.

Experiments : Sensitivity assessments of discovered T-FF using algorithm 1. We perform rigorous sensitivity assessments using feature map dropout experiments to demonstrate that our proposed *FF-RS* (ω) is able to quantify and discover *T-FF*. Feature map dropout suppresses (zeroing out) the resulting activations of the target feature maps. Particularly, feature map dropout of *T-FF* should satisfy the following sensitivity conditions:

1. Significant reduction in overall AP across ProGAN [37] and all unseen GANs [41,40,10,97,18,62] indicating poor cross-model forensic transfer.
2. Significant reduction in GAN /counterfeit detection accuracies across ProGAN [37] and all unseen GANs [41,40,10,97,18,62] compared to real image detection accuracies as ω is calculated for counterfeits.

Test bed details. We use the ForenSynths test set [87]. ω is calculated using 1000 ProGAN [37] counterfeits (validation set). We use the following experiment codes:

- top- k : Set of T-FF discovered using *FF-RS* (ω)
- random- k : Set of random feature maps used as a control experiment.
- low- k : Set of low-ranked feature maps corresponding to extremely small values of ω , i.e.: $\omega \approx 0$.

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 .

```

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

```

Results. We show the results in Table 1 for ResNet-50 and EfficientNet-B0 universal detectors. We clearly observe that feature map dropout of top- k features corresponding to T -FF satisfies both sensitivity conditions above indicating that our proposed FF -RS (ω) is able to quantify and discover *transferable forensic features*. We also observe that feature map dropout of low- k (low-ranked) forensic features has little / no effect on cross-model forensic transfer which further adds merit to our proposed FF -RS (ω).

5 Understanding Transferable Forensic Features (T-FF)

Given the successful discovery of T -FF using our proposed FF -RS (ω), in this section, we ask the following question: what counterfeit properties are detected by this set of T -FF? Though Wang *et al.* [87] hypothesize that universal detectors may learn low-level CNN artifacts for cross-model forensic transfer, no evidence is available to understand as to what features in counterfeits are being detected during cross-model forensic transfer.

5.1 LRP-max explanations of T-FF

We approach this problem from a visual interpretability perspective. In this section, we introduce a novel pixel-wise explanation method for feature map activations based on maximum spatial Layer-wise Relevance Propagation response

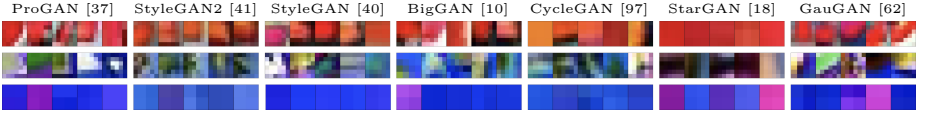


Fig. 3. Color is a critical T -FF in universal detectors: Large-scale study on visual interpretability of T -FF discovered through our proposed FF -RS (ω) reveal that color information is critical for cross-model forensic transfer. Each row represents a color-based T -FF and we show the LRP-max response regions for ProGAN and all 6 unseen GAN [41,40,10,97,18,62] counterfeits for our own version of EfficientNet-B0 [78] universal detector following the exact training / test strategy proposed by Wang *et al.* [87]. This detector is trained with ProGAN [37] counterfeits [87] and cross-model forensic transfer is evaluated on other unseen GANs. All counterfeits are obtained from the ForenSynths dataset [87]. The consistent color-conditional LRP-max response across all GANs for these T -FF clearly indicate that *color* is critical for cross-model forensic transfer in universal detectors. More visualizations are included in Supplementary G.

(LRP-max). The idea behind LRP-max is to independently visualize which pixels in the input space correspond to maximum spatial relevance scores for each T -FF. Particularly, instead of back-propagating using the detector logits, we back-propagate from the maximum spatial relevance neuron of each T -FF independently. LRP-max automatically extracts image regions for every T-FF and does not depend on external modules such as segmentation used in [7,8]. The pseudocode is included in 2.

Color is a critical T-FF in universal detectors: LRP-max visualizations of T -FF uncover the unexpected observation that a substantial amount of T -FF exhibits color-conditional activations. We show the LRP-max regions for ProGAN [37] and all unseen GANs [41,40,10,97,18,62] for ResNet-50 and EfficientNet-B0 universal detectors in Fig. 1 and 3 respectively. As one can observe, the consistent color-conditional LRP-max response across all GANs for these T-FF clearly indicate that color is critical for cross-model forensic transfer in universal detectors. This is notably surprising and observed for the first time in transferable image forensics research. In the next section, we conduct quantitative studies to rigorously verify that color is a critical T -FF in universal detectors.

5.2 Color Ablation Studies

In this section, we conduct 2 quantitative studies to show that *color* is a critical *transferable forensic feature* in universal detectors. Our studies measure the sensitivity of universal detectors before and after color ablation.

Study 1. We investigate the change in probability distribution of universal detectors when removing color information in counterfeits during cross-model forensic transfer. We specifically study the change in median counterfeit probability when removing color information (median is not sensitive to outliers). The results for both ResNet-50 and EfficientNet-B0 universal detectors are shown in Fig. 4. As one can clearly observe, color ablation causes the median probability predicted by the universal detector to drop by more than 89% across all un-

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:

$\hat{E}_{l_c}(x)$ 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 $\hat{E}_{l_c}(x).size()$ is (x_{height}, x_{width}) .

Every forensic feature map can be characterized by a unique set of l, c .

- 1 $z_{l_c}(x) \leftarrow LRP - FORWARD(M_{l_c}(x_i))$; */*(h, w) relevance scores*/*
 - 2 $h^*, w^* \leftarrow argmax(z_{l_c}(x))$; */*find index of max relevance*/*
 - 3 $z_{l_c}^{max}(x) \leftarrow z_{l_c}(x)[h^*, w^*]$; */*LRP-max response neuron*/*
 - 4 $E_{l_c}(x) \leftarrow LRP - BACKWARD(z_{l_c}^{max}(x))$; */*explain LRP-max neuron*/*
 - 5 $\hat{E}_{l_c}(x) \leftarrow \sum_{k=0}^3 (E_{l_c}(x)(k, x_{height}, x_{width}))$; */*spatial LRP-max*/*
 - 6 **return** $\hat{E}_{l_c}(x)$
-

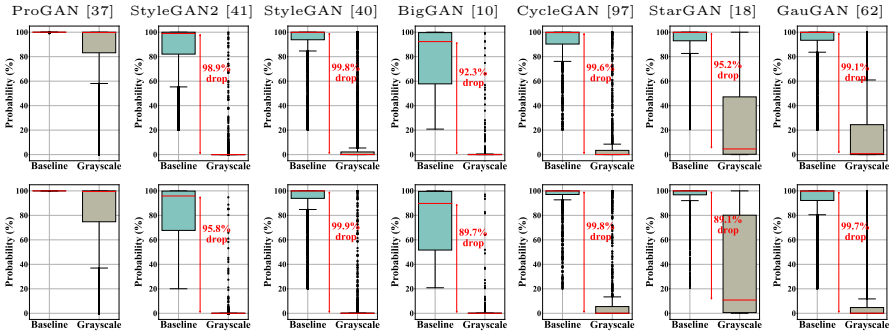


Fig. 4. *Color* is a critical T -FF in universal detectors: We show the box-whisker plots of probability (%) predicted by the universal detector for counterfeits before (Baseline) and after *color ablation* (Grayscale) for 7 GAN models. The red line in each box-plot shows the median probability. We show the results for the ResNet-50 universal detector [87] (top row) and our version of EfficientNet-B0 [78] universal detector following the exact training / test strategy proposed in [87] (bottom row). These detectors are trained with ProGAN counterfeits and cross-model forensic transfer is evaluated on other unseen GANs. All counterfeits are obtained from the ForenSynths dataset [87]. We clearly show that *color ablation* causes the median probability for counterfeits to drop by more than 89% across all unseen GANs. This is consistently seen across both universal detectors. These observations quantitatively show that *color* is a critical T -FF in universal detectors. AP and accuracies shown in Supplementary H.1.

seen GANs showing that *color* is a critical T -FF in universal detectors. This is observed in both ResNet-50 and EfficientNet-B0 universal detectors.

Study 2. In this study, we measure the percentage of T -FF that are color-conditional. Particularly, we conduct a statistical test to compare the maximum

Algorithm 3: Statistical test over maximum spatial activations for T - FF (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. T - FF set S **Output:** $p(l_c)$ where l, c indicates the layer and channel index of forensic feature maps. p indicates p -value of the statistical test.Every forensic feature map can be characterized by a unique set of l, c .

```

1 Set  $M$  to evaluation mode
2 for  $l', c'$  in  $S$  do
3    $A_b \leftarrow []$ ;                               /*store baseline counterfeits activations*/
4    $A_g \leftarrow []$ ;                               /*store grayscale counterfeits activations*/
5   for  $i$  in  $\{0, 1, \dots, n\}$  do
6      $a_b \leftarrow GLOBAL\_MAXPOOL(M_{l_c}(x_i))$ ;           /*baseline*/
7      $a_g \leftarrow GLOBAL\_MAXPOOL(M_{l_c}(grayscale(x_i)))$ ; /*grayscale*/
8      $A_b.append(a_b)$ 
9      $A_g.append(a_g)$ 
10  end
11   $p(l'_{c'}) \leftarrow MEDIAN - TEST(A_b, A_g)$ ;           /*median test*/
12 end
13 return  $p(l_c)$ 

```

globally pooled spatial activation distributions of each T - FF before and after color ablation. The intuition is that with color ablation, color-conditional T - FF will produce lower amount of activations for the same sample and we perform a hypothesis test to measure whether the maximum spatial activation distributions are statistically different before (Baseline) and after color ablation (Grayscale). Particularly, we use Mood's median test (non-parametric, low-power) with a significance level of $\alpha = 0.05$ in our study. The pseudocode is shown in Algorithm 3. The results for ResNet-50 and EfficientNet-B0 universal detectors are shown in Table 2 (rows 1, 2). Our results show that substantial amount of T - FF in universal detectors are color-conditional indicating that color is a critical T - FF . We also show the maximum spatial activation distributions for several color-conditional T - FF for ResNet-50 and EfficientNet-B0 universal detectors in Fig. 6. As one can observe, maximum spatial activations are suppressed for these T - FF across ProGAN [37] and all unseen GANs [41,40,10,97,18,62] when removing color information. This clearly suggests that these T - FF are color-conditional.

6 Applications : Color-Robust (CR) Universal Detectors

Reliance on substantial amount of color information for cross-model forensic transfer exposes universal detectors to attacks via color-ablated counterfeits. This is particularly unfavourable. In this section, we propose a data augmenta-

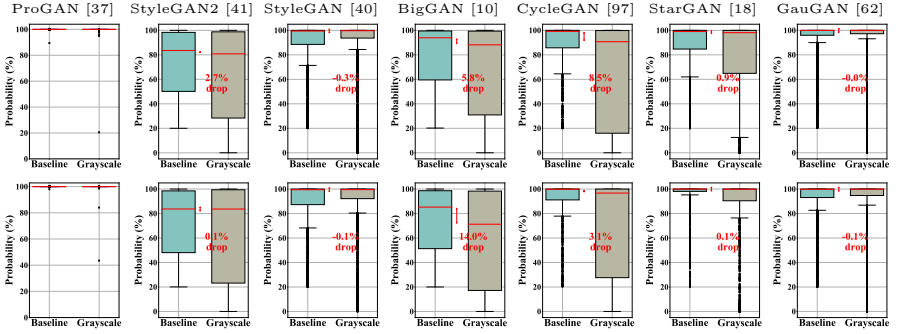


Fig. 5. *CR-universal detectors trained using our proposed data augmentation scheme (Sec. 6) are more robust to color ablation during cross-model forensic transfer:* These universal detectors are trained with data augmentation where color is ablated 50% of the time during training. This ensures that T -FF do not substantially rely on color information. We show the box-whisker plots of probability (%) predicted by the CR-universal detectors for counterfeits before (Baseline) and after *color ablation* (Grayscale) for 7 GAN models. The red line in each box-plot shows the median probability. We show the results for the ResNet-50 CR-universal detector [87] (top row) and EfficientNet-B0 [78] CR-universal detector (bottom row). We clearly observe that the median probability for counterfeits have similar values (compared to Fig. 4) before and after color ablation indicating CR-universal detectors are more robust to color-ablated counterfeit attacks. AP and accuracies shown in Supplementary H.2.

Table 2. Median Test Results. ① *Significant amount of T -FF are color-conditional (rows 1, 2):* We show the percentage(%) of color-conditional T -FF in ResNet-50 and EfficientNet-B0 universal detectors measured using Mood’s median test. We show the results for ProGAN [37] and all 6 unseen GANs [41,40,10,97,18,62]. Particularly, we consider a T -FF to be color conditional if the p -value of the median test is less than the significance level of $\alpha = 0.05$. As one can clearly observe, significant amount of T -FF are color-conditional. This quantitatively shows that color is a critical T -FF in universal detectors. ② *CR-universal detectors have lower amount of color-conditional T -FF (rows 3,4):* We clearly observe that training universal detectors using our proposed data augmentation scheme (Sec 6) results in detectors that contain noticeably lower amount of color-conditional T -FF.

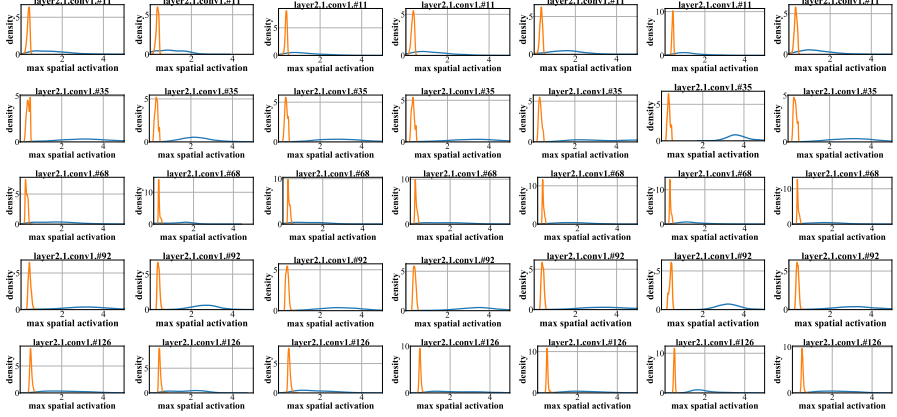
% Color-conditional	ProGAN [37]	StyleGAN2 [41]	StyleGAN [40]	BigGAN [10]	CycleGAN [97]	StarGAN [18]	GauGAN [62]
ResNet-50	85.1	74.6	73.7	68.4	86.8	71.1	70.2
Efficient-B0	51.9	48.1	40.7	40.7	44.4	44.4	37.0
CR-ResNet-50	55.3	33.3	48.2	31.6	56.1	48.2	39.5
CR-EfficientNet-B0	20.0	30.0	20.0	10.0	20.0	20.0	10.0

tion scheme to build Color-Robust (CR) universal detectors that do not substantially rely on color information for cross-model forensic transfer. The crux of the idea is to randomly remove color information from samples during training (both for real and counterfeit images). Particularly, we perform random Grayscale during training with 50% probability to maneuver universal detectors to learn T -FF that do not substantially rely on color information.

Results. Median probability analysis results for ResNet-50 and EfficientNet-B0 CR-universal detectors are shown in Fig. 5. We clearly observe that with our

ProGAN [37] StyleGAN2 [41] StyleGAN [40] BigGAN [10] CycleGAN [97] StarGAN [18] GauGAN [62]

ResNet-50



EfficientNet-B0

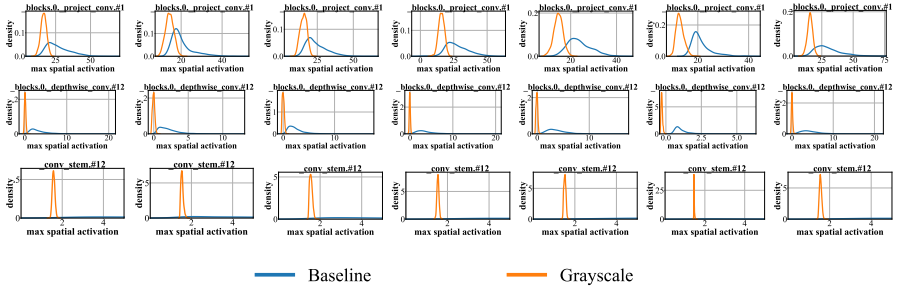


Fig. 6. Color-conditional T -FF in ResNet-50 and EfficientNet-B0: Each row represents a color-conditional T -FF. These are the exact same T -FF shown in Fig. 1 (ResNet-50) and Fig. 3 (EfficientNet-B0). We show the maximum spatial activation distributions for 7 GAN models before (Baseline) and after color ablation (Grayscale). We remark that for each counterfeit in the ForenSynths dataset [87], we apply global max pooling to the specific T -FF to obtain a *maximum spatial activation* value (scalar). We can clearly observe that these T -FF are producing noticeably lower spatial activations (max) for the same set of counterfeits after removing color information. This clearly indicates that these T -FF are color-conditional.

proposed data augmentation scheme, CR-universal detectors are more robust to color ablation during cross-model forensic transfer indicating that they learn T -FF that do not substantially rely on color information. We further show the percentage of color-conditional T -FF in CR-ResNet-50 and CR-EfficientNet-B0 in Table 2 (rows 3, 4), quantitatively showing that CR-universal detectors learn substantially lower amount of color-conditional T -FF.

T-FF in CR-Universal Detectors. We further discover T -FF in CR-universal detectors using our proposed FF -RS (ω). We show LRP-max visualization of T -FF in CR-ResNet-50 in Supplementary Fig. H.1. These T -FF largely correspond to patterns / artifacts (i.e.: wheels). We emphasize that our proposed method can identify different types of T -FF in addition to color.

7 Discussion and Conclusion

We conducted the *first analytical study to discover and understand transferable forensic features (T-FF) in universal detectors*. Our first set of investigations demonstrated that input-space attribution methods such as Guided-GradCAM [72] and LRP [5] are not informative to discover *T-FF* (Sec 4.1). In light of these observations, we study the forensic feature space of universal detectors. Particularly, we propose a novel *forensic feature relevance statistic (FF-RS)* to quantify and discover *T-FF* in universal detectors. Rigorous sensitivity assessments using feature map dropout convincingly show that our proposed FF-RS (ω) is able to successfully quantify and discover *T-FF* (Sec 4.2).

Further investigations on *T-FF* uncover an unexpected finding: *color* is a critical *T-FF* in universal detectors. We show this critical finding qualitatively using our proposed LRP-max visualization of discovered *T-FF* (Sec 5.1). Further we validate this finding quantitatively using median counterfeit probability analysis and statistical tests on maximum spatial activation distributions of *T-FF* based on color ablation (Sec 5.2). i.e.: We showed that $\approx 85\%$ of *T-FF* are color-conditional in the publicly released ResNet-50 universal detector [87]. Finally, we propose a simple data augmentation scheme to train Color-Robust (CR) universal detectors (Sec 6). We remark that color is not the only *T-FF*, but it is a critical *T-FF* in universal detectors. We also discuss computational complexity of FF-RS (ω) and LRP-max in Supplementary B. A natural question would be why is color a critical *T-FF*. Though this is not a straight-forward question to answer, we provide our perspective: Color distribution of real images is non-uniform, and we hypothesize that most GANs struggle to capture the diverse, multi-modal color distribution of real images. i.e.: low-density color regions. This may result in noticeable discrepancies between real and GAN images (counterfeits) in the color space which can be used as *T-FF* to detect counterfeits. To conclude, through this work we discover and understand *T-FF* in universal detectors for counterfeit detection, and hope that our contributions will inspire further research in image forensics and image synthesis methods.

Limitations / Broader Impact. With deepfakes-in-the-wild being generated using diverse techniques in addition to GAN-based methods including shallow methods (i.e.: Photoshop) and face-swapping frameworks (i.e.: DeepFaceLab [64]), studying transferable forensic features in such synthesis methods are essential to build robust general-purpose image forensics detectors. With increasing usage of machine learning methods in proliferating mis- and disinformation, we hope that our discovery on transferable forensic features can open-up more plausible research directions to combat the fight against visual disinformation.

Acknowledgements. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No.: AISG2-RP-2021-021; AISG Award No.: AISG-100E2018-005). This project is also supported by SUTD project PIE-SGP-AI-2018-01. Alexander Binder was supported by the SFI Visual Intelligence, project no. 309439 of the Research Council of Norway.

References

1. Synthetic media: How deepfakes could soon change our world, <https://www.cbsnews.com/news/deepfake-artificial-intelligence-60-minutes-2022-07-31/>
2. Abdollahzadeh, M., Malekzadeh, T., Cheung, N.M.: Revisit Multimodal Meta-Learning through the Lens of Multi-Task Learning. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021), <https://openreview.net/forum?id=V5prUH0rOP4>
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN (2017)
4. Arras, L., Osman, A., Müller, K.R., Samek, W.: Evaluating Recurrent Neural Network Explanations. In: ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 113–126. ACL (2019)
5. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
6. Balduzzi, D., Frean, M., Leary, L., Lewis, J.P., Ma, K.W., McWilliams, B.: The Shattered Gradients Problem: If ResNets are the answer, then what is the question? In: International Conference on Machine Learning (ICML). PMLR, vol. 70, pp. 342–350. PMLR (2017)
7. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6541–6549 (2017)
8. Bau, D., Zhu, J.Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A.: GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. In: International Conference on Learning Representations (2018)
9. Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R., Samek, W.: Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers. In: Villa, A.E., Masulli, P., Pons Rivero, A.J. (eds.) Artificial Neural Networks and Machine Learning – ICANN 2016. pp. 63–71. Springer International Publishing, Cham (2016)
10. Brock, A., Donahue, J., Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis. In: International Conference on Learning Representations (2019), <https://openreview.net/forum?id=B1xsqj09Fm>
11. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? understanding properties that generalize. In: European conference on computer vision. pp. 103–120. Springer (2020)
12. Chandrasegaran, K., Tran, N.T., Cheung, N.M.: A Closer Look at Fourier Spectrum Discrepancies for CNN-Generated Images Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7200–7209 (June 2021)
13. Chandrasegaran, K., Tran, N.T., Zhao, Y., Cheung, N.M.: Revisiting Label Smoothing and Knowledge Distillation Compatibility: What was Missing? In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 2890–2916. PMLR (17–23 Jul 2022)
14. Chen, C., Li, O., Tao, C., Barnett, A.J., Su, J., Rudin, C.: This Looks like That: Deep Learning for Interpretable Image Recognition, year = 2019. Curran Associates Inc., Red Hook, NY, USA

15. Chen, J., Song, L., Wainwright, M.J., Jordan, M.I.: L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. In: International Conference on Learning Representations (2019), <https://openreview.net/forum?id=S1E3Ko09F7>
16. Chen, L., Zhang, Y., Song, Y., Liu, L., Wang, J.: Self-supervised Learning of Adversarial Examples: Towards Good Generalizations for DeepFake Detections. In: CVPR (2022)
17. Choi, K., Grover, A., Singh, T., Shu, R., Ermon, S.: Fair generative modeling via weak supervision. In: International Conference on Machine Learning. pp. 1887–1898. PMLR (2020)
18. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
19. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: StarGAN v2: Diverse Image Synthesis for Multiple Domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
20. Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., Verdoliva, L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510 (2018)
21. Dan, V., Paris, B., Donovan, J., Hamelers, M., Roozenbeek, J., van der Linden, S., von Sikorski, C.: Visual Mis- and Disinformation, Social Media, and Democracy. *Journalism & Mass Communication Quarterly* **98**(3), 641–664 (2021). <https://doi.org/10.1177/10776990211035395>, <https://doi.org/10.1177/10776990211035395>
22. Desai, S.S., Ramaswamy, H.G.: Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) pp. 972–980 (2020)
23. Dhurandhar, A., Chen, P.Y., Luss, R., Tu, C.C., Ting, P., Shanmugam, K., Das, P.: Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 590–601. NIPS’18, Curran Associates Inc., Red Hook, NY, USA (2018)
24. Durall, R., Keuper, M., Keuper, J.: Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
25. Dzanic, T., Shah, K., Witherden, F.: Fourier Spectrum Discrepancies in Deep Network Generated Images. In: Thirty-fourth Annual Conference on Neural Information Processing Systems (NeurIPS) (December 2020)
26. Foley, J.: 14 deepfake examples that terrified and amused the internet (Apr 2022), <https://www.creativebloq.com/features/deepfake-examples>
27. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning. pp. 3247–3258. PMLR (2020)
28. Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., Li, B.: Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. In: BMVC (2020)
29. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Balcan, M.F., Weinberger, K.Q. (eds.)

- Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016), <https://proceedings.mlr.press/v48/gal16.html>
30. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*. vol. 27, pp. 2672–2680. Curran Associates, Inc. (2014), <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>
 31. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don’t lie: A generalisable and robust approach to face forgery detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5039–5049 (2021)
 32. Hao, K., Heaven, W.D.: The year deepfakes went mainstream (Dec 2020), <https://www.technologyreview.com/2020/12/24/1015380/best-ai-deepfakes-of-2020/>
 33. Harrison, E.: Shockingly realistic Tom Cruise deepfakes go viral on TikTok (Feb 2021), <https://www.independent.co.uk/arts-entertainment/films/news/tom-cruise-deepfake-tiktok-video-b1808000.html>
 34. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 9729–9738 (2020)
 35. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016)
 36. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: LayerCAM: Exploring Hierarchical Class Activation Maps For Localization. *IEEE Transactions on Image Processing* (2021)
 37. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: *International Conference on Learning Representations* (2018), <https://openreview.net/forum?id=Hk99zCeAb>
 38. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training Generative Adversarial Networks with Limited Data. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) *Advances in Neural Information Processing Systems*. vol. 33, pp. 12104–12114. Curran Associates, Inc. (2020), <https://proceedings.neurips.cc/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf>
 39. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* **34**, 852–863 (2021)
 40. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019)
 41. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and Improving the Image Quality of StyleGAN. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020)
 42. Khayatkhoei, M., Elgammal, A.: Spatial Frequency Bias in Convolutional Generative Adversarial Networks (Oct 2020)
 43. Kim, M., Tariq, S., Woo, S.S.: Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 1001–1012 (2021)

44. Koh, J.Y., Lee, H., Yang, Y., Baldrige, J., Anderson, P.: Pathdreamer: A World Model for Indoor Navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14738–14748 (October 2021)
45. Lee, K.S., Tran, N.T., Cheung, N.M.: Infomax-GAN: Improved adversarial image generation via information maximization and contrastive learning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3942–3952 (2021)
46. Li, H., Li, B., Tan, S., Huang, J.: Identification of deep network generated images using disparities in color components. *Signal Processing* **174**, 107616 (2020)
47. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face X-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5001–5010 (2020)
48. Lim, S.K., Loo, Y., Tran, N.T., Cheung, N.M., Roig, G., Elovici, Y.: DOPING: Generative Data Augmentation for Unsupervised Anomaly Detection with GAN. In: 18th IEEE International Conference on Data Mining, ICDM 2018. pp. 1122–1127. Institute of Electrical and Electronics Engineers Inc. (2018)
49. Lloyd, S.: Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**(2), 129–137 (1982). <https://doi.org/10.1109/TIT.1982.1056489>
50. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc. (2017), <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>
51. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16317–16326 (2021)
52. van der Maaten, L., Hinton, G.: Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**(86), 2579–2605 (2008), <http://jmlr.org/papers/v9/vandermaten08a.html>
53. Mahmud, A.H.: Deep dive into deepfakes: Frighteningly real and sometimes used for the wrong things (Oct 2021), <https://www.channelnewsasia.com/singapore/deepfakes-ai-security-threat-face-swapping-2252161>
54. Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L.: Detection of GAN-Generated Fake Images over Social Networks. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 384–389 (2018). <https://doi.org/10.1109/MIPR.2018.00084>
55. Marra, F., Gragnaniello, D., Verdoliva, L., Poggi, G.: Do GANs leave artificial fingerprints? In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 506–511. IEEE (2019)
56. McCloskey, S., Albright, M.: Detecting GAN-generated imagery using saturation cues. In: 2019 IEEE international conference on image processing (ICIP). pp. 4584–4588. IEEE (2019)
57. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-Wise Relevance Propagation: An Overview, pp. 193–209. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_10, https://doi.org/10.1007/978-3-030-28954-6_10
58. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition* **65**, 211–222 (2017). <https://doi.org/https://doi.org/10.1016/j.patcog.2016.11.008>, <https://www.sciencedirect.com/science/article/pii/S0031320316303582>

59. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2020)
60. Nataraj, L., Mohammed, T.M., Manjunath, B., Chandrasekaran, S., Flenner, A., Bappy, J.H., Roy-Chowdhury, A.K.: Detecting GAN generated fake images using co-occurrence matrices. *Electronic Imaging* **2019**(5), 532–1 (2019)
61. News, C.: Synthetic media: How deepfakes could soon change our world (Oct 2021), <https://www.cbsnews.com/news/deepfake-artificial-intelligence-60-minutes-2021-10-10/>
62. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)
63. Pearson, K.: LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572 (1901). <https://doi.org/10.1080/14786440109462720>
64. Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C.S., RP, L., Jiang, J., et al.: DeepFaceLab: Integrated, flexible and extensible face-swapping framework. arXiv preprint arXiv:2005.05535 (2020)
65. Pörner, N., Schütze, H., Roth, B.: Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In: Annual Meeting of the Association for Computational Linguistics (ACL). pp. 340–350. ACL (2018)
66. Razavi, A., van den Oord, A., Vinyals, O.: Generating Diverse High-Fidelity Images with VQ-VAE-2. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 32, pp. 14866–14876. Curran Associates, Inc. (2019), <https://proceedings.neurips.cc/paper/2019/file/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Paper.pdf>
67. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11 (2019)
68. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems* **28**(11), 2660–2673 (2017)
69. Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., Varshney, K.R.: Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development* **63**(4/5), 3:1–3:9 (2019). <https://doi.org/10.1147/JRD.2019.2945519>
70. SCHICK, N.: Deepfakes: The coming infocalypse. GRAND CENTRAL PUB (2021)
71. Schwarz, K., Liao, Y., Geiger, A.: On the Frequency Bias of Generative Models. *Advances in Neural Information Processing Systems* **34** (2021)
72. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
73. Shiohara, K., Yamasaki, T.: Detecting Deepfakes with Self-Blended Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18720–18729 (2022)
74. Simonite, T.: What Happened to the Deepfake Threat to the Election? (Nov 2020), <https://www.wired.com/story/what-happened-deepfake-threat-election/>

75. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for Simplicity: The All Convolutional Net. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015), <http://arxiv.org/abs/1412.6806>
76. Srinivas, S., Fleuret, F.: Full-gradient representation for neural network visualization. *Advances in neural information processing systems* **32** (2019)
77. Strumbelj, E., Kononenko, I.: An Efficient Explanation of Individual Classifications Using Game Theory. *J. Mach. Learn. Res.* **11**, 1–18 (mar 2010)
78. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)
79. Tan, S., Shen, Y., Zhou, B.: Improving the fairness of deep generative models without retraining. arXiv preprint arXiv:2012.04842 (2020)
80. Teo, C.T., Cheung, N.M.: Measuring fairness in generative models. arXiv preprint arXiv:2107.07754 (2021)
81. Tran, N.T., Bui, T.A., Cheung, N.: Dist-GAN: An Improved GAN Using Distance Constraints. In: ECCV (2018)
82. Tran, N.T., Tran, V.H., Nguyen, B.N., Yang, L., Cheung, N.M.M.: Self-supervised GAN: Analysis and Improvement with Multi-class Minimax Game. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 32. Curran Associates, Inc. (2019), <https://proceedings.neurips.cc/paper/2019/file/d04cb95ba2bea9fd2f0daa8945d70f11-Paper.pdf>
83. Tran, N.T., Tran, V.H., Nguyen, N.B., Nguyen, T.K., Cheung, N.M.: On data augmentation for GAN training. *IEEE Transactions on Image Processing* **30**, 1882–1897 (2021)
84. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. pp. 24–25 (2020)
85. Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., Liu, Y.: FakeSpotter: a simple yet robust baseline for spotting AI-synthesized fake faces. In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. pp. 3444–3451 (2021)
86. Wang, S.Y., Bau, D., Zhu, J.Y.: Rewriting Geometric Rules of a GAN. *ACM Trans. Graph.* **41**(4) (jul 2022). <https://doi.org/10.1145/3528223.3530065>, <https://doi.org/10.1145/3528223.3530065>
87. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020)
88. Xu, D., Yuan, S., Zhang, L., Wu, X.: FairGAN: Fairness-aware generative adversarial networks. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 570–575. IEEE (2018)
89. Yeom, S.K., Seegerer, P., Lapuschkin, S., Binder, A., Wiedemann, S., Müller, K.R., Samek, W.: Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recognition* **115**, 107899 (2021). <https://doi.org/https://doi.org/10.1016/j.patcog.2021.107899>, <https://www.sciencedirect.com/science/article/pii/S0031320321000868>

90. Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J.: LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. arXiv preprint arXiv:1506.03365 (2015)
91. Yu, N., Davis, L.S., Fritz, M.: Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7556–7566 (2019)
92. Zhang, X., Karaman, S., Chang, S.: Detecting and Simulating Artifacts in GAN Fake Images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6 (2019). <https://doi.org/10.1109/WIFS47025.2019.9035107>
93. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2185–2194 (2021)
94. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable Augmentation for Data-Efficient GAN Training. In: Conference on Neural Information Processing Systems (NeurIPS) (2020)
95. Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W.: Learning self-consistency for deepfake detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15023–15033 (2021)
96. Zhao, Y., Ding, H., Huang, H., Cheung, N.M.: A Closer Look at Few-shot Image Generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9140–9150 (2022)
97. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)

Supplementary Materials

Contents

This Supplementary provides additional experiments, analysis, discussion and code / reproducibility details to further support our findings. The Supplementary materials are organized as follows:

- Section A: A brief overview of the LRP-algorithm used
- Section B: Computational complexity of $FF-RS(\omega)$ / LRP-max.
- Section C: Non Color-conditional $T-FF$
- Section D: k hyper-parameter in top- k for $T-FF$
- Section E: Cross-model forensic transfer using BigGAN [10] pre-training dataset
- Section F: Is the performance degrade in universal detectors due to unseen corruptions (OOD)?
- Section G: Color-conditional $T-FF$ (Additional Results)
- Section H: CR-Universal Detectors (Additional Results)
- Section I: Pixel-wise explanations are not informative to discover $T-FF$ (Additional Results)
- Section J: Research Reproducibility / Code Details
- Section K: Future Work: Can we identify globally relevant channels for counterfeit detection in a Generator?

A A brief overview of the LRP-algorithm used

Layer-wise relevance propagation (LRP) [5] is a modified-gradient type algorithm for backward passes in neural networks and other models. LRP is based on the idea of replacing the partial derivatives, which are usually flowing back along the edges of a graph, by terms derived from Taylor decompositions for single layers [58] of the network. While the ϵ -LRP-rule is similar to gradient-times-input, other rules such as the β -rule [57] result in explanations which exhibit visually low noise and are robust to gradient shattering effects [6] common in deep neural networks due to its normalization properties. Consider a neuron y with inputs x_i , weights w_i , and a relevance score being already computed for its output being R_y . The relevance score R_y is the analogue for the total derivative $\frac{dz}{dy}$ in conventional backpropagation started at output logits, however computed using LRP. Then the relevance score for the input x_i according to the $\beta = 0$ -rule is given as

$$R_i = R_y \frac{(w_i x_i)_+}{\sum_k (w_k x_k)_+} \quad (1)$$

where $(\cdot)_+$ is the positive part. This measures the proportion of the positive part of the weighted input $(w_i x_i)_+$ for the input neuron i relative to the positive

weighted inputs from all inputs used to compute the value of neuron y . Therefore it redistributes relevance from an output to the inputs proportional to this fraction and proportional to the relevance R_y of the output neuron. We used the $\beta = 0$ -rule for all convolution layers and the ϵ -rule for the top-most fully connected layer. Before applying LRP, we fuse batchnorm layers into convolution layers and reset the batchnorm layers. The backpropagation in the resetted batchnorm layers uses the identity. Technically the base LRP algorithm is implemented in PyTorch as custom static autograd functions. This results for convolution layers in relevance scores having a shape of $(1, C, H, W)$ in the gradient field.

LRP scores computed in the input space of neural networks have been shown to perform well on metrics regarding the ordering of input space regions according to the computed explanation scores and the correlation of this ordering to changes in model output logits [68,65,4] when modifying the highest scoring regions.

B Computational Complexity of *FF-RS* (ω) / LRP-max

Both *FF-RS* (ω) and LRP-max require an additional forward and backward pass during computation. We emphasize that *FF-RS* (ω) and LRP-max are not used during training, and are only used for analysis / interpretability. Therefore, computational overhead is not substantial. All our experiments were performed using a single Nvidia RTX 3090 GPU.

C Non Color-conditional T-FF

There are a few *T-FF* that are not color-conditional. In this section, we show *non* color-conditional T-FF. We show LRP-max response image regions for ResNet-50 and EfficientNet-B0 in Fig. C.1 and C.3 respectively. We further show the maximum spatial activation distributions before and after color ablation for ResNet-50 and EfficientNet-B0 in Fig. C.2 and C.4 respectively. As one can observe using LRP-max response image regions, these *non* color-conditional *T-FF* contain frequency / texture artifacts. The maximum spatial activation distributions clearly show that these *non* color-conditional *T-FF* produce identical / similar distributions before and after color ablation.

D k hyper-parameter in top- k for *T-FF*

In this section, we include more discussion regarding the k hyper-parameter in top- k . We show that as we increase k , AP and GAN detection accuracies drop across ProGAN [37] and all unseen GANs [41,40,10,97,18,62]. For our analysis, we identify the *smallest* k with a substantial drop in cross-model forensic transfer as indicated by AP and GAN detection accuracies. The results for ResNet-50 and EfficientNet-B0 detectors are shown in Table D.1

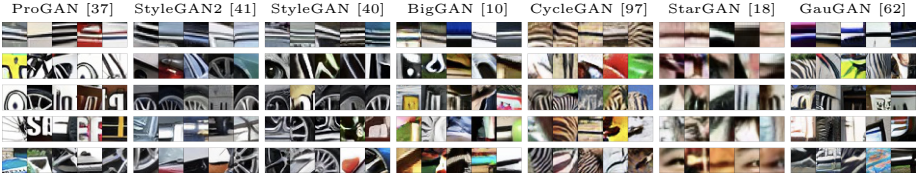


Fig. C.1. T -FF that are *not* color-conditional in ResNet-50 Universal detector: We show the LRP-max response regions for 5 *non* color-conditional T-FF for ProGAN [37] and all 6 unseen GANs [41,40,10,97,18,62]. Each row represents a *non* color-conditional T-FF. We emphasize that T -FF are discovered using our proposed *forensic feature relevance statistic* (FF-RS). This detector is trained with ProGAN [37] counterfeits [87] and cross-model forensic transfer is evaluated on other unseen GANs. All counterfeits are obtained from the ForenSynths dataset [87]. Visual inspection of LRP-max regions of *non* color-conditional T -FF shows frequency / texture artifacts. i.e.: rapid changes in pixel intensities. This shows that the universal detector also uses frequency / texture artifacts for cross-model transfer although *color* is a *critical* T -FF as $\approx 85\%$ of T -FF are color-conditional. We emphasize that our proposed method is capable of identifying different types of T -FF (i.e.: frequency / texture artifacts).

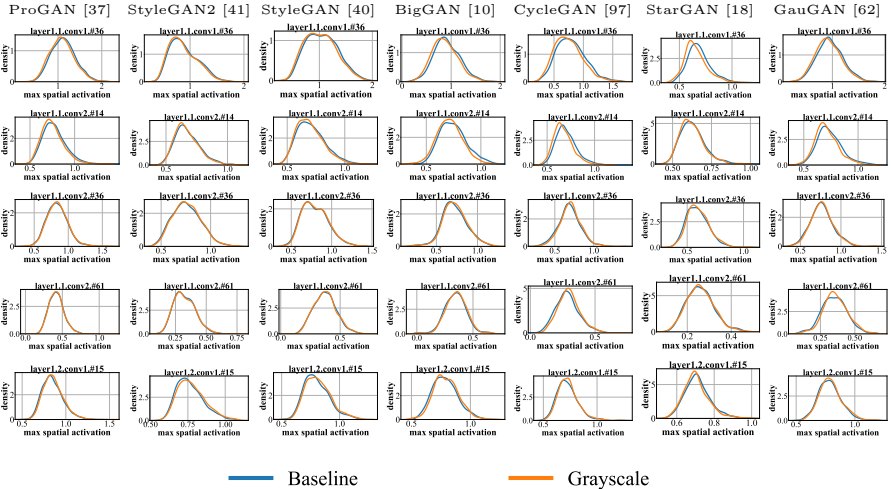


Fig. C.2. *Non Color-conditional* T -FF in ResNet-50: Each row represents a *non* color-conditional T -FF (exact same T -FF as shown in Fig. C.1), and we show the maximum spatial activation distributions for ProGAN [37], StyleGAN2 [41], StyleGAN [40], BigGAN [10], CycleGAN [97], StarGAN [18] and GauGAN [62] counterfeits before (Baseline) and after color ablation (Grayscale). We remark that for each counterfeit in the ForenSynths dataset [87], we apply global max pooling to the specific T-FF to obtain a *maximum spatial activation* value (scalar). We can clearly observe that these T -FF are producing identical / similar spatial activations (max) for the same set of counterfeits after removing color information which demonstrates that these T -FF do not respond to color information. This clearly indicates that these T -FF are *not* color-conditional (Confirmed by Mood’s median test).

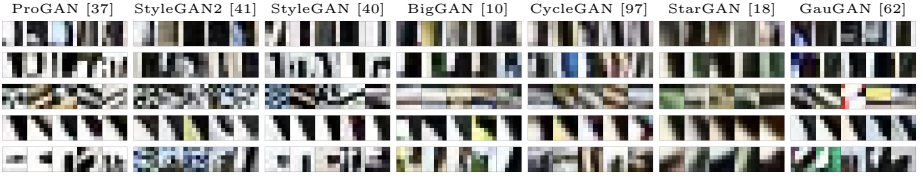


Fig. C.3. T -FF that are *not* color-conditional in EfficientNet-B0 Universal detector: We show the LRP-max response regions for 5 *non* color-conditional T-FF for ProGAN [37] and all 6 unseen GANs [41,40,10,97,18,62]. Each row represents a *non* color-conditional T-FF. We emphasize that T -FF are discovered using our proposed *forensic feature relevance statistic (FF-RS)*. This detector is trained with ProGAN [37] counterfeits [87] and cross-model forensic transfer is evaluated on other unseen GANs. All counterfeits are obtained from the ForenSynths dataset [87]. Visual inspection of LRP-max regions of *non* color-conditional T -FF shows frequency / texture artifacts. i.e.: rapid changes in pixel intensities. This shows that the universal detector also uses frequency / texture artifacts for cross-model transfer although *color is a critical T-FF* as $\approx 52\%$ of T -FF are color-conditional. We emphasize that our proposed method is capable of identifying different types of T -FF (i.e.: frequency / texture artifacts).

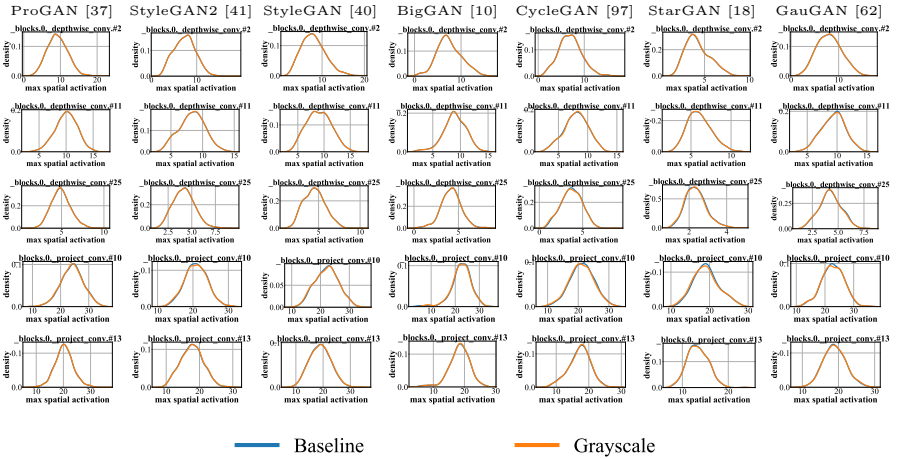


Fig. C.4. *Non Color-conditional T-FF in EfficientNet-B0*: Each row represents a *non* color-conditional T -FF (exact same T-FF as shown in Fig. C.3), and we show the maximum spatial activation distributions for ProGAN [37], StyleGAN2 [41], StyleGAN [40], BigGAN [10], CycleGAN [97], StarGAN [18] and GauGAN [62] counterfeits before (Baseline) and after color ablation (Grayscale). We remark that for each counterfeit in the ForenSynths dataset [87], we apply global max pooling to the specific T-FF to obtain a *maximum spatial activation* value (scalar). We can clearly observe that these T -FF are producing identical spatial activations (max) for the same set of counterfeits after removing color information which demonstrates that these T -FF do not respond to color information. This clearly indicates that these T -FF are *not* color-conditional. (Confirmed by our Mood’s median test).

Table D.1. Sensitivity assessments for different k values using feature map dropout of discovered T - FF : We show the results for the publicly released ResNet-50 universal detector [87] (top) and our own version of EfficientNet-B0 [78] universal detector (bottom) following the exact training / test strategy proposed in [87]. We show the AP, real and GAN detection accuracies for baseline [87] and different top- k forensic feature dropout. Feature map dropout is performed by suppressing (zeroing out) the resulting activations of target feature maps (i.e.: top- k). We can clearly observe that feature map dropout of top- k corresponding to T - FF results in substantial drop in AP and GAN detection accuracies across ProGAN and all 6 unseen GANs [41,40,10,97,18,62] as we increase k . Given that we aim to identify the *smallest* k , we identify $k = 114$ and $k = 27$ as the suitable k for ResNet-50 and EfficientNet-B0 universal detectors.

ResNet-50																						
		ProGAN [37]			StyleGAN2 [41]			StyleGAN [40]			BigGAN [10]			CycleGAN [97]			StarGAN [18]			GauGAN [62]		
		AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN
baseline [87]		100.0	100.0	100.0	99.3	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-29		98.6	99.9	40.7	84.9	89.2	62.3	84.9	92.9	52.4	66.8	85.1	35.4	76.9	89.4	42.2	87.7	98.2	30.4	85.6	94.0	45.6
top-57		96.8	99.9	26.3	84.0	91.1	54.9	84.0	92.4	50.6	63.2	83.3	30.9	71.4	88.9	30.6	86.0	98.1	29.0	82.4	92.7	41.2
top-114		69.8	99.4	3.2	56.6	89.4	11.3	56.6	90.6	13.7	55.4	86.3	18.3	61.2	91.4	17.4	72.6	89.4	35.9	71.0	95.0	18.8
top-228		58.6	99.3	2.3	49.2	29.2	76.6	49.2	24.5	76.2	51.6	48.1	50.6	50.2	83.0	16.2	59.3	46.7	66.4	60.7	65.5	52.5

EfficientNet-B0																						
		ProGAN [37]			StyleGAN2 [41]			StyleGAN [40]			BigGAN [10]			CycleGAN [97]			StarGAN [18]			GauGAN [62]		
		AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN	AP	Real	GAN
baseline [87]		100.	100.	100.	99.0	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-5		91.8	99.9	14.5	68.9	75.1	53.7	68.9	74.6	38.3	57.4	74.6	38.3	78.9	85.5	54.4	82.4	94.2	40.8	70.7	97.4	13.9
top-27		50.0	100.	0.0	52.1	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	46.2	100.	0.0
top-49		50.0	100.	0.0	50.0	100.	0.0	50.0	100.	0.0	50.0	100.	0.0	50.0	100.	0.0	50.0	100.	0.0	50.0	100.	0.0

E Cross-model forensic transfer using BigGAN [10] pre-training dataset

In this section, we show that color is a critical T - FF using an additional training dataset. We use BigGAN real / fake as second dataset with 1.04M images to train universal detectors following Wang *et al.* [87] and verify our findings. We remark that ForenSynths [87] uses ProGAN real / fake dataset. We perform large-scale experiments using EfficientNet-B0 universal detector. We report median counterfeit probability results for all 7 GANs [41,40,10,97,18,62] in Fig. E.1. Our results show on a second dataset that color ablation causes counterfeit probability to drop by $> 50\%$ for all unseen GANs. These results on another dataset further support that color is a critical T - FF in universal detectors for counterfeit detection.

F Is the performance degradation in universal detectors due to unseen corruptions (OOD)?

We remark that some performance degrade is due to CNNs' poor generalization to unseen corruptions / OOD (grayscale), but here we show that significant amount of degradation is due to color being a critical transferable forensic feature

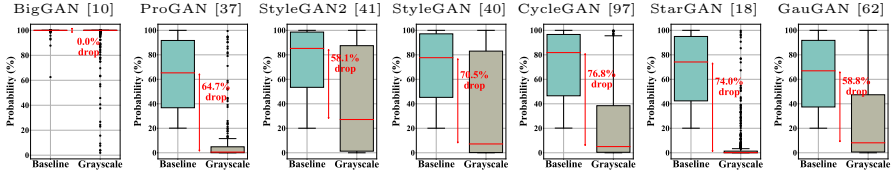


Fig. E.1. *Color* is a critical T -FF in *universal detectors* (Shown using BigGAN [10] pre-training dataset): We show the box-whisker plots of probability (%) predicted by the universal detector for counterfeits before (Baseline) and after *color ablation* (Grayscale) for BigGAN [10], ProGAN [37], StyleGAN2 [41], StyleGAN [40], CycleGAN [97], StarGAN [18] and GauGAN [62]. The red line in each box-plot shows the median probability. We show the results for the EfficientNet-B0 universal detector following the exact training / test strategy proposed in [87]. Using BigGAN real / fake dataset we verify that *Color* is a critical T -FF in Universal Detectors. We show that color ablation results in median probability for counterfeits drop by $> 58\%$ across all unseen GANs. Do note that median probability does not drop significantly for BigGAN during color ablation showing the importance of color for cross-model forensic transfer.

(T -FF) in the universal detector, therefore ablation of color (i.e., grayscale) leads to significant performance degrade. Specifically, we perform an experiment using official EfficientNet-B0 ImageNet classifier (architecture identical to our universal detector) under Grayscale (OOD) setup. We measure the median probability of the correct class before and after Grayscale (OOD) and observe only 17% drop due to Grayscale. Comparing the within-model OOD setup with the cross-model setup, the median probability drop during cross-model forensic transfer is much larger, i.e.: median probability drop during cross-model forensic transfer is $> 89\%$ (ProGAN pre-training, Fig. 4) and $> 58\%$ (BigGAN pre-training, Fig. E.1) for EfficientNet-B0 universal detector. This shows that color is critical in forensic transfer compared to within-model OOD setups. See row 1, col 1 in Fig. 4 and Fig. E.1, col 1 to verify that the median probability does not drop much for the GAN used to train universal detectors under Grayscale (OOD).

G Color-conditional T -FF (Additional Results)

In this section, we show more color-conditional T -FF to support our finding that *color* is a critical T -FF. We show LRP-max response image regions for ResNet-50 and EfficientNet-B0 in Fig. G.1 and G.3 respectively. We further show the maximum spatial activation distributions before and after color ablation for these color-conditional T -FF in Fig. G.2(ResNet-50) and Fig. G.4(EfficientNet-B0) respectively.

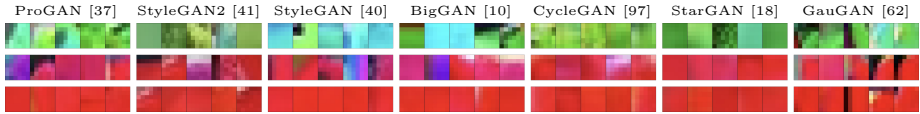


Fig. G.1. Additional results demonstrating that color is a critical *transferable forensic feature* ($T\text{-}FF$) in universal detectors (ResNet-50): Large-scale study on visual interpretability of $T\text{-}FF$ discovered through our proposed *forensic feature relevance statistic* ($FF\text{-}RS$), reveal that color information is critical for cross-model forensic transfer. Each row represents a color-conditional $T\text{-}FF$ and we show the LRP-max response regions for ProGAN [37], StyleGAN2 [41], StyleGAN [40], BigGAN [10], CycleGAN [97], StarGAN [18] and GauGAN [62] counterfeits for the publicly released ResNet-50 universal detector by Wang *et al.* [87]. This detector is trained with ProGAN [37] counterfeits [87] and cross-model forensic transfer is evaluated on other unseen GANs. All counterfeits are obtained from the ForenSynths dataset [87]. The consistent color-conditional LRP-max response across all GANs for these $T\text{-}FF$ clearly indicate that *color* is critical for cross-model forensic transfer in universal detectors.

H CR-Universal Detectors (Additional Results)

We show the AP, real and GAN detection accuracies for the universal Detectors in Table H.1 and CR-Universal Detectors trained using our proposed data augmentation scheme in Table H.2. As one can observe, our proposed CR-universal detectors are more robust and can avoid attacks from color-ablated counterfeits compared to the original detectors proposed by Wang *et al.* [87].

Table H.1. Universal detectors are more susceptible to color ablated counterfeit attacks as color is a critical $T\text{-}FF$: We show the results for the publicly released ResNet-50 universal detector [87] (top) and our own version of EfficientNet-B0 [78] universal detector (bottom) following the exact training and test strategy proposed in [87]. We show the AP, real and GAN image detection accuracies for Baseline and Grayscale (color ablated) images. As one can observe, AP and GAN detection accuracies drop *substantially* during cross-model transfer when removing color information from counterfeits.

ResNet-50																						
		ProGAN [37]			StyleGAN2 [41]			StyleGAN [40]			BigGAN [10]			CycleGAN [97]			StarGAN [18]			GauGAN [62]		
		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
Grayscale		99.9	100.0	81.5	89.1	92.7	61.9	96.7	94.6	84.8	75.2	85.8	48.8	84.2	94.5	41.0	89.2	93.4	60.7	97.6	97.7	78.8

EfficientNet-B0																						
		ProGAN [37]			StyleGAN2 [41]			StyleGAN [40]			BigGAN [10]			CycleGAN [97]			StarGAN [18]			GauGAN [62]		
		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	99.0	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
Grayscale		99.9	100.0	80.0	91.0	95.2	26.6	91.0	97.2	56.0	68.4	91.7	28.9	86.5	96.4	40.0	91.8	91.3	72.9	93.7	99.7	48.2

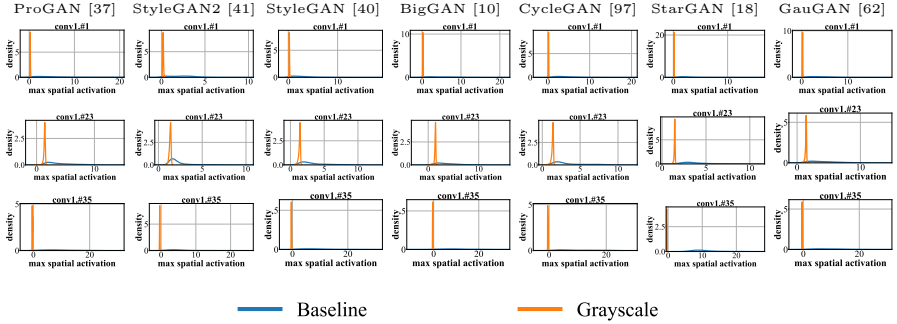


Fig. G.2. Additional results showing *Color-conditional T-FF* in *ResNet-50*: Each row represents a color-conditional *T-FF* (exact same *T-FF* as shown in Fig. G.1), and we show the maximum spatial activation distributions for ProGAN [37], StyleGAN2 [41], StyleGAN [40], BigGAN [10], CycleGAN [97], StarGAN [18] and GauGAN [62] counterfeits before (Baseline) and after color ablation (Grayscale). We remark that for each counterfeit in the ForenSynths dataset [87], we apply global max pooling to the specific *T-FF* to obtain a *maximum spatial activation* value (scalar). We can clearly observe that these *T-FF* are producing noticeably lower spatial activations (max) for the same set of counterfeits after removing color information. This clearly indicates that these *T-FF* are color-conditional (Confirmed by Mood’s median test).

Table H.2. CR-Universal detectors trained using our proposed data augmentation scheme are more robust to color ablated counterfeits: We show the results for the ResNet-50 universal detector [87] (top) and our own version of EfficientNet-B0 [78] universal detector (bottom) following the exact training / test strategy proposed in [87]. We show the AP, real and GAN image detection accuracies for Baseline and Grayscale (color ablated) images. As one can observe, AP and GAN detection accuracies remain similar during forensic transfer when removing color information from counterfeits.

CR-ResNet-50

	ProGAN [37]			StyleGAN2 [41]			StyleGAN [40]			BigGAN [10]			CycleGAN [97]			StarGAN [18]			GauGAN [62]		
	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	98.5	94.4	92.8	99.5	97.4	95.3	89.9	80.3	86.8	96.6	90.2	90.3	96.2	91.2	88.8	99.5	96.5	96.8
Grayscale	100.0	100.0	100.0	98.0	90.0	95.0	99.6	95.1	98.0	87.6	72.7	88.8	91.1	81.6	81.8	95.4	87.0	89.5	99.4	95.1	97.2

CR-EfficientNet-B0

	ProGAN [37]			StyleGAN2 [41]			StyleGAN [40]			BigGAN [10]			CycleGAN [97]			StarGAN [18]			GauGAN [62]		
	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	98.1	92.3	74.5	98.1	97.2	90.5	82.3	78.0	70.3	95.7	89.0	88.5	95.9	90.2	87.3	99.0	96.4	94.5
Grayscale	100.0	100.0	100.0	98.8	91.4	77.9	98.8	95.7	94.4	81.0	76.5	71.3	91.3	85.9	78.5	94.8	90.5	84.0	98.8	95.2	94.1

I Pixel-wise explanations are not informative to discover T - FF (Additional Results)

In this section, we show additional results to demonstrate that direct pixel-wise explanations of universal detector decisions are not informative to discover T - FF . Similar to main paper, we use 2 popular interpretation methods namely Guided-GradCAM [72] and LRP [9] to analyse the pixel-wise explanations of universal detector decisions. We show additional results for ResNet-50 detector in Fig. I.1. We also show results for EfficientNet-B0 in Fig. I.2 and I.3. As one can observe from Fig. I.1, I.2 and I.3 pixel-wise explanations of universal detector decisions are not informative to discover T - FF due to their focus on spatial localization.

J Research Reproducibility / Code Details

Code: Pytorch code is available at [here](#). Refer to README for step-by-step instructions. The codebase is clearly documented. The code is structured as follows:

- **lrp/**: Base Pytorch module containing LRP implementations for ResNet and EfficientNet architectures. This includes all Pytorch wrappers.
- **fmap_ranking/**: Pytorch module to calculate FF - RS (ω) for counterfeit detection.
- **sensitivity_assessment/**: Pytorch module to perform sensitivity assessments for T - FF and color ablation.
- **patch_extraction/**: Pytorch module to extract LRP-max response image regions for every T - FF .
- **activation_histograms/**: Pytorch module to calculate maximum spatial activation for images for every T - FF .
- **utils/**: Contains all utilities, helper functions and plotting functions.



Fig. G.3. Additional results demonstrating that color is a critical T - FF in universal detectors (EfficientNet-B0): Large-scale study on visual interpretability of T - FF discovered through our proposed FF - RS (ω) reveal that color information is critical for cross-model forensic transfer. Each row represents a color-based T - FF and we show the LRP-max response regions for ProGAN [37], StyleGAN2 [41], StyleGAN [40], BigGAN [10], CycleGAN [97], StarGAN [18] and GauGAN [62] counterfeits for our own version of EfficientNet-B0 [78] universal detector following the exact training / test strategy proposed by Wang *et al.* [87]. This detector is trained with ProGAN [37] counterfeits [87] and cross-model forensic transfer is evaluated on other unseen GANs. All counterfeits are obtained from the ForenSynths dataset [87]. The consistent color-conditional LRP-max response across all GANs for these T - FF clearly indicate that *color* is critical for cross-model forensic transfer in universal detectors.

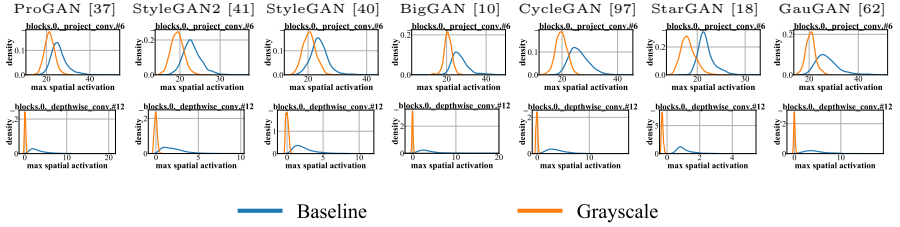


Fig. G.4. Additional results showing *Color-conditional T-FF in EfficientNet-B0*: Each row represents a color-conditional *T-FF* (exact same *T-FF* as shown in Fig. G.3), and we show the maximum spatial activation distributions for ProGAN [37], StyleGAN2 [41], StyleGAN [40], BigGAN [10], CycleGAN [97], StarGAN [18] and GauGAN [62] counterfeits before (Baseline) and after color ablation (Grayscale). We remark that for each counterfeit in the ForenSynths dataset [87], we apply global max pooling to the specific *T-FF* to obtain a *maximum spatial activation* value (scalar). We can clearly observe that these *T-FF* are producing noticeably lower spatial activations (max) for the same set of counterfeits after removing color information. This clearly indicates that these *T-FF* are color-conditional (Confirmed by Mood’s median test).

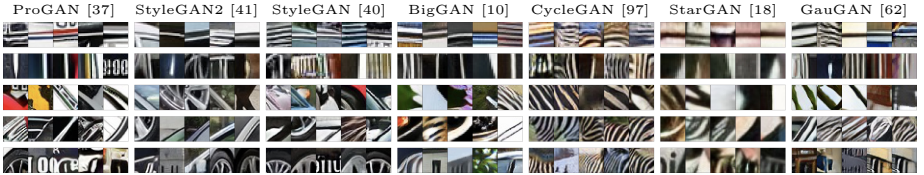


Fig. H.1. *T-FF in CR-ResNet-50*: Each row represents a *T-FF*. As visible, these *T-FF* are largely faintly colored. Notable patterns include wheels (row 5 in ProGAN [37], StyleGAN2 [41], StyleGAN [40]) and stripes in Zebra (rows 1-5 in CycleGAN [97]). We remark that CR-Universal detectors also contain a few color-conditional *T-FF*.

Pre-trained models: All pretrained models can be found at here. We provide both ResNet-50 and EfficientNet-B0 pretrained universal detectors. We also include CR-universal detector models. All our claims reported in Main / Supplementary can be reproduced using these checkpoints.

Docker information: For training /analysis in containerised environments (HPC, Super-computing clusters), please use nvr.io/nvidia/pytorch:20.12-py3 container.

Experiment details and hyper-parameters: For training universal detectors, we use the exact setup proposed in [87] with Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$), batch size of 64 and initial learning rate of $1e^{-4}$. For data augmentation, we use the exact setup proposed in [87] that includes random cropping (224x224), random horizontal flip and 50% JPEG + Blurring. All experiments were repeated 3 times. For LRP, we use $\beta = 0$ rule. For statistical tests, we use Mood’s median test with a significance level of $\alpha = 0.05$.

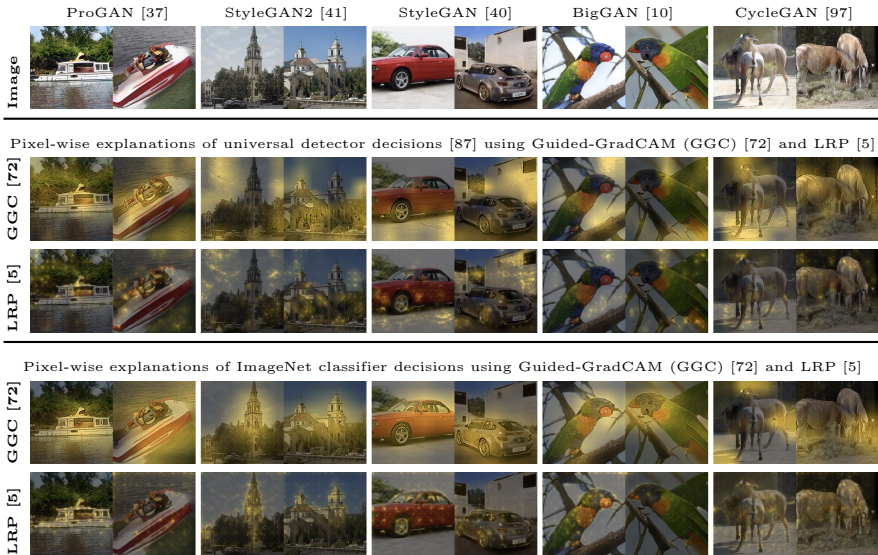


Fig. I.1. Additional results showing that pixel-wise explanations of universal detector decisions are not informative to discover $T\text{-}FF$: We show pixel-wise explanations using Guided-GradCAM (GGC) (row 2) [72] and LRP (row 3) [9] for the ResNet-50 universal detector [87] for ProGAN [37], CycleGAN [97], StarGAN [18], BigGAN [10] and StyleGAN2 [41]. The universal detector predicts probability $p \geq 95\%$ for all counterfeit images shown above. All these counterfeits are obtained from the ForenSynths dataset [87]. For LRP [9], we only show the positive relevances. We also show the pixel-wise explanations of ImageNet classifier decisions for the exact counterfeits using GGC (row 4) and LRP (row 5). This is shown as a control experiment to emphasize the significance of our observations. As one can clearly observe, pixel-wise explanations of universal detector decisions are not informative to discover $T\text{-}FF$ (row 2 and 3) as the explanations appear to be random and not reveal any meaningful visual features used for counterfeit detection. Particularly, it remains unknown as to why the universal detector outputs high detection probability ($p \geq 95\%$) for these counterfeits. On the other hand, pixel-wise explanations of ImageNet classifier decisions produce meaningful results. i.e.: The GGC (row 4) and LRP (row 5) explanation results for car samples (columns 5, 6) show that ImageNet uses features such as wheels/body to classify cars. This clearly shows that interpretability techniques such as GGC and LRP are not informative to discover $T\text{-}FF$ in universal detectors. In other words, we are unable to discover any forensic footprints based on pixel-wise explanations of universal detectors.

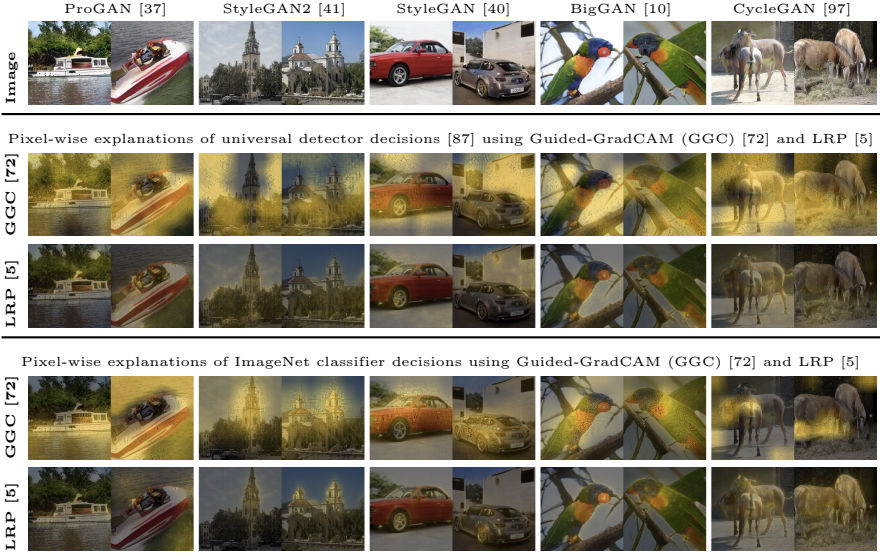


Fig. I.2. Additional results showing that pixel-wise explanations of universal detector decisions are not informative to discover $T\text{-}FF$ (EfficientNet-B0): We show pixel-wise explanations using Guided-GradCAM (GGC) (row 2) [72] and LRP (row 3) [9] for our version of EfficientNet-B0 universal Detector following the exact training / test strategy proposed in [87] for ProGAN [37], CycleGAN [97], StarGAN [18], BigGAN [10] and StyleGAN2 [41]. The universal detector predicts probability $p \geq 95\%$ for all counterfeit images shown above. All these counterfeits are obtained from the ForenSynths dataset [87]. For LRP [9], we only show the positive relevances. We also show the pixel-wise explanations of ImageNet classifier decisions for the exact counterfeits using GGC (row 4) and LRP (row 5). This is shown as a control experiment to emphasize the significance of our observations. As one can clearly observe, pixel-wise explanations of universal detector decisions are not informative to discover $T\text{-}FF$ (row 2 and 3) as the explanations appear to be random and not reveal any meaningful visual features used for counterfeit detection. Particularly, it remains unknown as to why the universal detector outputs high detection probability ($p \geq 95\%$) for these counterfeits. On the other hand, pixel-wise explanations of ImageNet classifier decisions produce meaningful results. i.e.: The GGC (row 4) and LRP (row 5) explanation results for car samples (columns 5, 6) show that ImageNet uses features such as wheels / body to classify cars. This clearly shows that interpretability techniques such as GGC and LRP are not informative to discover $T\text{-}FF$ in universal detectors. In other words, we are unable to discover any forensic footprints based on pixel-wise explanations of universal detectors.



Fig. I.3. Additional results showing that pixel-wise explanations of universal detector decisions are not informative to discover T -FF (EfficientNet-B0): We show pixel-wise explanations using Guided-GradCAM (GGC) (row 2) [72] and LRP (row 3) [9] for our version of EfficientNet-B0 universal Detector following the exact training / test strategy proposed in [87] for ProGAN [37], CycleGAN [97], StarGAN [18], BigGAN [10] and StyleGAN2 [41]. The universal detector predicts probability $p \geq 95\%$ for all counterfeit images shown above. All these counterfeits are obtained from the ForenSynths dataset [87]. For LRP [9], we only show the positive relevances. We also show the pixel-wise explanations of ImageNet classifier decisions for the exact counterfeits using GGC (row 4) and LRP (row 5). This is shown as a control experiment to emphasize the significance of our observations. As one can clearly observe, pixel-wise explanations of universal detector decisions are not informative to discover T -FF (row 2 and 3) as the explanations appear to be random and not reveal any meaningful visual features used for counterfeit detection. Particularly, it remains unknown as to why the universal detector outputs high detection probability ($p \geq 95\%$) for these counterfeits. On the other hand, pixel-wise explanations of ImageNet classifier decisions produce meaningful results. i.e.: The GGC (row 4) and LRP (row 5) explanation results for cat samples (columns 1, 2, 5, 6) show that ImageNet uses features such as eyes and whiskers to classify cats. This clearly shows that interpretability techniques such as GGC and LRP are not informative to discover T -FF in universal detectors. In other words, we can not discover any forensic footprints based on pixel-wise explanations of universal detectors.

K Future Work: Can we identify globally relevant channels for counterfeit detection in a Generator?

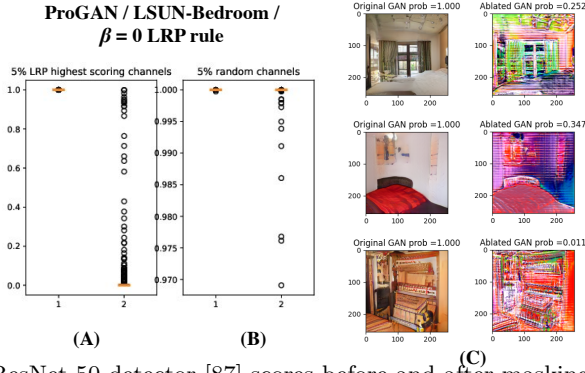


Fig. K.1. (A) ResNet-50 detector [87] scores before and after masking the 5% channels in the generator according to highest LRP scores computed for the generator. (B) ResNet-50 detector [87] scores before and after masking the 5% channels selected randomly in the generator. The orange line depicts the median of the box plot. Higher difference between both box plots within a subplot is better. Computed over 500 generated images trained over the LSUN-Bedrooms [90] class using a ProGAN [37]. One can see that masking 5% channels found by LRP in the generator leads to a very strong drop in detector scores (A) compared to masking 5% randomly selected channels results in a much smaller score decrease (B). (C) Original and ablated GAN samples with corresponding detector probabilities.

This section serves to motivate future directions from an image synthesis perspective. Particularly, we ask the question as to whether it's possible to identify feature maps in *GANs* that are responsible for generating forensic features that are detected by universal detectors.

In this section, we show preliminary results suggesting that it's possible to identify such globally relevant channels in a generator. Particularly, we perform LRP all the way into the Generator to identify the top highest scoring GAN channels that are responsible for counterfeit detection (i.e.: In the computational graph, the image is generated from a pre-trained ProGAN [37] model). We show that ablating these top-scoring GAN channels consequently results in large drop in probability predicted by the universal detector (We use the publicly released ResNet-50 in this experiment). This result is shown in Fig. K.1 that propagating LRP into the generator is able to identify the globally top-5% relevant channels for images. The box plot (A) shows a strong decrease after ablating these high-scoring GAN channels (though ablated GAN samples have poor visual quality). This can be compared to (B) where 5% of randomly selected GAN channels are ablated, which results in a very small decrease in counterfeit detection scores. These results show promising directions for understanding image synthesis methods, and we hope to explore this area in future work. We also hope to explore the properties of Fair Generative models [88,69,80,17,79], GANs / detectors trained using different techniques (regularization, knowledge transfer, pruning, few-shot learning, self-supervised learning) [13,2,89,34] and face-forgery detectors [43,16,73,93,95,47,31].