Discovering Transferable Forensic Features for CNN-generated Images Detection

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

[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 [61]. 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 [43,37,23,22,51]. With rapid improvements in CNN-based generative modelling [27,64,28,29,26,47,12,6,66,45,31,1,21,56,57,58,32,65], 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 [61]. i.e.: Publicly released ResNet-50 [24] universal detector by Wang *et al.* [61] trained only on ProGAN [26] counterfeits, surprisingly generalizes well to detect counterfeits from unseen GANs including StyleGAN2 [29], StyleGAN [28], BigGAN [6], CycleGAN [66], StarGAN [11] and GauGAN [44]. 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

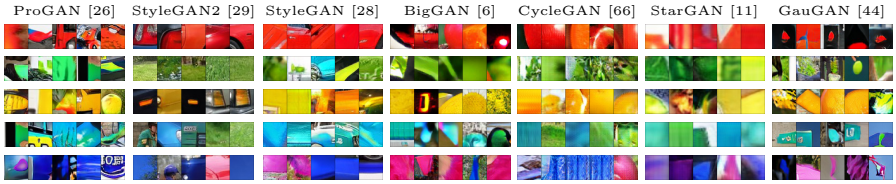

Fig. 1. Color is a critical *transferable forensic feature (T-FF)* in universal detectors: Large-scale study on visual interpretability of *T-FF* discovered through our proposed *forensic feature relevance statistic (FF-RS),* reveal that color information is critical for cross-model forensic transfer. Each row represents a color-conditional *T-FF* and we show the LRP-max response regions for different GANs counterfeits for the publicly released ResNet-50 universal detector by Wang *et al.* [61]. This detector is trained with ProGAN [26] counterfeits [61] and cross-model forensic transfer is evaluated on unseen GANs. All counterfeits are obtained from the ForenSynths dataset [61]. 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 [55] universal detector following the exact training / test strategy proposed by Wang *et al.* [61] in Fig. 3. More visualizations are included in Supplementary.

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

This work is motivated by a profound and challenging thesis statement: *What are the transferable forensic features (T-FF) in 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 from an unseen GAN (i.e.: StyleGAN2 [29]), what *T-FF* are used by the universal detector, such that it detects the synthetic car as counterfeit accurately? Though Wang *et al.* [61] hypothesize that universal detectors may learn low-level CNN artifacts for detection, no qualitative or 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 conduct 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 [50] and LRP [2] – 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 [24] 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 [61].

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 [16,17,63,8,49,30]. 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 [48,38,13,63,42,60,61] and frequency domain [17,16,8,18,35]. Particularly, notable number of works have proposed to use hand-crafted features for counterfeit detection [17,16,8,42]. Some recent works have also proposed methods to detect and attribute counterfeits to the generating architectures [62,39]. Anomaly detection techniques leveraging on pre-trained face recognition models have also been proposed [60].

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.* [13] and Zhang *et al.* [63] observed that counterfeit detectors generalized poorly during cross-model forensic transfer. In order to solve poor forensic transfer performance, Cozzolino *et al.* [13] proposed an autoencoder based

adaptation framework to improve cross-model forensic transfer. Using simple experiments, Mccloskey *et al.* [40] showed that detection based on the frequency of over-exposed pixels can provide good discrimination between real images and counterfeits. The work by Wang *et al.* [61] 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.* [7] showed that patch-based forensic 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 [46,36], or to explain single decisions of a neural network. The latter may follow very different lines of questioning, such as identifying similar training samples in k-NN and prototype CNNs [33,9], finding modified samples such as pertinent negatives [15], or model-based uncertainty estimates [20]. One class of algorithms aims at computing input space attributions. This includes Shapley values [54,34,10] 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 [52], Layer-wise Relevance Propagation (LRP) [2], Guided-GradCAM [50], Full-Grad [53], and class-attention-mapping inspired research [14,59,25,19,41]. Bau *et al.* proposed frameworks for interpreting representations at the feature map level used for GANs [3,4].

3 Dataset / Metrics

We use the ForenSynths dataset proposed by Wang *et al.* [61]. ForenSynths is the largest counterfeit benchmark dataset containing CNN-generated images from multiple generator architectures, datasets, loss functions and resolutions. In addition to ProGAN [26], we select 6 candidate GANs to comprehensively study cross-model forensic transfer in this work namely, StyleGAN2 [29], StyleGAN [28], BigGAN [6], CycleGAN [66], StarGAN [11] and GauGAN [44]. Following Wang *et al.* [61], 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.

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 [5,50,53,14,59,25,19,41]. 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) [50] and LRP (row 3) [2] for the ResNet-50 universal detector [61] for ProGAN [26], CycleGAN [66], StarGAN [11], BigGAN [6] and StyleGAN2 [29]. The universal detector predicts probability $p >= 95\%$ for all counterfeit images shown above. All these counterfeits are obtained from the ForenSynths dataset [61]. For LRP [2], 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* (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 ($p >= 95\%$) 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 shown in Supplementary.

We use 2 popular interpretation methods namely Guided-GradCAM [50] and LRP [2] to analyse the pixel-wise explanations of universal detector decisions. These methods were chosen due to their relatively low amount of gradient shattering noise. We show the pixel-wise explanation results of ResNet-50 universal detector [61] decisions for ProGAN [26] and 4 GANs not used for training – CycleGAN [66], StarGAN [11], BigGAN [6] and StyleGAN2 [29]– in Fig. 2. As one can observe in Fig. 2, pixel-wise explanations of universal detector decisions

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

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

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

are not informative to discover *T-FF* due to their focus on spatial localization. Particularly, we are unable to discover any forensic footprints based on pixel-wise explanations of universal detector decisions. This is consistently seen across both Guided-GradCAM [50] and LRP [2] methods. We remark that these observations do not indicate failure modes of Guided-GradCAM [50] or LRP [2] 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 very challenging problem as it requires quantifying the importance of every feature map in universal detectors for counterfeit detection. The ResNet-50 universal detector [61] 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

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 [61] (top) and our own version of EfficientNet-B0 [55] universal detector (bottom) following the exact training and test strategy proposed in [61]. We show the AP, real and GAN image detection accuracies for baseline [61], 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 [29,28,6,66,11,44] 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.

ResNet-50	**ProGAN** [26]			**StyleGAN2** [29]			**StyleGAN** [28]			**BigGAN** [6]			**CycleGAN** [66]			**StarGAN** [11]			**GauGAN** [44]		
$k = 114$	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**
baseline [61]	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**	**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**
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** [26]			**StyleGAN2** [29]			**StyleGAN** [28]			**BigGAN** [6]			**CycleGAN** [66]			**StarGAN** [11]			**GauGAN** [44]		
$k = 27$	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**	**AP**	**Real**	**GAN**
baseline	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

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 [26] 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 the ReLU operation).
- The relevance scores are calculated using LRP [2] (More details on LRP [2] in Supplementary). 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 [26] and all unseen GANs [29,28,6,66,11,44] indicating poor cross-model forensic transfer.
2. Significant reduction in GAN /counterfeit detection accuracies across ProGAN [26] and all unseen GANs [29,28,6,66,11,44] compared to real image detection accuracies as ω is calculated for counterfeits.

Test bed details: We use the ForenSynths test dataset proposed in [61]. ω is calculated using 1000 counterfeits from ProGAN [26] validation set in ForenSynths. 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.: $\omega \approx 0$.

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.* [61] hypothesize that universal detectors may learn low-level CNN artifacts for cross-model forensic transfer, no qualitative / quantitative evidence is available to understand as to what features in counterfeits are being detected during cross-model forensic transfer.

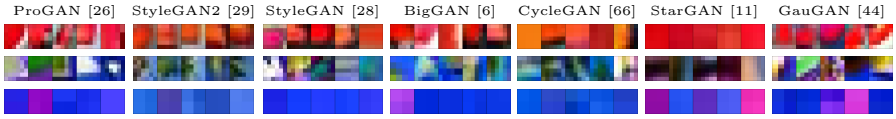

Fig. 3. Color is a critical *T-FF* in universal detectors: Large-scale study on visual interpretability of *T-FF* discovered through our proposed *FF-RS (ω)* reveal that color information is critical for cross-model forensic transfer. Each row represents a color-based *T-FF* and we show the LRP-max response regions for ProGAN and all 6 unseen GANs [29,28,6,66,11,44] counterfeits for our own version of EfficientNet-B0 [55] universal detector following the exact training / test strategy proposed by Wang *et al.* [61]. This detector is trained with ProGAN [26] counterfeits [61] and cross-model forensic transfer is evaluated on other unseen GANs. All counterfeits are obtained from the ForenSynths dataset [61]. 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.

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 (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. We remark that LRP-max does not depend on external modules such as segmentation used in Network Dissection [3] and GAN Dissection [4] methods. The pseudocode is included in Supplementary.

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 [26] and all unseen GANs [29,28,6,66,11,44] 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 very surprising and observed for the first time in contemporary image forensics research, yet shown qualitatively. 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.

Algorithm 2: 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.

Set M to evaluation mode
for l', c' *in* S **do**
  $A_b \leftarrow [\,]$; `/*store baseline counterfeits activations*/`
  $A_g \leftarrow [\,]$; `/*store grayscale counterfeits activations*/`
  for i *in* $\{0, 1, , ..., n\}$ **do**
    $a_b \leftarrow GLOBAL\_MAXPOOL(M_{l_c}(x_i))$; `/*baseline*/`
    $a_g \leftarrow GLOBAL\_MAXPOOL(M_{l_c}(grayscale(x_i)))$; `/*grayscale*/`
    $A_b.append(a_b)$
    $A_g.append(a_g)$
  end
  $p(l'_{c'}) \leftarrow MEDIAN - TEST(A_b, A_g)$; `/*median test*/`
end
return $p(l_c)$

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 unseen 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 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) with a significance level of $\alpha = 0.05$ in our study. The pseudocode is shown in Algorithm 2. The results for ResNet-50 and EfficientNet-B0 universal detectors are shown in Table 2. Our results show that substantial amount of *T-FF* in universal detectors

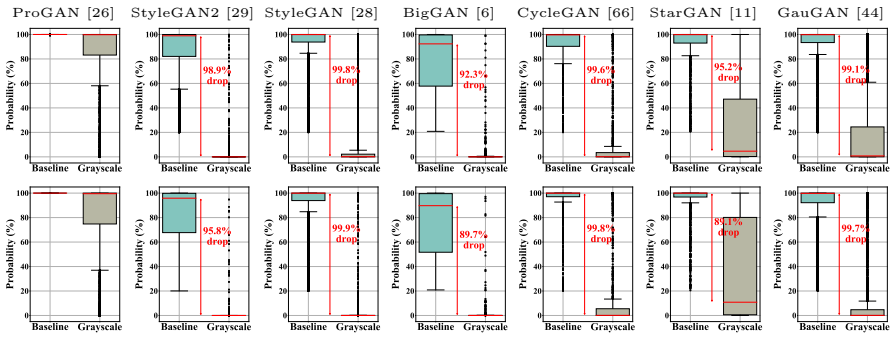

Fig. 4. *Color* is a critical *T-FF* in *universal detectors*: We show the box-whisker plots of probability (%) predicted by the universal detector for counterfeits before (Baseline) and after *color ablation* (Grayscale) for 7 GAN models. The red line in each box-plot shows the median probability. We show the results for the ResNet-50 universal detector [61] (top row) and our version of EfficientNet-B0 [55] universal detector following the exact training / test strategy proposed in [61] (bottom row). These detectors are trained with ProGAN [26] counterfeits [61] and cross-model forensic transfer is evaluated on other unseen GANs. All counterfeits are obtained from the ForenSynths dataset [61]. 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.

are color-conditional indicating that color is a critical *T-FF*. We also show the maximum spatial activation distributions for some color-conditional *T-FF* for ResNet-50 and EfficientNet-B0 universal detectors in Fig. 6 and 7 respectively. As one can observe maximum spatial activations are suppressed for these *T-FF* across ProGAN [26] and all other unseen GANs [29,28,6,66,11,44] 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 augmentation 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 Grayscaling 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. 4. We clearly observe that with our proposed data augmentation scheme, CR-universal detectors are more robust

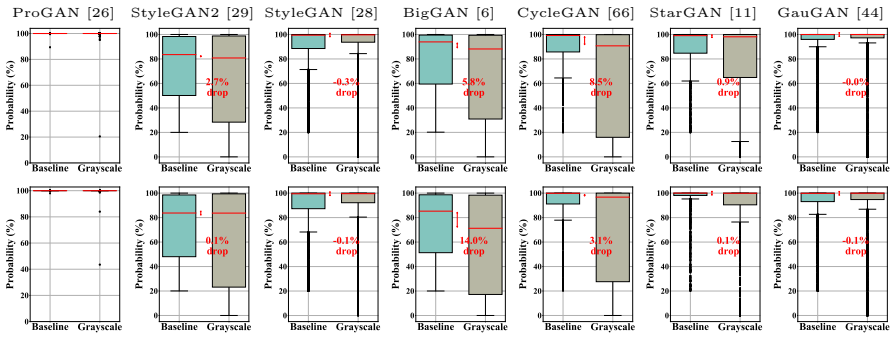

Fig. 5. *CR-universal detectors trained using our proposed data augmentation scheme (Sec. 6) are more robust to color ablation during cross-model forensic transfer:* These universal detectors are trained with data augmentation where color is ablated 50% of the time during training. This ensures that *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 [61] (top row) and EfficientNet-B0 [55] 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.

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 Table 3. With our proposed data augmentation scheme, we quantitatively show that CR-universal detectors contain substantially lower amount of color-conditional *T-FF*.

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 [50] and LRP [2] are not informative to discover *T-FF*. 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*.

Further investigations on the *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*, and quantitatively using median counterfeit probability analysis and statistical tests on

Table 2. *Significant amount of T-FF are color-conditional:* 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 [26] and all 6 unseen GANs [29,28,6,66,11,44]. 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.

% Color-conditional	ProGAN [26]	StyleGAN2 [29]	StyleGAN [28]	BigGAN [6]	CycleGAN [66]	StarGAN [11]	GauGAN [44]
ResNet-50	85.1	83.3	84.2	86.8	86.8	86.0	94.7
EfficientNet-B0	51.9	55.6	55.6	48.1	44.4	55.6	63.0

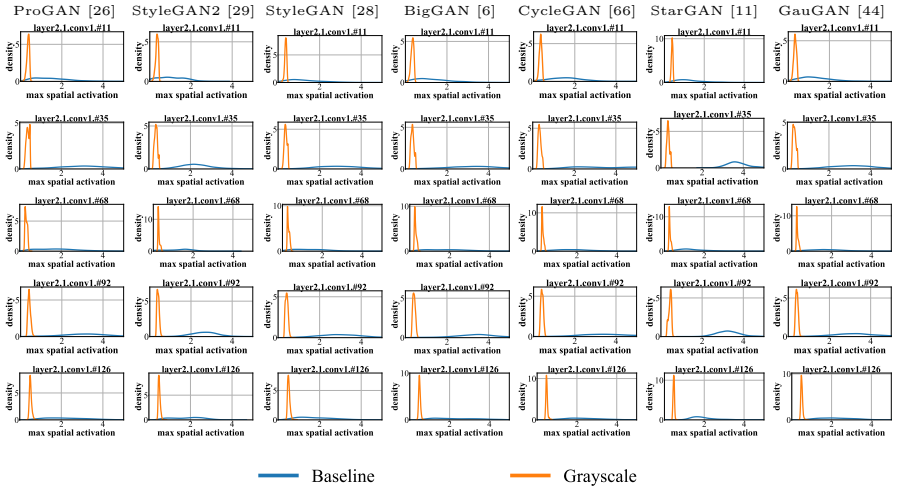

Fig. 6. *Color-conditional T-FF in ResNet-50:* Each row represents a color-conditional *T-FF* (exact same T-FF as shown in Fig. 1), and we show the maximum spatial activation distributions for 7 GAN models before (Baseline) and after color ablation (Grayscale). We remark that for each counterfeit in the ForenSynths dataset [61], 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.

maximum spatial activation distributions of *T-FF* based on color ablation. i.e.: We showed that $\approx 85\%$ of *T-FF* are color-conditional in the publicly released ResNet-50 universal detector [61]. Finally, we propose a simple data augmentation scheme to train Color-Robust (CR) universal detectors. We remark that color is not the only *T-FF*, but it is a critical *T-FF* in universal detectors. 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 GANs struggle

Table 3. *CR-universal detectors have noticeably lower amount of color-conditional T-FF:* We show the percentage(%) of color-conditional T-FF in ResNet-50 and EfficientNet-B0 CR-universal detectors measured using Mood's median test. We show the results for ProGAN [26] and all 6 unseen GANs [29,28,6,66,11,44]. 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 $\alpha = 0.05$. We clearly observe that training universal detectors using our proposed data augmentation scheme results in CR-universal detectors that contain noticeably lower amount of color-conditional *T-FF*.

% Color-conditional	ProGAN [26]	StyleGAN2 [29]	StyleGAN [28]	BigGAN [6]	CycleGAN [66]	StarGAN [11]	GauGAN [44]
ResNet-50	35.1	37.7	39.5	37.7	36.8	35.9	38.1
EfficientNet-B0	29.1	27.7	27.9	34.5	30.1	29.4	28.6

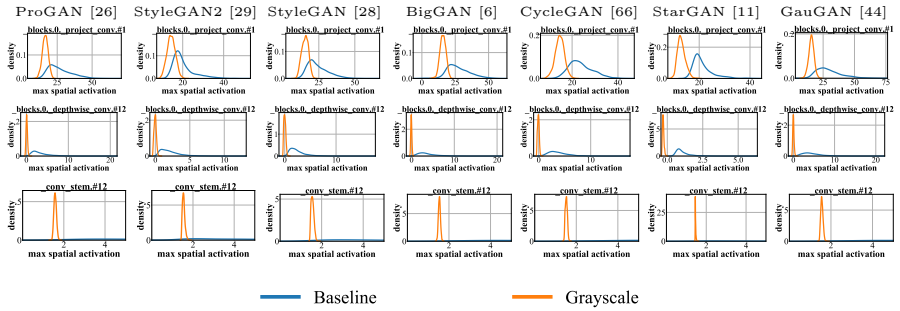

Fig. 7. *Color-conditional T-FF in EfficientNet-B0:* Each row represents a color-conditional *T-FF* (exact same T-FF as shown in Fig. 3), and we show the maximum spatial activation distributions for ProGAN [26] and all 6 unseen GANs [29,28,6,66,11,44] before (Baseline) and after color ablation (Grayscale). We remark that for each counterfeit in the ForenSynths dataset [61], 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.

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, and such discrepancies can be used as forensic features to discriminate between real and GAN images. We include additional experiments / analysis in Supplementary.

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. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan (2017)
2. 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)
3. 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)
4. 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)
5. 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)
6. 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`
7. 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)
8. 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)
9. Chen, C., Li, O., Tao, C., Barnett, A.J., Su, J., Rudin, C.: ¡i¿This¡/i¿ Looks like ¡i¿That¡/i¿: Deep Learning for Interpretable Image Recognition. Curran Associates Inc., Red Hook, NY, USA (2019)
10. 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`
11. 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)
12. 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)
13. 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)
14. 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)
15. 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)
16. 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)
17. 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)
18. 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)
19. 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)
20. 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
21. 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
22. 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/
23. 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
24. 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)
25. 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)
26. 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
27. 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
28. 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)

29. 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)
30. Khayatkhoei, M., Elgammal, A.: Spatial frequency bias in convolutional generative adversarial networks (Oct 2020)
31. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 105–114 (2017). https://doi.org/10.1109/CVPR.2017.19
32. 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)
33. 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
34. 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
35. 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)
36. 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/vandermaaten08a.html
37. 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
38. 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
39. 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)
40. 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)
41. 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)
42. 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)
43. 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/

44. 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)
45. 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 (CVPR) (June 2019)
46. 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
47. 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
48. 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)
49. Schwarz, K., Liao, Y., Geiger, A.: On the frequency bias of generative models. Advances in Neural Information Processing Systems **34** (2021)
50. 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)
51. Simonite, T.: What happened to the deepfake threat to the election? (Nov 2020), https://www.wired.com/story/what-happened-deepfake-threat-election/
52. 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
53. Srinivas, S., Fleuret, F.: Full-gradient representation for neural network visualization. Advances in neural information processing systems **32** (2019)
54. Strumbelj, E., Kononenko, I.: An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res. **11**, 1–18 (mar 2010)
55. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)
56. Tran, N.T., Bui, T.A., Cheung, N.: Dist-gan: An improved gan using distance constraints. In: ECCV (2018)
57. 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
58. 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)
59. 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)

60. 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)
61. 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)
62. 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)
63. 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
64. 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)
65. 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)
66. 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)