Supplementary Materials
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

Contents

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

A A brief overview over the LRP-algorithm used

Layer-wise relevance propagation (LRP) [2] 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[4] of the network. While the ϵ-LRP-rule is similar to gradient-times-input, other rules such as the β-rule[5] 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

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

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[7] [8] [9] when modifying the highest scoring regions.

B LRP-max pseudocode

In this section, we include the pseudo-code for obtaining LRP-max pixel-wise explanations. In particular, we study the LRP-max responses for *T-FF* in this

[4] Montavon et al.,: Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. In Pattern Recognition (2017)

[5] Montavon et al.,: Layer-Wise Relevance Propagation: An Overview. Book chapter in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (2019)

[6] Balduzzi et al.,: The Shattered Gradients Problem: If resnets are the answer, then what is the question?. In ICML (2017)

[7] Samek et al.,: Evaluating the Visualization of What a Deep Neural Network Has Learned. IEEE Transactions on Neural Networks and Learning Systems (2017)

[8] Pörner et al.,: Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In ACL (2018)

[9] Arras et al.,: Evaluating Recurrent Neural Network Explanations. In ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (2019)

work. The pseudo-code is shown in Algorithm 1. We remark that LRP-max is a procedure to *automatically extract image regions for every T-FF*.

Algorithm 1: 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.

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

C Computational Complexity of FF-RS / LRP-max

Both FF-RS and LRP-max require an additional forward and backward pass during computation. We emphasize that our proposed FF-RS and LRP-max are not used during training of universal detectors, but are only used for our analytical study. Therefore, we remark that there is no substantial amount of additional computational overhead.

D 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. D.1 and D.3 respectively. We further show the maximum spatial activation distributions before and after color ablation for ResNet-50 and EfficientNet-B0 in Fig. D.2 and D.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.

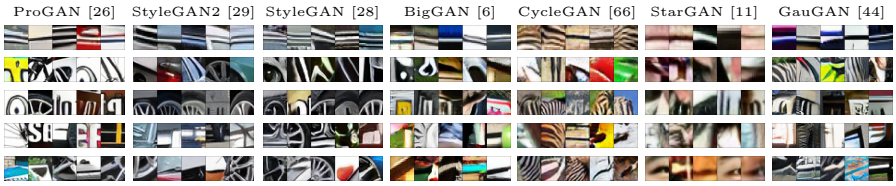

Fig. D.1. *T-FF* that are *not* color-conditional in ResNet-50 Universal detector: We show the LRP-max response regions for 5 *non* color-conditional T-FF for ProGAN [26] and all 6 unseen GANs [29,28,6,66,11,44]. Each row represents a *non* color-conditional T-FF. We emphasize that *T-FF* are discovered using our proposed *forensic feature relevance statistic (FF-RS)*. This detector is trained with ProGAN [26] counterfeits [61] and cross-model forensic transfer is evaluated on other unseen GANs. All counterfeits are obtained from the ForenSynths dataset [61]. 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 other *T-FF* (i.e.: frequency / texture artifacts).

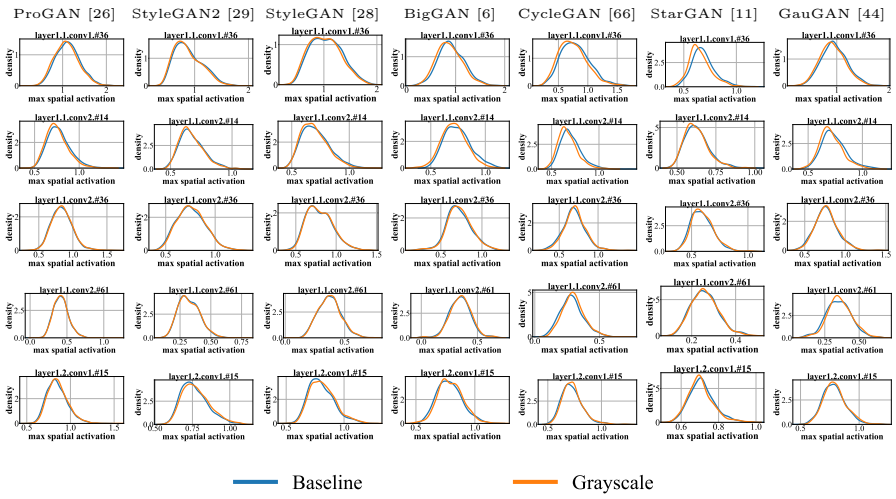

Fig. D.2. *Non Color-conditional T-FF in ResNet-50:* Each row represents a *non* color-conditional *T-FF* (exact same T-FF as shown in Fig. D.1), and we show the maximum spatial activation distributions for ProGAN [26], StyleGAN2 [29], StyleGAN [28], BigGAN [6], CycleGAN [66], StarGAN [11] and GauGAN [44] counterfeits 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 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 our Mood's median test).

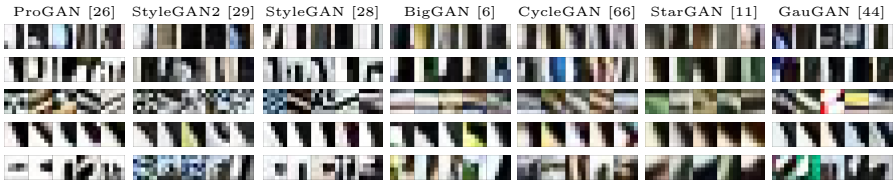

Fig. D.3. *T-FF* that are *not* color-conditional in EfficientNet-B0 Universal detector: We show the LRP-max response regions for 5 *non* color-conditional T-FF for ProGAN [26] and all 6 unseen GANs [29,28,6,66,11,44]. Each row represents a *non* color-conditional T-FF. We emphasize that *T-FF* are discovered using our proposed *forensic feature relevance statistic (FF-RS)*. This detector is trained with ProGAN [26] counterfeits [61] and cross-model forensic transfer is evaluated on other unseen GANs. All counterfeits are obtained from the ForenSynths dataset [61]. 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 other *T-FF* (i.e.: frequency / texture artifacts).

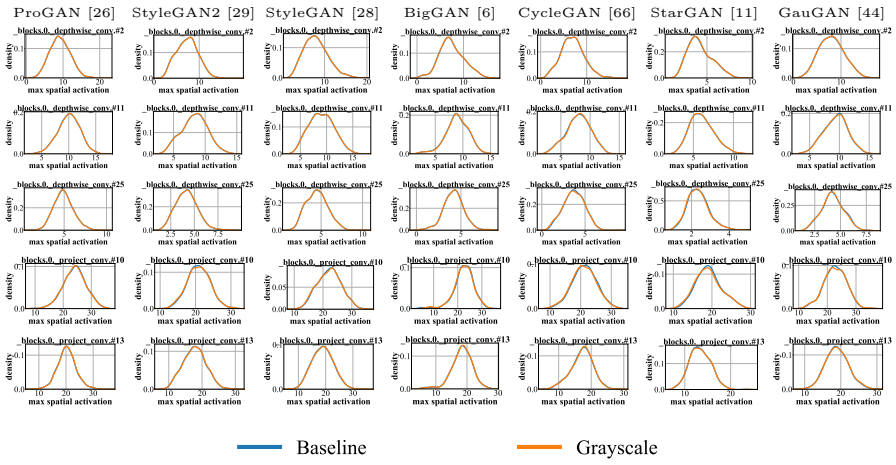

Fig. D.4. *Non Color-conditional T-FF in EfficientNet-B0:* Each row represents a *non* color-conditional *T-FF* (exact same T-FF as shown in Fig. D.3), and we show the maximum spatial activation distributions for ProGAN [26], StyleGAN2 [29], StyleGAN [28], BigGAN [6], CycleGAN [66], StarGAN [11] and GauGAN [44] counterfeits 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 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).

E *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, the AP and GAN detection accuracies drop across ProGAN [26] and all unseen GANs [29,28,6,66,11,44]. 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 E.1

Table E.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 [61] (top) and our own version of EfficientNet-B0 [55] universal detector (bottom) following the exact training / test strategy proposed in [61]. We show the AP, real and GAN detection accuracies for baseline [61] 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 topk-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] as we increase k. Given that we aim to identify the *smallestk*, we identify $k = 114$ and $k = 27$ as the suitable k for ResNet-50 and EfficientNet-B0 universal detectors.

ResNet-50	**ProGAN** [26]			**StyleGAN2** [29]			**StyleGAN** [28]			**BigGAN** [6]			**CycleGAN** [66]			**StarGAN** [11]			**GauGAN** [44]		
	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	99.3	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** [26]			**StyleGAN2** [29]			**StyleGAN** [28]			**BigGAN** [6]			**CycleGAN** [66]			**StarGAN** [11]			**GauGAN** [44]		
	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.	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**

F Cross-model forensic transfer using BigGAN [6] 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.* [61] and verify our findings. We remark that ForenSynths [61] uses ProGAN real / fake dataset. We perform large-scale experiments using EfficientNet-B0 universal detector. We report median counterfeit probability results for 6 GANs [29,28,6,66,11,44] in Fig. F.1. This shows on a second dataset that color ablation causes counterfeit probability to drop by $> 58\%$ for all unseen GANs. These results on another dataset further support that color is a critical *T-FF* in universal detectors for counterfeit detection.

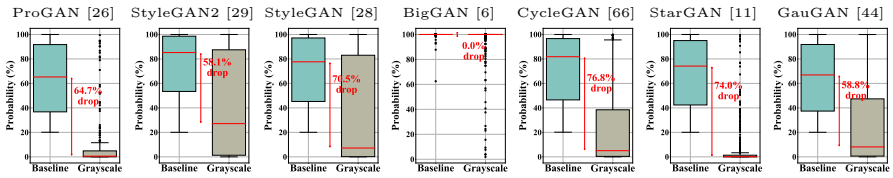

Fig. F.1. *Color* is a critical *T-FF* in *universal detectors*: We show the box-whisker plots of probability (%) predicted by the universal detector for counterfeits before (Baseline) and after *color ablation* (Grayscale) for ProGAN [26], StyleGAN2 [29], StyleGAN [28], BigGAN [6], CycleGAN [66], StarGAN [11] and GauGAN [44]. 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 [61]. 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.

G Is the performance degradation in universal detectors due to unseen corruptions?

We remark that some performance degrade is due to CNNs' poor generalization to unseen corruptions (grayscale), but here we show that significant amount of degradation is due to color being a critical transferable forensic feature (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, Fig. 4 main paper) and $> 58\%$ (BigGAN, Fig. F.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. F.1, Fig. 4, col 1 main paper: the median probability does not drop much for the GAN used to train universal detector under Grayscale (OOD).

H 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. H.1 and H.3 respectively. We further show the maximum spatial activation distributions before and after color ablation for

these color-conditional *T-FF* in Fig. H.2(ResNet-50) and Fig. H.4(EfficientNet-B0) respectively.

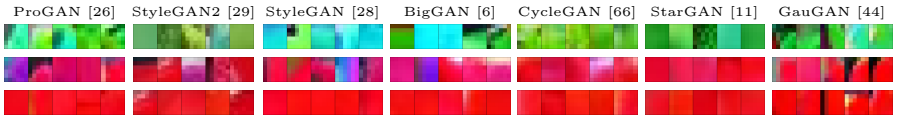

Fig. H.1. Additional results demonstrating that color is a critical *transferable forensic feature (T-FF)* in universal detectors (ResNet-50): Large-scale study on visual interpretability of *T-FF* discovered through our proposed *forensic feature relevance statistic (FF-RS),* reveal that color information is critical for cross-model forensic transfer. Each row represents a color-conditional *T-FF* and we show the LRP-max response regions for ProGAN [26], StyleGAN2 [29], StyleGAN [28], BigGAN [6], CycleGAN [66], StarGAN [11] and GauGAN [44] 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 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.

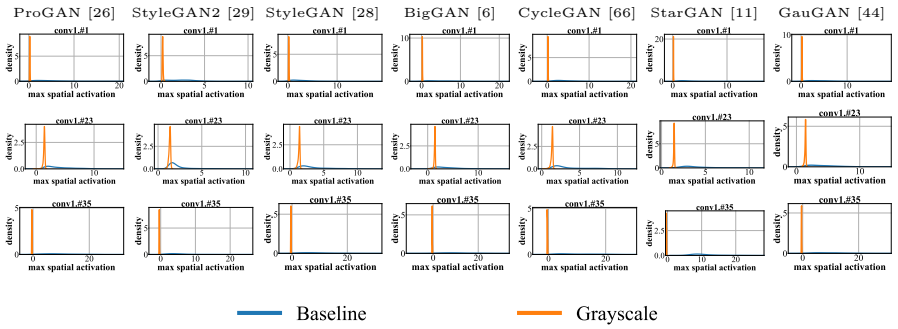

Fig. H.2. Additional results showing *Color-conditional T-FF in ResNet-50:* Each row represents a color-conditional *T-FF* (exact same T-FF as shown in Fig. H.1), and we show the maximum spatial activation distributions for ProGAN [26], StyleGAN2 [29], StyleGAN [28], BigGAN [6], CycleGAN [66], StarGAN [11] and GauGAN [44] counterfeits 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 (Confirmed by our Mood's median test).

I CR-Universal Detectors (Additional Results)

We show the AP, real and GAN detection accuracies for the universal Detectors in Table I.1 and for CR-Universal Detectors trained using our proposed data augmentation scheme in Table I.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 universal detectors proposed by Wang *et al.* [61].

Table I.1. Universal detectors are more susceptible to color ablated counterfeit attacks as color is a critical *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 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** [26]			**StyleGAN2** [29]			**StyleGAN** [28]			**BigGAN** [6]			**CycleGAN** [66]			**StarGAN** [11]			**GauGAN** [44]		
	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** [26]			**StyleGAN2** [29]			**StyleGAN** [28]			**BigGAN** [6]			**CycleGAN** [66]			**StarGAN** [11]			**GauGAN** [44]		
	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**

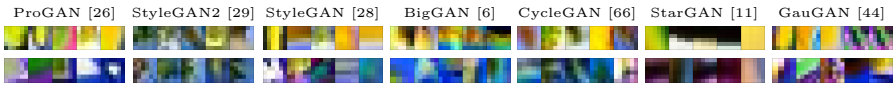

Fig. H.3. Additional results demonstrating that color is a critical *T-FF* in universal detectors (EfficientNet-B0): Large-scale study on visual interpretability of *T-FF* discovered through our proposed *FF-RS (ω)* reveal that color information is critical for cross-model forensic transfer. Each row represents a color-based *T-FF* and we show the LRP-max response regions for ProGAN [26], StyleGAN2 [29], StyleGAN [28], BigGAN [6], CycleGAN [66], StarGAN [11] and GauGAN [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.

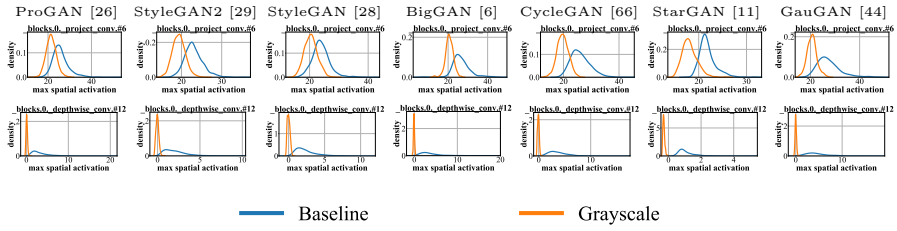

Fig. H.4. Additional results showing *Color-conditional T-FF in EfficientNet-B0:* Each row represents a color-conditional *T-FF* (exact same T-FF as shown in Fig. H.3), and we show the maximum spatial activation distributions for ProGAN [26], StyleGAN2 [29], StyleGAN [28], BigGAN [6], CycleGAN [66], StarGAN [11] and GauGAN [44] counterfeits 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 (Confirmed by our Mood's median test).

Table I.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 [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 and Grayscale (color ablated) images. As one can observe, AP and GAN detection accuracies remain similar during cross-model transfer when removing color information from counterfeits.

CR-ResNet-50	**ProGAN** [26]			**StyleGAN2** [29]			**StyleGAN** [28]			**BigGAN** [6]			**CycleGAN** [66]			**StarGAN** [11]			**GauGAN** [44]		
	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** [26]			**StyleGAN2** [29]			**StyleGAN** [28]			**BigGAN** [6]			**CycleGAN** [66]			**StarGAN** [11]			**GauGAN** [44]		
	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**

J 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 [50] and LRP [5] to analyse the pixel-wise explanations of universal detector decisions. We show additional results for ResNet-50 detector in Fig. J.1. We also show results for EfficientNet-B0 in Fig. J.2 and J.3. As one can observe from Fig. J.1, J.2 and J.3 pixel-wise explanations of universal detector decisions are not informative to discover T-FF due to their focus on spatial localization.

K Research Reproducibility / Code Details

Code: Pytorch code is available at `https://keshik6.github.io/transferable-forensic-features/`. 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.

Pre-trained models submission: All pretrained models can be found at `https://keshik6.github.io/transferable-forensic-features/`. 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 nvcr.io/nvidia/pytorch:20.12-py3 container.

Experiment details and hyper-parameters: For training universal detectors, we use the exact setup proposed in [61] 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 [61] that includes random cropping (224x224), random horizontal flip and 50% JPEG + Blurring. All experiments were repeated 3 times. For LRP, we use *beta*0 setting. For statistical tests, we use Mood's median test with a significance level of $\alpha = 0.05$.

Fig. J.1. Additional results showing that 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) [5] 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 [5], 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 ($p >= 95\%$) 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-FF* in universal detectors. In other words, we are unable to discover any forensic footprints based on pixel-wise explanations of universal detectors.

Fig. J.2. 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) [50] and LRP (row 3) [5] for our version of EfficientNet-B0 universal Detector following the exact training / test strategy proposed in [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 [5], 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 ($p >= 95\%$) 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-FF* in universal detectors. In other words, we are unable to discover any forensic footprints based on pixel-wise explanations of universal detectors.

L Broader Impact

The thesis of our work is to discover and understand *T-FF* in universal detectors, and we remark that our findings on color as a critical *T-FF* in universal detectors is very significant. Our findings suggest that contemporary CNN-based image synthesis methods may potentially struggle to capture the diverse,

Fig. J.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) [50] and LRP (row 3) [5] for our version of EfficientNet-B0 universal Detector following the exact training / test strategy proposed in [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 [5], 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 ($p >= 95\%$) 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.

multi-modal color distribution of real images thereby leaving detectable forensic footprints. We remark that this can inspire research to further improve image synthesis methods to avoid such color-based forensic footprints, making it potentially more difficult to detect visual counterfeits. In our opinion, we believe that image synthesis methods and our fight against visual disinformation will continue to parallely evolve in the foreseeable future.

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

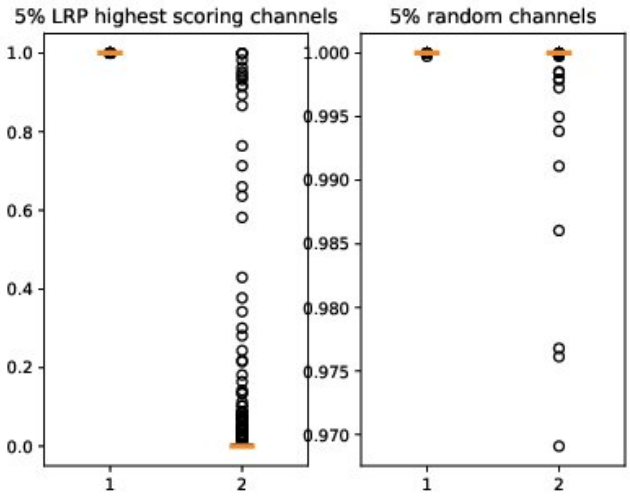

Fig. M.1. Left: ResNet-50 universal detector [61] scores before and after masking the 5% channels in the generator according to highest LRP scores computed for the generator. Right: ResNet-50 universal detector [61] scores before and after masking the 5% channels selected randomly in the generator. The 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 class using a ProGAN [26]. One can see that masking 5% channels found by LRP in the generator leads to a very strong drop in detector scores (Left) compared to masking 5% randomly selected channels results in a much smaller score decrease (Right).

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 the universal detector.

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 [26] 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. M.1 that propagating LRP into the generator is able to identify the globally top-5% relevant channels for images. The box plot on the left shows a strong decrease after ablating these

high-scoring GAN channels. This can be compared to the right figure 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 in understanding image synthesis methods, and we hope to explore this area in future work.

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)