SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
SINGAPORE INSTITUTE OF TECHNOLOGY
UNIVERSITY OF OSLO

ECCV TEL AVIV 2022

Discovering Transferable Forensic Features for CNN-generated Images Detection

European Conference on Computer Vision (ECCV) 2022 Oral

Keshigeyan Chandrasegaran

Ngoc-Trung Tran

Alexander Binder

Ngai-Man Cheung

Visual counterfeits are increasingly causing an existential conundrum in mainstream media

Samples generated using Zero-Insertion based Upsampling Architectural Variant (Chandrasegaran et al., 2021) of Progressive Growing of GAN (Karras et al., 2018)

Chandrasegaran, K., Tran, N. T., & Cheung, N. M. (2021). *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
Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.

With rapid improvements in generative modelling, detecting such counterfeits is increasingly becoming challenging and critical.

Samples generated using Zero-Insertion based Upsampling Architectural Variant (Chandrasegaran et al., 2021) of Progressive Growing of GAN (Karras et al., 2018)

Chandrasegaran, K., Tran, N. T., & Cheung, N. M. (2021). *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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.

However, a recent class of forensic detectors known as *universal detectors* (Wang et al., 2020) can surprisingly spot counterfeits regardless of generator architectures, loss functions, datasets or resolutions.

Samples generated using Zero-Insertion based Upsampling Architectural Variant (Chandrasegaran et al., 2021) of Progressive Growing of GAN (Karras et al., 2018)

Chandrasegaran, K., Tran, N. T., & Cheung, N. M. (2021). *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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.

Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). *CNN-generated images are surprisingly easy to spot... for now*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

This intriguing cross-model forensic transfer suggests the existence of *Transferable Forensic Features* (*T-FF*) in *universal detectors*.

Samples generated using Zero-Insertion based Upsampling Architectural Variant (Chandrasegaran et al., 2021) of Progressive Growing of GAN (Karras et al., 2018)

Chandrasegaran, K., Tran, N. T., & Cheung, N. M. (2021). *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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.

Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). *CNN-generated images are surprisingly easy to spot... for now*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

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

Samples generated using Zero-Insertion based Upsampling Architectural Variant (Chandrasegaran et al., 2021) of Progressive Growing of GAN (Karras et al., 2018)

Chandrasegaran, K., Tran, N. T., & Cheung, N. M. (2021). *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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.

Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). *CNN-generated images are surprisingly easy to spot... for now*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

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

Samples generated using Zero-Insertion based Upsampling Architectural Variant (Chandrasegaran et al., 2021) of Progressive Growing of GAN (Karras et al., 2018)

Chandrasegaran, K., Tran, N. T., & Cheung, N. M. (2021). *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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.

Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). *CNN-generated images are surprisingly easy to spot... for now*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

Are existing Interpretability methods capable of discovering *T-FF* ?

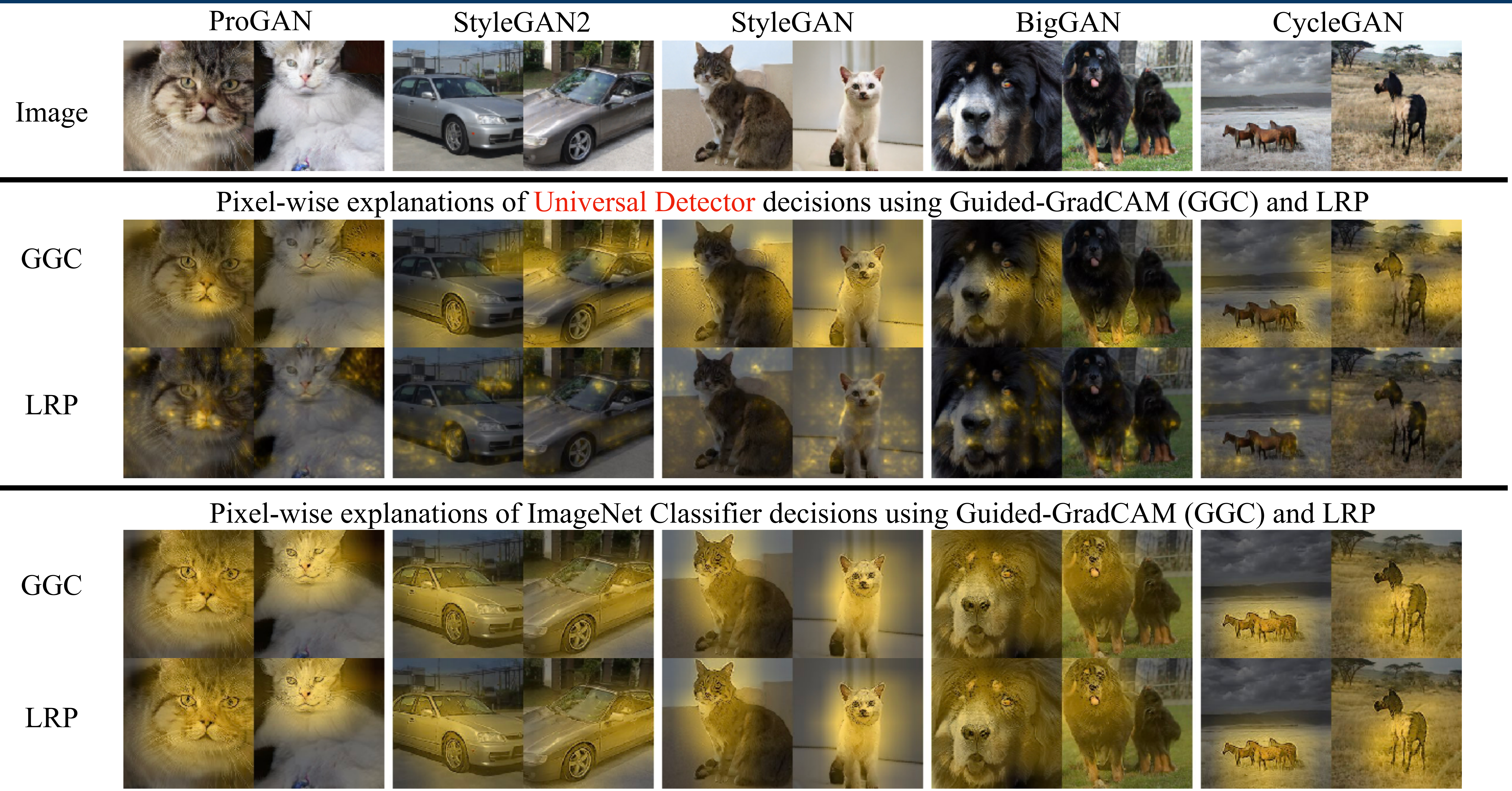
Samples generated using Zero-Insertion based Upsampling Architectural Variant (Chandrasegaran et al., 2021) of Progressive Growing of GAN (Karras et al., 2018)

Chandrasegaran, K., Tran, N. T., & Cheung, N. M. (2021). *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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.

Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). *CNN-generated images are surprisingly easy to spot... for now*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

Are existing Interpretability methods capable of discovering *T-FF* ?



$p_{counterfeit} \geq 95\ \%$

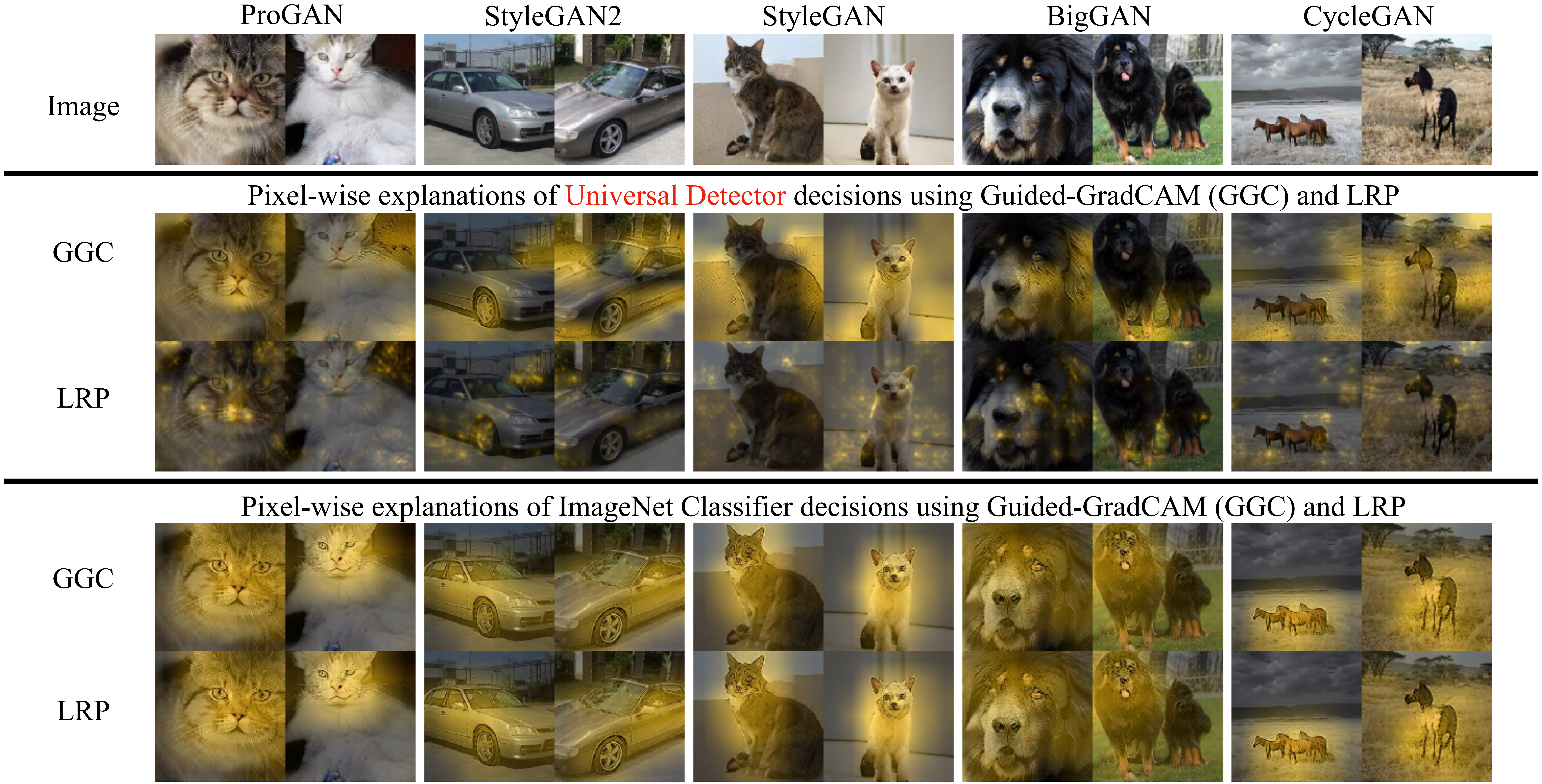
for all these counterfeits

Explanations are random and do not reveal any meaningful visual features

This is a control experiment

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.
Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and improving the image quality of stylegan*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Karras, T., Laine, S., & Aila, T. (2019). *A style-based generator architecture for generative adversarial networks*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Brock, A., Donahue, J., & Simonyan, K. (2018, September). *Large Scale GAN Training for High Fidelity Natural Image Synthesis.* In International Conference on Learning Representations.
Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). *Unpaired image-to-image translation using cycle-consistent adversarial networks.* In Proceedings of the IEEE international conference on computer vision
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization*. In Proceedings of the IEEE international conference on computer vision
Bach S, Binder A, Montavon G, Klauschen F, Müller KR, et al. (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE 10(7): e0130140.

Are existing Interpretability methods capable of discovering *T-FF* ?



$p_{counterfeit} \geq 95\ \%$ for all these counterfeits

Explanations are random and do not reveal any meaningful visual features

This is a control experiment

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.
Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and improving the image quality of stylegan*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Karras, T., Laine, S., & Aila, T. (2019). *A style-based generator architecture for generative adversarial networks*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Brock, A., Donahue, J., & Simonyan, K. (2018, September). *Large Scale GAN Training for High Fidelity Natural Image Synthesis.* In International Conference on Learning Representations.
Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). *Unpaired image-to-image translation using cycle-consistent adversarial networks.* In Proceedings of the IEEE international conference on computer vision
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization*. In Proceedings of the IEEE international conference on computer vision
Bach S, Binder A, Montavon G, Klauschen F, Müller KR, et al. (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE 10(7): e0130140.

Are existing Interpretability methods capable of discovering *T-FF* ?

ProGAN | StyleGAN2 | StyleGAN | BigGAN | CycleGAN

Image

$p_{counterfeit} \geq 95\ \%$

for all these counterfeits

Pixel-wise explanations of Universal Detector decisions using Guided-GradCAM (GGC) and LRP

GGC

LRP

Explanations are random and do not reveal any meaningful visual features

Pixel-wise explanations of universal detector decisions are not informative to discover *T-FF*

LRP

This is a control experiment

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.
Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and improving the image quality of stylegan*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Karras, T., Laine, S., & Aila, T. (2019). *A style-based generator architecture for generative adversarial networks*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Brock, A., Donahue, J., & Simonyan, K. (2018, September). *Large Scale GAN Training for High Fidelity Natural Image Synthesis.* In International Conference on Learning Representations.
Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). *Unpaired image-to-image translation using cycle-consistent adversarial networks.* In Proceedings of the IEEE international conference on computer vision
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization*. In Proceedings of the IEEE international conference on computer vision
Bach S, Binder A, Montavon G, Klauschen F, Müller KR, et al. (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE 10(7): e0130140.

We study the *Feature Space* of universal detectors

Samples generated using Zero-Insertion based Upsampling Architectural Variant (Chandrasegaran et al., 2021) of Progressive Growing of GAN (Karras et al., 2018)

Chandrasegaran, K., Tran, N. T., & Cheung, N. M. (2021). *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
Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.
Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). *CNN-generated images are surprisingly easy to spot... for now*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

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

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

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

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

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

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

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

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

$\boldsymbol{\omega}$ for a feature map quantifies $\dfrac{\textit{positive forensic relevance of the feature map}}{\textit{total unsigned forensic relevance of the entire layer}}$

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

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

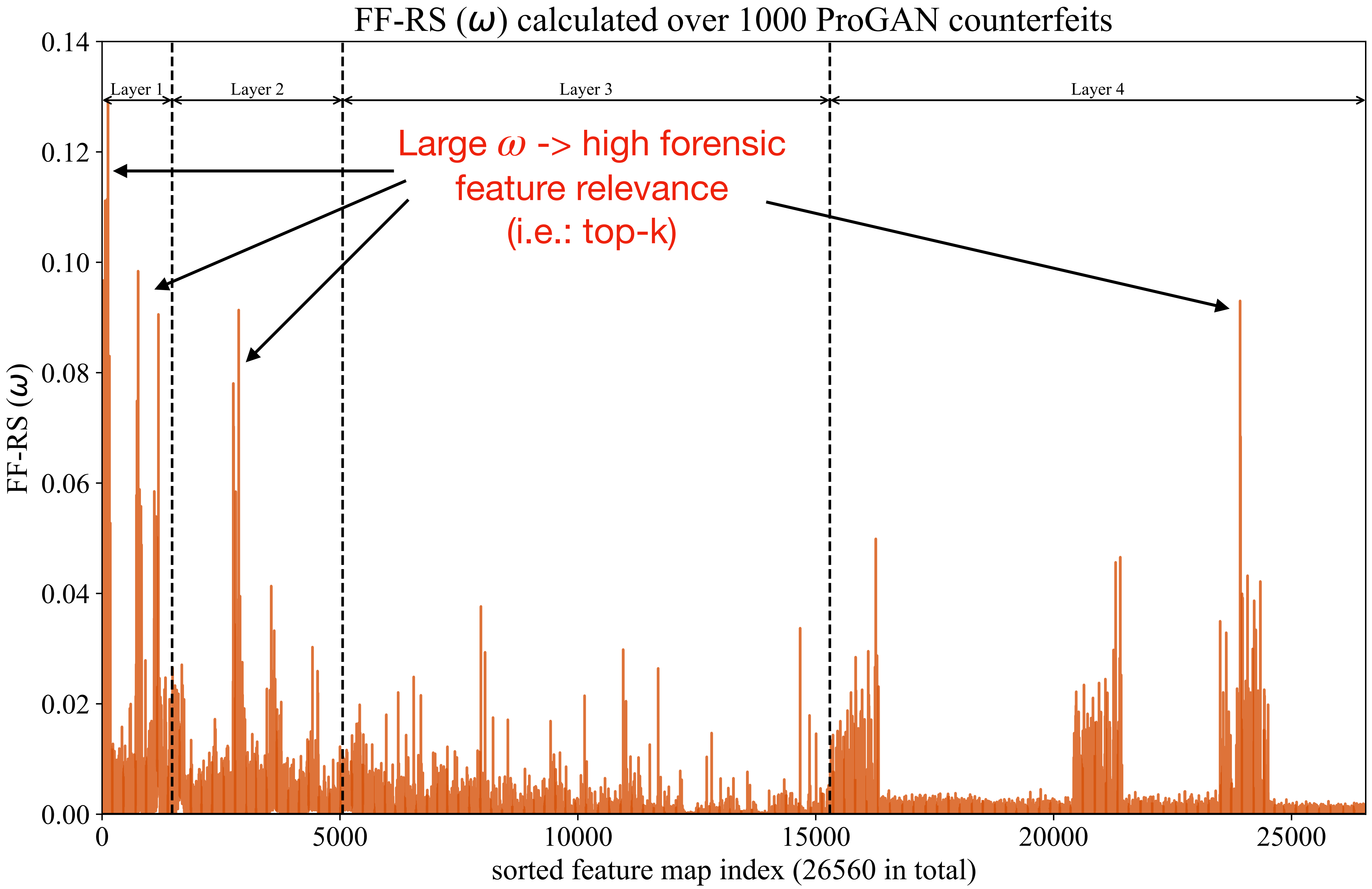
FF-RS ($\boldsymbol{\omega}$) is a scalar ([0,1]) that quantifies the forensic relevance of every feature map.

$$\boldsymbol{\omega} \text{ for a feature map quantifies } \frac{\textit{positive forensic relevance of the feature map}}{\textit{total unsigned forensic relevance of the entire layer}}$$

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**
            $\boxed{r_i(l', c', h, w) \leftarrow \frac{max(0, r_i(l', c', h, w))}{\sum_{c,h,w} ||r_i(l', c, h, w)||}}$
            $R.append(r_i)$; /*r_i.size():(layer, channel, height, width)*/
        end
    end
end
$\boxed{\omega(l_c) \leftarrow \sum_{h,w} \frac{1}{N} \sum_i^n R_i(l, c, h, w)}$; /*forensic feature relevance*/
return $\omega(l_c)$



Validation of our proposed *FF-RS* for discovering *T-FF*

top-k : Set of *T-FF* (top-ranked feature maps based on $\boldsymbol{\omega}$ values)

random-k : Set of random feature maps used as a control experiment

low-k : Set of low-ranked feature maps with very small $\boldsymbol{\omega}$ values

Validation of our proposed *FF-RS* for discovering *T-FF*

top-k : Set of *T-FF* (top-ranked feature maps based on $\boldsymbol{\omega}$ values)

random-k : Set of random feature maps used as a control experiment

low-k : Set of low-ranked feature maps with very small $\boldsymbol{\omega}$ values

Feature map dropout is performed by suppressing (zeroing out) the resulting activations of corresponding feature maps

Validation of our proposed *FF-RS* for discovering *T-FF*

top-k : Set of *T-FF* (top-ranked feature maps based on ω values)

random-k : Set of random feature maps used as a control experiment

low-k : Set of low-ranked feature maps with very small ω values

Feature map dropout is performed by suppressing (zeroing out) the resulting activations of corresponding feature maps

ResNet-50 feature map dropout results (Sensitivity assessments)

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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.
Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and improving the image quality of stylegan*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Karras, T., Laine, S., & Aila, T. (2019). *A style-based generator architecture for generative adversarial networks*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Brock, A., Donahue, J., & Simonyan, K. (2018, September). *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. In International Conference on Learning Representations.
Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). *Unpaired image-to-image translation using cycle-consistent adversarial networks.* In Proceedings of the IEEE international conference on computer vision
Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). *Stargan: Unified generative adversarial networks for multi-domain image-to-image translation*. In Proceedings of the IEEE conference on computer vision and pattern recognition
Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). *Semantic image synthesis with spatially-adaptive normalization*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

Validation of our proposed *FF-RS* for discovering *T-FF*

top-k : Set of *T-FF* (top-ranked feature maps based on ω values)

random-k : Set of random feature maps used as a control experiment

low-k : Set of low-ranked feature maps with very small ω values

Feature map dropout is performed by suppressing (zeroing out) the resulting activations of corresponding feature maps

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

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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.
Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and improving the image quality of stylegan*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Karras, T., Laine, S., & Aila, T. (2019). *A style-based generator architecture for generative adversarial networks*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Brock, A., Donahue, J., & Simonyan, K. (2018, September). *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. In International Conference on Learning Representations.
Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). *Unpaired image-to-image translation using cycle-consistent adversarial networks.* In Proceedings of the IEEE international conference on computer vision
Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). *Stargan: Unified generative adversarial networks for multi-domain image-to-image translation*. In Proceedings of the IEEE conference on computer vision and pattern recognition
Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). *Semantic image synthesis with spatially-adaptive normalization*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

What *counterfeit properties* are detected by this set of *T-FF* discovered using *FF-RS?*

Samples generated using Zero-Insertion based Upsampling Architectural Variant (Chandrasegaran et al., 2021) of Progressive Growing of GAN (Karras et al., 2018)

Chandrasegaran, K., Tran, N. T., & Cheung, N. M. (2021). *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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations.

Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). *CNN-generated images are surprisingly easy to spot... for now*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

What *counterfeit properties* are detected by this set of *T-FF*?

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

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

Maximum spatial relevance neuron

Back-propagate from maximum spatial relevance neuron during LRP

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

Input:
forensics detector M,
counterfeit image x where $x.size() = (3, x_{height}, x_{width})$,
forensic feature map l, c where l, c indicate layer and channel index respectively.

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

$z_{l_c}(x) \leftarrow LRP - FORWARD(M_{l_c}(x_i))$; `/*(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 $\hat{E}_{l_c}(x)$

Color is a critical *T-FF* (Qualitative Studies)

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

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

Color is a critical *T-FF* (Qualitative Studies)

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

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

We qualitatively show that color is a critical *T-FF* in universal detectors for *cross-model forensic transfer*

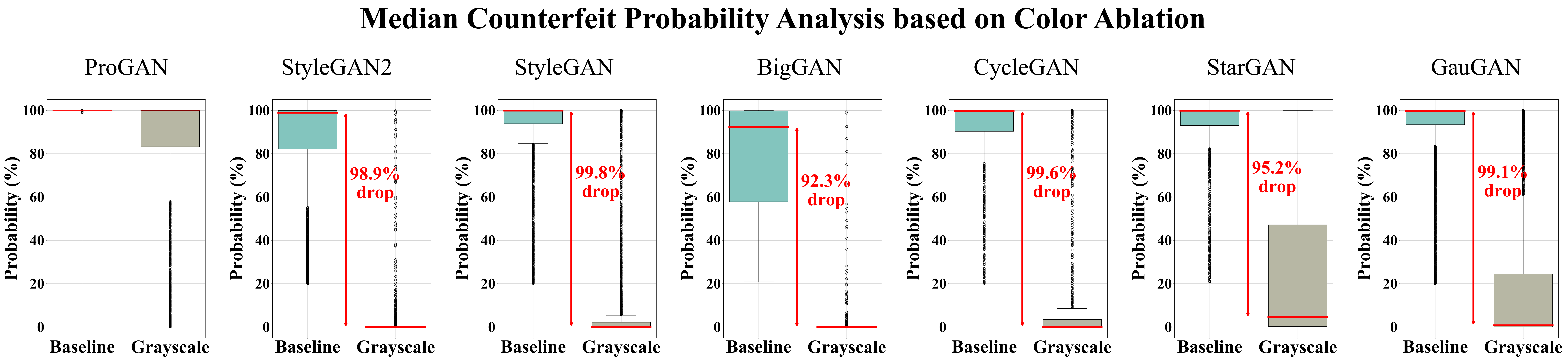
Color is a critical *T-FF* (Quantitative Studies)

Median Counterfeit Probability Analysis based on Color Ablation

We study the change in median counterfeit probability when removing color information (grayscale) from counterfeits.

$p_{counterfeit} = 0.97$ $p_{counterfeit} = 0.96$ $p_{counterfeit} = 0.95$ $p_{counterfeit} = 0.97$ $p_{counterfeit} = ?$ $p_{counterfeit} = ?$ $p_{counterfeit} = ?$ $p_{counterfeit} = ?$

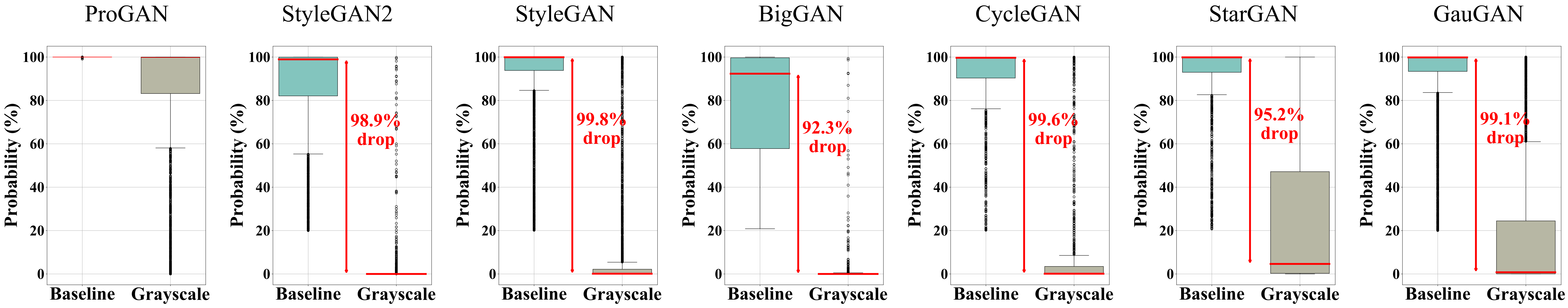
Color is a critical *T-FF* (Quantitative Studies I)



Color ablation causes the median probability predicted by universal detector to drop by > 89% across all unseen GANs

Color is a critical *T-FF* (Quantitative Studies II)

Median Counterfeit Probability Analysis based on Color Ablation



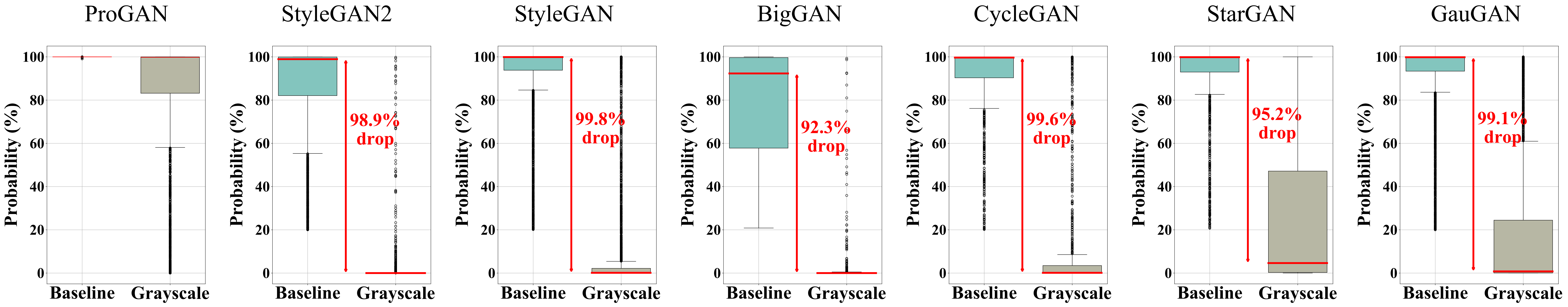
Color ablation causes the median probability predicted by universal detector to drop by > 89% across all unseen GANs

% Color-conditional *T-FF* (Mood's median test) based on maximum spatial activation distributions

% Color-conditional	ProGAN	StyleGAN2	StyleGAN	BigGAN	CycleGAN	StarGAN	GauGAN
ResNet-50	?	?	?	?	?	?	?

Color is a critical *T-FF* (Quantitative Studies II)

Median Counterfeit Probability Analysis based on Color Ablation



Color ablation causes the median probability predicted by universal detector to drop by > 89% across all unseen GANs

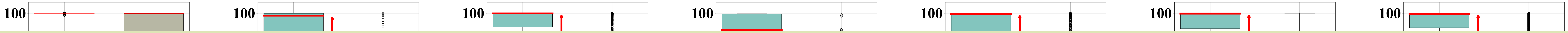
% Color-conditional *T-FF* (Mood's median test) based on maximum spatial activation distributions

% Color-conditional	ProGAN	StyleGAN2	StyleGAN	BigGAN	CycleGAN	StarGAN	GauGAN
ResNet-50	85.1	74.6	73.7	68.4	86.8	71.1	70.2

Color is a critical *T-FF* (Quantitative Studies)

Median Counterfeit Probability Analysis based on Color Ablation

ProGAN | StyleGAN2 | StyleGAN | BigGAN | CycleGAN | StarGAN | GauGAN



We quantitatively show that color is a critical *T-FF* in universal detectors for cross-model forensic transfer

$p_{counterfeit} = 0.97$ $p_{counterfeit} = 0.96$ $p_{counterfeit} = 0.95$ $p_{counterfeit} = 0.97$ $p_{counterfeit} = 0.04$ $p_{counterfeit} = 0.10$ $p_{counterfeit} = 0.00$ $p_{counterfeit} = 0.01$

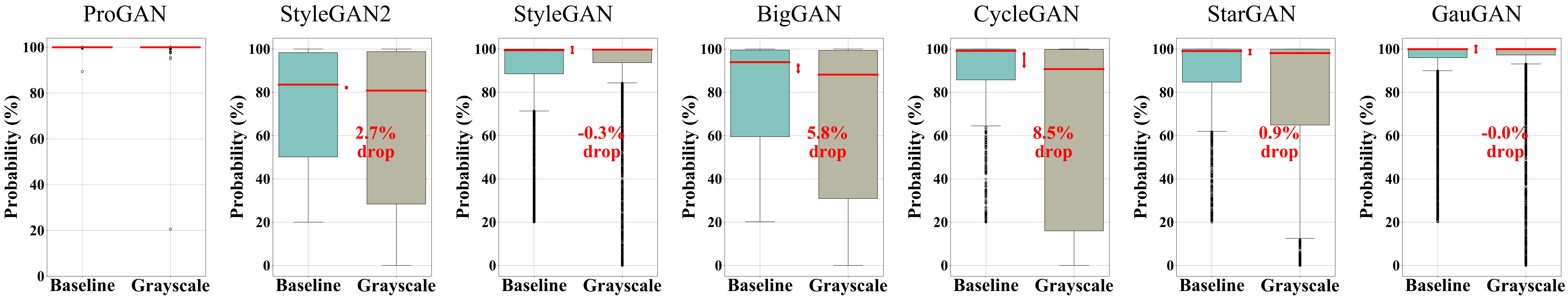
% Color-conditional *T-FF* (Mood's median test) based on maximum spatial activation distributions

% Color-conditional	ProGAN	StyleGAN2	StyleGAN	BigGAN	CycleGAN	StarGAN	GauGAN
ResNet-50	85.1	74.6	73.7	68.4	86.8	71.1	70.2

Applications : Color-Robust (CR) Universal Detectors

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

CR Detector



% Color-conditional *T-FF* (Mood's median test) based on maximum spatial activation distributions

% Color-conditional	ProGAN	StyleGAN2	StyleGAN	BigGAN	CycleGAN	StarGAN	GauGAN
ResNet-50	85.1	74.6	73.7	68.4	86.8	71.1	70.2
CR-ResNet-50	**55.3**	**33.3**	**48.2**	**31.6**	**56.1**	**48.2**	**39.5**

Key Takeaways

We propose a novel Forensic Feature Relevance Statistic (FF-RS) to quantify & discover Transferable Forensic Features *(T-FF)* in universal detectors for counterfeit detection.

Key Takeaways

We propose a novel Forensic Feature Relevance Statistic (FF-RS) to quantify & discover Transferable Forensic Features *(T-FF)* in universal detectors for counterfeit detection.

We qualitatively and quantitatively show that color is a critical Transferable Forensic Feature *(T-FF)* in universal detectors for counterfeit detection

Key Takeaways

We propose a novel Forensic Feature Relevance Statistic (FF-RS) to quantify & discover Transferable Forensic Features *(T-FF)* in universal detectors for counterfeit detection.

We qualitatively and quantitatively show that color is a critical Transferable Forensic Feature *(T-FF)* in universal detectors for counterfeit detection

Based on our findings, we propose a simple data augmentation scheme to train Color-Robust (CR) universal detectors

Code / Pre-trained models