TOWARDS HUMAN-UNDERSTANDABLE VISUAL EXPLANATIONS: HUMAN-IMPERCEPTIBLE CUES CAN BETTER BE REMOVED

Anonymous authors
Paper under double-blind review

ABSTRACT

Explainable AI (XAI) methods focus on explaining what a neural network has learned - in other words, identifying the features that are the most influential to the prediction. In this paper, we call them "distinguishing features". However, whether a human can make sense of the generated explanation also depends on the perceptibility of these features to humans. To make sure an explanation is human-understandable, we argue that the capabilities of humans, constrained by the Human Visual System (HVS) and psychophysics, need to be taken into account. We propose the *human perceptibility principle for XAI*, stating that, to generate human-understandable explanations, neural networks should be steered towards focusing on human-understandable cues during training. We conduct a case study regarding the classification of real vs. fake face images, where many of the distinguishing features picked up by standard neural networks turn out not to be perceptible to humans. By applying the proposed principle, a neural network with human-understandable explanations is trained which, in a survey, is shown to better align with human intuition. This is likely to make the AI more trustworthy and opens the door to humans learning from machines. In the case study, we specifically investigate and analyze the behaviour of the human-imperceptible high spatial frequency features in neural networks and XAI methods.

1 INTRODUCTION

Most of existing heatmap-based XAI methods such as LRP (Bach et al. (2015)), GradCAM (Selvaraju et al. (2017)) or, more recently, Vision Transformer (ViT)'s attention (Dosovitskiy et al. (2020)) have shown their ability on neural network explanation for traditional classification tasks, e.g. on CIFAR (Durall et al. (2019)), ImageNet (Deng et al. (2009)) or CUB (Welinder et al. (2010)). The generated explanations from these methods highlight the (most discriminative) regions related with the predicted class in the image and align well with human intuition. These popular classification datasets have been chosen to evaluate XAI because it is clear, for humans, which areas should be highlighted, providing a mechanism to evaluate the XAI methods. They have a relatively large inter-class visual variance (e.g. cat vs. car vs. fish in ImageNet) or, for fine-grained datasets like CUB, even though all the images are birds, the visual features defining the classes are known and determined, e.g. color, shape, and texture of particular parts. We refer to the features that define the difference between classes as *distinguishing features*. The main characteristic of these popular classification datasets is that *the distinguishing features are perceptible by humans.* In summary, humans can understand most of the explanations generated by existing methods on these datasets because *the distinguishing features are perceptible by both the current neural networks and humans.* This implies a common assumption in evaluating heatmap-based explanation methods, namely that an explanation that aligns well with human-understandable semantics is a better explanation of the model prediction. See the cat image in Fig. 1 for an example.

However, for other datasets, the visual inter-class variance is small, and the distinguishing features may not immediately be obvious to humans. For instance, consider the distinction between GAN-generated (fake) face images and real face images. It is under these circumstances, especially, that explanations of automated decisions become important. Yet unlike the cat example in Fig. 1, the fake face example shown below leads to an unintuitive explanation: the generated heatmap based

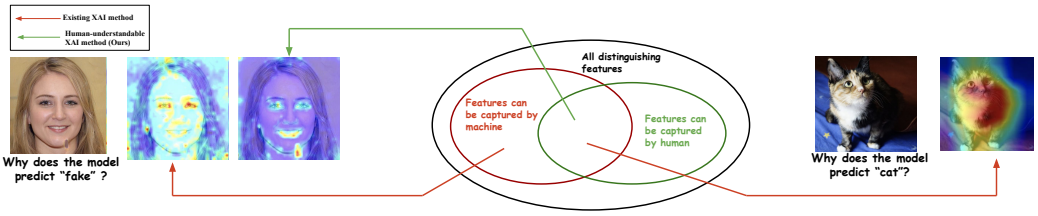

Figure 1: The perception mechanism for human and machine is different. Existing XAI methods focus on explaining the machine. Humans can only understand their explanation in as far as the model picks up distinguishing features that can be perceived by both machine and human. However, for some datasets distinguishing features used by the machine are not perceptible to humans. In this paper we propose a method to generate more human-understandable models and explanations on these datasets. Please zoom in to check the detail for the fake face example.

on the same method (left) highlights almost the entire image, leaving humans confused about the meaning of this explanation and leading to reduced rather than increased trust in the AI. This is in contrast to the heatmap generated using our method (right), which is more intuitive and helps humans to understand the decision.

We therefore introduce the **Human Perceptibility Principle for XAI**: *To generate human-understandable explanations, a neural network should be steered towards using features that are perceptible by humans.* We believe it is an important complement to current XAI methods, which mostly optimize their faithfulness w.r.t. to the model being explained and ignore the human factors which lead to the intelligibility of the provided explanation.

To this end, we revisit models of the Human Visual System and pyschophysics, in order to identify the main flaws of human vision. One of them is the poor perceptibility of high spatial frequency features (Pennebaker & Mitchell (1992)). On the contrary, neural networks, especially convolutional neural networks (CNN), have a tendency to focus on high spatial frequency features (Geirhos et al. (2019); Wang et al. (2020)). When a relevant distinguishing feature is related to high spatial frequency, the explanation from current methods can therefore be difficult for humans to understand.

We illustrate these ideas based on a case study: explaining the classification model for fake vs. real face images. We believe this is important and instructive since a more human-understandable explanation can help non-experts recognize fake face images in daily life while current explanation methods fail to do so. To evaluate our method, we conduct a survey, asking users' opinions on the generated explanations of our method and the method without applying our proposed principle (vanilla method).

In summary, this paper makes the following contributions: i) we focus on making XAI methods more human-understandable and propose the novel *Human perceptibility principle*, which, we believe, can be an important complement to current XAI research. ii) We theoretically and empirically show that human and neural networks have different preferences on capturing features from images, especially with respect to the high spatial frequency features, and propose a principle to generate human-understandable explanations based on the foundations of the Human Visual System and the psychophysics model. iii) We study the deepfake case where current XAI methods fail to generate human-understandable explanations. The qualitative results and our survey clearly show that explanations can be made more perceptible/understandable for humans by applying the proposed approach. In addition, we investigate and analyze the ability of recent vision transformers at learning/encoding and explaining high spatial frequency features.

2 Related Work

Faithful model explanation Input-modification-based XAI methods (Zeiler & Fergus (2014); Grün et al. (2016); Fong & Vedaldi (2017)), a.k.a. occlusion / perturbation-based methods, systematically occlude or modify parts of the input with the goal of modelling how modifications in the input space affect the network predictions. They operate under the assumption that there is no access to the internal states of the model to be explained, treating it as a black-box. As a consequence, they

possess reduced guarantees regarding to the faithfulness of the explanation w.r.t. the inner-workings of the model.

Based on the previous observation, another group of work (Simonyan et al. (2014); Smilkov et al. (2017); Zeiler & Fergus (2014); Selvaraju et al. (2017); Oramas M et al. (2019)) has been proposed with the goal of optimizing the faithfulness of the explanations w.r.t. the model being explained. By optimizing faithfulness, these methods successfully shed light into some of the internal decision-making processes of the networks. However they suffer from low-intelligibility and ambiguity in the produced explanations. All these works fixate on one direction, that is finding and visualizing the distinguishing features taken from the neural network when it is trained for a classification task. Different from them, we aim at finding a trade-off where the cues highlighted by the generated explanations are both important for the decision-making process of the network, and understandable by non-expert users. Moreover, while these methods operate in a post-hoc manner, the proposed principle aims at the design of models whose internal representation is interpretable-by-design.

Explanations through intermediate elements Following the previous efforts, a new group of methods has emerged which aim at generating explanations based on "concepts" derived from the model being explained. Towards this goal, Kim et al. (2018) train linear classifiers to derive concept vectors and link the importance of each concept with the classes of interest. While these concepts can be human-understandable, they have the requirement of being pre-defined. Furthermore, Ghorbani et al. (2019) compute superpixels from each image example. Then, concepts are identified by clustering all the superpixels, from the entire dataset, that are important for the model. Oramas M et al. (2019) identify sparse features encoded in the model that are relevant for the classes of interest. These features are then used as means for explanation. The methods listed above use intermediate concepts extracted from internal activations from the model. Therefore, while capable of providing an intuition, they do not necessarily possess a direct semantic meaning, thus, reducing their intelligibility.

To address this issue, Zhou et al. (2018) propose decomposing neural activations from images into semantically interpretable components, pre-trained from a large concept corpus. Then, explanations are obtained by projecting the feature vector into the learned interpretable basis. Very recently, Koh et al. (2020) have revisited the idea of predicting concepts, provided at training time, and then use these concepts to predict the label. These works enable attaching understandable text-based concepts as part of the output of a model. On the downside, this capability comes at the cost of additional semantic concept annotations for training. Moreover, for Zhou et al. (2018) expensive pixel-level annotations are also required.

Different from them, we investigate an orthogonal direction where we stress that distinguishing features must first be perceptible to humans in order to grant the characteristic of being intelligible.

3 METHODOLOGY

Problem Statement Existing heatmap-based model explanation methods try to explain the decision made by a pretrained neural network. In other words, the goal is to identify the most important features (for the addressed task) extracted by the neural network from the input image. Fig. 2 shows some explanations from ResNet50 with GradCAM (Selvaraju et al. (2017)) and ViT with attention-based method applied on the Imagenet and CUB dataset. People can easily understand the explanation since the distinguishing features are both perceptible by the machine and humans. Fig. 3 shows some examples of GradCAM and ViT (Dosovitskiy et al. (2020); Abnar & Zuidema (2020)) attention applied on a deepfake dataset (Karras et al. (2018)), where the model is trained to classify images as fake or real. Obviously, these methods fail. Not only do they generate very different explanations focusing on different image regions / distinguishing features, it is also difficult for humans to understand the generated explanations - or tell which one to trust more. Therefore, the main research question of this paper is: *For this type of scenario, is there a means to generate explanations that are more understandable by humans*

3.1 REVISITING HUMAN VISUAL SYSTEM AND PSYCHOPHYSICS MODEL

To answer the question, we need to identify the features that can / cannot be perceived by humans. Therefore, we revisit the Human Visual System (HVS) and Psychophysics model and discuss the

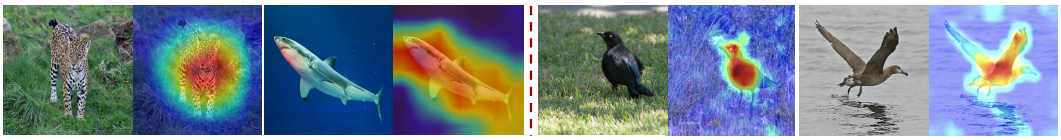

Figure 2: ResNet50 with GradCAM and ViT Attention-based method applied on the Imagenet dataset (left two) and CUB dataset (right two). People can easily understand the relationship between the heatmap and its corresponding class.

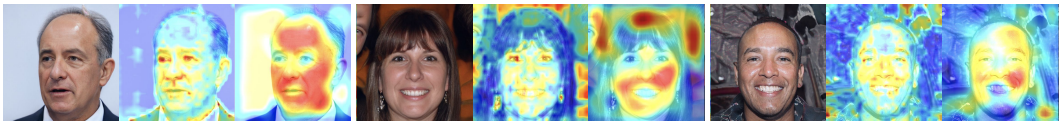

Figure 3: ViT attention (middle) and GradCAM visualization (right) for three styleGAN-generated fake images, based on the neural networks that classify the fake and real images. Note how two different XAI methods provide very different heatmaps/explanations. Can you understand why the network decided to focus on those regions?

main characteristics related to human perception on digital images. We limit ourselves to the essence and do not elaborate on the biology principles behind them.

Luminance and Color The HVS has more resolution on luminance than chroma information. This is one of the motivations of chroma subsampling in image compression field (S. Winkler & Kunt (2001)). In addition, due to the features of the cone, the HVS can perceive a limited variety of colors, estimated around 10 million. On the contrary, we know color can be represented as a numerical combination (e.g. RGB color space) in a machine, which leads to more than 16 million colors (with 8 bit per channel). The total number is clearly much larger than what can be perceived by humans.

Weber's Law Weber's Law describes that the Just-Noticeable Difference (JND) dS is proportional to the initial stimuli intensity S, $dS = KS$, where K is a constant. JND is the smallest change in stimuli that can be perceived by humans (Kandel Eric R. (2013)). In the context of digital images, the Weber's Law explains why humans are more sensitive to detailed regions, e.g. texture, edges, where there is a relative large difference around the neighbouring pixels, rather than flat regions. Likewise, structures in dark image areas are often missed.

Perceived Spatial Frequency Weber's Law indicates that humans are more sensible on detailed regions of an image. However, subtle differences in regions with very high spatial frequency are not perceptible by humans. For instance, JPEG (Pennebaker & Mitchell (1992)) takes advantage of this characteristic to quantize these high frequency components without a perceptible loss of quality by humans. In short, high frequency components are less perceptible to a human.

3.2 Revisiting Machine Vision

The most significant difference w.r.t. the HVS is that images are stored and processed as numbers in a machine. In other words, the perception mechanism is totally different from the HVS. Therefore, machines can easily distinguish very similar colors simply because the numbers of their RGB representation are different.

Here we focus the discussion on convolutional neural networks (CNN). We treat transformers as one kind of CNN since the convolutional blocks are still used in the architecture. Please refer to the original paper (Dosovitskiy et al. (2020)) for more technical details. Different from the HVS, CNNs have a good ability to capture high spatial frequency. (Geirhos et al. (2019); Liu et al. (2020); Wang et al. (2020)) have shown that CNNs leverage texture information significantly for classification tasks to the point of even being biased towards texture rather than shape (Geirhos et al. (2019)). Oramas M et al. (2019) also show that front layers of CNNs usually capture low-level features, e.g. color and texture.

3.3 PROPOSED APPROACH

In this section, we first describe the Human Perceptibility Principle for XAI in more detail. Given a dataset D and corresponding classification task T, there exists a set of distinguishing features $\Phi = [\phi_1, \phi_2, ..., \phi_n]$ based on which the classes in D can be distinguished (i.e. the classification task is feasible). In practice, we have a subset $\Phi^h \subseteq \Phi$ of features perceptible by humans and another subset $\Phi^m \subseteq \Phi$ of features perceptibly by machines. When training a neural network f on the dataset, it will use a subset Ψ^m of the distinguishing features in Φ^m to base its decisions on. If, however, $\Psi^m \cap \Phi^h = \emptyset$, the used features are not perceptible by humans and no human-understandable explanation can be generated. Therefore, the machine should be steered away from using features that are exclusively perceptible by machine $\Phi^m \setminus \Phi^h$, and stimulated to consider more human-understandable ones $\Phi^m \cap \Phi^h$.

To achieve this, there are three possible approaches: i) use a pre-processing technique to get rid of these imperceptible features at the image level (input), ii) use data-augmentation techniques to learn the network to become invariant to differences imperceptible by humans, or iii) during the training process, make the neural network focus less on these features, e.g. using additional loss terms penalizing the use of such features. Obviously, this is only possible if indeed, there exist human-understandable distinguishing features Φ^h for task T in D.

Based on the HVS and the psychophysics model, we identify the following characteristics that XAI methods should take into account in order to generate human-understandable explanations: i) Human eyes cannot distinguish very fine-grained color differences. ii) Human eyes can hardly perceive the difference in the high spatial frequency components. iii) Human eyes are sensitive to edge regions.

In our study, we focus on high spatial frequency components as the imperceptible distinguishing feature. We select the first approach and use a *bilateral filter* (Manduchi & Tomasi (1998)) to process the input images. It can smooth high frequency regions, narrowing the difference between the frequency distribution of fake and real images, as well as preserve the edges, where humans are sensitive to perceive, which exactly meets the requirement of the HVS. For popular classification datasets, the main difference between the classes usually lie on the shape, color and/or visible texture. This information is perceptible by both machines and humans, and we believe this is the reason why existing XAI methods can generate reasonable good explanations (to humans) on these datasets.

In this paper we do not focus on designing a new explanation method but rather focus on investigating means for making them more human-understandable.

4 EXPERIMENTS

4.1 DATASETS

FFHQ-HF-WS We start with a controlled experiment. To this end, we construct a two-class artificial dataset FFHQ-HF-WS, where we mimic two distinguishing features, with high and low spatial frequency, respectively, on the images of one of the classes. We sample 10K images from the FFHQ dataset (Karras et al. (2018)) and resize them to 224×224. For class 1, we add a high spatial frequency feature by periodically changing the intensity of the pixels row by row. More specifically, the RGB pixel value is set to 0.9 times the original value every two rows. We also add a low spatial frequency cue by putting a 15×15 white square on the image at a random location. Please note, there is no spatial frequency change inside of the white square. Some examples of this dataset are illustrated in Fig. 4. For class 2, we do not introduce any change. In total, there are 20K images, we use 16K for training and 4K for testing.

DeepFakes For the realistic GAN-generated deepfake images, we use a subset of Faces-HQ dataset (Durall et al. (2019)), which contains 10K 1024×1024 images from the FFHQ dataset as real ones and the same number, and resolution, of images from www.thispersondoesnotexist.com (TPDE) as fake ones. The fake images are generated by styleGAN (Karras et al. (2018)). In addition, we also use 10K CelebA-HQ (Lee et al. (2020)) images and 10K fake images generated by styleGAN trained on the CelebA-HQ dataset. We split them into training, validation and test splits with the proportion of $0.7 : 0.15 : 0.15$, respectively.

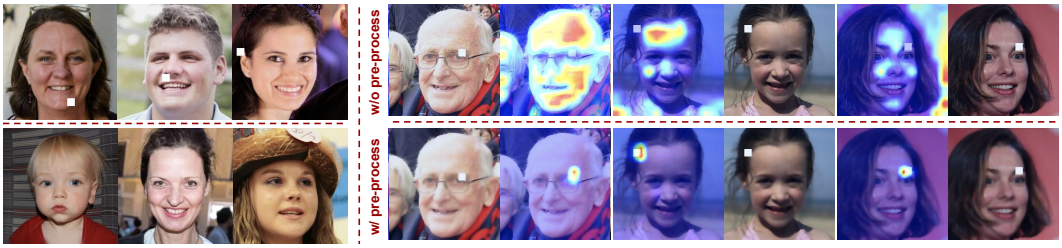

Figure 4: Examples for class-1 (top-left) and class-2 (bottom-left) from the FFHQ-HSF-WS dataset. The visualizations on top-right are computed by using the original images for training while for the ones below we apply the bilateral filter as pre-processing.

4.2 Controlled Experiment

In this section, we design a controlled experiment to show that by removing the human imperceptible features, machines can shift towards using human-understandable cues (if they exist).

We first use our FFHQ-HF-WS dataset to train a binary classifier based on a two-layer ViT-16 (Dosovitskiy et al. (2020)). Based on what we know about the HVS vs machine vision, white squares should be picked up by humans more easily while the neural network should prefer the high spatial frequency features. The classification accuracy is 1 since it is an easy task. We use the attention of the ViT as the explanation (Wiegreffe & Pinter (2019); Chefer et al. (2020)). To visualize it we follow the method proposed by Abnar & Zuidema (2020) to roll out the attention. We show some visualizations on the top-right part of Fig. 4. It shows that the heatmaps focus mostly (albeit not exclusively) on some large regions, i.e. face, background, where the high spatial frequency feature is applied and clear, rather than the white square region. This verifies that indeed for CNNs, (ViT here), high spatial frequency features are easier to capture.

Then, in order to make the neural network take the cue that is preferably captured by humans, we apply the bilateral filter on the images as a preprocessing step. After processing the data, we conduct a similar experiment. The classification accuracy is 99.8%, slightly lower than the previous one. Similarly, we visualize the attention learned by the neural network in the bottom-right of Fig. 4. Now we can see the attention successfully shifts from the previous high spatial frequency region to the white square region, which aligns better with human perception and, consequently, seems more intuitive to humans.

To quantitatively evaluate the attention heatmap, we calculate its intersection over union (IoU) with the corresponding ground truth (GT) mask. For the high spatial frequency feature, the GT mask is the whole image except the white square while the white square region is the GT mask for the low spatial frequency feature. Inspired by (Choe et al. (2020); Oramas M et al. (2019)), we use 100 thresholds between 0 and 1 to binarize the generated heatmap and calculate the area under the curve (AUC) of the IoU curve. Results presented in Table 1 indicate that indeed the attention heatmap shifts from the high spatial frequency region to the white square region which is more human-understandable, after applying the bilateral filter on images.

Experiment	*AUC-IoU HSF*	*AUC-IoU WS*
w/o pre-process filter	0.18	0.00
with pre-process filter	0.01	0.27

Table 1: AUC-IoU result for the two experiments. AUC-IoU HSF indicates the generated heatmap with high spatial frequency mask while AUC-IoU WS refers to the one with the white square mask.

4.3 Real Case Study: Deepfakes

Here we focus on the detection of deepfake images. More specifically, the images generated by styleGAN (Karras et al. (2018)). Fig. 3 has shown that current model explanation methods fail to provide human-understandable explanations. We try to analyze it and generate more human-understandable explanations for these images. We consider this important and instructive, since a

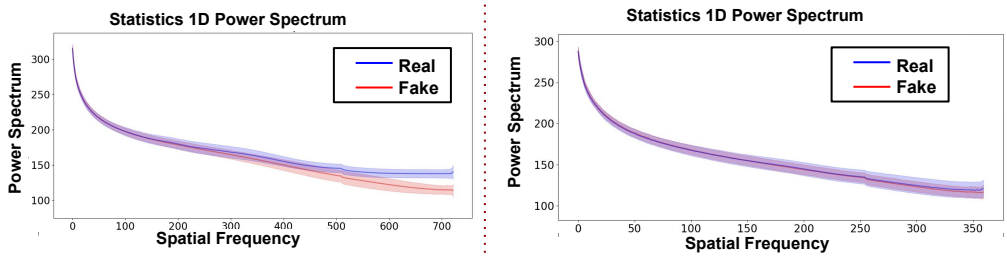

Figure 5: Spatial frequency distributions for real and fake images. Distribution from original images (left) and from images after‘ the pre-processing (right).

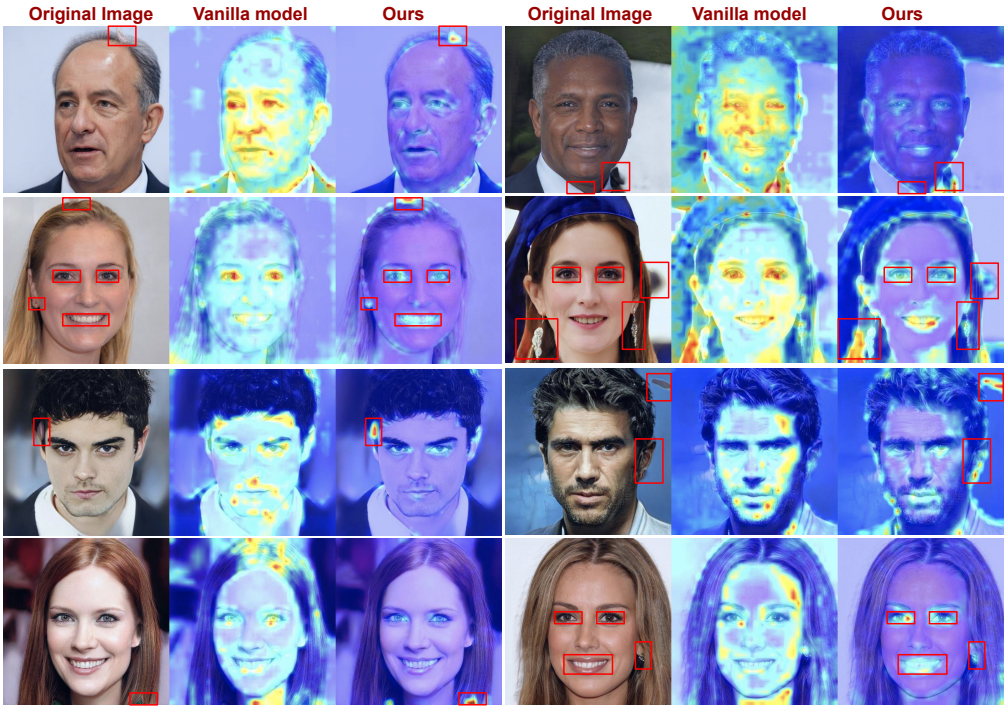

Figure 6: Qualitative comparison for the explanation of *fake* images generated by ViT-16 attention. Examples from the TPDE dataset (row 1-2) and styleGAN trained by celebA-HQ dataset (row 3-4). Do you feel it makes more sense by looking at the right explanation? Please zoom in or refer to Appendix for more details.

more human-understandable explanation can help non-experts have a better ability at recognizing these fake face images in daily life. For the sake of simplicity, we mainly use FFHQ/TPDE images for most of our experiments, except for the qualitative results. Since the results are very similar, we report the experiment based on the CelebA-HQ in Appendix.

What are the unique features in fake images? Recently (Durall et al. (2019; 2020); Liu et al. (2020); Frank et al. (2020)) showed that the spatial frequency component can be used to distinguish GAN-generated images from real images, especially the high spatial frequency part. Durall et al. (2019) even use a simple support vector machine to successfully classify the real/fake images by extracting and using their frequency components. This can explain what happens in Fig. 3: the heatmap might actually be covering the high frequency regions. However, it is difficult for humans to perceive the high frequency differences and, consequently, understand the explanation. Similar to Durall et al. (2019), we analyze the spatial frequency distributions of the real and fake images. For each image, we take Discrete Fourier Transform to get the amplitude spectrum, then the azimuthal average is applied on the 2D amplitude spectrum and the final 1D spatial frequency distribution is obtained. Fig. 5 (left) shows the distribution of 100 random images from the real and fake classes.

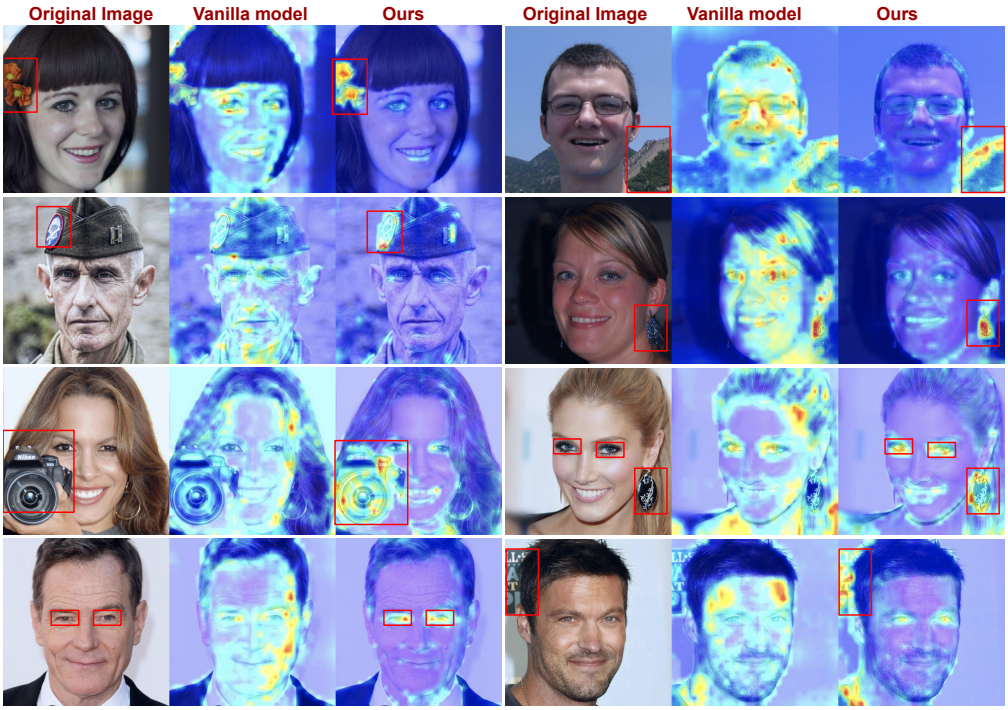

Figure 7: Qualitative comparison for the explanation of *real* images generated by ViT-16 attention. The first two rows are from FFHQ while the last two rows are from celebA-HQ dataset. Do you feel it makes more sense by looking at the right explanation? Please zoom in or refer to Appendix for more details.

It is clear that the fake and real images can be distinguished by observing the high spatial frequency part.

For the human factor, Liu et al. (2020) did a user study, where they asked participants their criteria for fake images. The result shows that users normally take cues like "asymmetrical eyes", "irregular teeth" etc., i.e. the shape and color artifacts, rather than very high spatial frequency pattern. Recently, Guo et al. (2021) showed that irregular pupil shape is also an important clue for detecting GAN-generated face images. These findings are in line with what we concluded from HVS.

Generating Human-understandable explanations According to our approach, we need to reduce the effect of the high frequency feature on the neural network when it is trained. Here we simply use a bilateral filter to pre-process the data. Fig. 5 (right) shows the spatial frequency distribution of these pre-processed images. Please note that the difference in the high frequency part is significantly reduced[1]. Here we use the ViT-16 model with 12 layers to train the classifier and use the attention as the explanation. In addition, in order to accelerate the training process, the model we use is pre-trained on the Imagenet dataset (Dosovitskiy et al. (2020)). The reason we choose Vision transformer is based on the fact that it is patch-based and the human-understandable cues are more local. We also train a classifier using the unprocessed images as a reference model, we refer to it as the vanilla method.

Fig. 6 and Fig. 7 show several explanation examples for fake and real face images, respectively[2]. We use red bounding boxes to indicate some human-understandable regions on these images. Compared with the explanations from the vanilla approach, our method can indeed localize these human-understandable cues, as suggested by Guo et al. (2021) and Liu et al. (2020), such as asymmetrical earrings, weird teeth and eyes, and the bubble artifacts for fake images, versus accessories and particular background features for real images.

[1]The number of bins in the figures are different since we resize the images to 512×512 in order to fit them into the GPU.

[2]Please refer to Appendix for the faithfulness validation of the generated heatmap.

Then, we calculate the prediction accuracy on the testing split. Table 2 indicates that the vanilla method achieves slightly better performance (+1.4 pp), which means that for this type of dataset, high spatial frequency feature is a useful cue for the neural network. This observation is also in line with the works (Geirhos et al. (2019); Liu et al. (2020); Wang et al. (2020)) discussed before. In addition, to verify the dependency of the vanilla model on high frequency features, we measure the performance of the vanilla model (i.e. trained with original images) when classifying filtered images. In this setting, the prediction accuracy drops to 50%, which confirms that indeed the vanilla model takes the spatial frequency feature as the most important cue - without it the model reaches chance levels.

Survey In order to get a human assessment on the explanation heatmaps, we build a website and conduct a survey. For each fake image, we show the original image and two explanations: one generated by the vanilla model and one based on our model. The main question asked in the survey is "*Which explanation heatmap is closer to your idea that the image is fake?*" We use 100 fake images and its corresponding two explanations, from where we randomly pick 20 images to show to the participants each time. After finishing the questionnaire, we also asked two questions for feedback: *Q1: Do you feel you got better at recognizing deepfakes based on the test?* and *Q2: Do you feel the heatmaps helped in this process?*

In total, 84 participants joined the study, 69 of them (82.1%) consider our explanation is closer to what they think (i.e., they selected our method more often than the vanilla method) , 13 of them (15.5%) prefer the explanation from the vanilla method while there are 2 (2.4%) users that consider both explanations equally good (The same number of votes for both methods). On average, our explanations received 71.0% votes while 29.0% votes went to the vanilla model. Regarding the last two questions, 73.2% of the participants feel they got better at recognizing the fake images after the test, 91.5% of them feel the heatmaps they chose were helpful. We list the statistics from the study in Table 5. The survey suggests that indeed by removing the imperceptible high spatial frequency feature, the explanations become more human-understandable. Also, it is interesting to see, that most of the participants feel that they are better at recognizing fake images after doing the test[3].

Experiment	Accuracy (%)
Vanilla	99.97
Ours	98.57
Vanilla model with filtered images	50.00

Table 2: Fake vs real face image prediction accuracy of vanilla model and our model.

Explanation	Participants (%)	Votes (%)
Vanilla	15.5	29.0
Ours	**82.1**	**71.0**
Feedback Questions		Yes (%)
Q1: Better at recognizing fake images?		73.2
Q2: Explanation helpful?		91.5

Table 3: Survey statistics.

Discussion One could argue that that a text-based explanation, such as 'this image differs significantly in its high frequency spectrum from real images', might be enough. Indeed this might be sufficient for a technical user familiar with signal processing. However, this is not necessarily the case for an average user. For average users, a visual heatmap can provide more direct explanation by highlighting the real/fake cues. Therefore, in order to stress the improved intelligibility goal of our work, and increase the accessibility of the generated explanations, we focus on visual heatmaps.

5 Conclusion

We propose the *Human Perceptibility Principle for XAI*. This principle aims at the generation of human understandable explanations by steering the network to use features that are perceptible by humans. We believe it can fill a gap in the current model explanation literature. Results from our evaluation suggest that model explanations are indeed more human-understandable when the proposed principle is applied. We illustrate this on a toy problem as well as on the task of deep fake detection. In addition, the results of our survey suggest that the explanations produced from our method are more understandable to average people and have the potential to effectively serve as a guide on how to address the task on their own.

[3]But obviously, this would have to be evaluated separately.

6 ETHICS STATEMENT

- Based on our generated heatmap visualizations, researchers can improve the quality of GAN-generated images, eliminating the human-understandable flaws. On the contrary, a more advanced deep fake detection method can be proposed. It will be a positive feedback process.
- The generated visual explanations can reveal, to an attacker, the features considered by the detector to detect altered inputs. This could serve as means to improve the attacking method so that it exploits other features not considered by the detector. This is an issue that is common on any method aiming at providing explainability/intepretability capabilities and of any system ensuring transparency.
- For the case study, the feature that is exclusively captured by machine is the high frequency (HF) feature. The high frequency feature is caused by styleGAN. In other words, we train our model on this certain distribution. The bilateral filtering is not applicable on a dataset whose machine-exclusive distinguishing feature is not HF related.
- This work puts forward the overarching goal of investigating means for generating *human-understandable* explanations. However, having explanations that are human-understandable is something subjective in nature. Here we have reported results from a survey that we conducted to validate our method. While preliminary, our results indicate a trend that the explanation visualizations generated via the proposed methods are more perceptible and understandable for average people. Worth noting is that there is a current bias in the age-group of the participants of our survey, which is more focused on young people. Besides addressing this issue, future evaluations should focus on well-studied survey methods.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv:2005.00928*, 2020.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 2015.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. *arXiv preprint arXiv:2012.09838*, 2020.

Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *CVPR*, 2020.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020.

Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv:1911.00686*, 2019.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. *CVPR*, 2020.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, pp. 3449–3457. IEEE Computer Society, 2017.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 13–18 Jul 2020.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*, 2019.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 2019.

Felix Grün, Christian Rupprecht, Nassir Navab, and Federico Tombari. A taxonomy and library for visualizing learned features in convolutional neural networks. *arXiv*, abs/1606.07757, 2016.

Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. *arXiv:2109.00162*, 2021.

Schwartz James H. Siegelbaum Steven A. Hudspeth A. J. Kandel Eric R., Jessell Thomas M. *Principles of neural science*. Springer, 2013. ISBN 9780071390118.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CVPR*, 2018.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2673–2682. PMLR, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zhengzhe Liu, Xiaojuan Qi, and Philip Torr. Global texture enhancement for fake face detection in the wild. *CVPR*, 2020.

R. Manduchi and C. Tomasi. Bilateral filtering for gray and color images. In *ICCV*, 1998.

José Oramas M, Kaili Wang, and Tinne Tuytelaars. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In *ICLR*, 2019.

William B. Pennebaker and Joan L. Mitchell. *JPEG Still Image Data Compression Standard*. Van Nostrand Reinhold, New York, 1992.

C. J. van den Branden Lambrecht S. Winkler and M. Kunt. *Vision models and applications to image and video processing*. Springer, 2001. ISBN 978-0-7923-7422-0.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *ICCV*, abs/1610.02391, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv*, abs/1706.03825, 2017.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *EMNLP*, abs/1908.04626, 2019.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *ECCV*. Springer International Publishing, 2014.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *ECCV (8)*, volume 11212 of *Lecture Notes in Computer Science*, pp. 122–138. Springer, 2018.

7 APPENDIX

In this section we will mainly

- validate the faithfulness of the generated explanation via ViT.
- show the experiment results on celebA-HQ (Lee et al. (2020)) (real) and fake images generated by styleGAN (Karras et al. (2018)) which is trained on celebA-HQ.
- show additional qualitative results and
- show more details on our survey.

7.1 VALIDATING THE GENERATED EXPLANATIONS

In this section, we conduct a perturbation test to study the explanation heatmap and its corresponding model. For each image I, we obtain the attention heatmap A via the model f: $A{=}f(I)$. Then we sort the pixel-level weights in A in descending order. We gradually remove the top-k most important pixels of I according to the relevance-rank indicated in A and send the perturbed input image I_p to f, obtaining a new prediction. We expect that performance will drop as relevant pixels/regions are removed. For reference, we also randomly remove the same amount of pixels on I, noted as I_p^r.

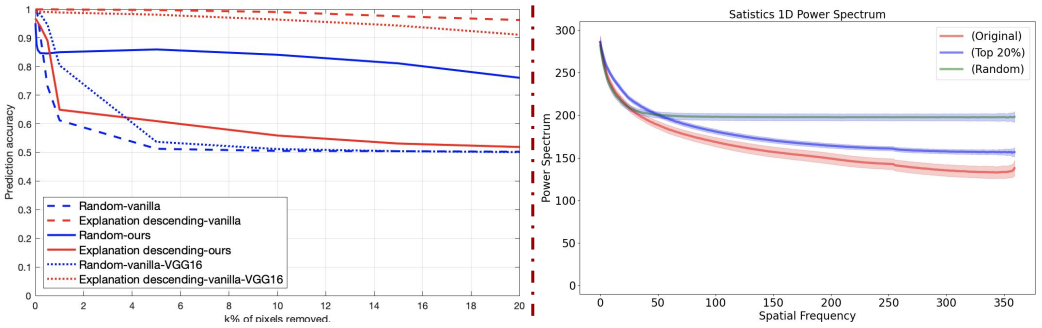

Figure 8: Left: Perturbation test for the vanilla models and our model. Right: Spatial frequency distribution for I (red), I_p (blue) and I_p^r (green).

Fig 8 (left) shows the perturbation test result for the vanilla method (dashed lines) and our human-understandable method (solid lines). For our method, the trend is clear, the performance of the model decreases significantly as more important pixels are removed. The performance drops to nearly a random guess when the top-20% of the important pixels are removed while the number is around 75% for random removal.

It is interesting to see that in the very beginning, the random removal influences the model (to be explained) more than our explanation method. We think it is because of the architecture of the ViT model whose input is based on several $16{\times}16$ patches. In the very beginning the random-picked pixels are distributed more separated, which means it can influence more patches at the same time, while the explanation generated by our method is more localized, i.e. less patches are influenced. When k reaches 0.5%, the performance of our model drops dramatically while the vanilla model keeps stable. This experiment verifies that the attention heatmap of ViT can be regarded as explanation.

For the vanilla model, surprisingly, we observe an opposite trend. When $k{=}20\%$, the random removal experiment has reached random guess performance while performance is still quite high (around 99%) when top 20% of the most important pixels are removed. Does it indicate the explanation generated from the vanilla model is useless? How should we interpret it?

To answer this question we go back to the spatial frequency of the images. We empirically verify that the vanilla model mainly takes the spatial frequency feature to do classification. The spatial frequency feature here is more global, occurs over the whole image. Randomly removing pixels can destroy the spatial frequency distribution since these pixels are more separated around the whole image. On the contrary, the affected pixels are more assembled when top k of the most important

ones are removed, less patches are influenced and the spatial frequency distribution does not change significantly. Fig. 8 (right) shows the spatial frequency distribution of the two cases as well as the original images, which implies our analysis is correct. In addition, we train a traditional CNN VGG16 in the vanilla manner and obtain a GradCAM explanation. We repeat the perturbation test on it and observe a similar trend (see the dotted lines in Fig. 8 left).

Therefore, the explanation generated for the vanilla method can only have the concept level meaning for the global high spatial frequency feature. The traditional perturbation test that removing pixels according to the weight of explanation map cannot be applied here.

7.2 Experiment on celebA-HQ-based Data

Fig.9 shows the spatial frequency distribution for the original images (top) and the pre-processed images (bottom). It has the very similar trend with FFHQ-based images in Fig. 7 in the manuscript.

Table 4 lists the prediction performance on celebA-HQ-based data. Similarly, the vanilla method has slightly better performance and when the prediction performance of the vanilla model decreases to 50% when we use pre-processed images. This further confirms our previous observations in the manuscript.

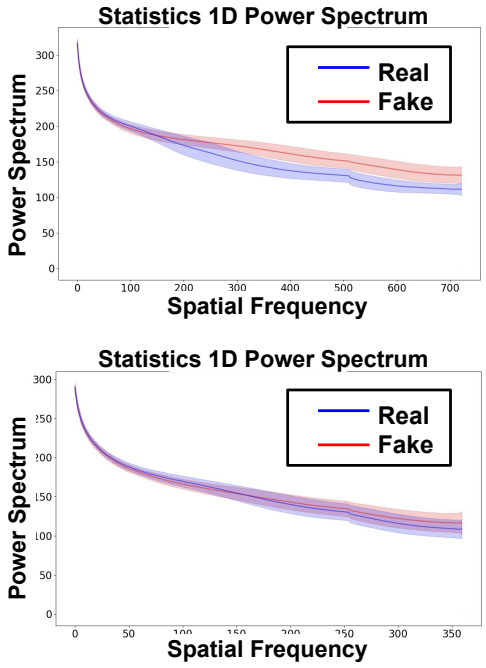

Figure 9: Spatial frequency distributions for real and fake images. Distribution from original images (top) and from images after the pre-processing (down)

Experiment	Accuracy (%)
Vanilla	99.77
Ours	99.23
Vanilla model with filtered images	50.00

Table 4: CelebA-HQ-based Fake vs real face image prediction accuracy of vanilla model and our model.

Explanation	Participants (%) (CS / Non-CS)	Votes (%) (CS / Non-CS)
Vanilla	17.5 / 11.1	27.6 / 34.8
Ours	**82.4 / 81.5**	**72.4 / 65.2**
Explanation		Yes (%) (CS / Non-CS)
Q1: Better at recognizing fake images?		82.5 / 61.5
Q2: Explanation helpful?		93.3 / 85.7

Table 5: Survey statistics.

7.3 Additional Qualitative Results

In this section, we display additional qualitative results for both fake and real images. Fig. 13 and Fig. 14 show additional explanations on fake images. These images are generated by styleGAN trained with FFHQ Karras et al. (2018) and celebA-HQ, respectively. Fig. 15 and Fig. 16 show the FFHQ and celebA-HQ images, respectively.

7.4 More Details on Survey

For each questionnaire, the participants can answer up to 20 questions. We consider a participant prefers one explanation over the other if that type of explanation receives more than 50% of the votes from that participant. We also calculate the average vote, by summing up all the votes for the two explanations and then compute the proportion.

In the survey, we also asked participants about their background. More specifically, we asked if they had studies related to computer science or informatics. There are 57 participants answering Yes and we refer to them as the "CS" group, while 27 participants belong to the "Non-CS" group since they answered No. Table 5 shows the statistics based on this split. It is clear that independent of whether participants had experience or not with informatics they preferred our method. This is motivating since shows the potential of our methods towards layman users.

Fig. 10 shows the histogram of the selection rate of our explanation in each questionnaire. Here the selection rate is defined as the amount of our explanation selected by the participant over all questions in one questionnaire. It indicate that for the participants who prefer our explanation in the questionnaire, more than half of them (56.3%) select our explanation for over 80% of the questions.

Fig. 11 displays two failure cases where the vanilla model receives more votes than ours. We can observe that the face skin looks quite weird on the original images. For the images with low quality, the fake region is more perceptible by humans.

Fig. 12 shows the online interface used for the survey.

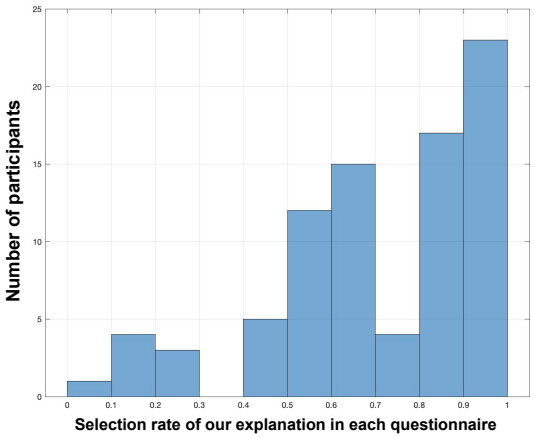

Figure 10: Histogram of the selection rate of our explanation in each questionnaire.

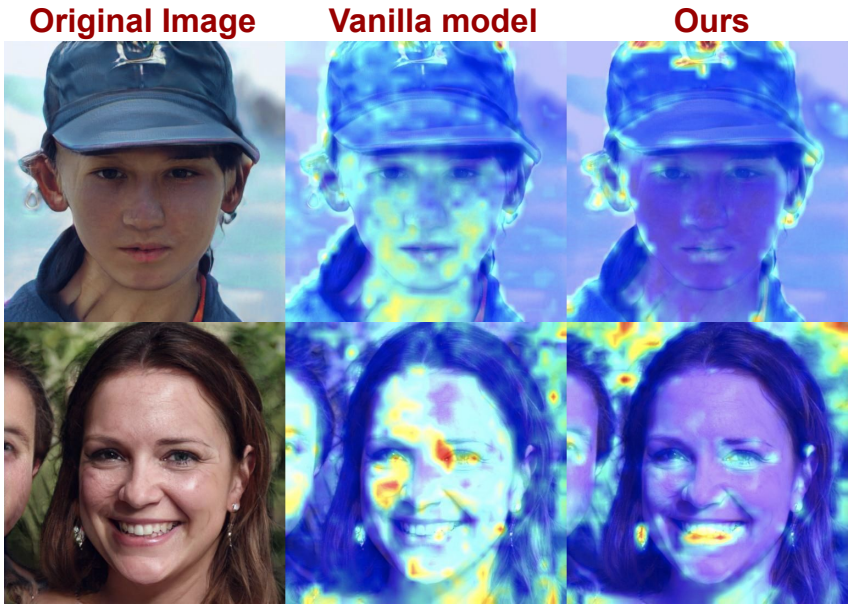

Figure 11: Failure cases where the vanilla model receive more votes.

Figure 12: A (part of) the screen shot of our online survey.

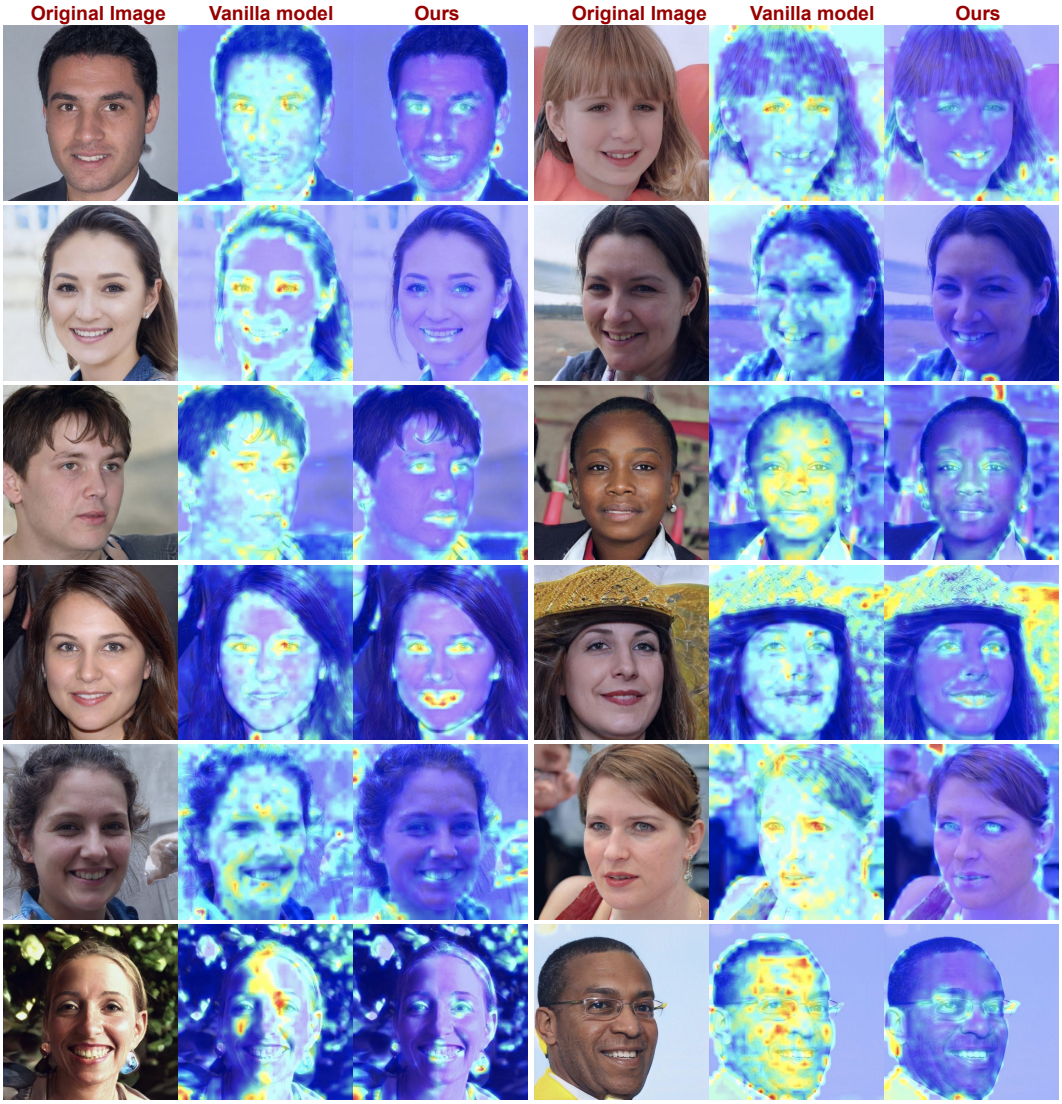

Figure 13: Qualitative comparison for the explanation of *fake* images generated by ViT-16 attention. They are generated by styleGAN trained with **FFHQ** dataset. Do you feel it makes more sense by looking at the right explanation?

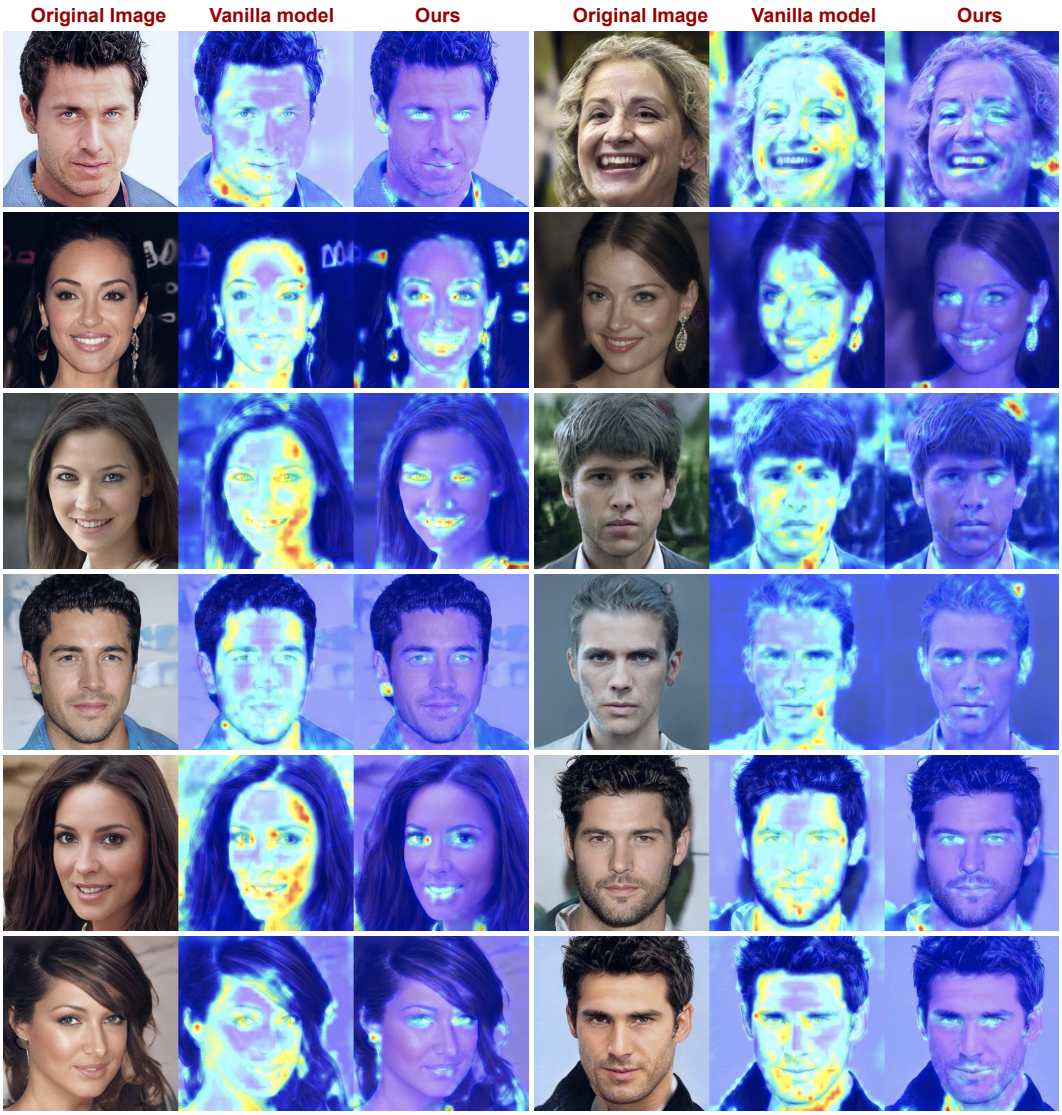

Figure 14: Qualitative comparison for the explanation of *fake* images generated by ViT-16 attention. They are generated by styleGAN trained with **celebA-HQ** dataset. Do you feel it makes more sense by looking at the right explanation?

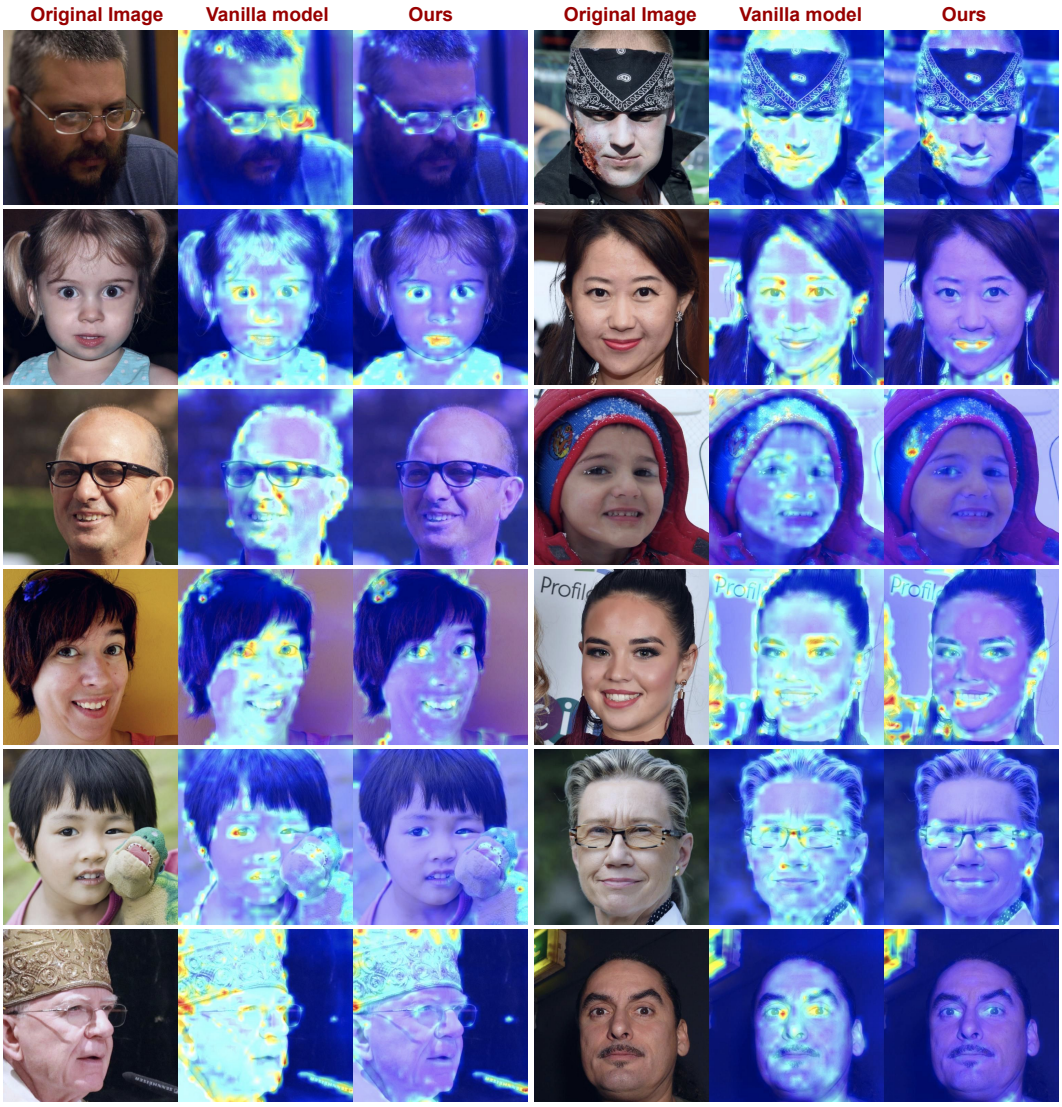

Figure 15: Qualitative comparison for the explanation of *real* images generated by ViT-16 attention. They are from **FFHQ** dataset. Do you feel it makes more sense by looking at the right explanation?

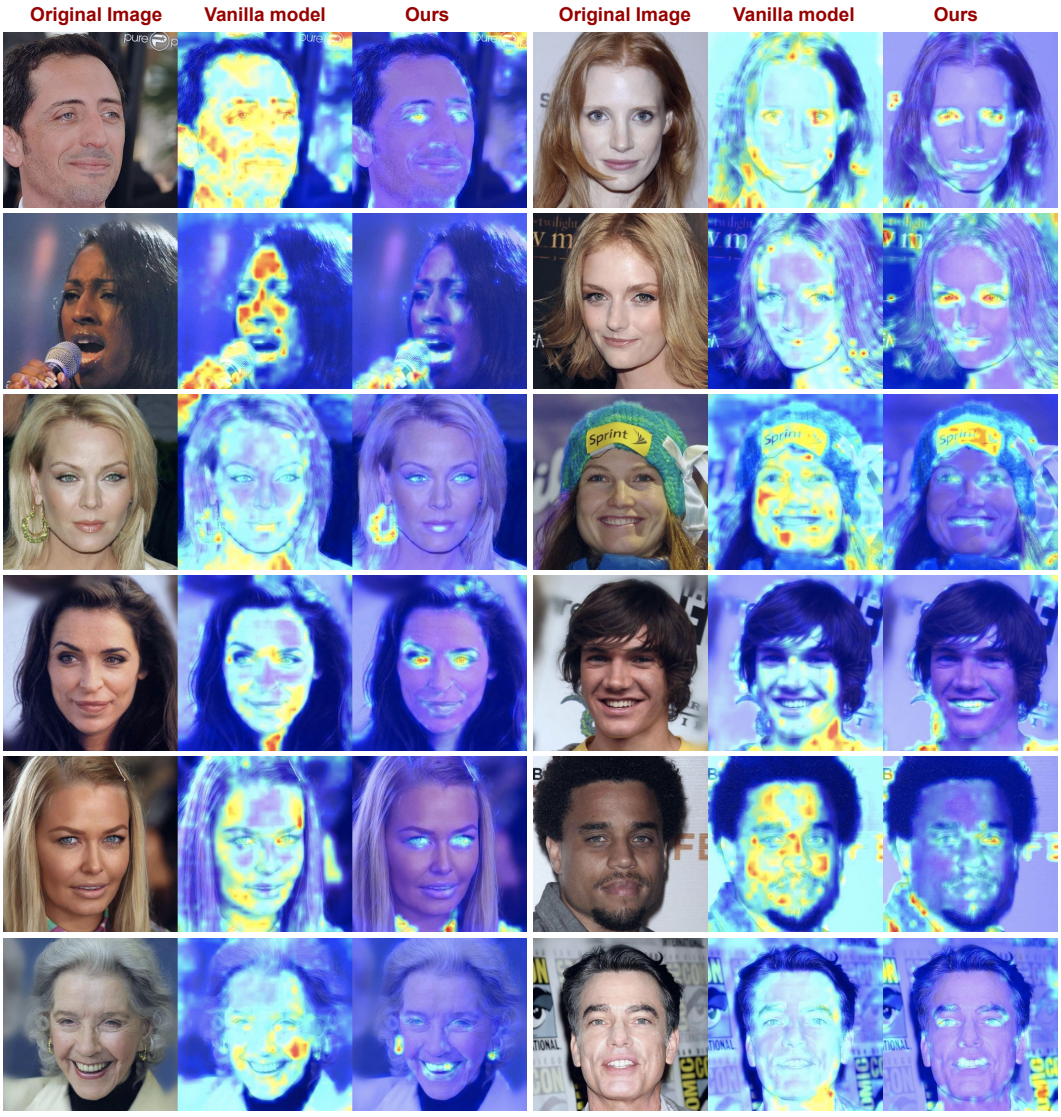

Figure 16: Qualitative comparison for the explanation of *real* images generated by ViT-16 attention. They are from **celebA-HQ** dataset. Do you feel it makes more sense by looking at the right explanation?