Gradient Frequency Modulation for Visually Explaining Video Understanding Models

Xinmiao Lin
xl3439@rit.edu
Wentao Bao
wb6219@rit.edu
Matthew Wright
Matthew.Wright@rit.edu
Yu Kong
Yu.Kong@rit.edu

Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY

Abstract

In many applications, it is essential to understand why a machine learning model makes the decisions it does, but this is inhibited by the black-box nature of state-of-the-art neural networks. Because of this, increasing attention has been paid to explainability in deep learning, including in the area of video understanding. Due to the temporal dimension of video data, the main challenge of explaining a video action recognition model is to produce spatiotemporally consistent visual explanations, which has been ignored in the existing literature. In this paper, we propose Frequency-based Extremal Perturbation (F-EP) to explain a video understanding model's decisions. Because the explanations given by perturbation methods are noisy and non-smooth both spatially and temporally, we propose to modulate the frequencies of gradient maps from the neural network model with a Discrete Cosine Transform (DCT). We show in a range of experiments that F-EP provides more spatiotemporally consistent explanations that more faithfully represent the model's decisions compared to the existing state-of-the-art methods.

1 Introduction

Since the first major success of deep learning in 2012, there has been an explosion of applications of these models, such as in image classification [8], machine translation [5] and tumor detection [30]. As people interact more frequently with these models, and they are applied in critical applications like medical diagnosis [25] and terrorism detection [41], it is becoming increasingly important to understand their decisions. Otherwise, it can be hard for people to trust the models, since blind trust could result in catastrophic consequences. A plethora of evidence shows the importance of explanation towards understanding and building trust in cognitive psychology [21], philosophy [38] and machine learning [10, 20].

Numerous works have emerged, especially in the computer vision field, to explain the decisions of deep neural networks. Researchers have come to a consensus that model explanation should be interpretable and faithful to the model [14, 28, 37]. A common type

of model explanation in the image domain takes the form of a heatmap or *s*aliency map that localizes the salient areas of an input for a model's decision. Among these approaches, perturbation-based methods [11, 12, 42] produce explanations that are more faithful and fine-grained than CAM- and backpropagation-based methods [22, 28, 35, 48], but also produce noise that appears to humans as randomly selected, which hurts their interpretability. See Fig. 6 for a comparison of these explanation methods.

Another challenge arises when the aforementioned methods for the image domain are leveraged to the video domain: the explanations fail to capture the motion dynamics. We argue that *spatiotemporal consistency* should be added as another criterion for model explanation in the video domain. A spatiotemporally consistent explanation should focus on the target object/scene and follow its spatiotemporal trajectory, such as the gymnast in Fig. 6. This greatly helps the interpretability of the explanation. At the same time, the video understanding models may indeed not pay attention to the target object/scene at each frame because of the moving edges, shot changes, and other issues [39, 40]. Spatiotemporally consistent explanations help the end-users or the developers of models to better assess the model's understanding of the dataset and improve upon it. If the explanations are faithful to the model/data, they may not be spatiotemporally consistent; and if they are more spatiotemporally consistent, they may lose faithfulness. Thus, the challenge resides in finding a balance between the faithfulness of an explanation and its interpretability.

Therefore, we propose the Frequency-based Extremal Perturbation method (F-EP) which extends the EP method [11] to the video domain. F-EP, illustrated in Fig. 1, aims to find the salient areas of the input by perturbing a mask, where the perturbations are the gradients of the output label with respect to the input. The masks, comprised of the aggregated perturbations, become the explanations for the video input. At each iteration of the optimization process, F-EP transforms the gradients into the frequency space using a common Fourier Transform: Discrete Cosine Transform (DCT). Then, F-EP modulates the frequency signals and transforms them back to the gradient space to update the masks.

This approach is inspired from works [43, 45] showing that neural networks not only learn the low-frequency components of an image, such as faces and shapes, but also the high-frequency components that appear like noise to humans. Therefore, the gradients of the salient high-frequency components will be high and make the gradient maps noisy. To overcome this problem, F-EP proposes to transform the gradients into the frequency domain and selects a combination of low and high frequency components of the gradients to achieve semantically meaningful explanations that are also faithful to the model. Meanwhile, because the explanations focus on the semantic features of each frame, such as a person performing an action, and the majority of frames contain these features, the explanations will naturally become spatiotemporally consistent.

Since no existing metrics address spatiotemporal consistency of explanations, we propose the *Spatiotemporal Consistency (STC)* metric to measure how well the explanations align with the ground truth bounding boxes. Because bounding boxes indicate the spatial position of the foreground object/scene, spatiotemporal consistent explanations will have high alignment with them. We experimentally show that F-EP achieves superior performance in terms of spatiotemporal consistency and faithfulness than the state-of-the-art methods. In summary, our contributions in this paper are as follows.

- We propose the F-EP method, which can simultaneously reduce the noise in the explanations and make them spatiotemporally consistent by modulating the gradients in the frequency space.

- We propose a new metric, Spatiotemporal Consistency (STC), to accurately assess the spatiotemporal consistency of an explanation in the video domain.
- In the ablation studies section, we show that F-EP using low frequency gradients is able to achieve state-of-the-art performance, while the addition of high-frequency gradients can improve its performance even further.

2 Related Work

Model explainability is increasingly popular in a diverse range of deep learning applications. In this section, we present only the methods for image and video recognition models.

Class Activation Map (CAM)-based. CAM-based methods, such as CAM [48], Grad-CAM [28], Grad-CAM++ [4], and Score-CAM [44], use weighted sums of either activation maps or gradients of the class label with respect the input to produce a heatmap. Although CAM-based methods are efficient [9, 27], usually requiring just a few forward and backward passes, the resulting heatmaps are coarse due to the upsampling process. The Saliency Tubes approach [36] adapts Grad-CAM to the video domain.

Backprop-based. Backprop-based methods [2, 13, 22, 31, 33, 37, 46, 47] compute saliency scores for each input pixel by starting with the final layer and, one layer at a time, inferring the importance of the inputs to the outputs of the layer until the input is reached. For example, Layerwise Relevance Backpropagation (LRP)-based methods [13, 22] assign a relevance score for each neuron and backpropagate the relevance score of each neuron to the input layer. The explanations of backprop-based methods have been criticized for being similar to the results of an edge detector [1] and insensitive to randomization of the model's parameters, which should intuitively alter the results [32].

Perturbation-based. Perturbation-based explanation methods [11, 12, 18, 23, 24, 31, 42] add perturbations to the input features to quantify the importance of each pixel. For example, RISE [23] measures the importance of pixels by randomly masking areas of input features and observing changes to the output. Meaningful Perturbation (MP) [12] and Extremal Perturbation (EP) [11] methods optimize masks as explanations to localize salient areas. STEP [18] extends EP to the video domain by adding spatiotemporal smoothness constraints, and it is the current state-of-the-art for explaining video understanding models. Although perturbation-based methods [11, 18] are able to produce more fine-grained explanations and are more sensitive to changes in the model's weights compared to the CAM-based and backpropagation-based methods [4, 28, 31, 37], their explanations are still noisy and not spatiotemporally consistent (see Fig. 6). Our proposed F-EP approach probes the frequency signals in the gradients in order to denoise the explanations and make them focus more tightly on the target object in every frame.

Fourier Transforms. Fourier transforms decompose a data point in the spatial or temporal domain into a combination of functions in the corresponding frequency domain. Discrete Cosine Transform (DCT) is a type of Fourier Transform using only real numbers. DCT was first proposed for efficient image compression, where only the content-defining features are preserved [43]. DCT is used in machine learning for image compression with SVM [17], image classification [26], and more. Recently, [15, 29] explore the effectiveness of using DCT on crafting adversarial samples. Although F-EP is also using perturbation to produce explanations, the perturbations are added *on the masks* and not the inputs as in [15, 29]. The optimization objectives are also different: adversarial attacks aim to generate samples

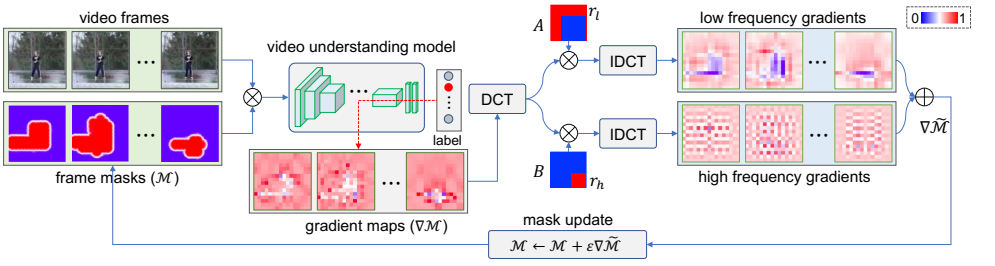


Figure 1: Frequency-based Extremal Perturbation (F-EP). The video frames perturbed with masks M are fed to the video understanding model to compute the gradients ∇M. ∇M is transformed to the frequency domain using DCT. r_l and r_h are the ratios of low- and high- frequency signals to keep in the gradients. The modulated gradients are then transformed back to the video domain using IDCT and combined together to get $(\nabla\tilde{M})$, which is then used to update the masks.

to change the model's prediction, while explanations aim to faithfully represent a model's decision. Also, the adversarial samples are sparse, noisy and hard to interpret, while the explanations given by F-EP focus tightly on the target object/scene spatially and temporally.

3 Methodology

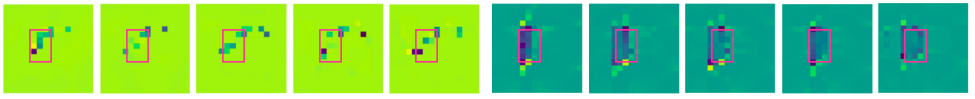
In this paper, F-EP is experimented on video classification models as in EP [11]. A future research direction could be to evaluate F-EP on other types of video understanding models. Denote the video classification model by Φ, the original input video clip as $X \in \mathbb{R}^{T\times C\times H\times W}$ and the predicted label $y = \Phi(X)$, where y is among a set of Y classes, T is the number of frames in a video clip, H and W are the height and width of the frame/mask, and C is the number of channels. We first briefly introduce the baseline method, Extremal Perturbation (EP) [11], and then explain our proposed Frequency-based Extremal Perturbation (F-EP) method, which is shown in Fig. 1.

3.1 Extremal Perturbation

Extremal Perturbation (EP) was proposed to explain image understanding models and aims to localize the salient areas in the input for a model's decision through perturbations on the masks. Let the masks be $M \in \mathbb{R}^{T\times 1\times H\times W}$, the optimization objective of EP is:

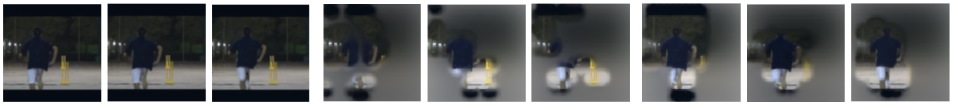
$$M_a^* = \arg\max_M \Phi_y(M \otimes X) - \lambda R_a(M). \tag{1}$$

The left term in Eq. (1) maximizes the classification confidence of the model Φ to the perturbed input $(M \otimes X)$. The operator $\otimes$ is the perturbation operator which applies local Gaussian blurring to each pixel in M. The second term in Eq. (1) is $R_a(M) = \|\text{vecsort}(M) - r_a\|^2$, which regularizes the area of the mask M with respect to the area constant a. λ is the regularizer constant. r_a is a binarized vector with a ones and $(1-a)$ zeros, and vecsort(M) is sorted vector of the masks M. Thus, $R_a(M)$ computes the loss of the area of masks with respect to the area constraint a. To solve the optimization problem (1), we can derive the gradients of the output class label with respect to the masks where ε is the learning rate of



(a) EP [11] Gradient Maps (b) F-EP Gradient Maps

Figure 2: **Illustration of GFM Effect**. We visualize the gradient maps of five consecutive video frames by EP [11] and our proposed F-EP method. Red boxes are foreground objects. They clearly show that the gradients of our method are more spatiotemporal consistent.



(a) Original, CricketBowling, 0.59 (b) EP [11], CricketBowling, 1.0 (c) F-EP, CricketBowling, 0.85

Figure 3: **Interpretability vs Faithfulness**. Under each figure is the method, predicted class and probability. Without GFM, EP [11] produces explanations that are noisy and not spatiotemporal consistent with attention on the background and part of human motion. While F-EP gives more interpretable explanations, the faithfulness can sometimes be compromised, i.e., accuracy of 0.85 vs 1.0.

the optimization (details in [11]):

$$\nabla M = \frac{\partial(\Phi_y(M \otimes X) - \lambda R_a(M))}{\partial M}, \tag{2}$$

The optimal solution M^* thus can be approximated by an iterative gradient ascent process. In each iteration, the masks M are updated as follows:

$$M^{f+1} = M^f + \varepsilon \nabla M, \tag{3}$$

3.2 Frequency-based Extremal Perturbation

Motivation. A model explanation is expected to be interpretable by people and faithfully represent a model's inference process [14]. This becomes a challenge when the neural network not only uses low-frequency features for learning larger shapes and objects like faces, but also high-frequency features that look like noise [45]. Because EP aims to achieve higher model confidence regardless of the the frequencies chosen, thus the high frequency features are included if they are more salient than the low frequency features which make explanations noisy and not spatiotemporally consistent.

Because the gradients encode the importance of each pixel (Eq. (2)), important high-frequency features would have high gradients in ∇M, making ∇M noisy. A noisy ∇M leads to noisy explanations M, since the latter is iteratively updated with the former, as shown in Eq. (3). We propose to transform the gradients into the frequency domain using the Discrete Cosine Transform (DCT), such that the signals related to the content-defining features, which reside in the lower end of the spectrum of the frequencies [43], are preserved. High-frequency signals are not only associated with noise, but also the fine-grained details such as textures/edges. Thus, the low frequency signals allow the explanations to focus on the target object/person and become spatiotemporal consistent, while the high frequency features make the explanations more faithful to the model and fine-grained, see Fig. 3. The Fig.

2 shows that the frequency modulated gradient maps focus more on the foreground object and become spatiotemporal consistent when consecutive frames contain the target object. In the following section, we propose our method Frequency-based Extremal Perturbation (F-EP) and theoretically demonstrate that modulating gradients is equivalent to modulating the masks.

Gradient Frequency Modulation. F-EP is presented in Fig. 1. At each iteration of the optimization, F-EP transforms the gradient maps ∇M into the frequency domain using DCT. DCT is chosen over Discrete Fourier Transform (DFT) because the basis functions used in DFT are complex-valued, while ∇M only contains real numbers, meaning that DFT would compute useless transformations [3]. Because ∇M is three dimensional, the DCT is also three dimensional. Denote the DCT as $\mathcal{H}$ and the frequency map of the gradient ∇M as $G \in \mathbb{R}^{H\times W\times T}$, the (i,j,k)-th spatio-temporal entry of G ($i \in \{1,\dots,H\}$, $j \in \{1,\dots,W\}$, and $k \in \{1,\dots,T\}$) is computed by the three-dimensional DCT:

$$\begin{aligned} G_{i,j,k} = \mathcal{H}(\nabla M)_{i,j,k} &= a \sum_{z=0}^{T-1} \sum_{y=0}^{W-1} \sum_{x=0}^{H-1} c_x c_y c_z \nabla M_{x,y,z} d_{x,i}^{(H)} d_{y,j}^{(W)} d_{z,k}^{(T)} \\ &= a h_k^T \left(h_j^T \left(h_i^T \nabla M\right)\right) \end{aligned} \tag{4}$$

where $a = \left(\frac{2}{H}\right)^{\frac{1}{2}} \left(\frac{2}{W}\right)^{\frac{1}{2}} \left(\frac{2}{T}\right)^{\frac{1}{2}}$ is a constant and $d_{x,i}^{(H)} = cos\left[(2x+1)i\pi/(2H)\right]$ is the cosine basis function, $c_x = 1/\sqrt{2}$ if $x = 0$, otherwise $c_x = 1$. $h_i \in \mathbb{R}^{H\times 1}$ is the vector form of the element-wise product between the vector c_x and $d_{x,i}^{(H)}$ along x-axis, i.e., $h_i = c_x \odot d_{x,i}^{(H)}$. The definitions of $\{c_y, c_z\}$, $\{d_{y,j}^{(W)}, d_{z,k}^{(T)}\}$, and $\{h_j, h_k\}$ are similar to c_x, $d_{x,i}^{(H)}$, and h_i. Thus, Eq. (4) shows that the DCT operation is an intrinsically linear system, and according to the gradient ascent rule in Eq. (3), we have the following equality:

$$\mathcal{H}(M^{f+1}) = \mathcal{H}(M^f) + \varepsilon \mathcal{H}(\nabla M). \tag{5}$$

This equation shows that applying frequency modulation on the gradient map ∇M is equivalent to frequency modulation on the optimal mask M^*. We provide a more detailed derivation in the supplementary material. Therefore, it is straightforward to linearly modulate the gradient frequency map G. To this end, we perform the frequency modulation on G as follows:

$$\nabla \tilde{M} = \tilde{\mathcal{H}}(A \odot G) + \tilde{\mathcal{H}}(B \odot G), \tag{6}$$

where $\tilde{\mathcal{H}}$ is the inverse DCT (IDCT) function and $\odot$ is the element-wise product. Note both $\tilde{\mathcal{H}}$ are linear systems. The low frequency mask $A \in \mathbb{R}^{H\times W\times T}$ is defined as $A_{i,j,k} = 1$ if $i \le r_l * H, j \le r_l * W, k \le r_l * T$, otherwise zero, and the high frequency mask $B \in \mathbb{R}^{H\times W\times T}$ is defined as $B_{i,j,k} = 0$ if $i \le (1-r_h) * H, j \le (1-r_h) * W, k \le (1-r_h) * T$, otherwise one. See the A and B matrices in Fig. 1. r_l and r_h denote the ratios of low- and high-frequency components to be preserved, respectively. Note that $r_l, r_h \ge 0$ and $r_l + r_h \le 1$.

Eq. (6) first selects a ratio r_l of low-end frequencies and a ratio r_h of high-end frequencies from G, then performs IDCT on the selected high and low frequencies separately. By summing up the gradients maps modulated at different frequencies, we ensure that the gradient information contains both the low-end and high-end features.

Finally, instead of using the raw gradient ∇M in equation (3), we propose to use the frequency-modulated gradient $\nabla \tilde{M}$ such that:

$$M^{f+1} = M^f + \varepsilon \nabla \tilde{M}. \tag{7}$$

Since DCT and IDCT are computationally efficient, our proposed modulation does not incur too much computational cost.

4 Experiment

4.1 Experimental Settings

Evaluation Metrics. Evaluation has always been one of the most challenging parts of studying model explainability. Due to the lack of ground truth explanations, existing methods can only be evaluated in a post hoc manner. We follow prior work in using two standard metrics that also apply to static images. First, **Drop in Confidence** (**DC**) measures the decrease in model's confidence on the explanations compared to the original input [4] (smaller DC is better). It is represented as $\sum_i^N \max(0, y - y_e)/N$ where $y_e = \Phi(X_e)$ and $X_e = M \odot X$ as in Eq. (1). Second, **Accuracy** (**Acc.**) measures the model's classification accuracy on the explanations X_e.

In addition, since no existing evaluation metric is suitable to assess the spatiotemporal consistency of an explanation, we propose a new metric called **Spatiotemporal Consistency** (**STC**). STC calculates the overlap between the explanations and the ground truth bounding boxes. Mathematically, we define $STC = \sum_{i,j,k}^{H,W,T} \mathbb{1}_{\{O_{i,j,k}=1, M_{i,j,k} \geq \tau\}}$, where $O \in \mathbb{R}^{T \times 1 \times H \times W}$ are the ground truth bounding boxes where $O_{i,j,k} = 1$ for all pixels inside the bounding boxes, and 0 otherwise. τ denotes the threshold of masks M during evaluation because M are continuous values within range $[0,1]$. Because the bounding boxes focus tightly on the foreground object/person in each frame, a more spatiotemporally consistent explanation will have higher overlap with them.

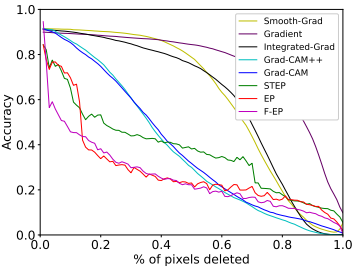
Experimental Details We evaluate F-EP against CAM-based methods: Grad-CAM [28] and Grad-CAM++ [4], and backpropagation-based methods: Gradients [31], Integrated Grad [37] and Smooth Grad [33]. We extend the original implementation of these methods[1] to the video domain. EP [11] was extended to the video domain in STEP [18] [2]. The target models to be explained are R(2+1)D [40] and TSM [19].[3] For video datasets, we use UCF101-24 [34] and Epic-Kitchens-Object [7]. UCF101-24 is a video understanding dataset of 24 actions with annotated bounding boxes and Epic-Kitchens-Object records cooking activities from a first-person point of view. The hyperparameters used are the same as in STEP [18], except the step size is 13 and sigma is 23. The parameters r_l and r_h are selected based on the best performance on the validation set.

4.2 Quantitative Results

UCF101-24 Results. Table 1 presents F-EP against SOTA methods on the UCF101-24 [34] dataset based on two target models, R(2+1)D [40] and TSM [19]. The outperformance of F-EP ($r_l = 0.5$, $r_h = 0.2$) on the R(2+1)D model shows that it is able to produce more faithful and spatiotemporal consistent explanations than all the other methods, including the baselines EP [11] and STEP [18]. On TSM, F-EP ($r_l = 0.5$, $r_h = 0.1$) achieves best on the STC metric, and second best on the DC and Acc. metrics. The first reason is that

[1] https://github.com/jacobgil/pytorch-grad-cam
[2] https://github.com/shinkyo0513/Video-Visual-Explanations
[3] https://github.com/open-mmlab/mmaction2

Figure 4: DAUC of all the compared methods, which plots the model's confidence drop with respect to the percentage of pixels deleted. Note that the most salient pixels are deleted first. **Lower Area Under the Curve (AUC) is better.** F-EP has the lowest AUC compared to other methods, which indicates that the features contained in the explanations are more salient.



Method	R(2+1)D			TSM		
	DC (↓)	Acc. (↑)	STC (↑)	DC (↓)	Acc. (↑)	STC (↑)
Gradients [31]	89.4	6.0	0.4	90.8	2.1	0.6
Integrated Grad [37]	89.5	6.9	0.5	89.5	4.0	1.2
Smooth Grad [33]	90.7	1.0	1.8	90.7	2.7	1.2
Grad-CAM [28]	42.7	58.1	28.9	88.6	11.2	0.1
Grad-CAM++ [4]	42.3	58.3	35.1	24.1	74.8	17.4
EP [11]	37.9	61.7	63.8	40.0	63.0	70.6
STEP [18]	34.2	67.1	63.8	39.9	65.2	73.9
F-EP (ours)	32.9	73.4	67.0	33.3	71.0	74.6

Table 1: Results (%) on UCF101-24 [34] with R(2+1)D [40] and TSM [19] models. The best and second best performing methods are shown in red and blue. Our method F-EP shows consistent improvement over the state-of-the-art methods on different models.

the UCF101-24 dataset has high scene representation bias [6], and the model needs not to look at the actual activity to give a correct prediction. Because the explanations given by F-EP focus more on the person performing the activity, which are at the lower end of the frequency domain, the model will have lower confidence on these explanations. Second, R(2+1)D uses 3D convolution filters to extract spatiotemporal features, while TSM uses 2D convolution filters that capture less temporal information. Thus, a more spatiotemporally consistent explanation will have higher model confidence in R(2+1)D than TSM.

Fig. 4 illustrates the Deletion metric [23], which measures the drop in model's confidence when the most important pixels are removed. The importance of pixels are given by the saliency maps. A lower Area Under the Curve (AUC) implies the explanation method is better and the saliency features contained are more faithful to the model. We see that F-EP has the lowest AUC compared to other explanation methods.

Epic-Kitchens-Objects Results. Table 2 reports the results of R(2+1)D on the Epic-Kitchens-Object [7] dataset. F-EP ($r_l = 0.7$, $r_h = 0$) performs slightly less well on DC and Acc than

Table 2: Results (%) on Epic-Kitchens [7] with R(2+1)D [40]. The best and second performance are shown in red and blue. The explanations given by F-EP are more spatiotemporal consistent than SOTA method while being comparable in the metrics of faithfulness (DC and Acc.).

Method	DC (↓)	Acc. (↑)	STC (↑)
Gradients [31]	48.1	16.0	0.4
Integrated Grad [37]	49.0	5.5	0.4
Smooth Grad [33]	49.3	6.1	0.8
Grad-CAM [28]	28.3	48.0	27.2
Grad-CAM++ [4]	54.8	42.1	30.6
EP [11]	34.2	43.4	58.0
STEP [18]	32.6	43.8	61.0
F-EP (ours)	32.4	44.4	67.8

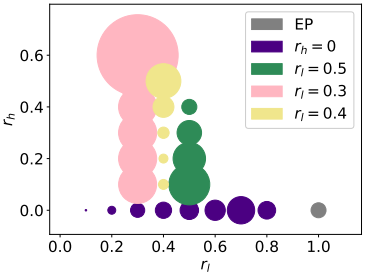


Figure 5: Performance comparison of different r_l and r_h. The vertical series of circles show the performance of $r_l = \{0.3, 0.4, 0.5\}$ with increasing r_h. Note that $r_l + r_h \leq 1$. The horizontal series of circles show when $r_h = 0$ with increasing r_l. EP [11] is the lower right circle.

Grad-CAM, but much better in STC (Qualitative results are shown in the supplemental material). The explanations given by F-EP have lower model confidence because videos in the Epic-Kitchens-Object dataset do not contain the target object in every frame but F-EP produces explanation for each frame. Therefore, non-target objects are included in the explanations which have lower model confidence. A future research direction is to have the optimization algorithm attend only to salient frames for the output decision.

Tables 1 and 2 show that backpropagation-based methods, Gradients [31], Integrated Grad [37] and SmoothGrad [33], perform poorly compared to CAM- and perturbation-based methods. For example, Fig. 6's second row (more results in the supplement) shows that Integrated Grad's explanations consist of sparse pixels that align poorly with the ground truth. These explanations are often perceived by the model as adversarial samples and have lower model confidence or even incorrect predicted labels [16]. SmoothGrad and Gradients produce similar explanations and hence face the same problems.

4.3 Ablation Studies

Without High Frequency. In Fig. 5, when $r_h = 0$ and as r_l increases, the STC performance increases because the higher frequency signals encode the details such as edges/textures to increase faithfulness of explanations. There is a sharp drop in performance after $r_l = 0.7$, because as ∇M contains more and more higher frequency signals, it is overwhelmed with the excess amount of details. The explanations also become noisier and less spatiotemporal consistent, because higher frequency features are preferred over the low frequency features if they could increase more model confidence.

Combination of High and Low Frequencies. We conduct experiments with varying r_h and deterministic $r_l = \{0.4, 0.5, 0.6\}$. When $r_l = 0.3$, as r_h increases, STC increases because the amount of frequency signals on the lower end frequency spectrum is not enough and higher frequency signals add valuable information to the explanations. However, when $r_l = 0.4$, STC performance increases less quickly than $r_l = 0.3$. This could mean that $r_l = 0.4$ is a threshold at which the utility of high frequency signals decreases. Finally, when $r_l = 0.5$, STC actually decreases when r_h increases. Also, F-EP is able to outperform the baseline method EP at multiple combination of frequencies which confirms our hypothesis that not all frequency signals are salient. One future work direction is to search the appropriate ratios r_l and r_h for each sample because samples vary across dataset and actions.

4.4 Qualitative Results

In Fig. 6, we see that the explanations produced by F-EP are much easier to interpret than those of the prior works. Integrated Grad explanations are noisy, sparse and difficult to

Figure 6: **Visual comparison of explanation methods on the UCF101-24 dataset [34].** The input (first row) contains consecutive frames from the activity FloorGymnastics. On the right of each method, the first word is the predicted label on the explanation where FG = FloorGymnastics, BB = BalanceBeam, WB = WritingOnBoard, and the number denotes the predicted probability.

interpret, because it directly uses gradients, which contain a substantial amount of noise. Grad-CAM produces explanations that fail to capture the activity dynamics. EP and STEP has difficulty in correctly capturing the person's movement in the consecutive frames. In comparison, F-EP is able to produce explanations that focus on the person performing the activity on each frame (more qualitative results are in the supplement).

5 Conclusion

In this paper, we propose an explanation method for video understanding models called Frequency-based Extremal Perturbation (F-EP) which aims to produce spatiotemporal consistent explanations that are faithful to the model and interpretable by humans. F-EP transforms the gradients to the frequency domain and modulates the frequency signals in order to preserve those gradient components that help to explain a model's decision. We experiment F-EP on various datasets and models and show that F-EP is able to outperform existing state-of-the-art works.

Acknowledgments. This work is supported in part by the John S. and James L. Knight Foundation, the Miami Foundation through the Ethics and Governance of the Artificial Intelligence Initiative, the National Science Foundation under award No. 2040209, and the Army Research Office under grant number W911NF-21-1-0236.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018.

[2] Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff. Excitation backprop for rnns. In *CVPR*, 2018.

[3] J.F. Blinn. What's that deal with the dct? *IEEE Computer Graphics and Applications*, 13, 1993.

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, 2018.

[5] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://www.aclweb.org/anthology/W14-4012.

[6] Jinwoo Choi, Chen Gao, C. E. Joseph Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *NeurIPS*, 2019.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] Saurabh Desai and Harish G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *WACV*, 2020.

[10] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud.*, 58, 2003.

[11] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, 2019.

[12] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017.

[13] Jindong Gu, Yinchong Yang, and Volker Tresp. Understanding individual decisions of cnns via contrastive backpropagation. In *ACCV*, 2019.

[14] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40, 2019. URL https://ojs.aaai.org/index.php/aimagazine/article/view/2850.

[15] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In *Proceedings of The Uncertainty in Artificial Intelligence Conference*, 2020.

[16] Christian Szegedy Ian Goodfellow, Jon Shlens. Explaining and harnessing adversarial examples. In *ICLR*, 2014.

[17] Azam Karami, Soosan Beheshti, and Mehran Yazdi. Hyperspectral image compression using 3d discrete cosine transform and support vector machine learning. In *ISSPA*, 2012.

[18] Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. Towards visually explaining video understanding networks with perturbation. In *WACV*, 2021.

[19] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019.

[20] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16, 2018.

[21] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10, 2006.

[22] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65, 2017.

[23] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[25] Jonathan G. Richens, CiarÃ¡n M. Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11, 2020.

[26] Aniket Roy, Rajat Subhra Chakraborty, Udaya Sameer, and Ruchira Naskar. Camera source identification using discrete cosine transform residue features and ensemble classifier. In *CVPRW*, 2017.

[27] Xiaohu Dong Yulan Guo Yinghui Gao Biao Li Ruigang Fu, Qingyong Hu. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. In *BMVC*, 2020.

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[29] Yash Sharma, Gavin Weiguang Ding, and Marcus A. Brubaker. On the effectiveness of low frequency perturbations. In *IJCAI*, 2019.

[30] Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*, 9, 2019.

[31] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

[32] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why modified BP attribution fails. In *ICML*, 2019.

[33] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. In *ICML Workshop*, 2017.

[34] Khurram Soomro, Amir Roshan Zamir, Mubarak Shah, Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[35] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR Workshop*, 2015.

[36] A. Stergiou, G. Kapidis, G. Kalliatakis, C. Chrysoulas, R. Veltkamp, and R. Poppe. Saliency tubes: Visual explanations for spatio-temporal convolutions. In *ICIP*, 2019.

[37] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

[38] Lombrozo Tania. The instrumental value of explanations. *Philosophy Compass*, 6, 2011.

[39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2014.

[40] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

[41] H. M. Verhelst, A. W. Stannat, and G. Mecacci. Machine learning against terrorism: How big data collection and analysis influences the privacy-security dilemma. *Science and Engineering Ethics*, 26, 2020.

[42] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *CVPR*, 2019.

[43] G.K. Wallace. The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38, 1992.

[44] Haofan Wang, Mengnan Du, Fan Yang, and Zijian Zhang. Score-cam: Improved visual explanations via score-weighted class activation mapping. In *CVPRW*, 2020.

[45] Haohan Wang, Xindi Wu, Pengcheng Yin, and Eric P. Xing. High frequency component helps explain the generalization of convolutional neural networks. In *CVPR*, 2020.

[46] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[47] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *IJCV*, 126, 2017.

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.