Quantitative Metrics for Evaluating Explanations of Video DeepFake Detectors
Federico Baldassarre¹, Quentin Debard², Gonzalo Fiz Pontiveros², Tri Kurniawan Wijaya²

Introduction
- DeepFake generation and detection are booming.
- However, explainability is often left behind.
- Existing explanation metrics measure faithfulness and correctness with respect to the model but ignore the user perspective, which is left to subjective qualitative evaluation.
- We introduce quantitative metrics for evaluating explanations from the human perspective, both visual quality and informativeness.
- Using these metrics, we compare existing approaches to improve explanation heatmaps and discuss their effectiveness.

Proposed metrics
Visual quality
How interpretable is the explanation heatmap to humans?

Smoothness
High-frequency cues, e.g. texture imperfections, are harder to perceive. An explanation $h$ with low Total Variation appears smoother (low-freq).

$$TV(h) = \int \| \nabla h \|_1 d\lambda$$

Spatial locality
Explanations that focus on many spatially-distant details are ambiguous. We express the locality of an explanation through its spatial covariance.

$$\sigma = |\det(\Sigma)| = |\det\left(\mathbb{E}[p_p^T] - \mathbb{E}[p][\mathbb{E}[p]]^T\right)||$$

Sparsity
A few highly important regions are more informative than many mildly important ones. The Gini Index is used to measure such sparsity:

$$\text{Gini}(h) = \frac{2}{THW} \sum_{i,j} h(p_i) h(p_j) - \frac{THW + 1}{THW}$$

Manipulation detection
Does the explanation focus on the forged parts?
Explanations should overlap with manipulated areas. For better control, we recombine (real, fake) pairs and limit the forgery to a specific area. Then we can evaluate weakly-supervised manipulation detection.

Evaluation
Overview of existing techniques
Let's evaluate them using our metrics

Observations
- Low-pass filtering the input videos does not improve explanation smoothness, contrary to what observed for images in previous work.
- Regularizing activations yields smoother (low TV) and sparser (high Gini Index) explanations, but hinders classification accuracy.
- Cutout augmentation results in better generalization from DFDC to DFD and better manipulation detection. Little effect on other metrics.
- Compared to the CNN baseline, the MViT transformer produces smoother and sparser explanations that also perform well for manipulation detection. Little effect on spatial locality.

User presentation
How to present a heatmap to users? Rare, medium, or well-done?
In a small study, users preferred the most structured visualizations (blob detection, semantic aggregation).