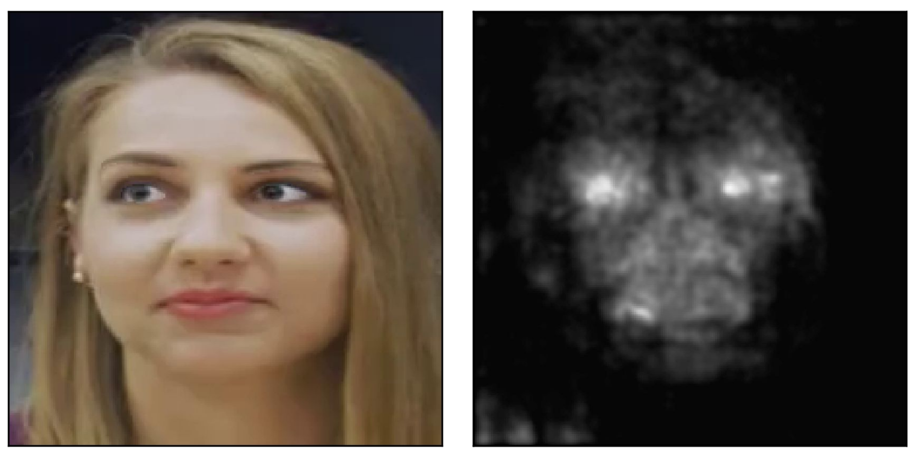
Quantitative Metrics for Evaluating Explanations of Video DeepFake Detectors

Federico Baldassarre[1], Quentin Debard[2], Gonzalo Fiz Pontiveros[2], Tri Kurniawan Wijaya[2]

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).

$$\mathrm{TV}(h) = \int_{\mathcal{G}} \|\nabla h\|_1 \, d\lambda$$

Spatial locality

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

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

Sparsity

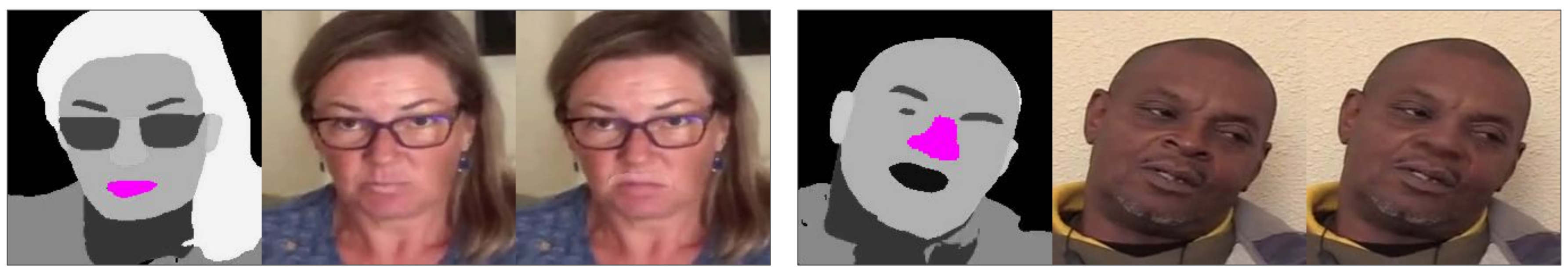
A few highly important regions are more informative than many mildly important ones. The Gini Index is used to measure such sparsity:

$$\mathrm{Gini}(h) = \frac{2}{THW}\frac{\sum_i i \cdot h(\rho_i)}{\sum_i h(\rho_i)} - \frac{THW+1}{THW}$$

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.



Two examples of part-specific manipulation:semantic parsing, real video, fake video

Evaluation

Overview of existing techniques

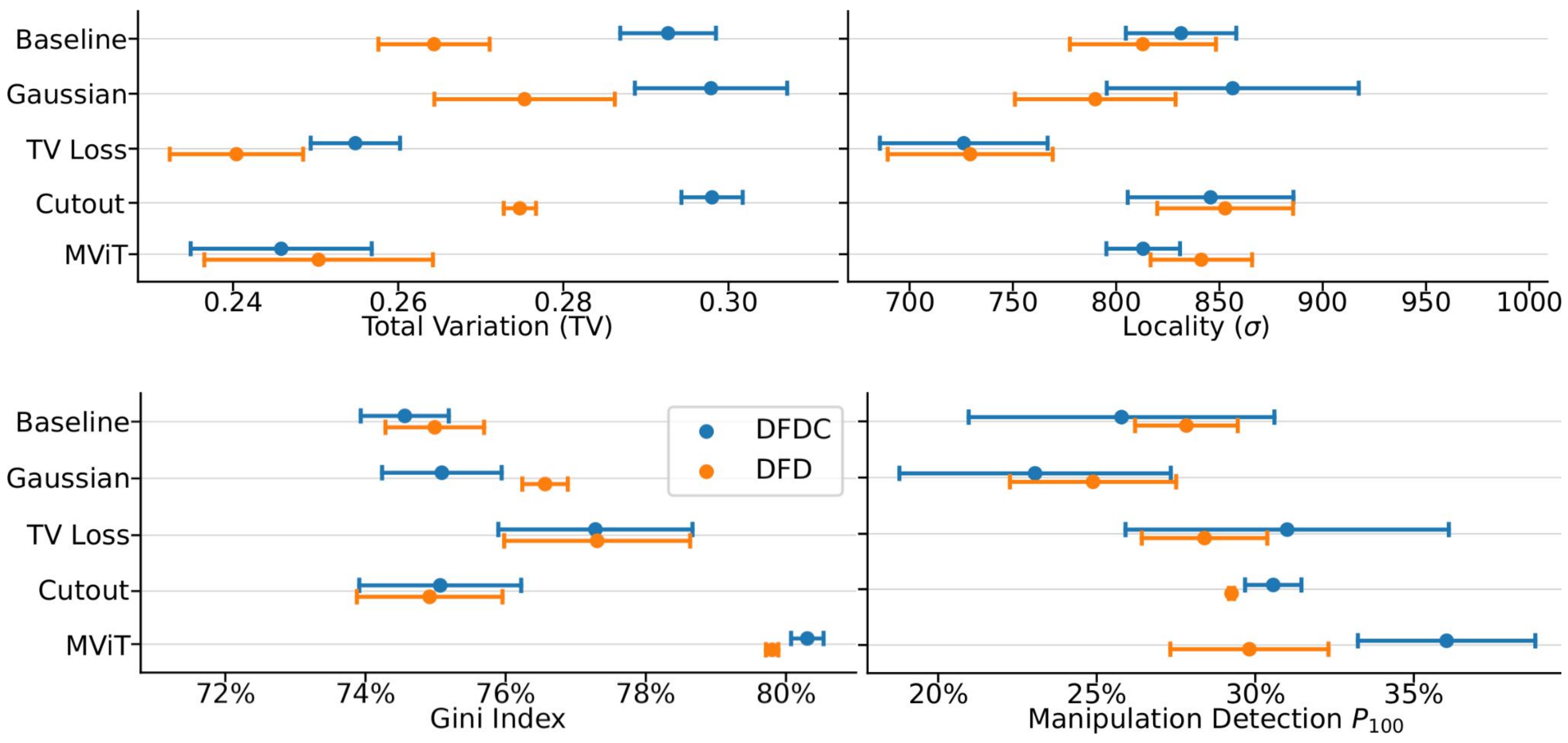
Input preprocessing
Gaussian filtering to remove high-frequency artifacts.

Activation regularization
Total Variation loss to induce smooth neuron activations.

Data augmentation
Cutout augmentation to capture more diverse manipulation cues.

Architecture design
Different inductive biases in CNN and transformers.

Let's evaluate them using our metrics



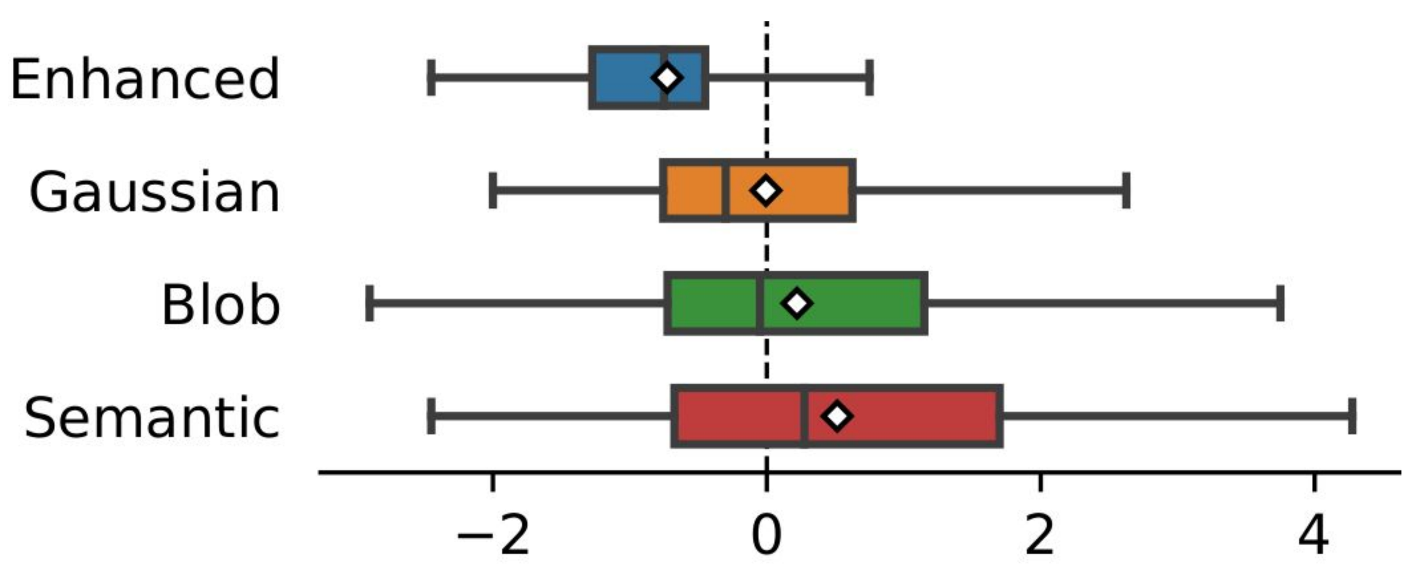
Observations

- **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).



The post-processing techniques in our user study. From top to bottom: a simple blur filter to smoothen the heatmap, a single gaussian approximation, the largest blobs of relevance, an aggregation based on semantic face parsing,

[1] KTH - Royal institute of Technology, Stockholm, Sweden
[2] Huawei Ireland Research Center, Dublin, Ireland