Information Technologies Institute

Combatting video-borne disinformation & Increasing trust in AI methods for combatting disinformation

Vasileios Mezaris

Research Director, Head of the Intelligent Digital Transformation Laboratory

CERTH-ITI, Thermi-Thessaloniki, Greece

Part1: Combatting video-borne disinformation

- Video is a powerful & persuasive medium…
- …also a “useful” medium for spreading misinformation / disinformation

Various possible routes for detecting video-borne misinformation / disinformation:

- Detecting manipulations in the visual modality (incl. deep fakes, synthetic videos)
- Detecting manipulations in the audio modality
- Detecting misalignments between modalities (incl. lip syncing, video-text misalignment)
- Finding the original version of a manipulated video
- Finding the original source of a non-manipulated video (just used out-of-context)

Part1: Combatting video-borne disinformation

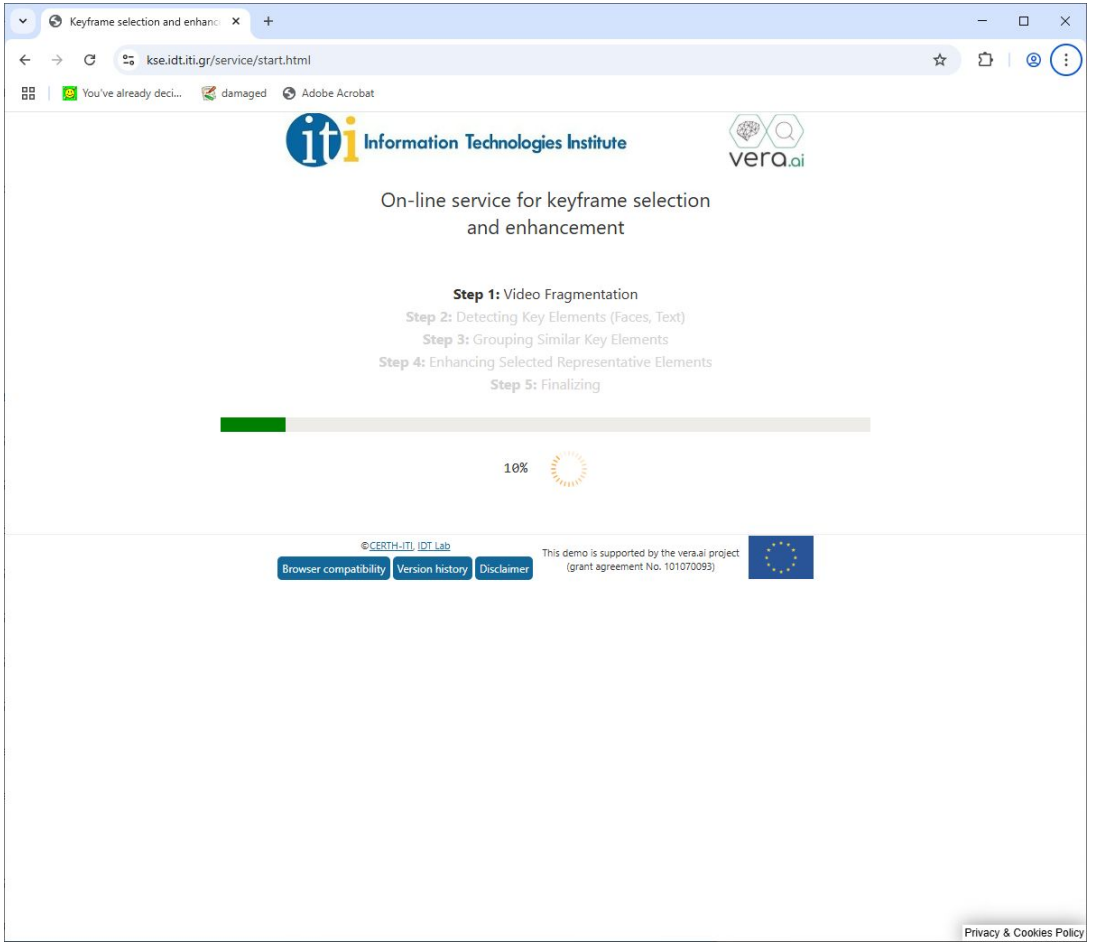
We will discuss tools and related methods for finding the original posting / version of a video:

- The KSE (Keyframe Selection and Enhancement) service of vera.ai (also integrated in the Verification Plugin*)
- The RVS (Reverse Video Search) functionality of AI4TRUST
- The RL-DiVTS method for keyframe selection

* D. Teyssou, J.-M. Leung, E. Apostolidis, K. Apostolidis, S. Papadopoulos, M. Zampoglou, O. Papadopoulou, V. Mezaris, "The InVID Plug-in: Web Video Verification on the Browser", Proc. Int. Workshop on Multimedia Verification (MuVer 2017) at ACM Multimedia 2017, Mountain View, CA, USA, October 2017.
https://chromewebstore.google.com/detail/fake-news-debunker-by-inv/mhccpoafgdgbhnjfhkcmgknndkeenfhe

Reverse Video Search: the KSE Service, https://kse.idt.iti.gr/

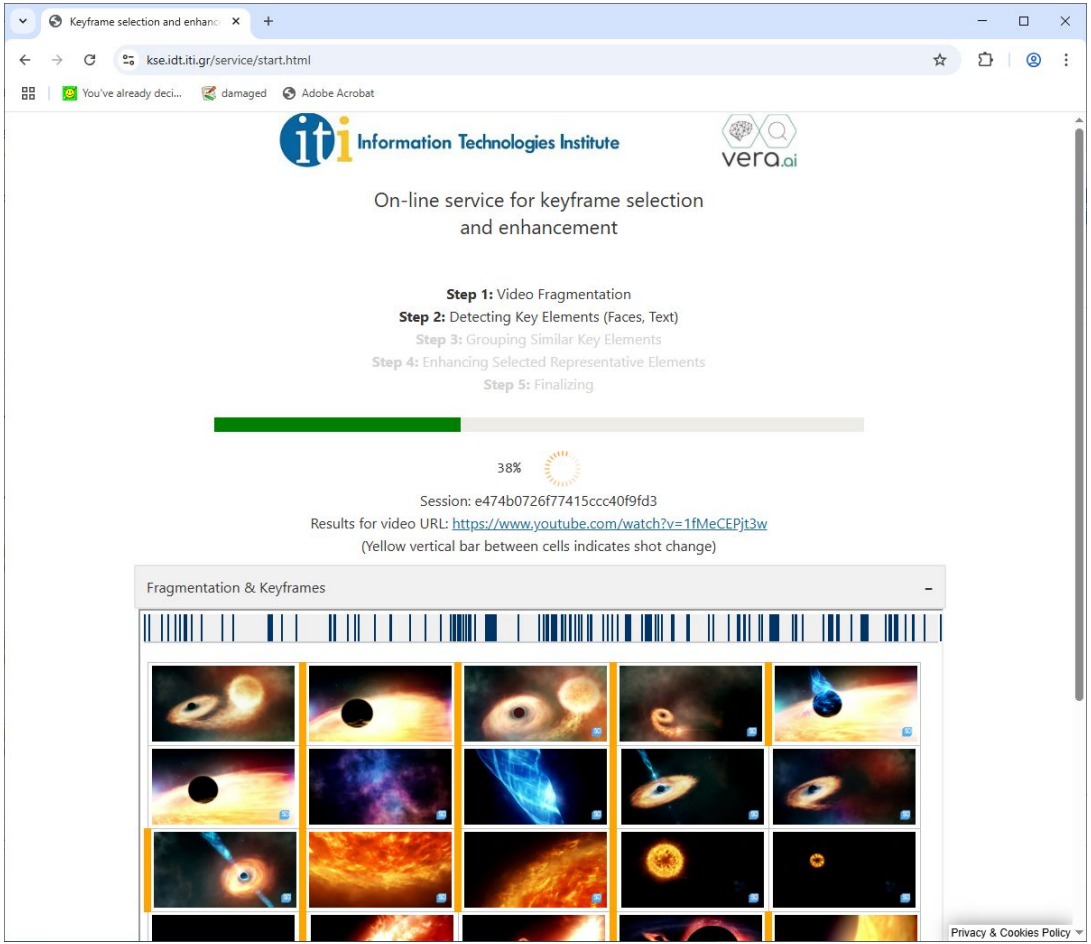
- Video temporal fragmentation; extract keyframes for reverse image search (many frames; user to select)
- Detect key elements (faces, text)
- Group similar key elements, and select the least blurry one from each group
- Employ super-resolution techniques to enhance the selected key elements



Video source: https://www.youtube.com/watch?v=1fMeCEPjt3w

Reverse Video Search: the KSE Service, https://kse.idt.iti.gr/

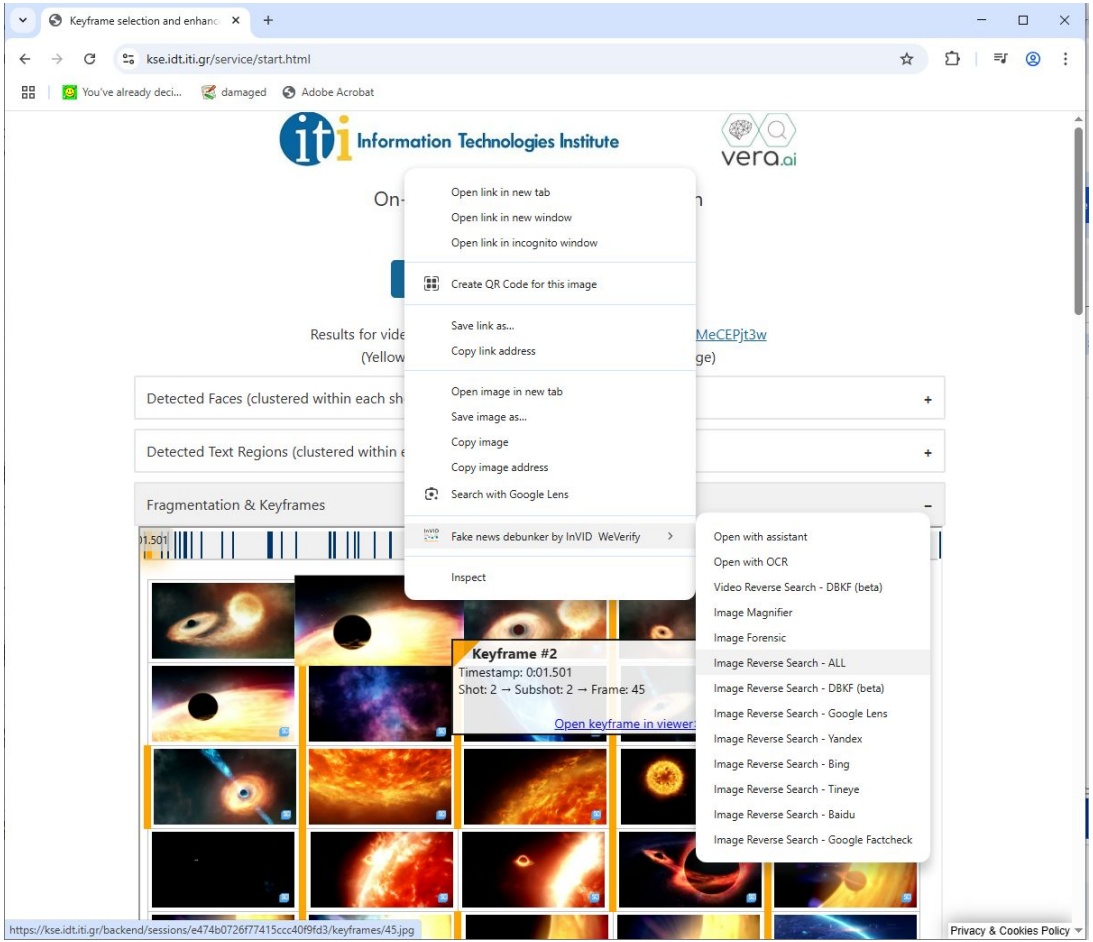
- Video temporal fragmentation; extract keyframes for reverse image search (many frames; user to select)
- Detect key elements (faces, text)
- Group similar key elements, and select the least blurry one from each group
- Employ super-resolution techniques to enhance the selected key elements



Video source: https://www.youtube.com/watch?v=1fMeCEPjt3w

Reverse Video Search: the KSE Service, https://kse.idt.iti.gr/

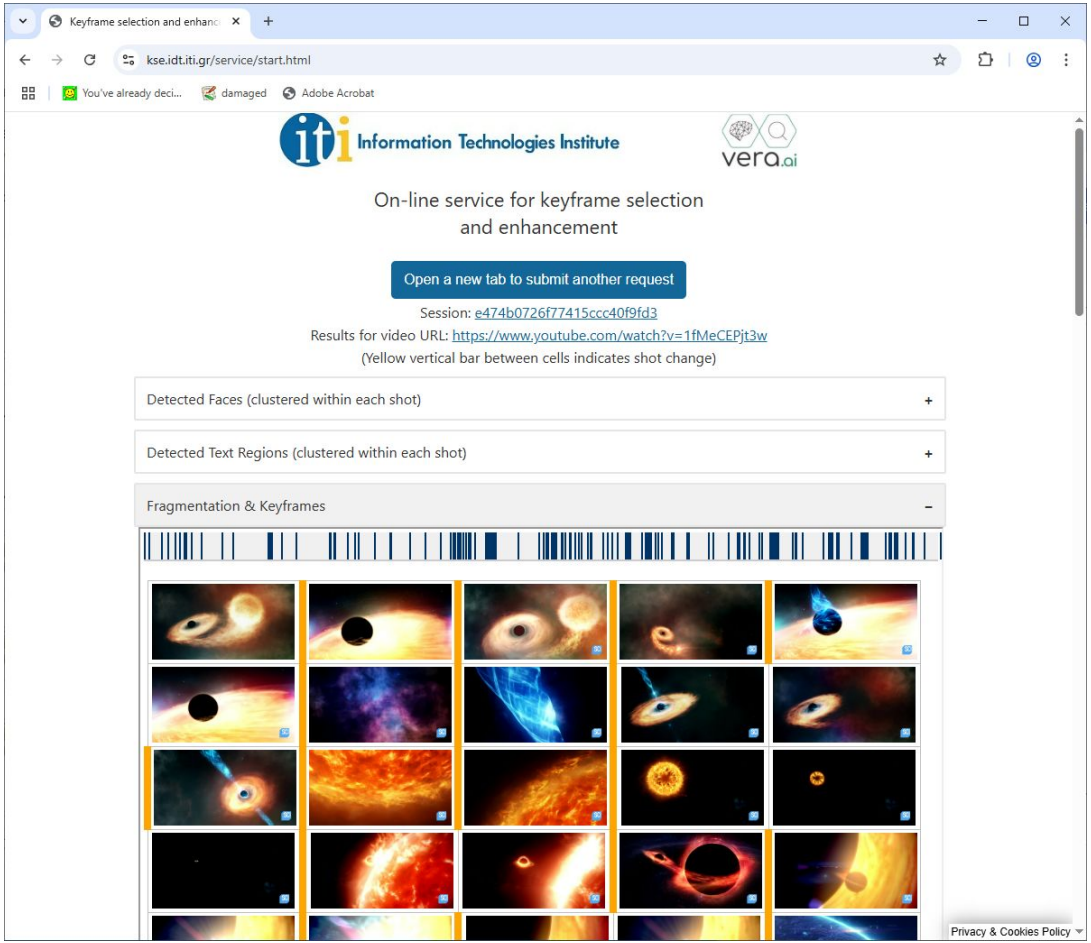
- Video temporal fragmentation; extract keyframes for reverse image search (many frames; user to select)
- Detect key elements (faces, text)
- Group similar key elements, and select the least blurry one from each group
- Employ super-resolution techniques to enhance the selected key elements



Video source: https://www.youtube.com/watch?v=1fMeCEPjt3w

Reverse Video Search: the KSE Service, https://kse.idt.iti.gr/

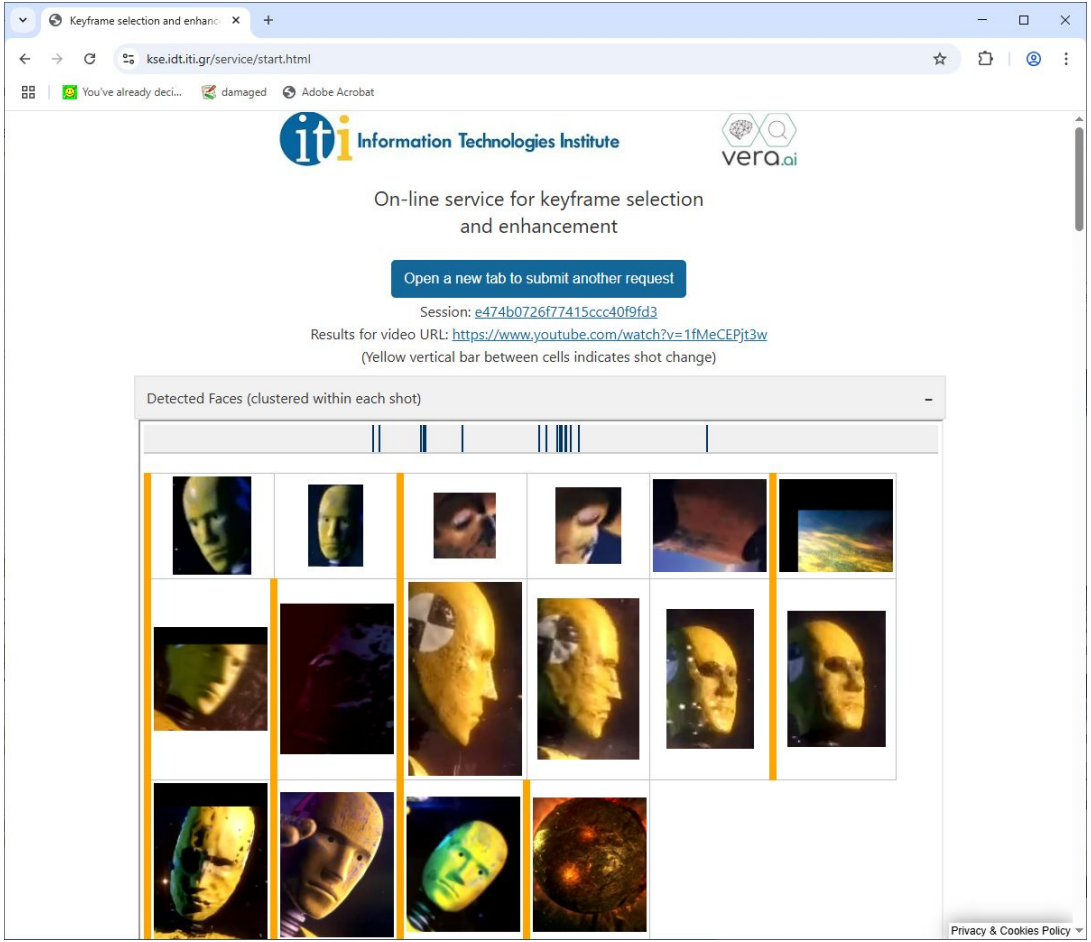
- Video temporal fragmentation; extract keyframes for reverse image search (many frames; user to select)
- Detect key elements (faces, text)
- Group similar key elements, and select the least blurry one from each group
- Employ super-resolution techniques to enhance the selected key elements



Video source: https://www.youtube.com/watch?v=1fMeCEPjt3w

Reverse Video Search: the KSE Service, https://kse.idt.iti.gr/

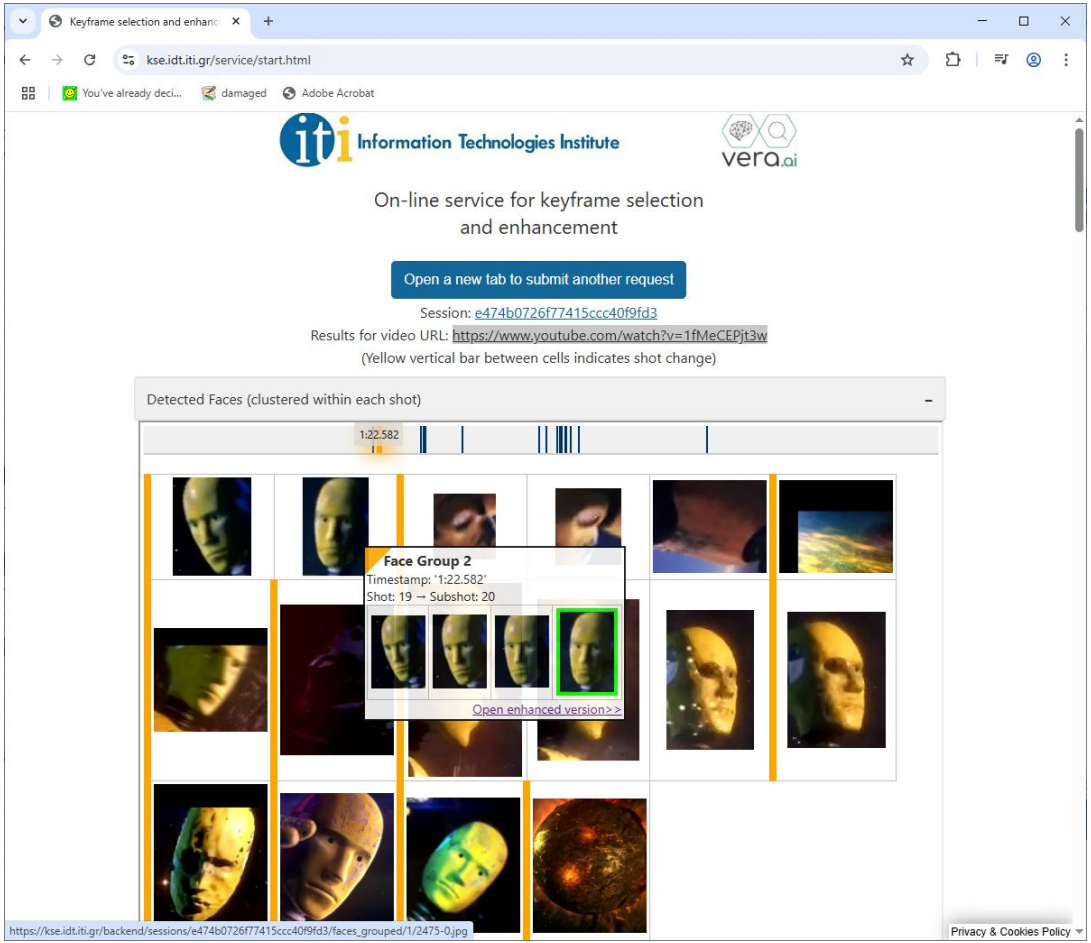
- Video temporal fragmentation; extract keyframes for reverse image search (many frames; user to select)
- Detect key elements (faces, text)
- Group similar key elements, and select the least blurry one from each group
- Employ super-resolution techniques to enhance the selected key elements



Video source: https://www.youtube.com/watch?v=1fMeCEPjt3w

Reverse Video Search: the KSE Service, https://kse.idt.iti.gr/

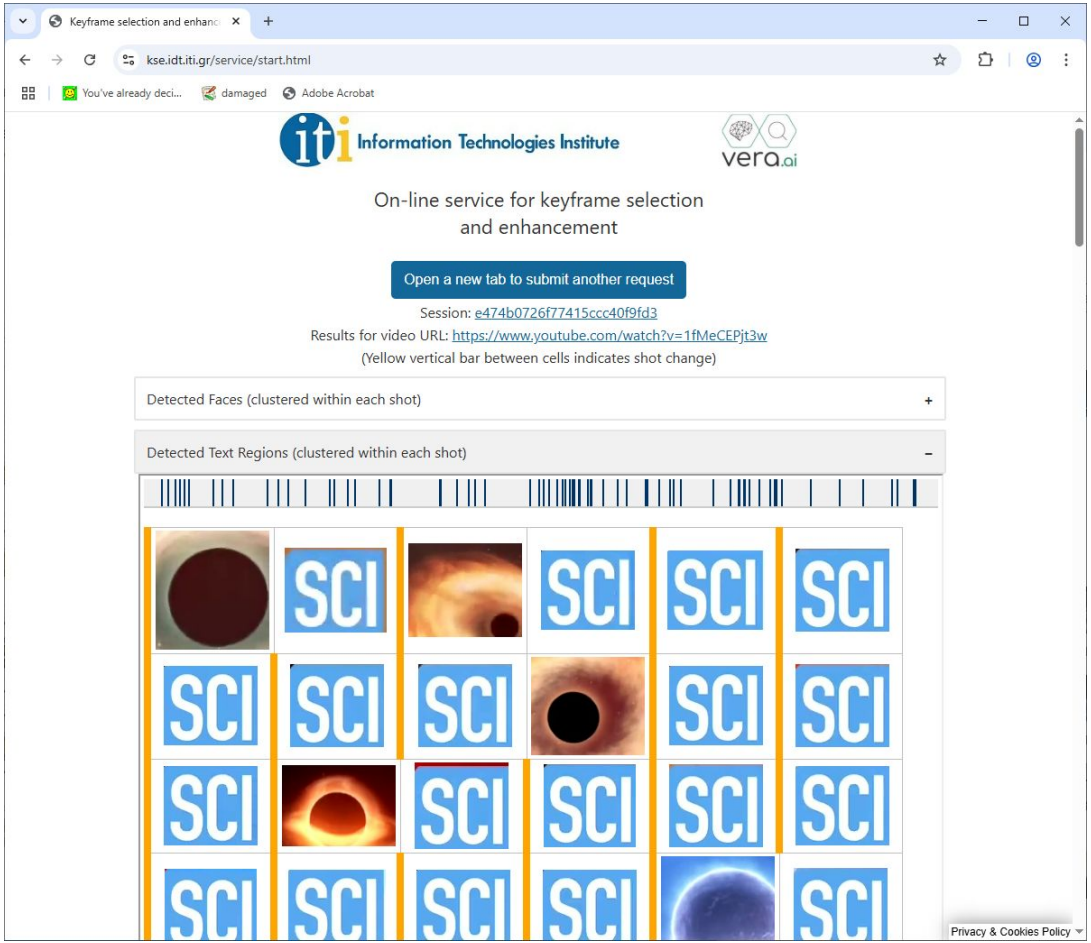
- Video temporal fragmentation; extract keyframes for reverse image search (many frames; user to select)
- Detect key elements (faces, text)
- Group similar key elements, and select the least blurry one from each group
- Employ super-resolution techniques to enhance the selected key elements



Video source: https://www.youtube.com/watch?v=1fMeCEPjt3w

Reverse Video Search: the KSE Service, https://kse.idt.iti.gr/

- Video temporal fragmentation; extract keyframes for reverse image search (many frames; user to select)
- Detect key elements (faces, text)
- Group similar key elements, and select the least blurry one from each group
- Employ super-resolution techniques to enhance the selected key elements



Video source: https://www.youtube.com/watch?v=1fMeCEPjt3w

Reverse Video Search

The KSE Service:

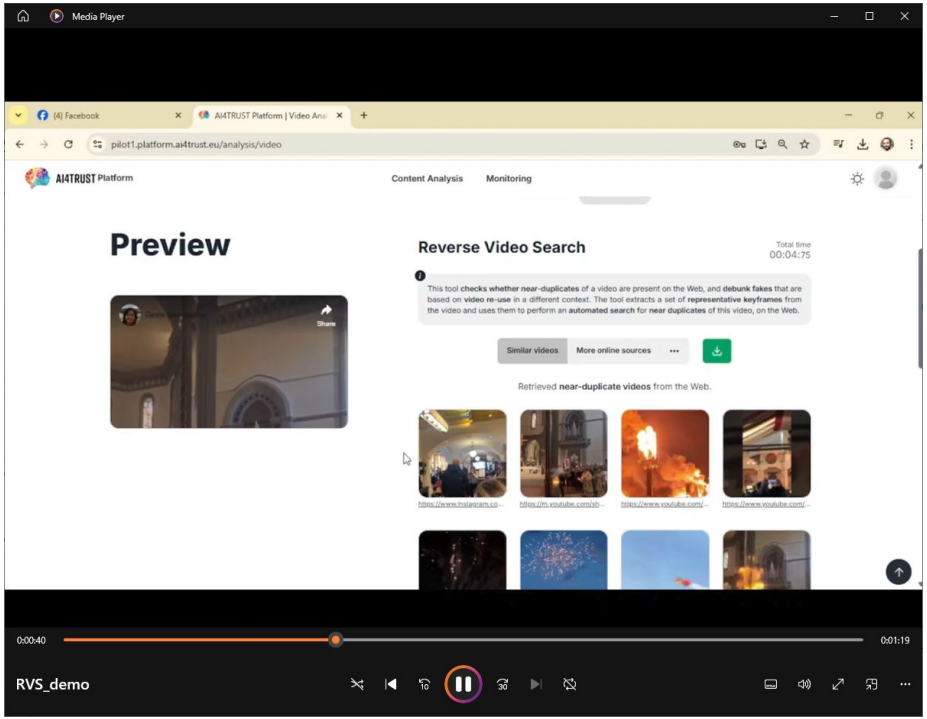
- Facilitates visual inspection of the content for interesting cues
- Gives the user a wide choice of keyframes for manually initiating reverse search on the web
- Requires the user to both initiate (typically, multiple) reverse searches and to manually inspect the results of each search -> great flexibility, but also requires a lot of user interaction (and effort)

How about making a different compromise between offered flexibility and required interaction?

- Automatically select just a few keyframes (Why few? Otherwise, cost and latency increase)
- Automatically run reverse search using these keyframes (on whichever web search engine(s)) and aggregate the results, to present a single list of the few most relevant findings to the user

Reverse Video Search: the RVS functionality of AI4TRUST

- Video demo:



Reverse Video Search: the RVS functionality of AI4TRUST

The key methodological challenge for realizing the RVS functionality of AI4TRUST:

- Selecting just very few, representative keyframes, to automatically perform reverse search on the web

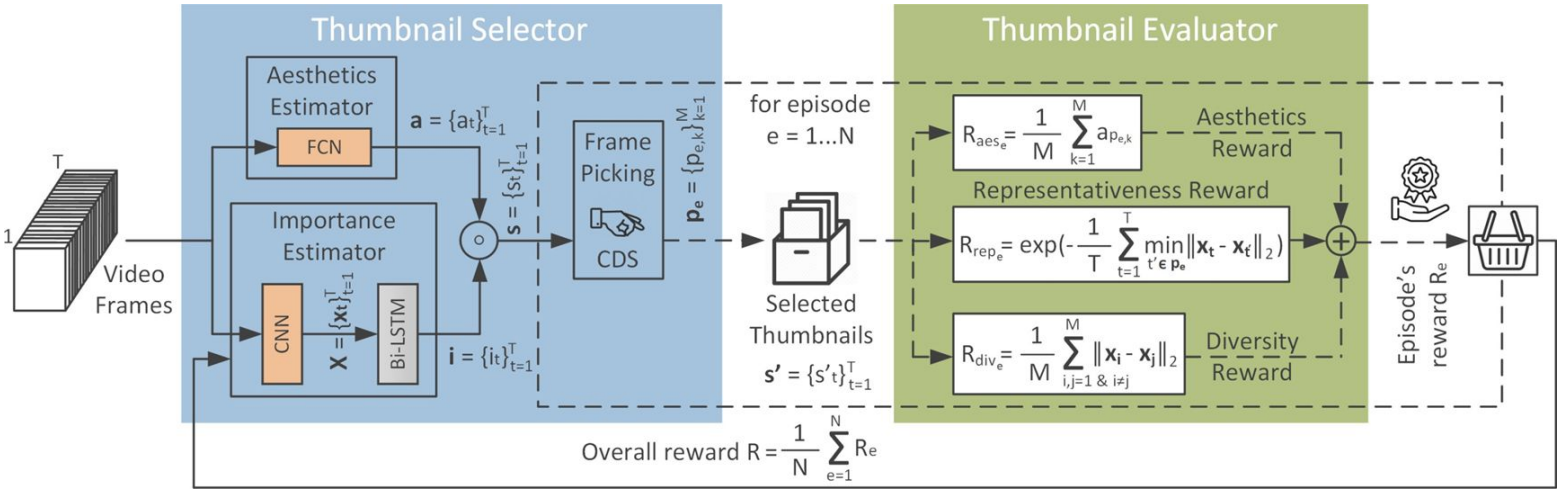
Our solution:

- Use our RL-DiVTS method *

* E. Apostolidis, G. Balaouras, V. Mezaris, I. Patras, "Selecting a Diverse Set of Aesthetically-pleasing and Representative Video Thumbnails using Reinforcement Learning", IEEE Int. Conf. on Image Processing (ICIP 2023), Kuala Lumpur, Malaysia, Oct. 2023. DOI:10.1109/ICIP49359.2023.10222743.

Keyframe Selection: RL-DiVTS

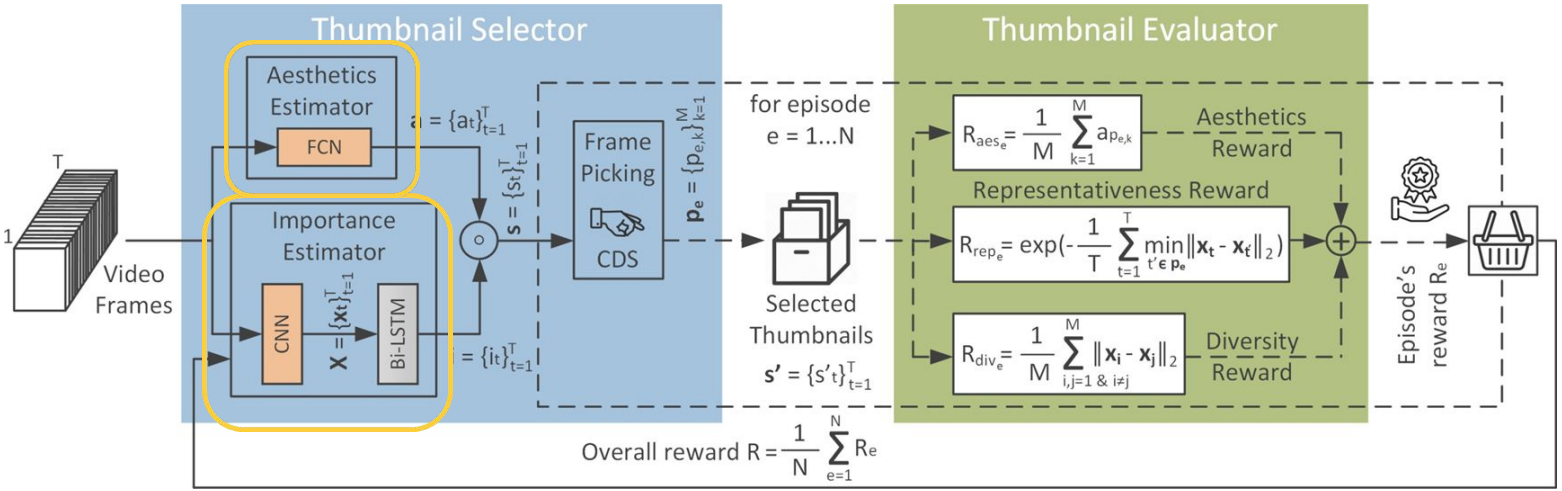
RL-DiVTS main concept: A Thumbnail Selector learns the task with the help of a Thumbnail Evaluator and after using its feedback on various features as a reward for reinforcement learning



Keyframe Selection: RL-DiVTS

Thumbnail Selector

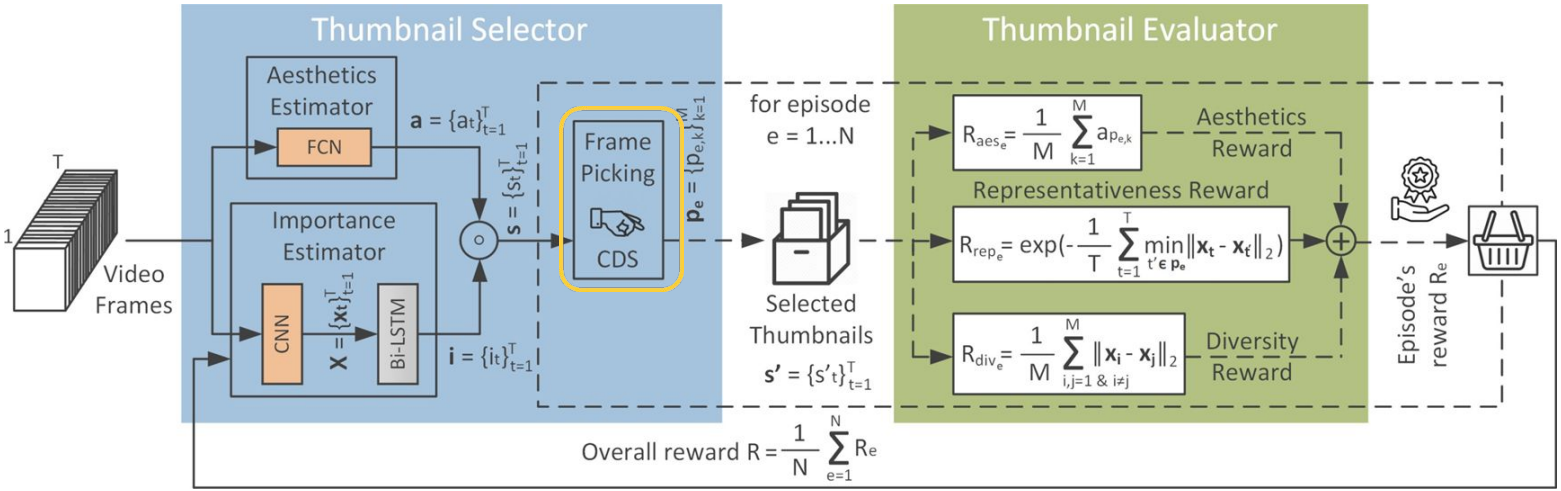
- Aesthetic Estimator: model of a Fully Convolutional Network trained on AVA dataset
- Importance Estimator: pre-trained CNN on ImageNet & trainable bi-directional LSTM



Keyframe Selection: RL-DiVTS

Thumbnail Selector

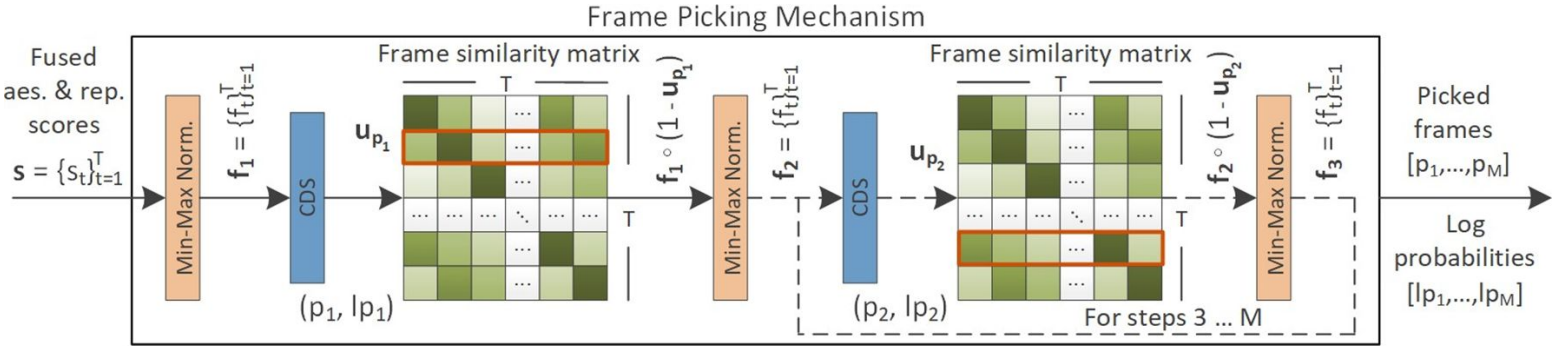
- The computed scores (frames' aesthetic quality, importance) are fused via their Hadamard product; the resulting sequence of scores is used by the Frame Picking mechanism



Keyframe Selection: RL-DiVTS

Thumbnail Selector

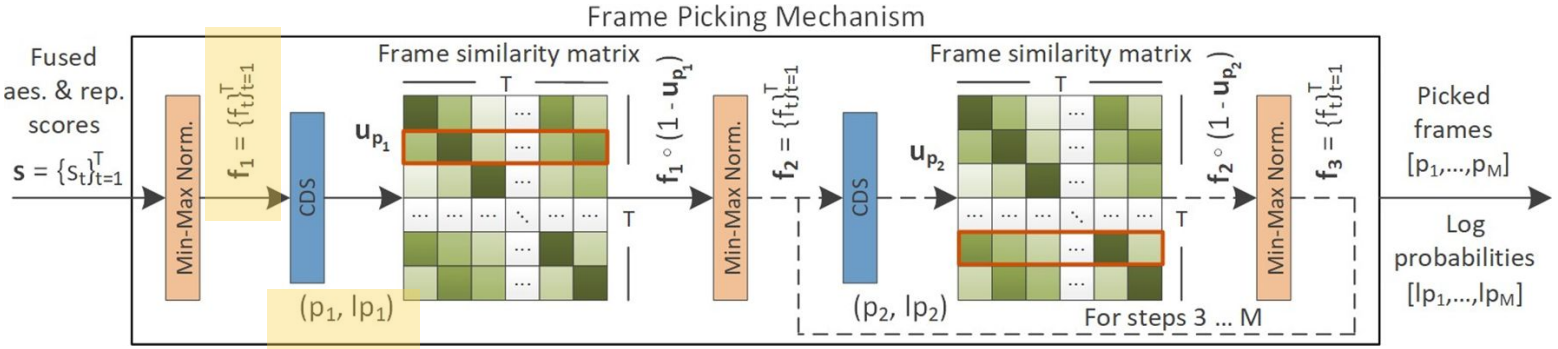
- Frame Picking Mechanism: Categorical Distribution Sampler that picks frames in a sequential manner and applies a re-weighting process to demote the selection of frames that are visually-similar with the previously selected ones



Keyframe Selection: RL-DiVTS

Thumbnail Selector

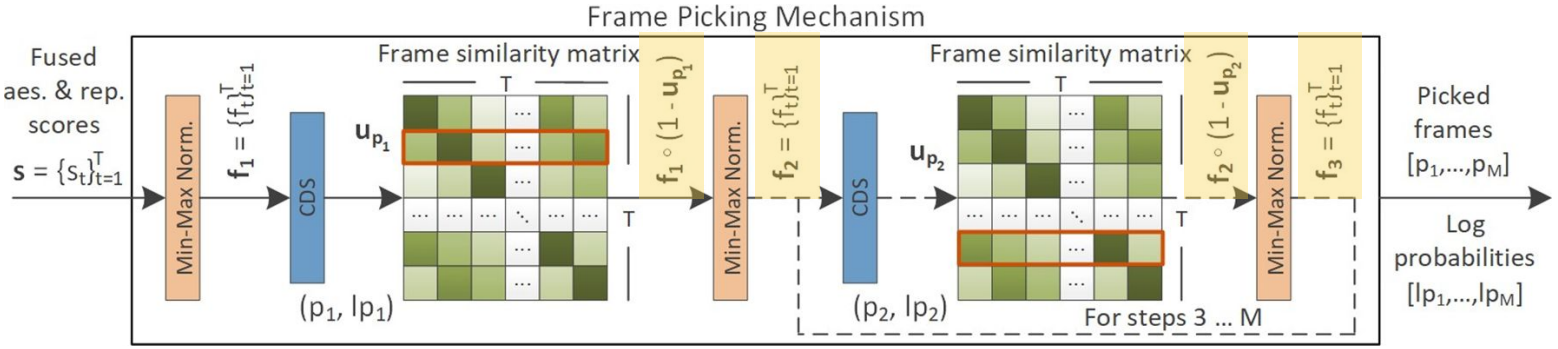
- At the first step, the distribution is based on the normalized version of the fused aesthetic and representativeness scores, and the sampling process results in the first picked frame and a log probability of picking this sample from the distribution



Keyframe Selection: RL-DiVTS

Thumbnail Selector

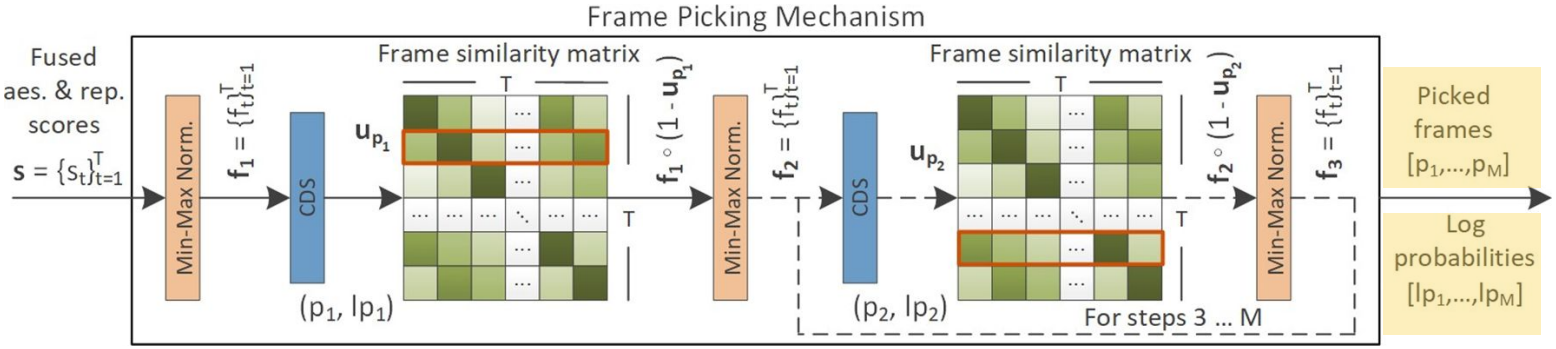
- At each subsequent step, the distribution is based on the normalized version of the distribution from the previous step after applying a re-weighting process that demotes the selection of frames that are visually-similar to the already picked ones



Keyframe Selection: RL-DiVTS

Thumbnail Selector

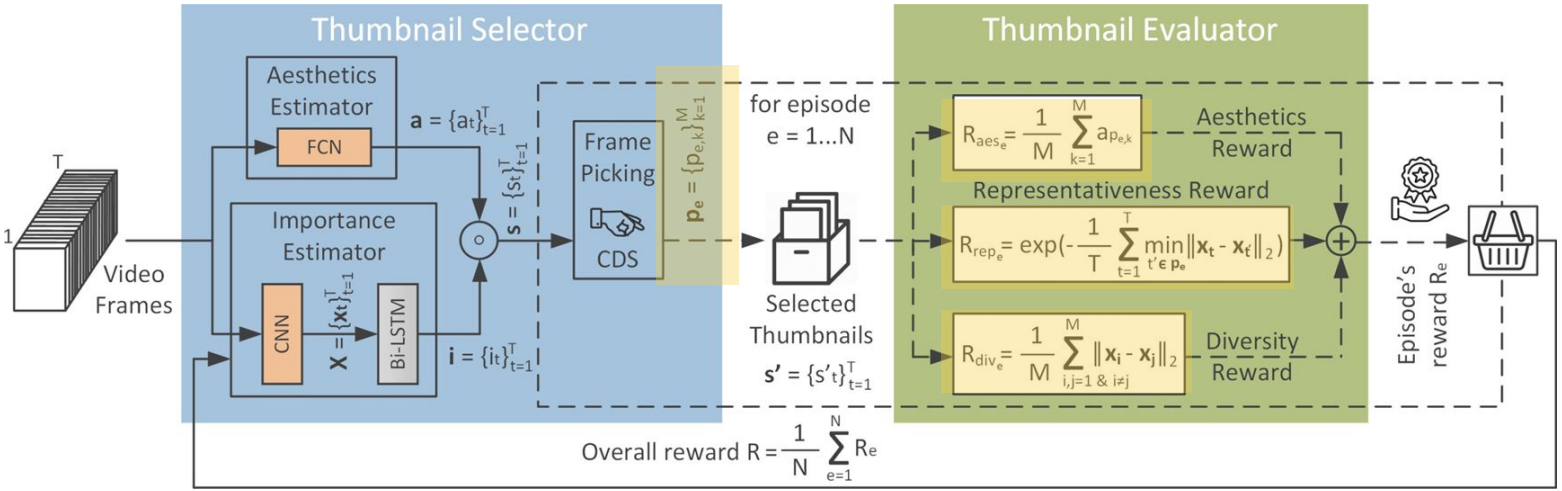
- After the end of the M steps the Frame Picking mechanism defines a set of picked frames and a set of log probabilities that are used to compute the expected reward in the context of episodic reinforcement learning



Keyframe Selection: RL-DiVTS

Thumbnail Evaluator

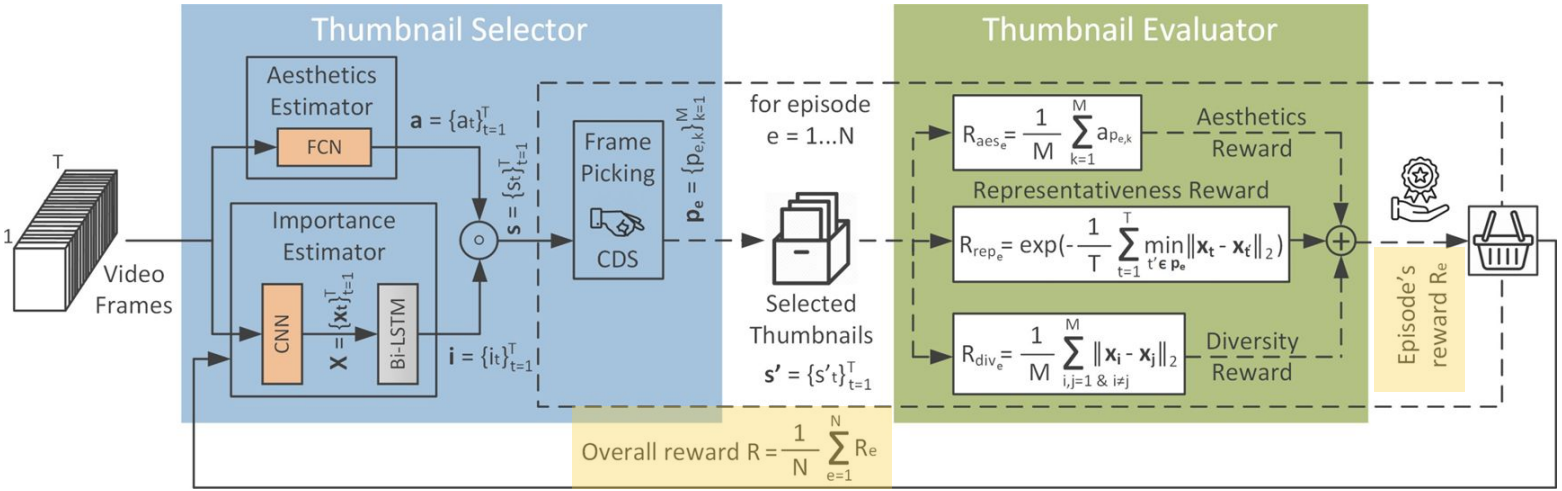
- Evaluates the visual content of the selected thumbnails in terms of aesthetic quality, representativeness and diversity, and returns a reward for reinforcement learning



Keyframe Selection: RL-DiVTS

Thumbnail Evaluator

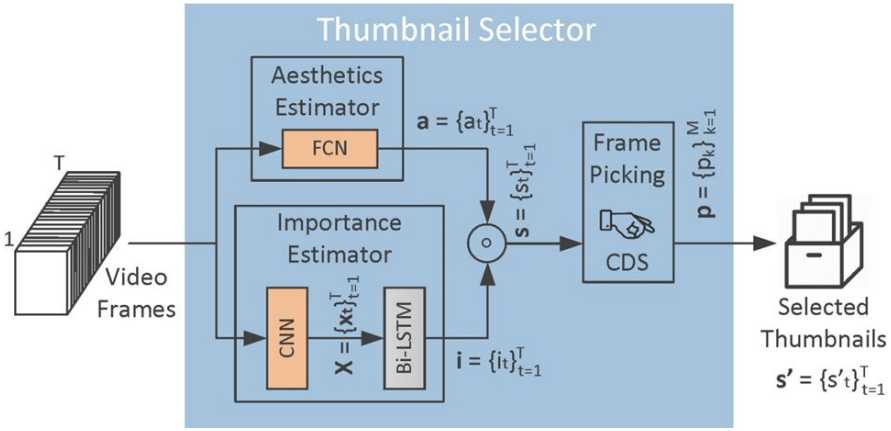
- Overall reward for an episode: weighted sum; Average reward across episodes: the feedback of the Thumbnail Evaluator for the current training sample



Keyframe Selection: RL-DiVTS

Inference Stage

- Thumbnail Selector assesses the entire frame sequence in terms of aesthetic quality and visual importance and uses its mechanism to pick three video thumbnails



- Aesthetically-pleasing
- Representative
- Visually-diverse

RL-DiVTS Experiments: Datasets

Open Video Project (OVP) (https://sites.google.com/site/vsummsite/download)

- 50 videos of various genres (e.g. documentary, educational, historical, lecture)
- Video length: 46 sec. to 3.5 min.
- Annotation: keyframe-based video summaries (5 per video)

Youtube (https://sites.google.com/site/vsummsite/download)

- 50 videos of diverse content (e.g. news, TV-shows, sports, commercials)
- Video length: 9 sec. to 11 min.
- Annotation: keyframe-based video summaries (5 per video)

RL-DiVTS Experiments: Evaluation Protocol

- Ground-truth thumbnails: top-3 selected keyframes by human annotators
- Similarity estimation between ground-truth and machine-selected keyframes: Structural Similarity Index Measure (SSIM); call it match is SSIM > 0.7
- Evaluation measure: Top-3 matching
  - Overlap between top-3 machine- and human-selected thumbnails per video
  - Expressed as percentage (%)
- Run experiments using 5 different randomly-created splits of the used data (80% training; 20% testing) and 5 random seeds for network initialization, and report the average performance and the standard deviation over these runs

RL-DiVTS Experimental Comparisons

Thumbnail selection performance
(Top-3 matching %)

	OVP	YouTube
Baseline (random)	8.63 ± 2.50	4.41 ± 1.77
AC-SUM-GAN [1]	7.87 ± 3.41	7.33 ± 0.70
CA-SUM [2]	7.60 ± 2.85	8.00 ± 3.56
Hecate-VTS [3]	11.72	16.47
ReconstSum [4]	12.18	**18.25**
ARL-VTS [5]	12.50 ± 3.37	7.83 ± 1.49
RL-DiVTS (proposed)	**25.33 ± 3.97**	17.50 ± 2.57

Training time and memory footprint

	Training time (sec/epoch)		# Parameter (in Millions)
	OVP	YouTube	
ARL-VTS [5]	38.41	62.43	28.36
RL-DiVTS (proposed)	2.33	2.70	**12.60**

[1] E. Apostolidis, et al., (2021). AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization. IEEE Trans. CSVT , vol. 31, no. 8, pp. 3278-3292, Aug. 2021.
[2] E. Apostolidis et al., (2022). Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames. ACM ICMR 2022
[3] Y. Song, et al., (2016). To Click or Not To Click: Automatic Selection of Beautiful Thumbnails from Videos. CIKM 2016
[4] H. Gu et al., (2018). From Thumbnails to Summaries - A Single Deep Neural Network to Rule Them All. IEEE ICME 2018
[5] E. Apostolidis et al., (2021). Combining adversarial and reinforcement learning for video thumbnail selection. ACM ICMR 2021

RL-DiVTS Ablation study

Contribution of the adopted video thumbnail evaluation criteria (Top-3 matching %)

Variants	OVP	YouTube
RL-DiVTS w/o AES	14.13 ± 2.96	10.33 ± 1.73
RL-DiVTS w/o REP	20.53 ± 1.91	13.17 ± 1.09
RL-DiVTS w/o DIV	**26.40 ± 1.30**	14.33 ± 1.49
RL-DiVTS w/o CDS	24.67 ± 3.16	15.00 ± 1.44
RL-DiVTS (proposed)	25.33 ± 3.97	17.50 ± 2.57

The removal of either of the used criteria and the integrated CDS-based Frame Picking mechanism results in performance degradation in, at least, one of the used datasets

Impact of the number of picked frames during training (Top-3 matching %)

Frames	OVP	YouTube
3	20.80 ± 1.66	13.67 ± 1.73
6	**25.33 ± 3.97**	**17.50 ± 2.57**
9	19.33 ± 1.94	11.67 ± 3.12

Picking fewer or more than 6 frames during training leads to reduced performance in both datasets

RL-DiVTS Conclusions

- Video thumbnail selection is learned using a pair of Thumbnail Selector and Evaluator
- Thumbnail Selector picks thumbnails based on their aesthetic quality and visual importance and using a diversity-aware frame picking mechanism
- Thumbnail Evaluator assesses picked thumbnails using tailored reward functions and his feedback is used for reinforcement learning
- Experiments on two benchmark datasets (OVP and Youtube) showed the advanced performance of RL-DiVTS against SoA video thumbnail selection or summarization approaches
- Ablations documented the significance of each of the utilized criteria for video thumbnail selection and the contribution of the integrated frame picking mechanism

Part1: Combatting video-borne disinformation - Conclusions

- Trying to find the original posting / version of the video is probably the first thing to do when assessing the veracity of a video
  - One of the first functionalities integrated in the Verification Plugin*
  - The most widely-used functionality of the Plugin (accounts for >20% of the actions taken by the >100.000 active users of the Plugin)
- Automatically selecting just a few representative keyframes is key to automating reverse video search

* D. Teyssou, J.-M. Leung, E. Apostolidis, K. Apostolidis, S. Papadopoulos, M. Zampoglou, O. Papadopoulou, V. Mezaris, "The InVID Plug-in: Web Video Verification on the Browser", Proc. Int. Workshop on Multimedia Verification (MuVer 2017) at ACM Multimedia 2017, Mountain View, CA, USA, October 2017.
https://chromewebstore.google.com/detail/fake-news-debunker-by-inv/mhccpoafgdgbhnjfhkcmgknndkeenfhe

Part 2: Increasing trust in AI methods for combatting disinformation

Part2: Increasing trust in AI methods for combatting disinformation

- Deepfake detectors (and many other detectors / classifiers in the disinformation domain) produce binary decisions or confidence scores
- Typically, they are far from reaching 100% accuracy (especially in the wild)

How can a user trust and make good use of such imperfect decisions?

- AI Explainability techniques can provide more insight on how a decision was made
- We will discuss our method for explaining deepfake detector outputs *

* K. Tsigos, E. Apostolidis, V. Mezaris, "Improving the Perturbation-Based Explanation of Deepfake Detectors Through the Use of Adversarially-Generated Samples", Proc. IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW 2025), Tucson, AZ, USA, pp. 658-667, Feb. 2025. DOI:10.1109/WACVW65960.2025.00080.

Deepfakes: definition and current status

Definition:

- Deepfakes are AI manipulated media in which, a person's face or body is digitally swapped to alter their identity or reenacted according to a driver video

Current status:

- Advancement of Generative AI allows to create deepfakes that are increasingly difficult to detect
- Over the last years deepfakes have been used as a means for spreading disinformation
- Increasing need for effective solutions for deepfake detection

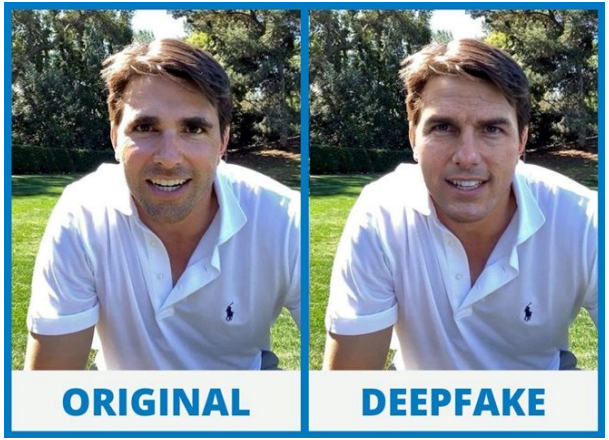


Image source: https://malcomvetter.medium.com/deep-deep-fakes-d4507c735f44

How to detect them?

Through human inspection

- An investigator carefully checks for inconsistencies or artifacts in the image or video, e.g. unnatural lighting and facial movements, or mismatched audio

Using trained deepfake detectors

- An investigator analyses the image or video using a trained deepfake detector and takes into account the output of the analysis for making a decision

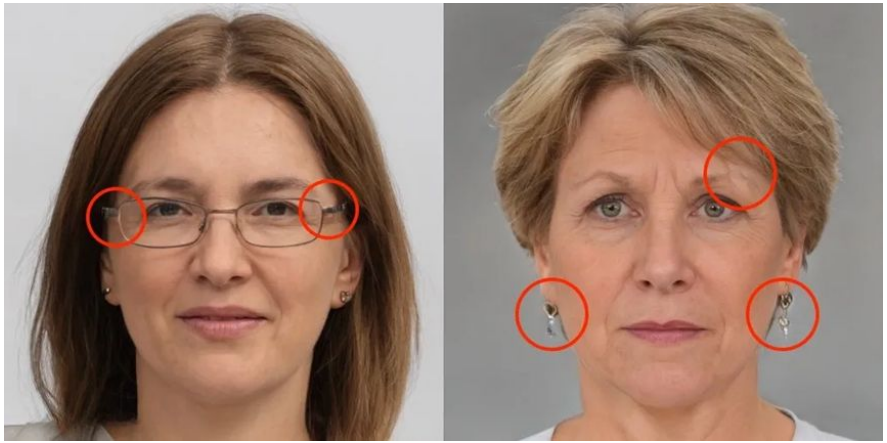


Image source: https://bdtechtalks.com/2023/05/12/detect-deepfakes-ai-generated-media

Why explainable deepfake detection?

- The decision mechanism behind trained deepfake detectors is neither visible to the user nor straightforward to understand
- Enhancing deepfake detectors with explanation mechanisms about their outputs would significantly improve the users' trust in them

Visual explanations could provide

- Insights about the applied manipulation for creating the detected deepfake
- Clues about the trustworthiness of the detector’s decision

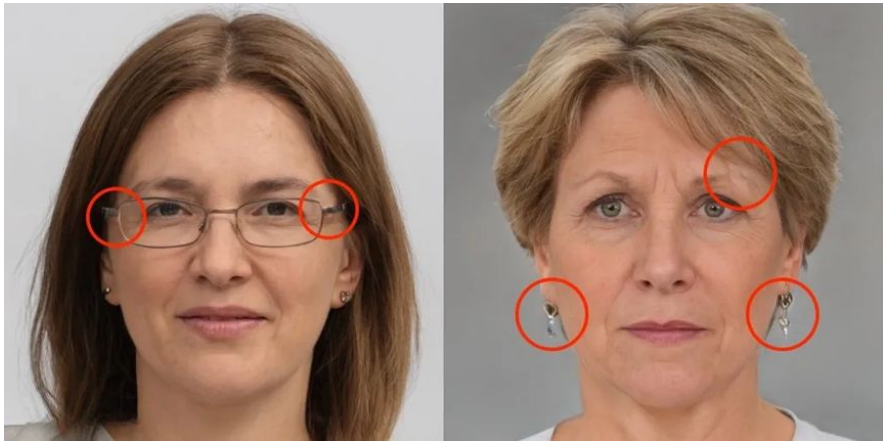


Image source: https://bdtechtalks.com/2023/05/12/detect-deepfakes-ai-generated-media

Related work

Categorization	XAI methods	Approach
Scope	Global	Provide a complete description of the model (SHAP, SOBOL, Global Surrogate Models)
	Local	Focus on explaining predictions made by a specific instance or input of the model (LIME, SHAP, RISE, Anchor, LRP)
Stage	Ante-hoc	Employed during the training and development stage (Decision Trees, Linear/Logistic Regression, Rule-based Models)
	Post-hoc	Employed after the training process (LIME, SHAP, SOBOL, RISE, LRP)
Methodology	Perturbation-based	Operate by modifying input data and observing changes in the model's output (LIME, SHAP, SOBOL, RISE)
	Gradient-based	Operate by computing gradients of the model's predictions for input data (Grad-CAM, Grad-CAM++, LRP, SmoothGrad)

M. Mersha, K. Lam, J. Wood, A. K. Al-Shami, J. Kalita. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. Neurocomputing, 599:128111, 2024.

Related work

Work	Approach
Malolan et al., 2020	Use of **LIME** and **LRP** to explain an XceptionNet deepfake detector; **quantitative** evaluation on a few samples focusing on their robustness against affine transformations or Gaussian blurring of the input
Pino et al. 2021	Use of adaptations of **SHAP**, **Grad-CAM** & **self-attention methods**, to explain deepfake detectors; **quantitative evaluation** taking into account low-level features of visual explanations
Xu et al., 2022	Production of **heatmap visualizations** and **UMAP topology** explanations using the learned features of a linear deepfake detector; **qualitative** evaluation on some examples and examining the manifolds
Silva et al., 2022	Use of **Grad-CAM** to explain an ensemble of CNNs and an attention-based model for deepfake detection; **qualitative** evaluation using a few examples
Jayakumar et al., 2022	Use of **Anchors** and **LIME** to explain an EfficientNet deepfake detector; **qualitative** evaluation with human participants and extraction of metrics for quantitative evaluation
Aghasanli et al., 2023	Use of **support vectors/prototypes** of an SVM and **xDNN classifier** to explain a ViT deepfake detector; **qualitative** evaluation using a few examples
Haq et al., 2023	Production of **textual explanations** for a neurosymbolic method that detects emotional inconsistencies in manipulated faces using a deepfake detector; **evaluation discussed theoretically**
Gowrisankar et al., 2024	**Quantitative evaluation framework** that takes into account the drop in the detector's accuracy after adversarial attacks on regions of fake images by leveraging the produced explanations of their non-manipulated counterparts
Tsigos et al., 2024	**Quantitative evaluation framework** (following the idea of the above) which uses the produced explanation after detecting a deepfake image and does not require access to its original counterpart.

Related work

Categorization	XAI methods	Approach
Scope	Global	Provide a complete description of the model (SHAP, SOBOL, Global Surrogate Models)
	Local	Focus on explaining predictions made by a specific instance or input of the model (LIME, SHAP, RISE, Anchor, LRP)
Stage	Ante-hoc	Employed during the training and development stage (Decision Trees, Linear/Logistic Regression, Rule-based Models)
	Post-hoc	Employed after the training process (LIME, SHAP, SOBOL, RISE, LRP)
Methodology	Perturbation-based	Operate by modifying input data and observing changes in the model's output (LIME, SHAP, SOBOL, RISE)
	Gradient-based	Operate by computing gradients of the model's predictions for input data (Grad-CAM, Grad-CAM++, LRP, SmoothGrad)

M. Mersha, K. Lam, J. Wood, A. K. Al-Shami, J. Kalita. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. Neurocomputing, 599:128111, 2024.

Remarks on perturbation-based methods

- **Various types of perturbations** have been used on explainable image/video classifiers, including: occlusion of input features, replacement with fixed/random values, blurring, or Gaussian noise
- **Less suitable** for explaining deepfake detectors, as they might result in **images outside training data distribution**, giving rise to the OOD issue & leading to unexpected model behavior
- The XAI method **cannot accurately detect** whether the observed change in the detector's output relates to the modification of important input features or with the shift in the data distribution

original labels: *"watch", "pen"*

replace with background to explain label *"pen"*

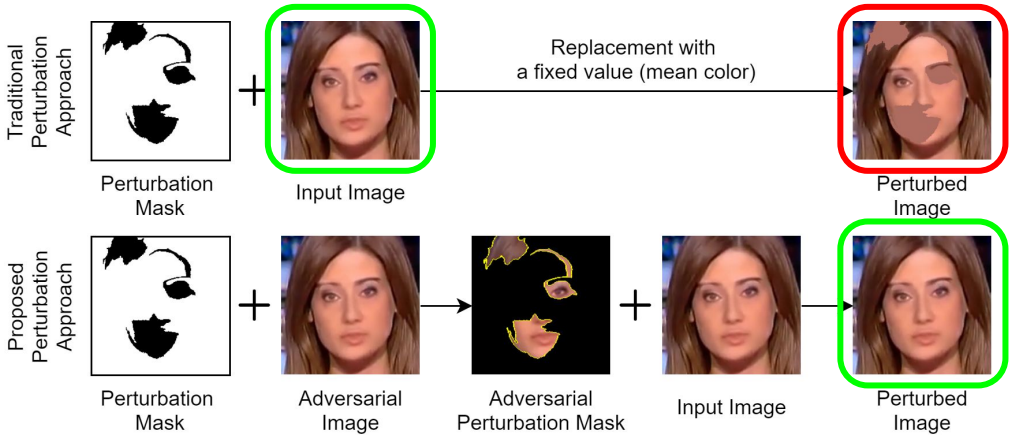
replace with background to explain label *"watch"*

(a) Original (b) Occlude (c) Replace (d) Blurring (e) Gaussian noise

Our main idea

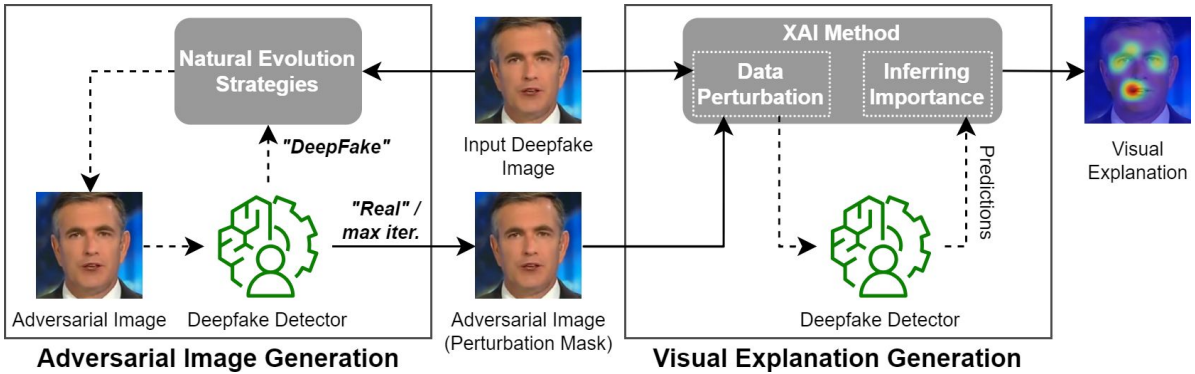
"Use an adversarially-generated sample of the input deepfake image that flips the detector's decision, to form perturbation masks for inferring the importance of different features"

- Leads to perturbed instances that are visually-similar with the input image (see lower part)
- Avoids OOD issues raised by traditional perturbation approaches (see upper part)
- Allows to infer the importance of different features more effectively



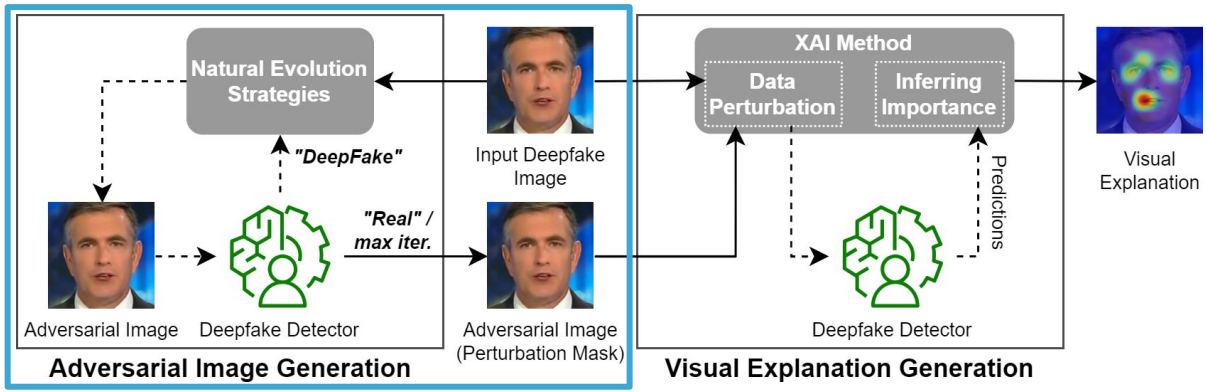
Processing pipeline

- Produces a visual explanation after a detector classifies an input image as a deepfake, providing clues about regions of the image that were found to be manipulated
- Adversarial image generation: creates an adversarial sample of the input image that can fool the detector to classify is as “real”
- Visual explanation generation: uses the generated sample to form perturbation masks, infer the importance of different parts of the input image and generate the visual explanation



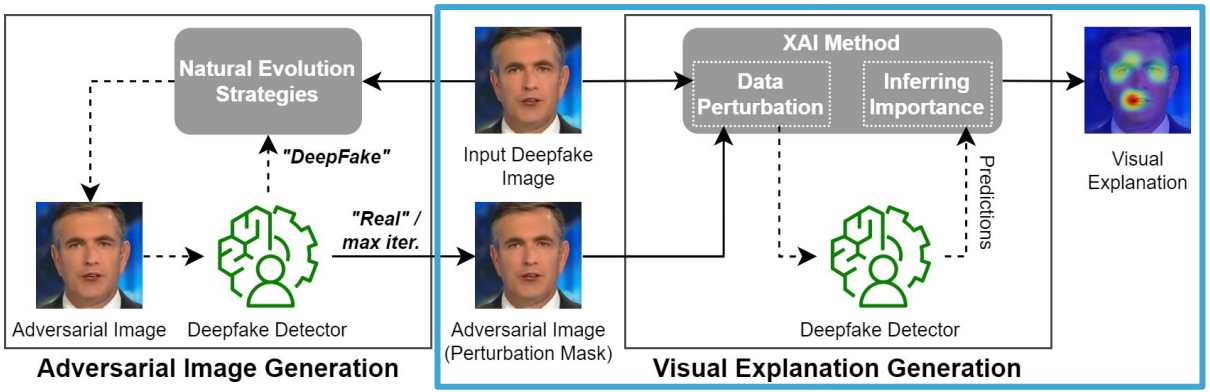
Processing pipeline - Adversarial image generation

- Implemented through an iterative process that stops when the detector is fooled to classify the generated adversarial sample as “real” or a maximum number of iterations is reached
- Performed by progressively adding a small magnitude of Gaussian noise to the entire image using Natural Evolution Strategies and the computed gradients based on the detector’s output
- The applied NES try to find a minimal change in the pixels’ values that will cause the detector to mis-classify the generated image as “real”



Processing pipeline - Visual explanation generation

- Can include any perturbation-based explanation method (e.g., LIME, SHAP, etc.)
- Uses the adversarial image to form perturbation masks for the input deepfake image
- Produces a number of perturbed images that are analyzed by the detector
- Takes into account the applied perturbations and the corresponding predictions of the detector, to infer the importance of different input features and create the visual explanation



Experimental setup - Explanation methods

- **LIME**: replaces portions of the input image with the mean pixel value and approximates the model's behavior by fitting perturbation data and the model's outputs into a simpler model
- **SHAP**: constructs an additive feature attribution model that attributes an effect to each input feature and sums the effects (Shapley values) as a local approximation of the output
- **SOBOL**: performs blurring-based perturbations and uses the relationship of perturbation masks and model's predictions to estimate the order of Sobol' indices and each region's importance
- **RISE**: produces binary masks to occlude image regions, uses the model's predictions to weight the corresponding mask, and aggregates the weighted masks together to form the explanation

We combine each of these methods with the proposed adversarial-based perturbation approach and name the modified versions as LIME_{adv}, SHAP_{adv}, SOBOL_{adv} and RISE_{adv}, respectively

Experimental setup - Dataset

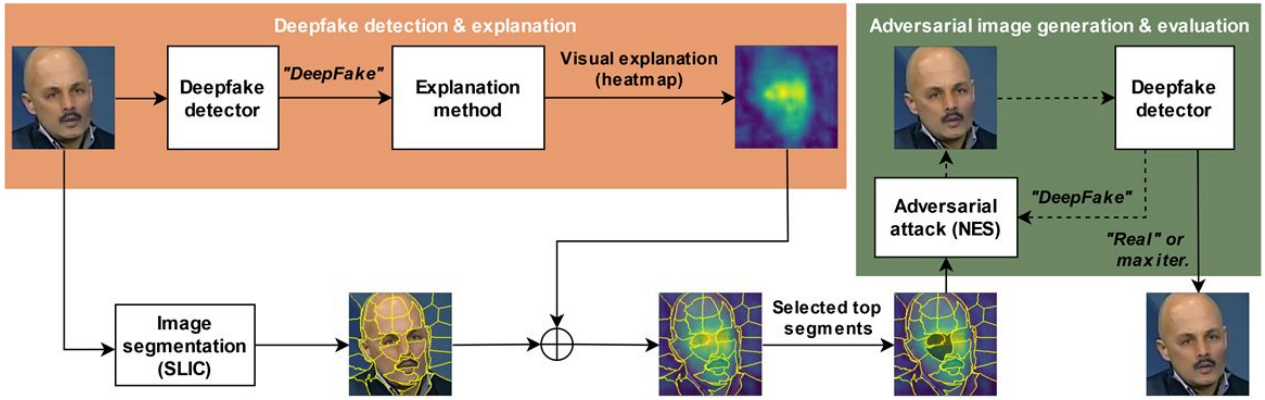
FaceForensics++
(https://github.com/ondyari/FaceForensics)

- Contains 1000 original videos and 4000 fake videos
- 4 fake video classes: FaceSwap (FS), DeepFakes (DF), Face2Face (F2F), NeuralTextures (NT)
- 720 videos for training, 140 for validation and 140 for testing, respectively
- Used 127 videos from each different class of the test set and sampled 10 frames per video, creating four sets of 1270 images

Image Source:
https://github.com/ondyari/FaceForensics

Experimental setup - Evaluation protocol *

- Assesses the performance of an XAI method by examining the extent to which the image regions that were found as the most important ones, can be used to flip the deepfake detector's decision
- Regions defined using SLIC and scored by averaging pixel-level scores from visual explanation
- Alteration of top-3 scoring (most important) regions performed via adversarial attacks using NES

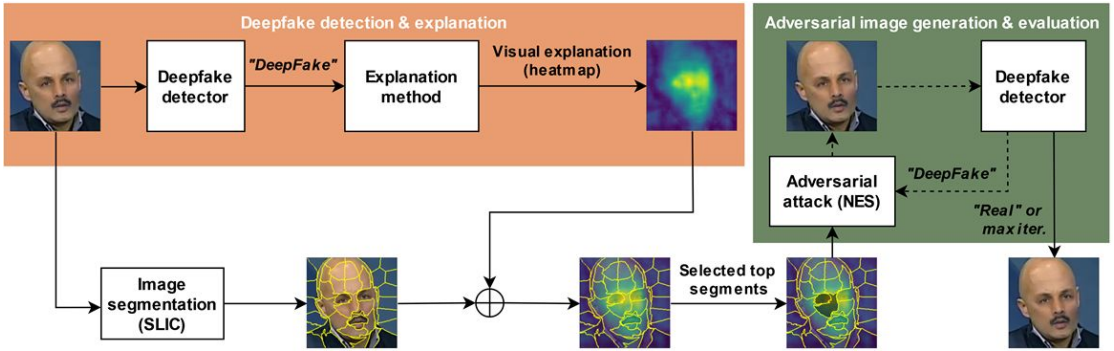


* K. Tsigos, E. Apostolidis, S. Baxevanakis, S. Papadopoulos,V. Mezaris. (2024). Towards quantitative evaluation of explainable AI methods for deepfake detection. Proc. of the 3rd ACM Int. Workshop on Multimedia AI against Disinformation, MAD @ACM MM 2024

Experimental setup - Evaluation protocol

Measures

- **(drop in) Detection accuracy** after affecting the top-3 scoring regions by the XAI method; the lower the accuracy scores, the higher the ability of a method to spot the most important regions
- **Explanation sufficiency**: difference in detector's output after affecting the top-3 scoring regions by the XAI method; high scores indicate high impact of the top-3 scoring regions on the detector's output and high sufficiency for the produced visual explanation



Experimental results - Quantitative analysis

	DF			F2F			FS			NT		
Original	0.978			0.977			0.982			0.924		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
LIME	0.735	0.440	0.245	0.803	0.633	0.484	0.864	0.698	0.559	0.579	0.340	0.197
LIME_{adv}	**0.673**	**0.363**	**0.205**	**0.784**	**0.568**	**0.392**	**0.790**	**0.538**	**0.378**	**0.411**	**0.155**	**0.056**
SHAP	0.813	0.609	0.450	0.846	0.739	0.637	0.876	0.702	0.543	0.686	0.497	0.344
SHAP_{adv}	0.706	0.434	0.249	0.811	0.647	0.482	0.820	0.611	0.430	0.485	0.238	0.129
SOBOL	0.750	0.591	0.490	0.816	0.653	0.512	0.874	0.703	0.574	0.621	0.417	0.313
SOBOL_{adv}	0.696	0.507	0.383	0.828	0.680	0.555	0.831	0.627	0.506	0.478	0.252	0.152
RISE	0.877	0.766	0.686	0.843	0.710	0.622	0.896	0.809	0.734	0.783	0.637	0.513
RISE_{adv}	0.769	0.679	0.618	0.876	0.824	0.784	0.917	0.867	0.830	0.515	0.378	0.322

Detection accuracy for the different types of fakes, on the original images and on variants of them after adversarial attacks on image regions corresponding to the top-1, top-2 and top-3 scoring segments by the different XAI methods. Best (lowest) scores in bold and second best scores underlined.

Experimental results - Quantitative analysis

	DF			F2F			FS			NT		
Original	0.978			0.977			0.982			0.924		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
LIME	0.735	0.440	0.245	0.803	0.633	0.484	0.864	0.698	0.559	0.579	0.340	0.197
LIME_{adv}	**0.673**	**0.363**	**0.205**	**0.784**	**0.568**	**0.392**	**0.790**	**0.538**	**0.378**	**0.411**	**0.155**	**0.056**
SHAP	0.813	0.609	0.450	0.846	0.739	0.637	0.876	0.702	0.543	0.686	0.497	0.344
SHAP_{adv}	0.706	0.434	0.249	0.811	0.647	0.482	0.820	0.611	0.430	0.485	0.238	0.129
SOBOL	0.750	0.591	0.490	0.816	0.653	0.512	0.874	0.703	0.574	0.621	0.417	0.313
SOBOL_{adv}	0.696	0.507	0.383	0.828	0.680	0.555	0.831	0.627	0.506	0.478	0.252	0.152
RISE	0.877	0.766	0.686	0.843	0.710	0.622	0.896	0.809	0.734	0.783	0.637	0.513
RISE_{adv}	0.769	0.679	0.618	0.876	0.824	0.784	0.917	0.867	0.830	0.515	0.378	0.322

Detection accuracy for the different types of fakes, on the original images and on variants of them after adversarial attacks on image regions corresponding to the top-1, top-2 and top-3 scoring segments by the different XAI methods. Best (lowest) scores in bold and second best scores underlined.

- Modified LIME is the top-performing method for all types of fakes and experimental settings

Experimental results - Quantitative analysis

	DF			F2F			FS			NT		
Original	0.978			0.977			0.982			0.924		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
LIME	0.735	0.440	0.245	0.803	0.633	0.484	0.864	0.698	0.559	0.579	0.340	0.197
LIME_{adv}	**0.673**	**0.363**	**0.205**	**0.784**	**0.568**	**0.392**	**0.790**	**0.538**	**0.378**	**0.411**	**0.155**	**0.056**
SHAP	0.813	0.609	0.450	0.846	0.739	0.637	0.876	0.702	0.543	0.686	0.497	0.344
SHAP_{adv}	0.706	0.434	0.249	0.811	0.647	0.482	0.820	0.611	0.430	0.485	0.238	0.129
SOBOL	0.750	0.591	0.490	0.816	0.653	0.512	0.874	0.703	0.574	0.621	0.417	0.313
SOBOL_{adv}	0.696	0.507	0.383	0.828	0.680	0.555	0.831	0.627	0.506	0.478	0.252	0.152
RISE	0.877	0.766	0.686	0.843	0.710	0.622	0.896	0.809	0.734	0.783	0.637	0.513
RISE_{adv}	0.769	0.679	0.618	0.876	0.824	0.784	0.917	0.867	0.830	0.515	0.378	0.322

Detection accuracy for the different types of fakes, on the original images and on variants of them after adversarial attacks on image regions corresponding to the top-1, top-2 and top-3 scoring segments by the different XAI methods. Best (lowest) scores in bold and second best scores underlined.

- Modified SHAP is the second most competitive one, while modified RISE is the weakest

Experimental results - Quantitative analysis

	DF			F2F			FS			NT		
Original	0.978			0.977			0.982			0.924		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
LIME	0.735	0.440	0.245	0.803	0.633	0.484	0.864	0.698	0.559	0.579	0.340	0.197
LIME_{adv}	**0.673**	**0.363**	**0.205**	**0.784**	**0.568**	**0.392**	**0.790**	**0.538**	**0.378**	**0.411**	**0.155**	**0.056**
SHAP	0.813	0.609	0.450	0.846	0.739	0.637	0.876	0.702	0.543	0.686	0.497	0.344
SHAP_{adv}	0.706	0.434	0.249	0.811	0.647	0.482	0.820	0.611	0.430	0.485	0.238	0.129
SOBOL	0.750	0.591	0.490	0.816	0.653	0.512	0.874	0.703	0.574	0.621	0.417	0.313
SOBOL_{adv}	0.696	0.507	0.383	0.828	0.680	0.555	0.831	0.627	0.506	0.478	0.252	0.152
RISE	0.877	0.766	0.686	0.843	0.710	0.622	0.896	0.809	0.734	0.783	0.637	0.513
RISE_{adv}	0.769	0.679	0.618	0.876	0.824	0.784	0.917	0.867	0.830	0.515	0.378	0.322

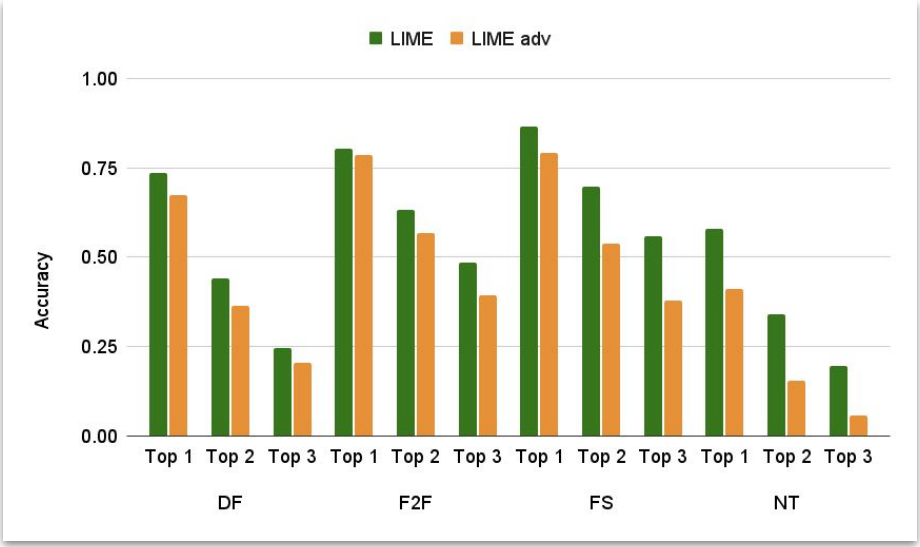
Detection accuracy for the different types of fakes, on the original images and on variants of them after adversarial attacks on image regions corresponding to the top-1, top-2 and top-3 scoring segments by the different XAI methods. Best (lowest) scores in bold and second best scores underlined.

- Pairwise comparisons show that our perturbation approach leads to better scores in most cases

Experimental results - Quantitative analysis

- Modified version of LIME performs consistently better than LIME
- Leads to further drop in detection accuracy; 10.5% on average, up to 18.5% in some cases
- Similar observations can be made for SHAP and SOBOL, while RISE shows some mixed results

"The proposed data perturbation approach has a positive contribution on the performance of most methods"



Detection accuracy for LIME and its modified version; the lower the accuracy, the higher the ability of the method to spot the most important image regions for the detector's output

Experimental results - Quantitative analysis

	DF			F2F			FS			NT		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
LIME	0.195	0.408	0.539	0.121	0.238	0.334	0.087	0.189	0.262	0.233	0.363	0.431
LIME_{adv}	**0.252**	**0.464**	**0.564**	**0.136**	**0.276**	**0.379**	**0.134**	**0.278**	**0.367**	**0.326**	**0.455**	**0.501**
SHAP	0.137	0.300	0.402	0.092	0.158	0.222	0.073	0.181	0.269	0.167	0.282	0.357
SHAP_{adv}	0.215	0.423	0.534	0.112	0.224	0.324	0.114	0.240	0.334	0.280	0.405	0.463
SOBOL	0.166	0.277	0.352	0.108	0.212	0.296	0.078	0.180	0.259	0.198	0.302	0.362
SOBOL_{adv}	0.215	0.349	0.426	0.105	0.201	0.270	0.110	0.221	0.286	0.282	0.393	0.443
RISE	0.087	0.162	0.219	0.091	0.173	0.223	0.060	0.114	0.157	0.115	0.204	0.273
RISE_{adv}	0.155	0.220	0.266	0.066	0.100	0.128	0.043	0.073	0.097	0.244	0.315	0.351

Explanation sufficiency for the different types of fakes, after adversarial attacks on image regions corresponding to the top-1, top-2 and top-3 scoring regions by the XAI methods. Best (highest) scores in bold and second best scores underlined

Experimental results - Quantitative analysis

	DF			F2F			FS			NT		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
LIME	0.195	0.408	0.539	0.121	0.238	0.334	0.087	0.189	0.262	0.233	0.363	0.431
LIME_{adv}	**0.252**	**0.464**	**0.564**	**0.136**	**0.276**	**0.379**	**0.134**	**0.278**	**0.367**	**0.326**	**0.455**	**0.501**
SHAP	0.137	0.300	0.402	0.092	0.158	0.222	0.073	0.181	0.269	0.167	0.282	0.357
SHAP_{adv}	0.215	0.423	0.534	0.112	0.224	0.324	0.114	0.240	0.334	0.280	0.405	0.463
SOBOL	0.166	0.277	0.352	0.108	0.212	0.296	0.078	0.180	0.259	0.198	0.302	0.362
SOBOL_{adv}	0.215	0.349	0.426	0.105	0.201	0.270	0.110	0.221	0.286	0.282	0.393	0.443
RISE	0.087	0.162	0.219	0.091	0.173	0.223	0.060	0.114	0.157	0.115	0.204	0.273
RISE_{adv}	0.155	0.220	0.266	0.066	0.100	0.128	0.043	0.073	0.097	0.244	0.315	0.351

Explanation sufficiency for the different types of fakes, after adversarial attacks on image regions corresponding to the top-1, top-2 and top-3 scoring regions by the XAI methods. Best (highest) scores in bold and second best scores underlined

- Modified LIME exhibits consistently better performance than the other modified methods

Experimental results - Quantitative analysis

	DF			F2F			FS			NT		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
LIME	0.195	0.408	0.539	0.121	0.238	0.334	0.087	0.189	0.262	0.233	0.363	0.431
LIME$_{adv}$	**0.252**	**0.464**	**0.564**	**0.136**	**0.276**	**0.379**	**0.134**	**0.278**	**0.367**	**0.326**	**0.455**	**0.501**
SHAP	0.137	0.300	0.402	0.092	0.158	0.222	0.073	0.181	0.269	0.167	0.282	0.357
SHAP$_{adv}$	0.215	0.423	0.534	0.112	0.224	0.324	0.114	0.240	0.334	0.280	0.405	0.463
SOBOL	0.166	0.277	0.352	0.108	0.212	0.296	0.078	0.180	0.259	0.198	0.302	0.362
SOBOL$_{adv}$	0.215	0.349	0.426	0.105	0.201	0.270	0.110	0.221	0.286	0.282	0.393	0.443
RISE	0.087	0.162	0.219	0.091	0.173	0.223	0.060	0.114	0.157	0.115	0.204	0.273
RISE$_{adv}$	0.155	0.220	0.266	0.066	0.100	0.128	0.043	0.073	0.097	0.244	0.315	0.351

Explanation sufficiency for the different types of fakes, after adversarial attacks on image regions corresponding to the top-1, top-2 and top-3 scoring regions by the XAI methods. Best (highest) scores in bold and second best scores underlined

- Modified LIME exhibits consistently better performance than the other modified methods
- Outperforms its original counterpart in all cases

Experimental results - Quantitative analysis

	DF			F2F			FS			NT		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
LIME	0.195	0.408	0.539	0.121	0.238	0.334	0.087	0.189	0.262	0.233	0.363	0.431
LIME_{adv}	**0.252**	**0.464**	**0.564**	**0.136**	**0.276**	**0.379**	**0.134**	**0.278**	**0.367**	**0.326**	**0.455**	**0.501**
SHAP	0.137	0.300	0.402	0.092	0.158	0.222	0.073	0.181	0.269	0.167	0.282	0.357
SHAP_{adv}	0.215	0.423	0.534	0.112	0.224	0.324	0.114	0.240	0.334	0.280	0.405	0.463
SOBOL	0.166	0.277	0.352	0.108	0.212	0.296	0.078	0.180	0.259	0.198	0.302	0.362
SOBOL_{adv}	0.215	0.349	0.426	0.105	0.201	0.270	0.110	0.221	0.286	0.282	0.393	0.443
RISE	0.087	0.162	0.219	0.091	0.173	0.223	0.060	0.114	0.157	0.115	0.204	0.273
RISE_{adv}	0.155	0.220	0.266	0.066	0.100	0.128	0.043	0.073	0.097	0.244	0.315	0.351

Explanation sufficiency for the different types of fakes, after adversarial attacks on image regions corresponding to the top-1, top-2 and top-3 scoring regions by the XAI methods. Best (highest) scores in bold and second best scores underlined

- Pairwise comparisons of original and modified explanation methods document the mostly positive impact of our perturbation approach in the deepfake explanation performance

Experimental results - Quantitative analysis

- Studied the introduced complexity by the proposed perturbation approach
- Counted the needed time and model inferences for producing an explanation
- Observed (the expected) increase in complexity
- Focusing on the best-performing LIME method, we argue that:

	Computing time per image (in sec.)		Number of inferences per image	
	Original	Modified	Original	Modified
LIME	4.4	20.8	200	2081
SHAP	2.6	18.7	45	1928
SOBOL	0.4	16.9	17	1898
RISE	1.8	18.8	32	1913

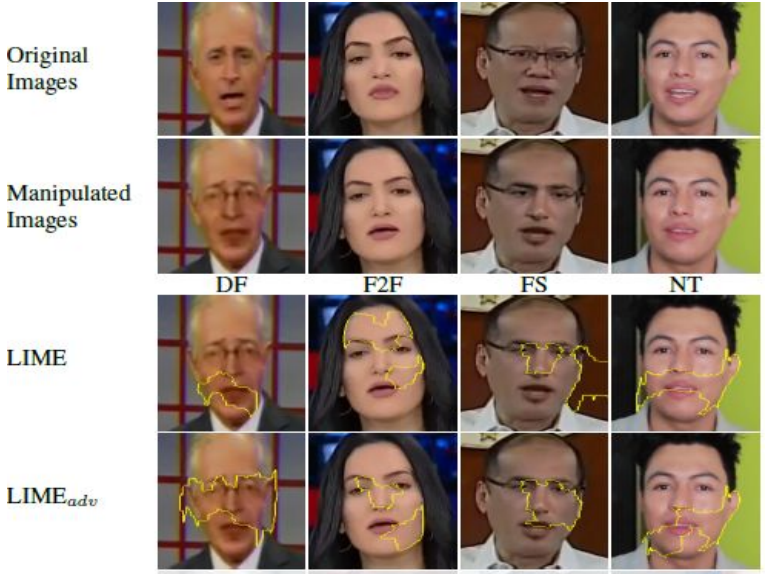
Computational complexity of the original and modified explanation methods

"The introduced computation overhead is balanced by the observed performance gains and does not restrict the use of this method to obtain explanations during real-time deepfake detection tasks"

Experimental results - Qualitative analysis

The modified version of LIME:

- Defines more completely the manipulated area in the case of the DF sample
- Spots more accurately the regions close to the eyes, mouth and chin in the case of the F2F sample
- Demarcates more accurately the modified region in the case of FS
- Puts more focus on the manipulated chin of the NT sample, that was missed by LIME

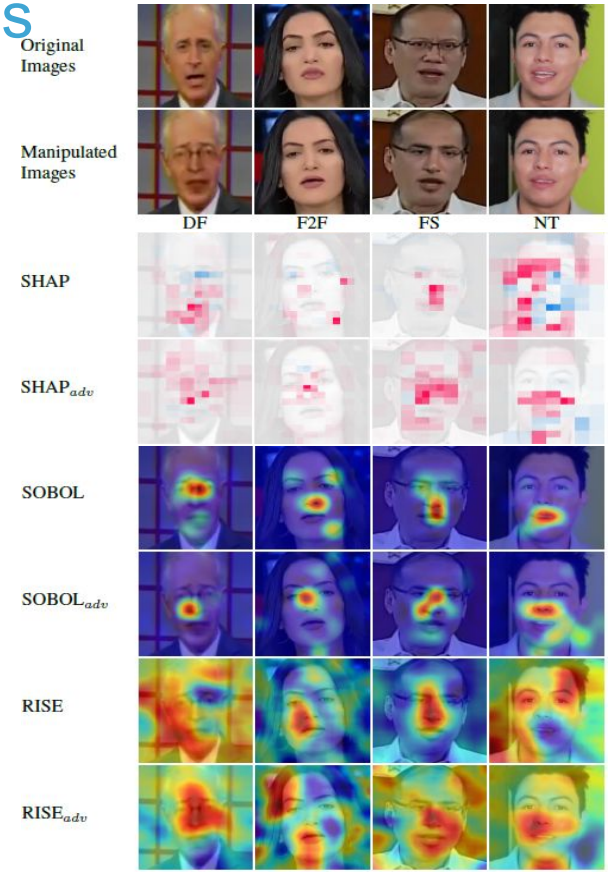


Experimental results - Qualitative analysis

Similar remarks can be made for other methods:

- Modified SHAP appears to produce more complete and better-focusing explanations than SHAP
- Modified SOBOL seems to define more accurately the manipulated regions than SOBOL
- Modified RISE exhibits higher sufficiency in demarcating the manipulated regions than RISE

“Our observations document the improved performance of the modified explanation methods and the positive impact of the proposed perturbation approach”



Part2: Increasing trust in AI methods for combatting disinformation - Conclusions

- Presented our approach to improving the performance of perturbation-based methods when explaining deepfake detectors
- Suggested the use of adversarially-generated samples of the input deepfake images, to form perturbation masks for inferring the importance of input features
- Integrated the proposed approach in four SOTA perturbation-based explanation methods from the literature (LIME, SHAP, SOBOL, RISE)
- Evaluated the performance of the resulting modified methods using a benchmarking dataset (FaceForensics++) and an evaluation protocol
- Documented the positive contribution of the proposed perturbation approach, quantified the gains in the performance of most of these methods, and demonstrated the ability of the modified methods to produce more accurate explanations

Thank you for your attention!

Vasileios Mezaris, bmezaris@iti.gr
with the contribution of Evlampios Apostolidis, apostolid@iti.gr

Code and models available at:
RL-DiVTS (Keyframe Selection): https://github.com/e-apostolidis/RL-DiVTS
Explainable deepfaked detection: https://github.com/IDT-ITI/Adv-XAI-Deepfakes
More materials: https://www.iti.gr/~bmezaris; https://idt.iti.gr/

These works have been funded by the EU as part of the Horizon 2020 and Horizon Europe Framework Programs, under grant agreements 951911 (AI4Media), 101070093 (vera.ai) and 101070190 (AI4TRUST)