Panchromatic Imagery Copy-paste Localization Through Data-driven Sensor Attribution

E.D. Cannas (†), J. Horvàth (*), S. Baireddy (*),
P. Bestagini (†), E.J. Delp (*), S. Tubaro (†)

(†) Image and Sound Processing Lab (ISPL)
Politecnico di Milano, Milan, Italy

(*) Video and Image Processing Laboratory
Purdue University, West Lafayette, IN, USA

Motivation

Overhead images, i.e., images captured by a platform such as an aircraft or satellite, have a strategic role in numerous fields:

1. Land-cover mapping;
2. Earth monitoring;
3. Military and intelligence applications [*].

[*] https://www.space.com/russia-ukraine-invasion-satellite-photos

Motivation

Many online websites offer overhead images for free → easily accessible, but also **easy to forge [*]**

We need forensic instruments to determine the authenticity of overhead images

[*] https://www.bellingcat.com/news/2014/11/14/russian-state-television-shares-fake-images-of-mh17-being-attacked/

Goal

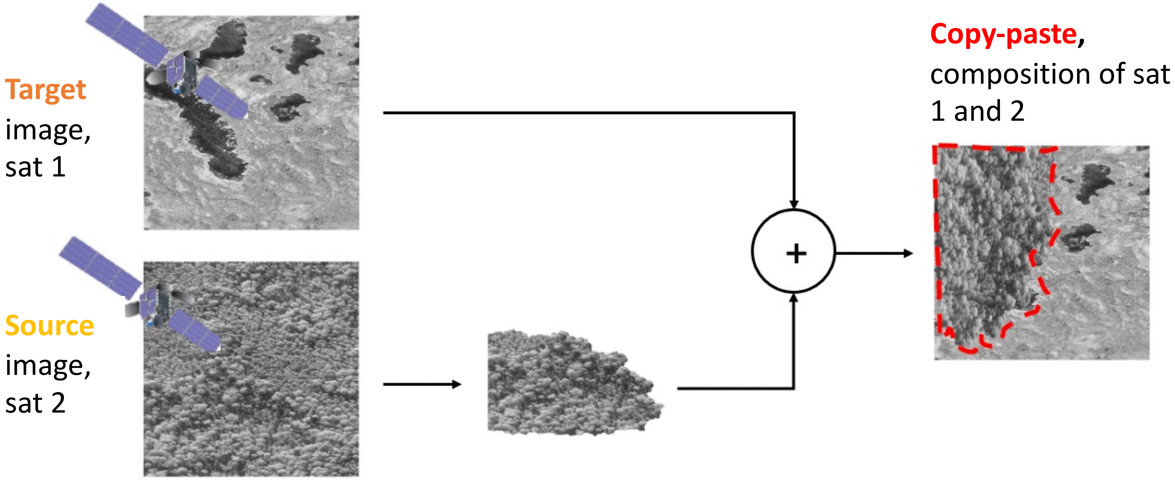
Given a panchromatic sample, we want to localize doctored pixel regions coming from a image generated by a different satellite → **copy-paste attacks**

RGB image

VS

Panchromatic image

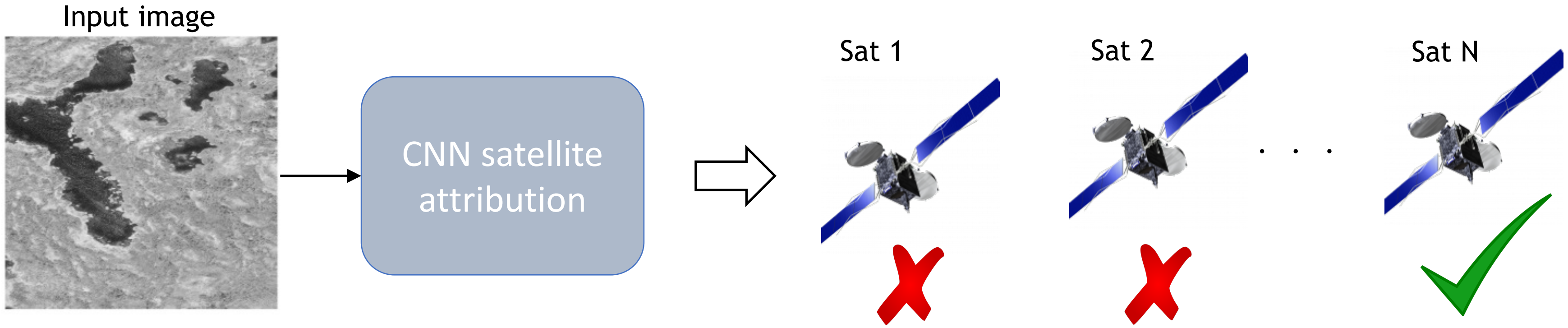
Copy-paste attack



Methodology

We leverage **sensor attribution traces** extracted by an ensemble of **Convolutional Neural Networks (CNNs):**

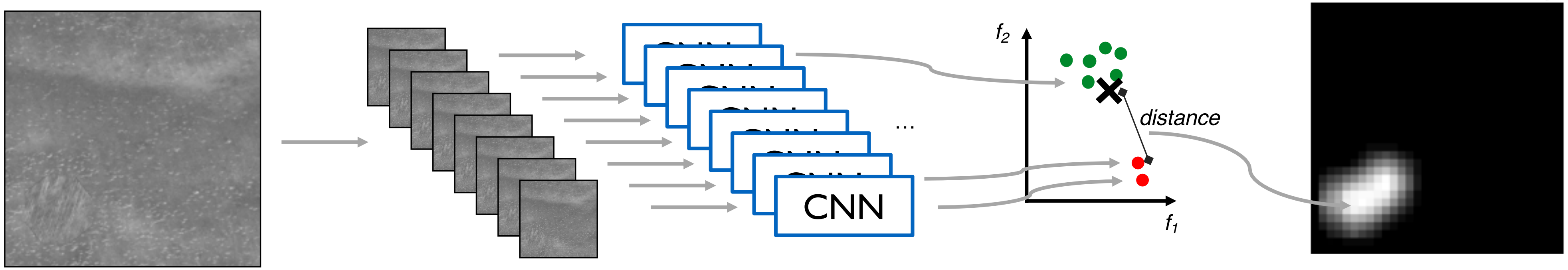
1. We purposely **train** a CNN to predict the satellite that generated a panchromatic image;
2. At **test time**, we localize copy-paste attacks as inconsistencies in the attribution traces, i.e., pixel regions coming from a satellite different from the rest of the image, in a **patch-wise manner.**



Methodology

We leverage **sensor attribution traces** extracted by an ensemble of **Convolutional Neural Networks (CNNs):**

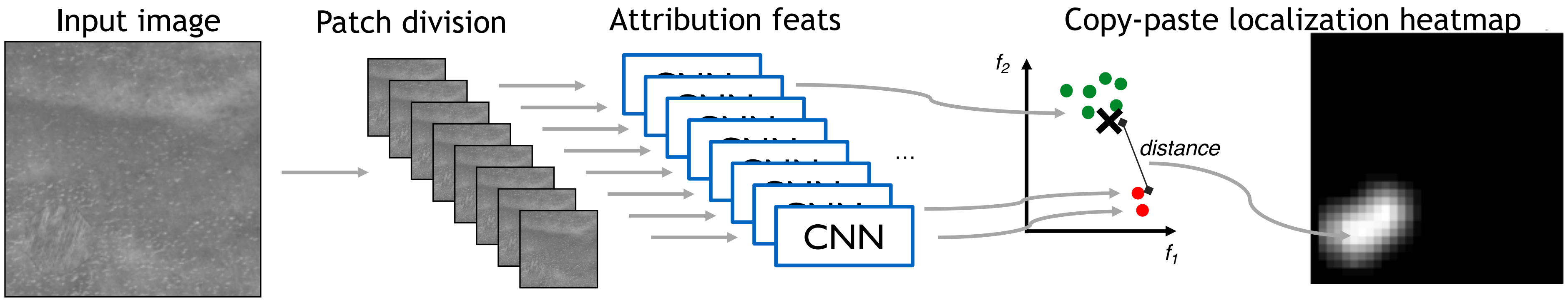
1. We purposely **train** a CNN to predict the satellite that generated a panchromatic image;
2. At **test time,** we localize copy-paste attacks as inconsistencies in the attribution traces, i.e., pixel regions coming from a satellite different from the rest of the image, in a **patch-wise manner.**



Methodology

Given an image under analysis:

1. **Split it into patches** (128x128 resolution, 32x32 stride);
2. Extract a feature vector $\boldsymbol{y_i}$ **per patch** with the CNN;
3. Compute the **average feature vector** $\hat{\boldsymbol{y}} = \sum_{i=1}^{N} \boldsymbol{y_i}/\boldsymbol{N}$;
4. Compute **distance** $\boldsymbol{d_i} = |\boldsymbol{y_i} - \hat{\boldsymbol{y}}|$ between each feature vector and the average one;
5. Attribute the distance value to each patch as **copy-paste heatmap.**



Experimental setup

Dataset:

1. **8-bit panchromatic** images from Maxar technologies;
2. **5** satellites (GE01, QB02, WV01-2-3), **5** geographical regions (barren, field, forest, snow, urban);
3. **1024x1024** patches extracted from the images, for a total of 100’000 samples divided into:
    1. **Training set:** 50’000 images for training the satellite attribution CNN;
    2. **Test set:** 50’000 images further elaborated to create **copy-paste** attacks.

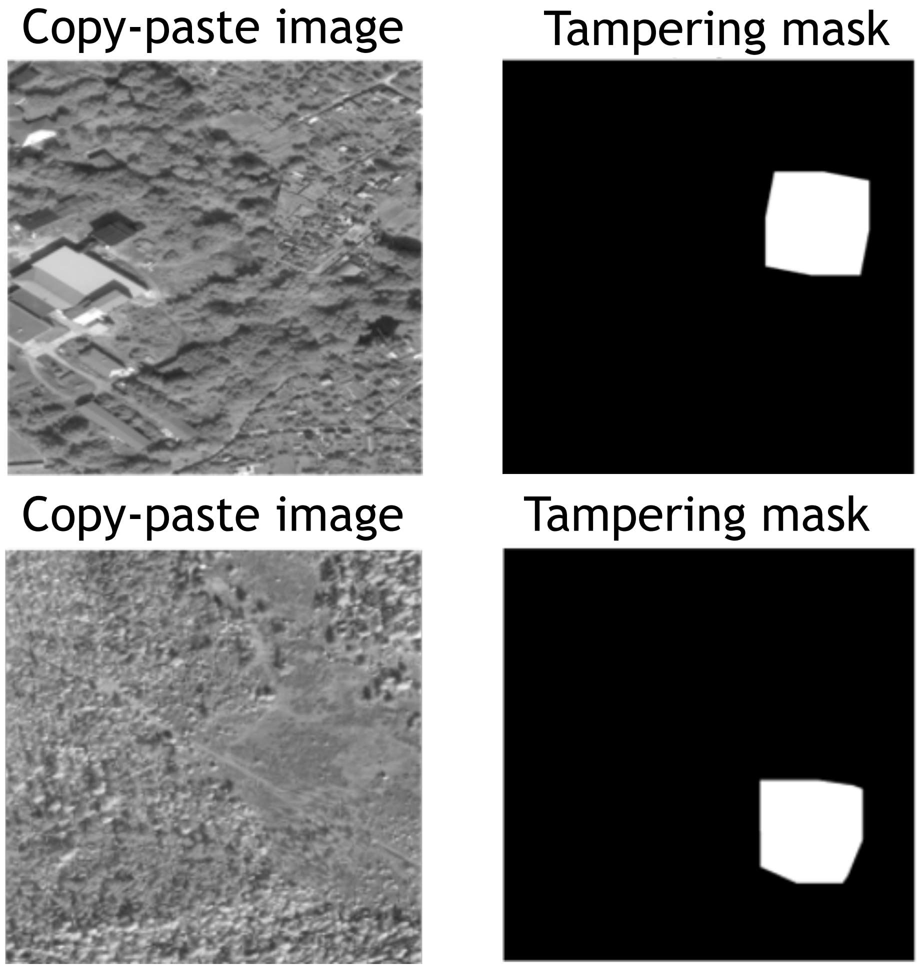
Experimental setup

Test set → 50'000 images further elaborated to create **copy-paste attacks**

Semantically coherent copy-paste attacks (i.e., same geographical regions) with **source** and **target** sample coming from **different satellites**

Different editing operations applied to make the attack more plausible (e.g., blurring, resizing, affine transforms, etc.)

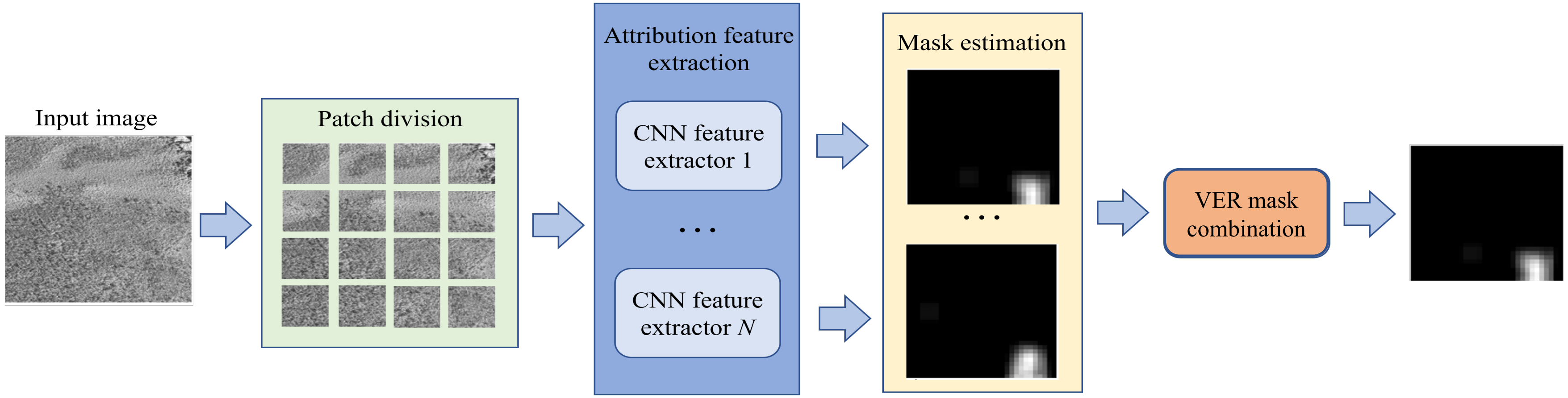
Tampering mask highlights the region under attack



Experimental setup

Satellite attribution CNN:

1. **EfficientNetB0** trained as a M-class satellite classifier;
2. **Model ensembling:**
   1. **Trained 5 different** networks on subset of the available satellites;
   2. **Combine the responses** from each element of the ensemble using the **Variance to Entropy Ratio (VER).**



Results (1)

Comparison with State-Of-The-Art (SOTA):

1. **Noiseprint [*]** and **Splicebuster [**]** as baselines;
2. **Test set comprehending:**
   1. 50 **copy-paste** per **source satellite** and **geographical region**;
   2. 3 types of editing (i.e., Gaussian blurring, resizing, affine transform);
   3. **3750** total test samples, with **fixed tampered area** of **256x256** pixels.
3. Localization results computed as Receiving Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values from the generated copy-paste heatmaps against the binary tampering masks.

[*] D. Cozzolino and L. Verdoliva, ''Noiseprint: A CNN-based camera model fingerprint'', IEEE TIFS 2020
[**] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector", IEEE WIFS 2015

Results (1)

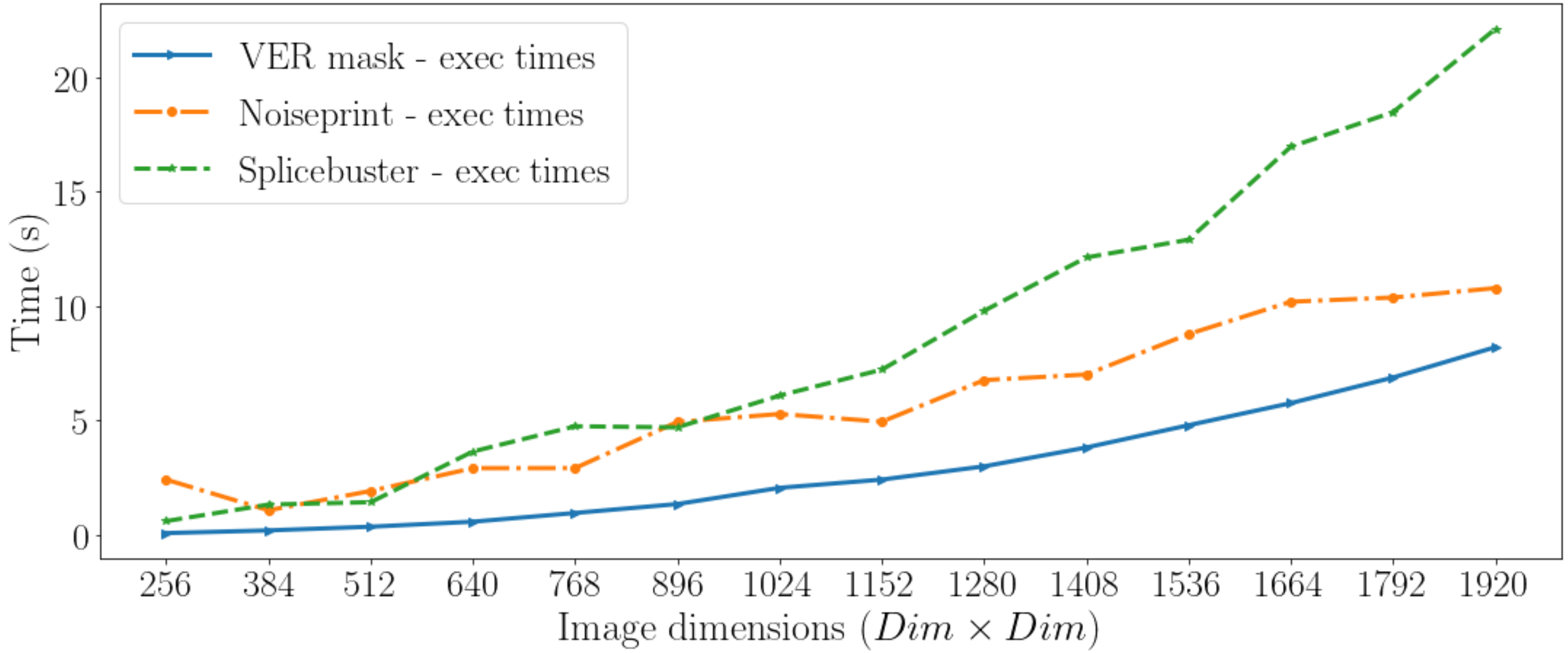
Comparison with State-Of-The-Art (SOTA)

Region	Method	AUC
	Noiseprint	0.949
Barren	Splicebuster	0.976
	VER mask (ours)	0.977
	Noiseprint	0.961
Field	Splicebuster	0.976
	VER mask (ours)	0.977
	Noiseprint	0.960
Forest	Splicebuster	0.974
	VER mask (ours)	0.978
	Noiseprint	0.930
Snow	Splicebuster	0.950
	VER mask (ours)	0.952
	Noiseprint	0.928
Urban	**Splicebuster**	**0.970**
	VER mask (ours)	0.963

Satellite	Method	AUC
	Noiseprint	0.939
GE01	Splicebuster	0.969
	VER mask (ours)	0.991
	Noiseprint	0.952
QB02	Splicebuster	0.970
	VER mask (ours)	0.978
	Noiseprint	0.948
WV01	Splicebuster	0.965
	VER mask (ours)	0.981
	Noiseprint	0.944
WV02	**Splicebuster**	**0.964**
	VER mask (ours)	0.927
	Noiseprint	0.946
WV03	**Splicebuster**	**0.978**
	VER mask (ours)	0.974

Results (1)

Comparison with State-Of-The-Art (SOTA)



Results (2)

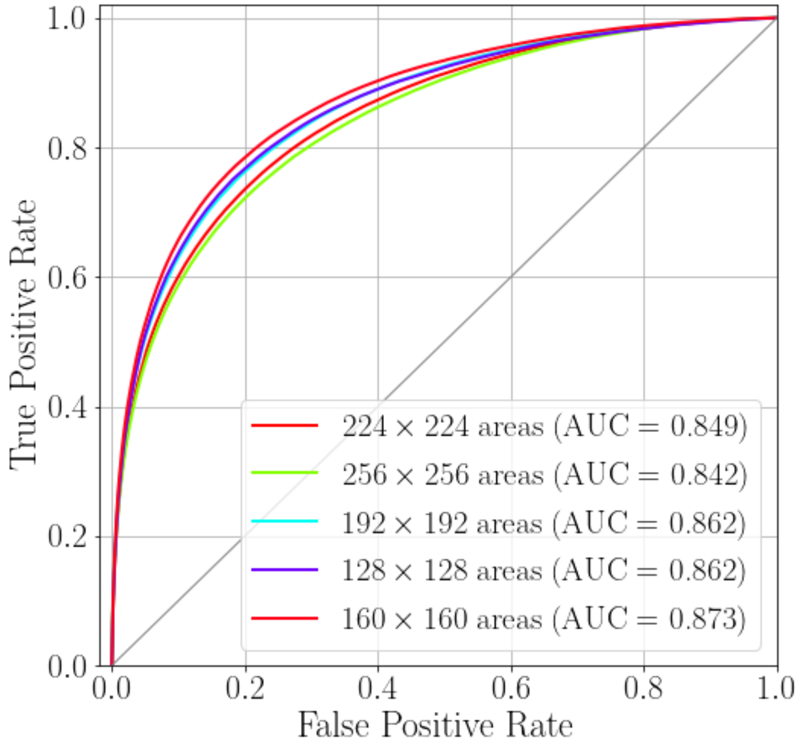
"In the wild" dataset

1. More post-processing operations:
   1. 4 different types of blurring (i.e., Gaussian blur, motion blur, median blur, and average blur);
   2. 3 types of affine transforms (i.e., random rotation and resize, piece-wise affine transform and a prospective transform);
   3. 2 types of contrast enhancement (i.e, logarithmic and sigmoid contrast).
2. Different resolutions of the tampered area (from 128x128 to 256x256);
3. **22'500 samples in total** equally distributed across source satellite and geographical region.

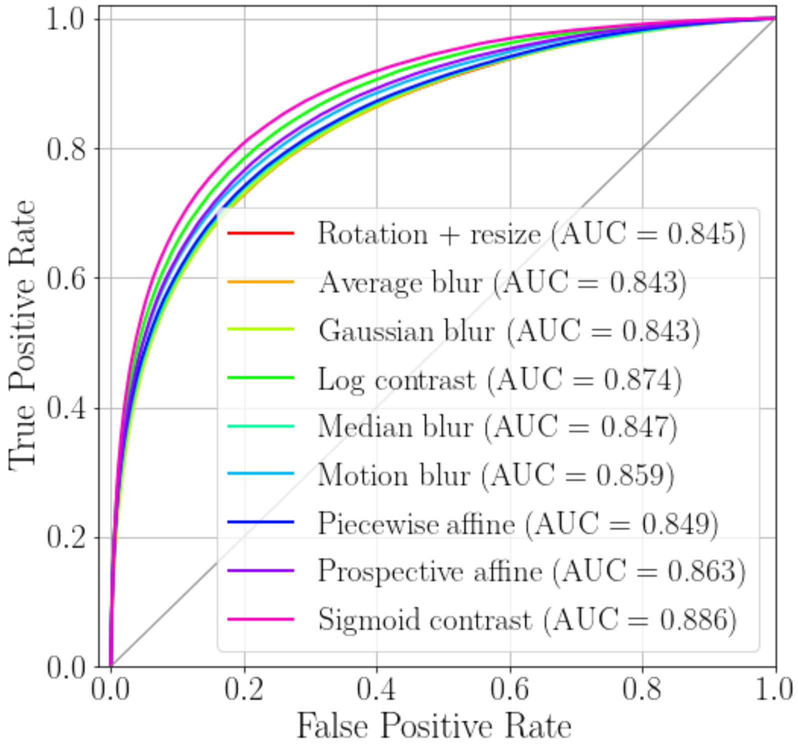
Results (2)

“In the wild” dataset

Results divided by tampered area resolution



Results divided by editing operation applied



Conclusions

We investigated the problem of copy-paste localization in the context of panchromatic imagery

Our solution does not require training on copy-paste samples and is considerably faster than SOTA techniques for natural imagery

Future works will be devoted to the integration of open-set recognition and uncertainty estimation techniques for handling the case where both the source and target satellites are unknown by the networks ensemble

Panchromatic Imagery Copy-paste Localization Through Data-driven Sensor Attribution

E.D. Cannas (†), J. Horvàth (*), S. Baireddy (*),
Paolo Bestagini (†), E.J. Delp (*), Stefano Tubaro (†)

(†) Image and Sound Processing Lab (ISPL)
Politecnico di Milano, Milan, Italy

ISPL
Image and Sound Processing Lab

VIPER
Video and Image Processing Laboratory

(*) Video and Image Processing Laboratory
Purdue University, West Lafayette, IN SA