Non-Semantic Evaluation of Image Forensics Tools: Methodology and Database

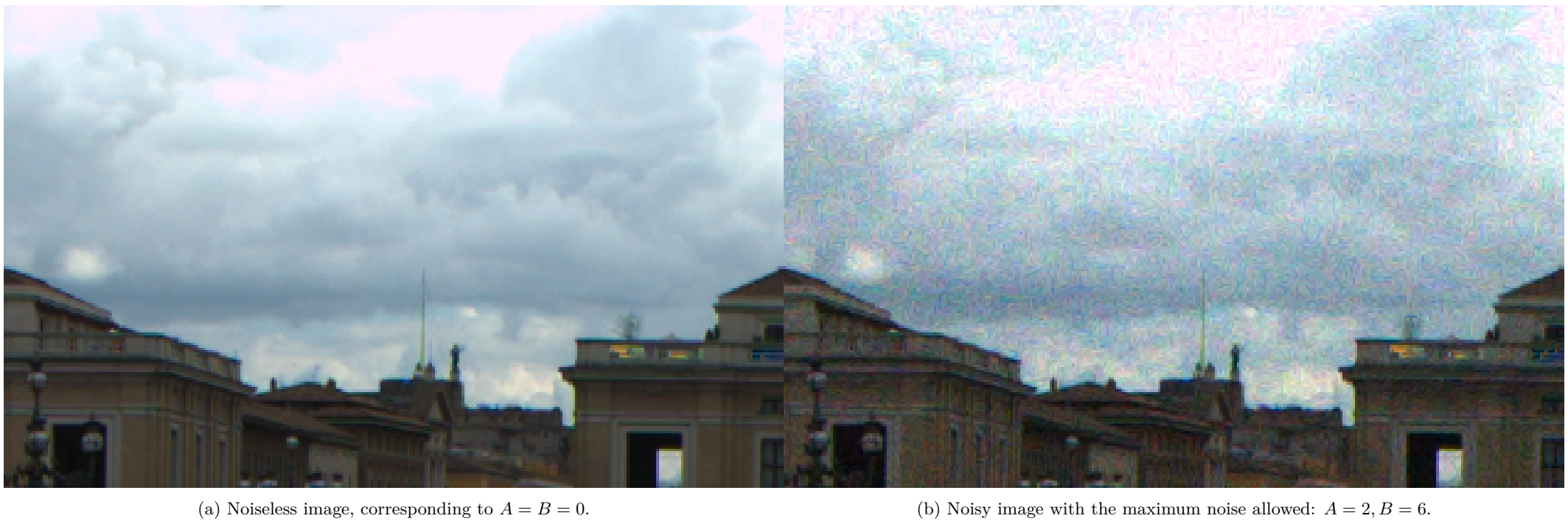
Quentin Bammey, Tina Nikoukhah, Marina Gardella,
Rafael Grompone von Gioi, Miguel Colom, Jean-Michel Morel
Centre Borelli – École Normale Supérieure Paris-Saclay – Université Paris-Saclay

Supplementary Materials

In these supplementary materials, we provide additional details on the creation of the database, the distribution of sizes of the forgery masks, show the scores of the different methods on our datasets with other metrics, and details on the scoring.

1 Dataset creation

Raw Noise Level dataset In this dataset, the added Poisson noise's variance follows a linear relation given by $\sigma^2 = A + Bu$, where A and B are constants and u is the noiseless image's intensity. As explained in Section 3 of the main article, this is a model of the camera noise at image acquisition. We randomly select two pairs of parameters A and B, one for the authentic region and another for the forged region. A is chosen between 0 and 2, and B between 0 and 6. These parameters are typical of cameras, and the maximum noise allowed already corresponds to a highly unfavourable scenario, as seen in Figure 1



(a) Noiseless image, corresponding to $A = B = 0$.

(b) Noisy image with the maximum noise allowed: $A = 2, B = 6$.

Figure 1: An image from our database (cropped) in the two extreme scenarios of the model: Figure 1a corresponds to a noiseless image, whereas Figure 1b has the maximum allowed noise.

JPEG quality and JPEG grid datasets In these two datasets, the image is JPEG compressed with a quality chosen randomly between 75 and 100, with 100 being the highest possible quality for JPEG compression. At a quality factor of 75, the image is already strongly degraded visually, and blocking artefacts from the compression can even be seen to the naked eye, as seen in Fig 2. Even the most basic JPEG analysis methods are easily able to detect and analyse the JPEG compression at such a low quality factor. There is no need for even more compressed images in our database, since they would be even easier to detect.

Figure 2: Compressed with a quality factor of 75, JPEG blocking artefacts are already visible to the naked eye, and can easily be detected even by the most basic JPEG analysis methods. There is thus no need to feature even more compresed images in the dataset, since such forgeries would be trivial to detect.

Overall, parameter bounds for these datasets were chosen so that the resulting images would be typical of what can be expected from cameras, and varied enough to feature both difficult-to-analyze images (a very low noise or a mild JPEG compression are harder to detect than strong noise or compression).

Hybrid dataset To create this dataset, we adopt the following procedure for each image:

1. We randomly choose whether to modify two or three steps of the pipeline (added noise, demosaicing grid/method, JPEG grid/quality). If we only change two, we select which steps to change.
2. For JPEG and CFA modifications, we select whether we only change the CFA and JPEG grids, or if we change the demosaicing methods, the JPEG quality factor and potentially the CFA and JPEG grids. The decision is made jointly for JPEG and CFA, as the CFA and JPEG Grid datasets mimic artefacts commonly found in internal copy-move forgeries, whereas the CFA Algorithm and JPEG Quality datasets represent inconsistencies more typical of splicing.

3. Finally, for each different change, we select its parameters in the same way as for the specific datasets.

In short, each image has a minimum of two parameters that vary between the two parts. At the maximum, all studied parameters in this article (raw noise level, demosaicing pattern and algorithm, JPEG grid and quality factor) can vary between the two regions.

2 Mask size distribution

The distribution of forgery masks' sizes can be seen in Figure 3.

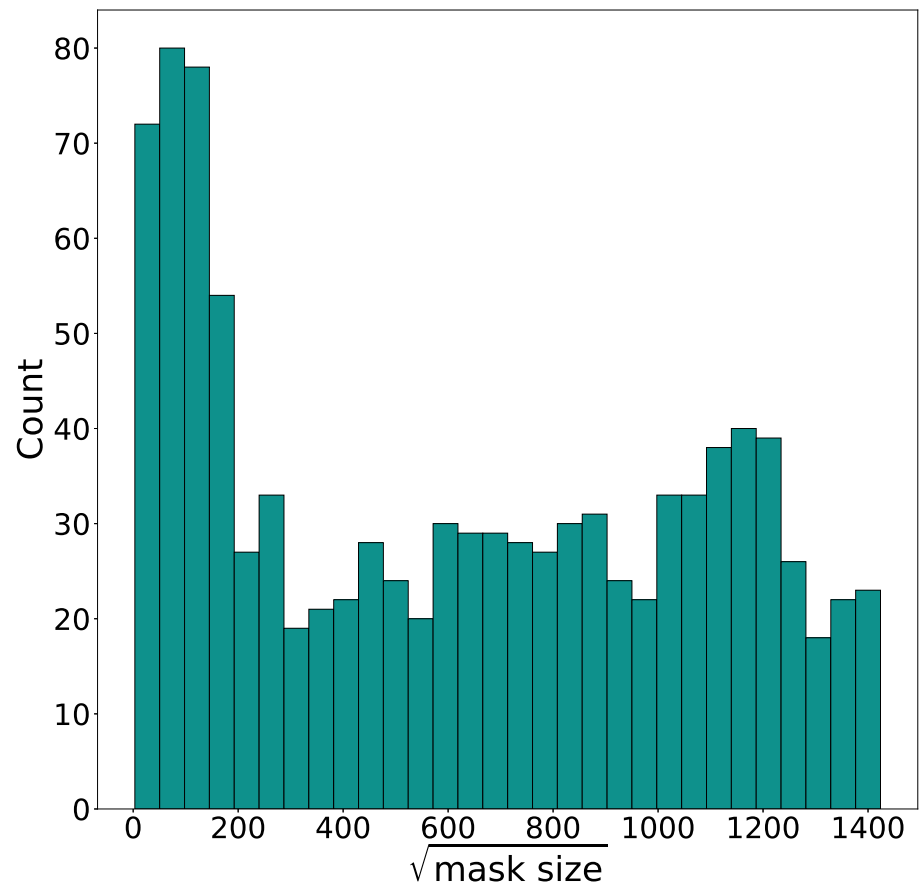


Figure 3: Distribution of the mask sizes. We show the square root of the size, which represents the side of its equivalent square. For comparison, the images in the dataset are all either 2155×1434 or 2474×1640 in size.

3 Scores

In the main article, we only showed the results of the Matthews correlation coefficient (MCC). In these supplementary materials, we also provide results with the Intersection over Union (IoU) in Table 1 and the F_1 score in Table 2

For a given image with a binary detection output, if TP, FP, FN, TN respectively represent the number of true positives, false positives, false negatives and true negatives, the intersection over union is defined as the size of the intersection of the ground truth mask and the detection mask divided by the size of their union, in other words:

$$IoU = \frac{TP}{TP + FN + FP}$$

The F_1 score is the harmonic mean of the precision and recall, and can be written with the confusion matrix as

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FN + FP)}$$

Finally, the Matthews correlation coefficient is defined as

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

As most forgery detection methods provide not a heatmap, but an output, we instead use a weighted confusion matrix. If a pixel is detected as forged with a probability $0 < p < 1$, then:

- If this pixel is truly forged, it counts towards true positives with a weight of p, and towards false negatives with a weight of $1 - p$,
- If this pixel is authentic, it counts towards false positives with a weight of p, and towards true negatives with a weight of $1 - p$.

In short, if $\mathcal{M}$ is the forgery mask and Y the continuous output, with values between 0 and 1:

- $TP = \sum Y \odot \mathcal{M}$
- $FP = \sum Y \odot \bar{\mathcal{M}}$
- $FN = \sum (1 - Y) \odot \mathcal{M}$
- $TN = \sum (1 - Y) \odot \bar{\mathcal{M}}$

where $\odot$ represents pointwise product.

In the case of the Lyu and Mahdian algorithms, a specific transformation is needed to score their results. Those methods do not act directly as detectors, but rather locally estimate and output the noise level. To turn their outputs into a heatmap detection, we compute and normalize the distance of the output to its median; if N is the noise level estimation and M its median, the heatmap is defined as

$$Y = \frac{|Y - N|}{\max(N, 1 - N)}.$$

		Dataset					
		Noise Level	CFA Grid	CFA Algorithm	JPEG Grid	JPEG Quality	Hybrid
Noise-based	Noisesniffer	0.093 (0.150) 0.067 (0.123)	0.010 (0.029) 0.010 (0.032)	0.036 (0.101) 0.020 (0.060)	0.013 (0.034) 0.013 (0.038)	0.056 (0.113) 0.033 (0.082)	0.076 (0.130) 0.053 (0.110)
	Lyu	0.065 (0.078) 0.065 (0.090)	0.072 (0.088) 0.078 (0.110)	0.072 (0.089) 0.078 (0.110)	0.078 (0.091) 0.088 (0.110)	0.079 (0.093) 0.086 (0.114)	0.075 (0.088) 0.079 (0.105)
	Mahdian	0.056 (0.086) 0.064 (0.110)	0.069 (0.087) 0.081 (0.118)	0.089 (0.104) 0.101 (0.135)	0.088 (0.095) 0.097 (0.113)	0.106 (0.113) 0.116 (0.136)	0.097 (0.116) 0.106 (0.139)
CFA-based	Bammey	0.020 (0.057) 0.032 (0.107)	0.617 (0.318) 0.603 (0.327)	0.456 (0.373) 0.446 (0.374)	0.119 (0.111) 0.117 (0.114)	0.121 (0.112) 0.118 (0.111)	0.196 (0.243) 0.193 (0.243)
	Shin	0.081 (0.086) 0.081 (0.090)	0.152 (0.144) 0.149 (0.143)	0.142 (0.142) 0.140 (0.142)	0.098 (0.089) 0.096 (0.089)	0.098 (0.089) 0.096 (0.088)	0.107 (0.105) 0.105 (0.103)
	Choi	0.032 (0.014) 0.036 (0.010)	0.573 (0.155) 0.548 (0.149)	0.393 (0.155) 0.362 (0.156)	0.004 (0.001) 0.004 (0.001)	0.003 (0.001) 0.005 (0.000)	0.152 (0.085) 0.140 (0.086)
JPEG-based	Zero	0.000 (0.000) 0.000 (0.000)	0.000 (0.000) 0.000 (0.000)	0.000 (0.000) 0.000 (0.000)	0.764 (0.311) 0.727 (0.350)	0.709 (0.355) 0.683 (0.378)	0.616 (0.418) 0.600 (0.426)
	CAGI	0.109 (0.102) 0.108 (0.101)	0.109 (0.102) 0.108 (0.102)	0.109 (0.102) 0.110 (0.104)	0.114 (0.108) 0.104 (0.103)	0.116 (0.110) 0.107 (0.104)	0.111 (0.105) 0.102 (0.100)
	FDF-A	0.118 (0.123) 0.110 (0.121)	0.099 (0.100) 0.096 (0.109)	0.100 (0.100) 0.096 (0.107)	0.191 (0.174) 0.188 (0.186)	0.194 (0.175) 0.191 (0.191)	0.190 (0.173) 0.183 (0.186)
	I-CDA	0.117 (0.110) 0.117 (0.110)	0.114 (0.109) 0.114 (0.109)	0.114 (0.109) 0.114 (0.109)	0.387 (0.336) 0.391 (0.339)	0.389 (0.331) 0.385 (0.333)	0.366 (0.335) 0.371 (0.339)
	CDA	0.114 (0.108) 0.110 (0.108)	0.112 (0.108) 0.112 (0.115)	0.112 (0.108) 0.111 (0.113)	0.423 (0.314) 0.392 (0.318)	0.417 (0.318) 0.389 (0.320)	0.370 (0.319) 0.348 (0.312)
	BAG	0.004 (0.005) 0.005 (0.013)	0.016 (0.055) 0.033 (0.131)	0.016 (0.055) 0.032 (0.131)	0.257 (0.319) 0.252 (0.321)	0.256 (0.315) 0.250 (0.317)	0.200 (0.305) 0.195 (0.303)
Multi-purpose tools	Noiseprint	0.103 (0.114) 0.103 (0.130)	0.037 (0.044) 0.041 (0.060)	0.071 (0.089) 0.073 (0.101)	0.059 (0.065) 0.063 (0.083)	0.157 (0.160) 0.143 (0.164)	0.138 (0.145) 0.133 (0.153)
	ManTraNet	0.047 (0.053) 0.042 (0.051)	0.027 (0.029) 0.027 (0.033)	0.071 (0.106) 0.060 (0.100)	0.022 (0.018) 0.022 (0.019)	0.076 (0.107) 0.073 (0.112)	0.080 (0.112) 0.079 (0.114)
	Self-Consistency	0.161 (0.231) 0.216 (0.314)	0.121 (0.178) 0.164 (0.270)	0.128 (0.188) 0.170 (0.277)	0.115 (0.174) 0.153 (0.255)	0.160 (0.238) 0.218 (0.309)	0.198 (0.275) 0.270 (0.334)
	Splicebuster	0.082 (0.112) 0.091 (0.137)	0.035 (0.052) 0.046 (0.099)	0.074 (0.098) 0.080 (0.130)	0.038 (0.052) 0.045 (0.080)	0.083 (0.116) 0.091 (0.144)	0.086 (0.114) 0.094 (0.137)

Table 1: Intersection over Union (IoU) scores. The mean of the IoU scores over each image of the dataset, as well as the standard deviation in parentheses, are shown for the exogenous mask and endogenous mask datasets. Grayed-out numbers represent results of methods on datasets that are irrelevant to said methods.

		Dataset					
		Noise Level	CFA Grid	CFA Algorithm	JPEG Grid	JPEG Quality	Hybrid
Noise-level-based	Noisesniffer	0.142 (0.210) 0.106 (0.177)	0.019 (0.052) 0.018 (0.054)	0.056 (0.137) 0.034 (0.094)	0.024 (0.058) 0.024 (0.063)	0.089 (0.162) 0.055 (0.122)	0.118 (0.185) 0.084 (0.159)
	Lyu	0.114 (0.125) 0.111 (0.136)	0.123 (0.139) 0.129 (0.160)	0.123 (0.139) 0.128 (0.160)	0.132 (0.143) 0.145 (0.161)	0.134 (0.144) 0.142 (0.165)	0.129 (0.137) 0.132 (0.155)
	Mahdian	0.096 (0.135) 0.105 (0.159)	0.118 (0.136) 0.132 (0.166)	0.148 (0.158) 0.161 (0.187)	0.149 (0.149) 0.160 (0.164)	0.175 (0.169) 0.185 (0.190)	0.159 (0.173) 0.169 (0.193)
CFA-based	Bammey	0.035 (0.091) 0.048 (0.141)	0.702 (0.316) 0.687 (0.326)	0.526 (0.397) 0.515 (0.397)	0.196 (0.172) 0.192 (0.174)	0.199 (0.173) 0.194 (0.170)	0.273 (0.275) 0.269 (0.275)
	Shin	0.139 (0.139) 0.138 (0.142)	0.239 (0.207) 0.234 (0.205)	0.223 (0.203) 0.221 (0.203)	0.166 (0.144) 0.164 (0.145)	0.167 (0.145) 0.163 (0.144)	0.179 (0.161) 0.176 (0.159)
	Choi	0.049 (0.0.025) 0.052 (0.020)	0.630 (0.172) 0.603 (0.165)	0.444 (0.183) 0.408 (0.184)	0.008 (0.002) 0.007 (0.002)	0.006 (0.002) 0.008 (0.001)	0.180 (0.105) 0.163 (0.109)
JPEG-based	Zero	0.000 (0.000) 0.000 (0.000)	0.000 (0.000) 0.000 (0.000)	0.000 (0.000) 0.000 (0.000)	0.811 (0.315) 0.770 (0.358)	0.754 (0.364) 0.725 (0.389)	0.650 (0.434) 0.632 (0.442)
	CAGI	0.181 (0.161) 0.180 (0.160)	0.181 (0.161) 0.181 (0.161)	0.182 (0.162) 0.183 (0.163)	0.188 (0.167) 0.174 (0.160)	0.191 (0.169) 0.178 (0.161)	0.184 (0.164) 0.172 (0.157)
	FDF-A	0.190 (0.183) 0.178 (0.179)	0.167 (0.156) 0.159 (0.167)	0.168 (0.156) 0.159 (0.165)	0.287 (0.231) 0.279 (0.244)	0.291 (0.235) 0.280 (0.251)	0.286 (0.231) 0.271 (0.242)
	I-CDA	0.192 (0.172) 0.192 (0.172)	0.187 (0.171) 0.187 (0.171)	0.187 (0.170) 0.188 (0.170)	0.472 (0.357) 0.475 (0.358)	0.477 (0.353) 0.472 (0.354)	0.451 (0.351) 0.455 (0.353)
	CDA	0.188 (0.169) 0.183 (0.168)	0.186 (0.168) 0.184 (0.174)	0.186 (0.168) 0.183 (0.172)	0.521 (0.336) 0.486 (0.339)	0.513 (0.340) 0.483 (0.341)	0.461 (0.340) 0.440 (0.334)
	BAG	0.008 (0.009) 0.009 (0.021)	0.026 (0.083) 0.043 (0.155)	0.027 (0.082) 0.043 (0.155)	0.319 (0.359) 0.312 (0.359)	0.319 (0.357) 0.311 (0.356)	0.247 (0.344) 0.240 (0.341)
Multi-purpose tools	Noiseprint	0.169 (0.170) 0.165 (0.188)	0.068 (0.076) 0.073 (0.100)	0.121 (0.137) 0.122 (0.153)	0.104 (0.108) 0.108 (0.130)	0.241 (0.221) 0.219 (0.227)	0.216 (0.205) 0.206 (0.215)
	ManTraNet	0.086 (0.086) 0.076 (0.084)	0.051 (0.052) 0.050 (0.058)	0.117 (0.149) 0.101 (0.140)	0.043 (0.034) 0.041 (0.036)	0.126 (0.149) 0.120 (0.155)	0.134 (0.149) 0.130 (0.153)
	Self-Consistency	0.221 (0.284) 0.266 (0.349)	0.179 (0.232) 0.211 (0.308)	0.187 (0.242) 0.217 (0.315)	0.171 (0.228) 0.201 (0.296)	0.218 (0.286) 0.270 (0.349)	0.259 (0.320) 0.328 (0.371)
	Splicebuster	0.136 (0.164) 0.143 (0.187)	0.064 (0.087) 0.076 (0.134)	0.124 (0.146) 0.127 (0.176)	0.069 (0.088) 0.077 (0.114)	0.135 (0.165) 0.141 (0.194)	0.142 (0.166) 0.149 (0.186)

Table 2: F_1 scores. The mean of the F_1 scores over each image of the dataset, as well as the standard deviation in parentheses, are shown for the exogenous mask and endogenous mask datasets. Grayed-out numbers represent results of methods on datasets that are irrelevant to said methods.

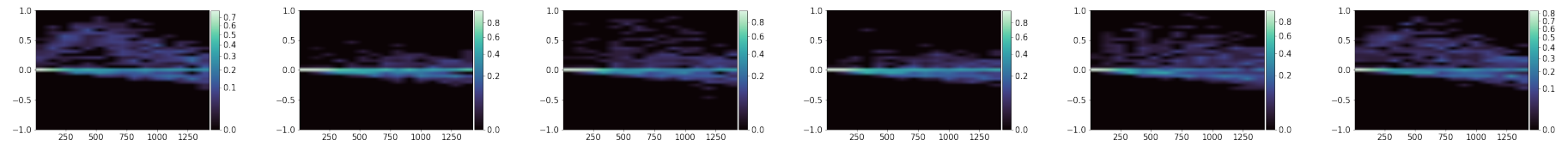
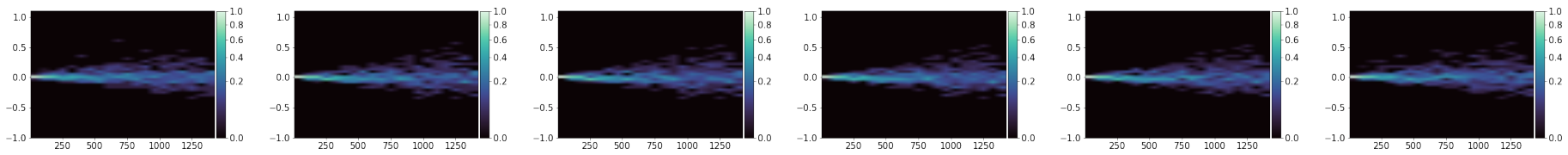
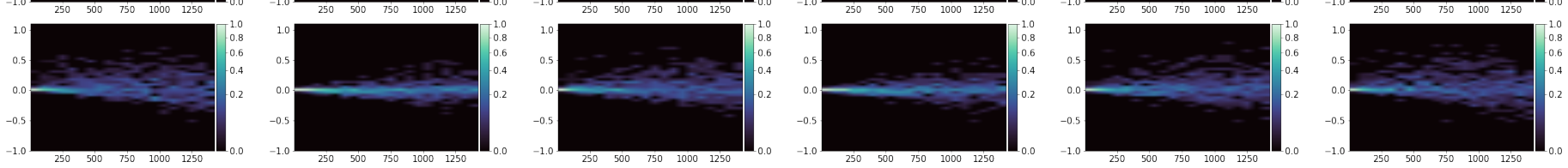
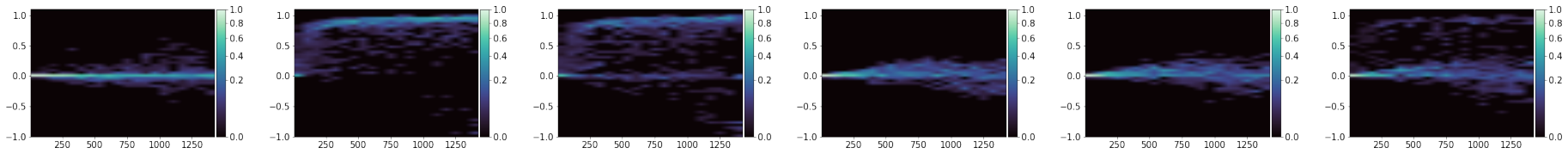
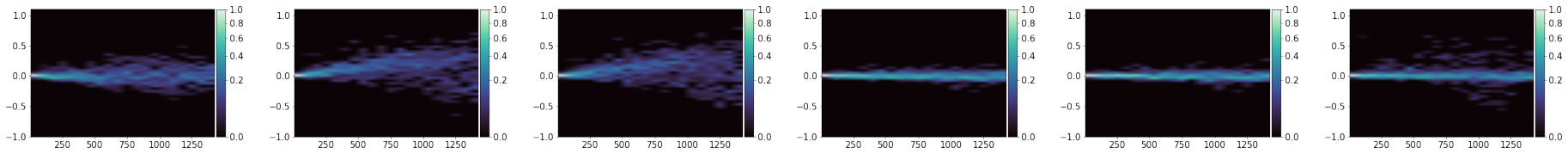
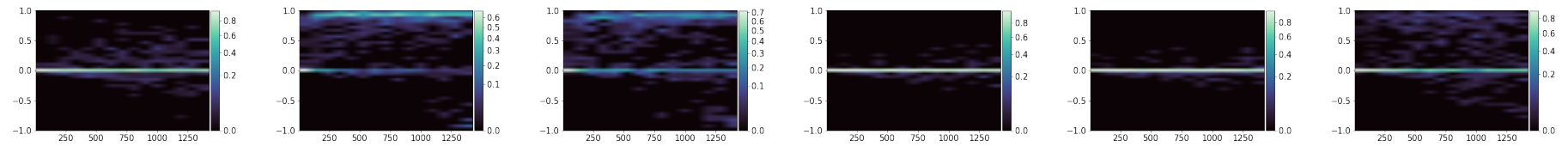
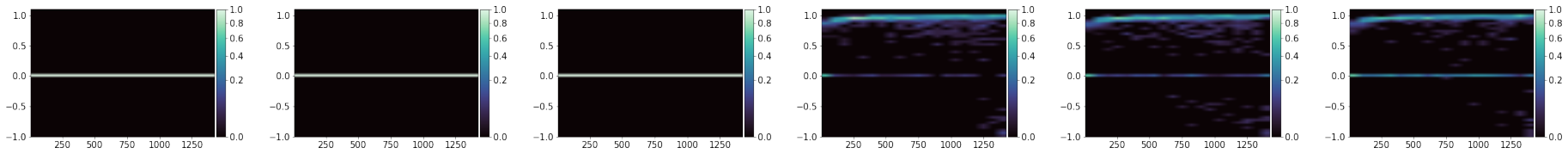
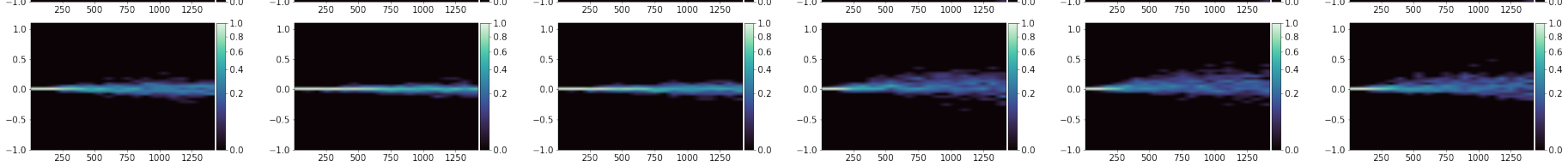
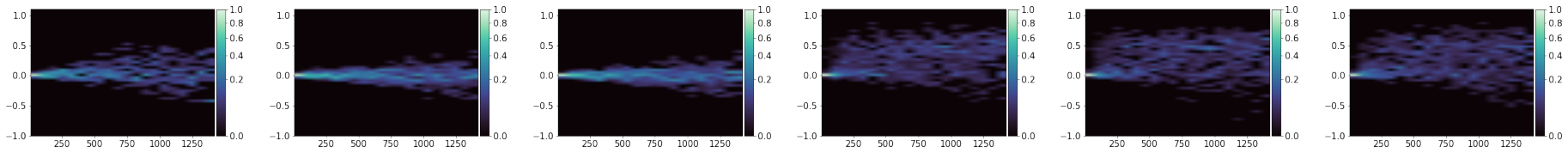
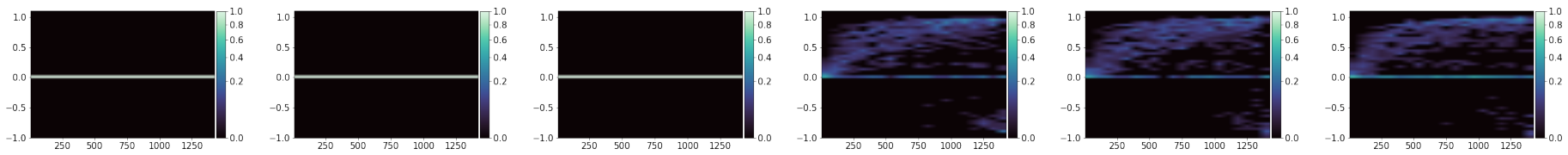
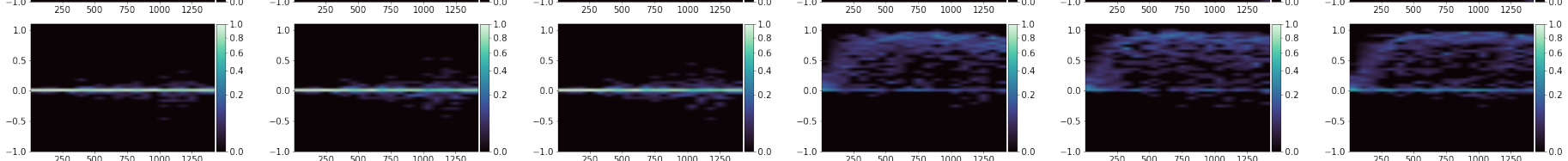
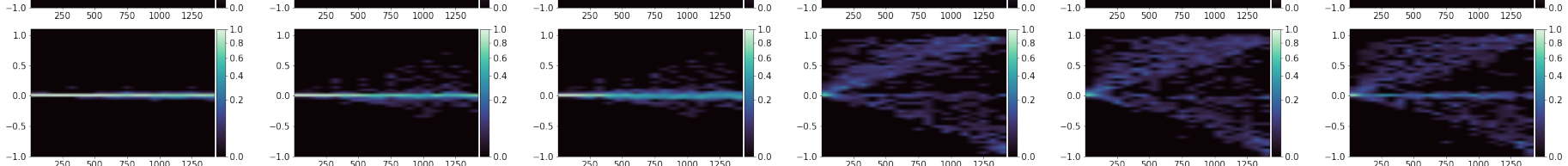
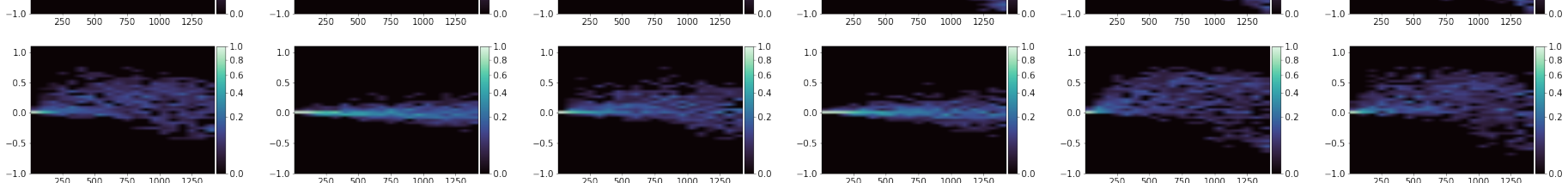
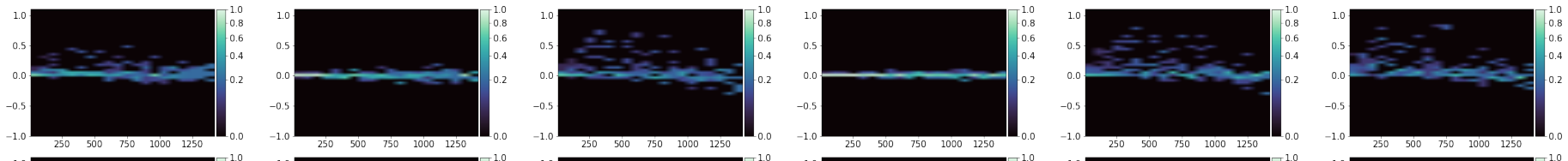
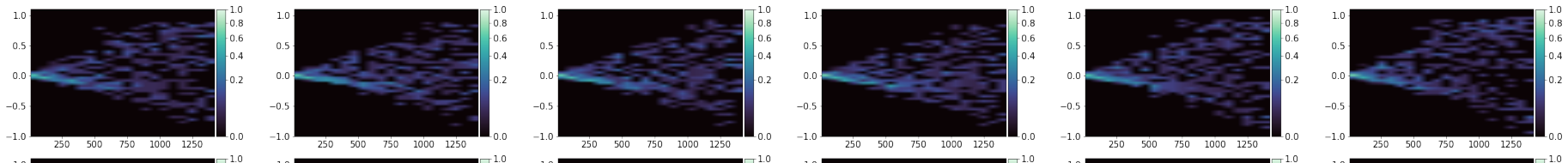
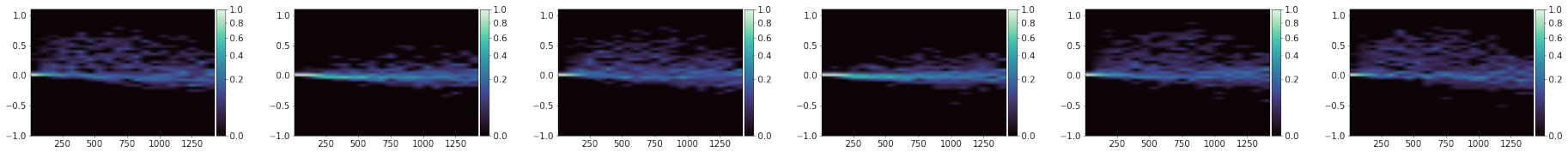
		Exomask dataset					
		Noise Level	**CFA Grid**	**CFA Algorithm**	**JPEG Grid**	**JPEG Quality**	**Hybrid**
Noise-based Methods	Noisesniffer						
	Lyu						
	Mahdian						
CFA-based Methods	Bammey						
	Shin						
	Choi						
JPEG-based Methods	Zero						
	CAGI						
	FDF-A						
	I-CDA						
	CDA						
	BAG						
Multi-purpose methods	Noiseprint						
	ManTraNet						
	Self-Consistency						
	Splicebuster						

Table 3: 2D histograms representing the repartition of MCC scores for each method on each dataset with exomasks, depending on the mask size. Each bin is identified by its mask size (X axis, shown is the square root of that size in pixels) and the MCC score (Y axis). Its value represents the percentage of images in that size bin whose score was in that score bin, out of all the images in that size bin

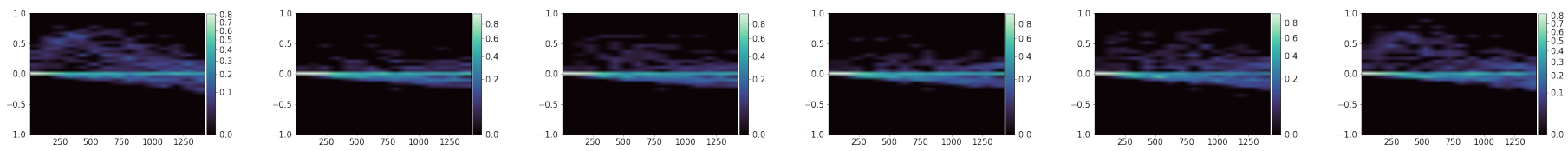
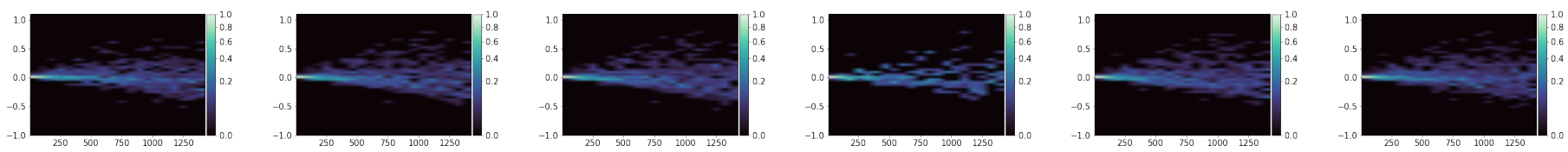
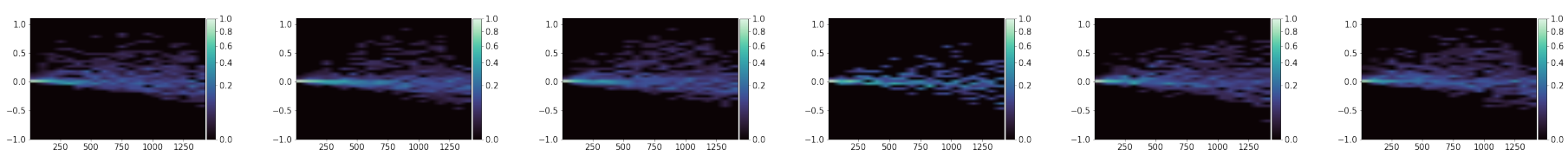
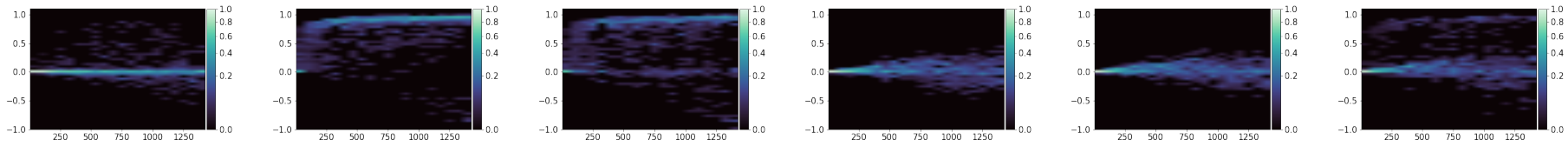
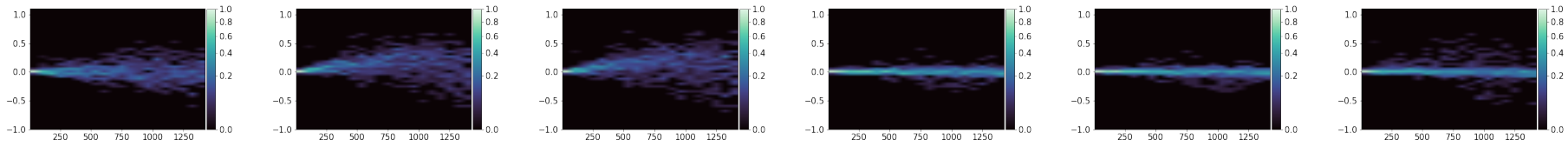
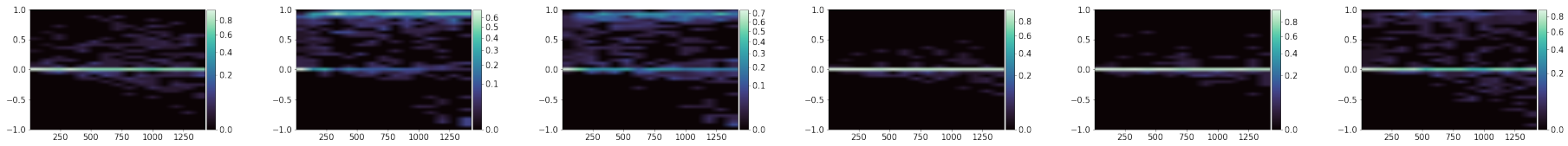
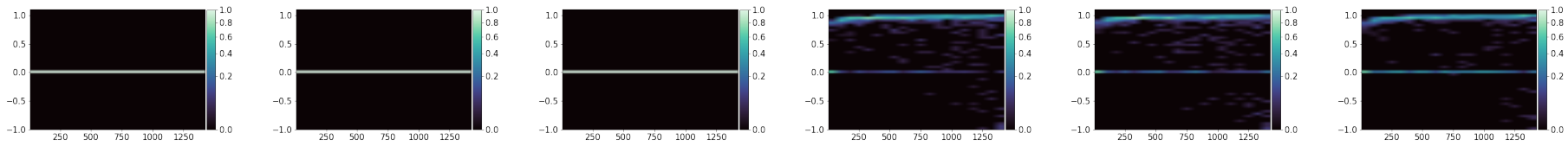
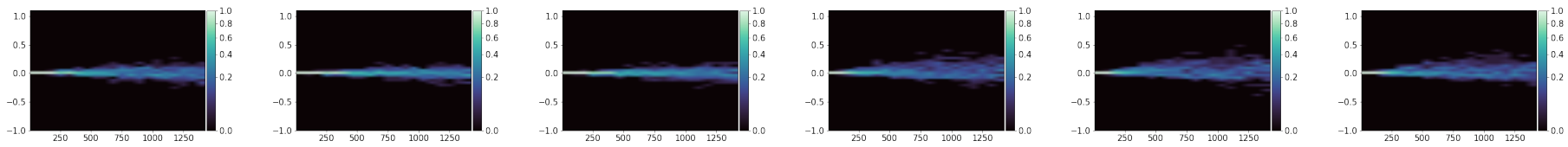
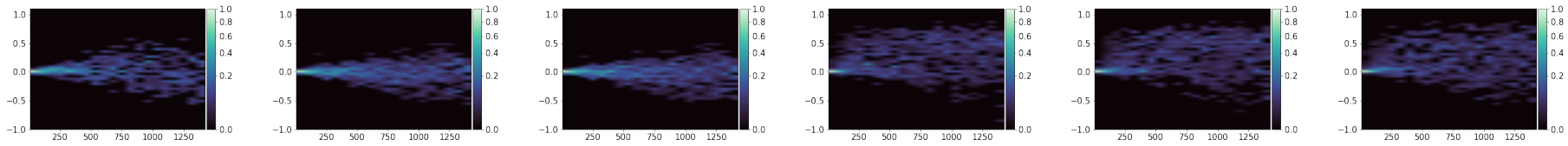
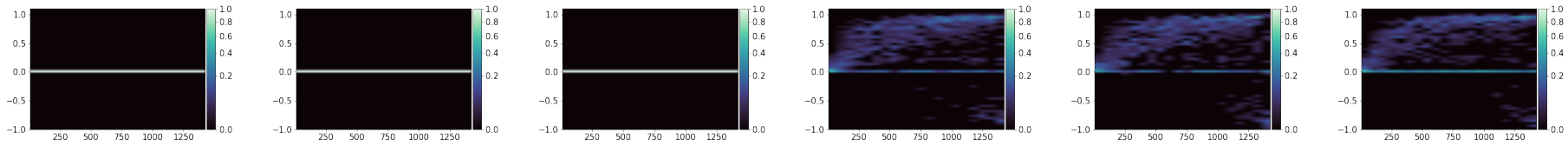
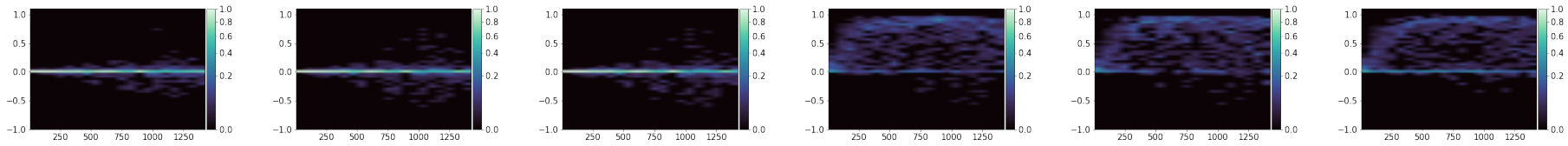
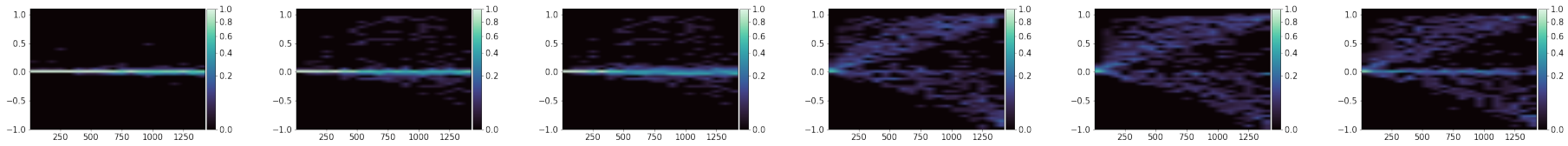
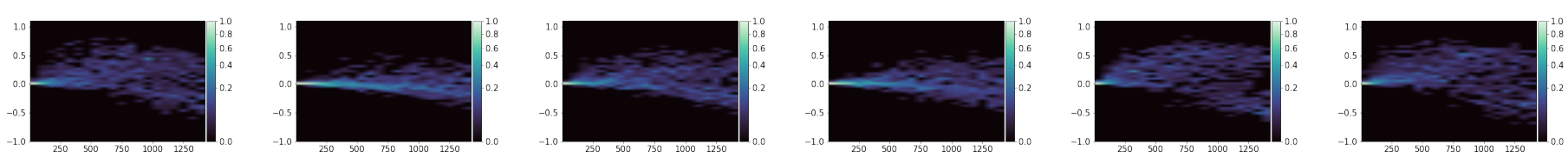
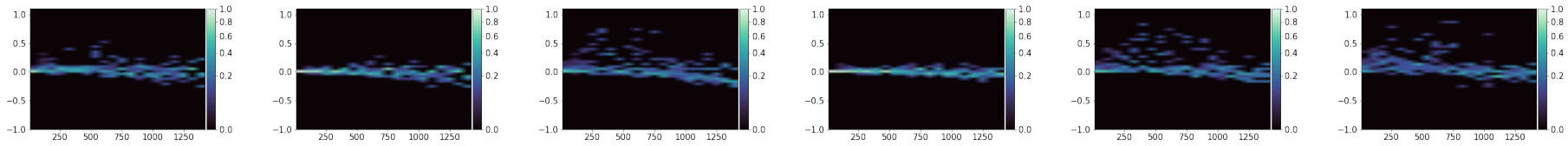
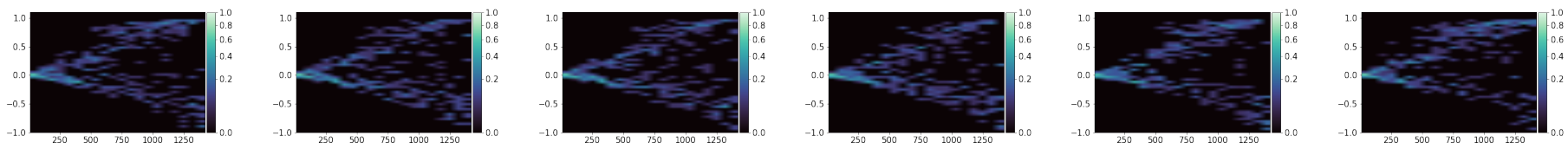
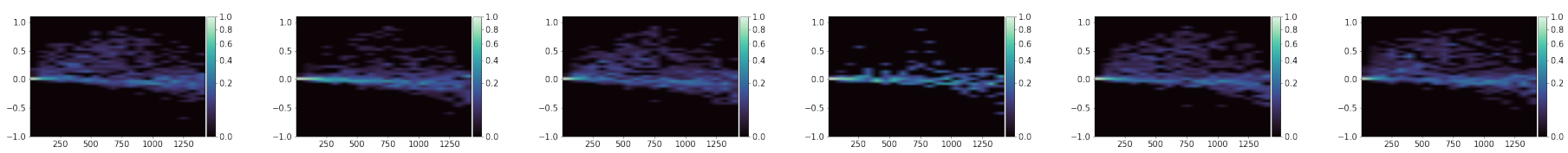
		Endomask dataset					
		Noise Level	**CFA Grid**	**CFA Algorithm**	**JPEG Grid**	**JPEG Quality**	**Hybrid**
Noise-based Methods	Noisesniffer						
	Lyu						
	Mahdian						
CFA-based Methods	Bammey						
	Shin						
	Choi						
JPEG-based Methods	Zero						
	CAGI						
	FDF-A						
	I-CDA						
	CDA						
	BAG						
Multi-purpose methods	Noiseprint						
	ManTraNet						
	Self-Consistency						
	Splicebuster						

Table 4: 2D histograms representing the repartition of MCC scores for each method on each dataset with endomasks, depending on the mask size. Each bin is identified by its mask size (X axis, shown is the square root of that size in pixels) and the MCC score (Y axis). Its value represents the percentage of images in that size bin whose score was in that score bin, out of all the images in that size bin