Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Do Pretrained Encoders Absorb Thermodynamic Priors? A Controlled Study of Physics-Inspired Features for Image Splicing Localization

Saleema K

sleemak@uoc.ac.in

University of Calicut

Lajish V. L.

University of Calicut

Research Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11659905/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Do Pretrained Encoders Absorb Thermodynamic Priors? A Controlled Study of Physics-Inspired Features for Image Splicing Localization

Saleema K[1,2*] and Lajish V. L.[1]

[1*]Department of Computer Science, University of Calicut, Thenhipalam, Malappuram, 673635, Kerala, India.

[2]Department of Computer Science, MES Keveeyam College, Valanchery, Malappuram, 676552, Kerala, India.

*Corresponding author(s). E-mail(s): sleemak@uoc.ac.in;

Abstract

Localization networks for image splicing often add an entropy or noise stream to an RGB backbone, yet few studies test whether such physics-inspired priors add information a high-capacity encoder cannot learn. We model local image statistics as a canonical ensemble: patch Shannon entropy splits into mean pixel entropy and mixing entropy, residual variance estimates a local noise temperature, and a splice boundary acts as an interface between regions of different Helmholtz free energy. We test this thermodynamic prior under four controls. Used alone, fixed thermodynamic maps localize splicing no better than an uninformative all-forged mask on five of six benchmarks. A linear probe shows that a frozen ImageNet-pretrained HRNet-W32 already encodes patch entropy ($R^2 = 0.83$) and noise temperature ($R^2 = 0.71$), versus 0.56 and 0.44 for a random encoder. Across encoder capacities, the prior helps only weak encoders (trained from scratch or on 10% of the data), mainly on Columbia ($+0.056 \pm 0.016$ F_1 over three seeds), but not fully trained pretrained ones (-0.005 mean zero-shot F_1). Removing a learnable thermodynamic layer from a pretrained SegFormer never lowers F_1 by more than 0.004 on eight benchmarks and cuts training cost by 63%. Scale-matched controls show that the maps' boundary evidence is largely explained by patch brightness and contrast; noise temperature beats these controls by only 0.015–0.023 at boundaries on two datasets. Uncertainty-based loss weighting lets the physics terms starve the segmentation loss. Trained only on CASIA v2, our baseline reaches zero-shot F_1 of 0.766 on CASIA v1, 0.349 on NIST16 and 0.568 on IMD2020.

Keywords: Image forensics, Splicing localization, Physics-informed neural networks, Entropy, Linear probing, Negative results

1 Introduction

Digital images serve as evidence in legal proceedings, journalism and public communication, yet modern editing tools can composite foreign image regions within seconds. Splicing localization remains difficult because pasted regions are semantically plausible and are usually followed by boundary feathering, resampling and JPEG recompression that hide the seams [1, 2].

State-of-the-art localization networks typically pair an RGB encoder with an auxiliary stream that exposes low-level artifacts: noise residuals from high-pass or constrained convolutions [3, 4], JPEG quantization traces [5], or learned camera fingerprints [6]. Entropy descriptors have also been used for tampering detection [7]. The gains reported for such streams, however, usually come from single training runs. Two questions remain open: what a splice changes in the statistics of local pixel ensembles, and whether a hand-crafted physical stream adds information that a high-capacity encoder cannot extract from RGB by itself.

Statistical mechanics offers a framework for the first question. By Jaynes' maximum entropy principle, the Boltzmann distribution is the least-biased description of a system with a mean energy constraint [8], which applies directly to intensity distributions in local patches. An authentic photograph comes from one acquisition and demosaicing chain, so its noise temperature is uniform and its statistics vary smoothly. A splice introduces a second chain and hence an interface across which entropy, noise temperature and the Helmholtz free energy $F = U - TS$ [9] change abruptly.

The second question needs controls that forensic studies often omit: an auxiliary stream can appear useful because of a lucky seed, an under-trained baseline, or because it merely re-encodes local brightness and contrast. We therefore test thermodynamic priors in four controlled experiments: (i) the raw maps used alone, against an uninformative all-forged mask; (ii) linear probes testing whether frozen ImageNet features already encode the thermodynamic variables; (iii) the gain from the maps across scratch, low-data and fully pretrained encoders with multiple seeds; and (iv) boundary evidence of the maps against scale-matched brightness and contrast controls (Fig. 1).

Our contributions are: (1) a statistical-mechanical model of local image statistics in which patch entropy decomposes exactly into mean pixel entropy and mixing entropy, and residual variance acts as a local noise temperature (Section 3); (2) two architectures that inject these quantities, fixed maps in a CNN (ThermoDiff-F) and a learnable thermodynamic layer in a transformer (ThermoDiff++) (Section 4); (3) evidence of representation absorption: ImageNet pretraining already makes entropy and temperature linearly decodable, and the benefit of the maps vanishes as encoder capacity grows (Sections 6.2–6.4); (4) a scale-matched boundary analysis showing that the maps' edge alignment is mostly explained by brightness and contrast (Section 6.5); and (5) evidence that uncertainty-based loss weighting lets physics terms starve the segmentation loss (Section 6.6).

2 Related Work

Splicing localization. Early methods used hand-crafted traces such as noise inconsistencies [10] and camera-response or geometry irregularities [11]. Deep methods replaced these features: RGB-N adds steganalysis rich models to a two-stream Faster R-CNN [3, 12], constrained convolutions learn prediction-error filters [4], and ManTra-Net treats localization as anomaly detection [13]. MVSS-Net combines multi-view noise features with edge supervision [14], CAT-Net models JPEG compression artifacts [5], and ObjectFormer [15], PSCC-Net [16], TruFor [6] and IML-ViT [17] lead recent benchmarks. IMDL-BenCo standardized training and evaluation and showed that many earlier gains came from inconsistent protocols [18]; we adopt its CASIA v2 protocol.

Diffusion and noise residuals. Perona–Malik diffusion smooths within regions while preserving high-gradient boundaries [19]; its forward–backward behaviour [20], regularized well-posed forms [21], links to robust statistics [22] and total-variation formulations [23, 24] are well studied. In forensics, denoising residuals underpin camera fingerprints such as Noiseprint [25].

Statistical mechanics and physics-informed learning. Gibbs random fields brought statistical mechanics into image modelling [26], and physics-informed networks embed physical laws in the training objective [27] as observational, inductive or learning biases [28]. Since image formation has no closed governing equation, we use an inductive bias (a thermodynamic input branch) and a learning bias (a free-energy interface loss). The closest forensic work, Jalab et al. [7], trained support vector machines on fractional-entropy descriptors for whole-image classification, without localization or end-to-end learning.

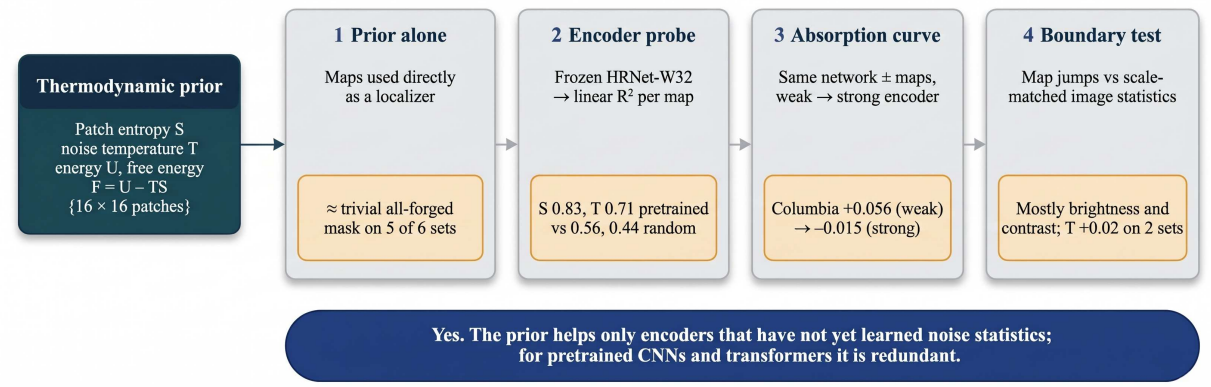

Fig. 1 Overview of the study. A fixed thermodynamic prior is tested in four controlled regimes: alone as a localizer, as a linear-probe target for a frozen encoder, as an extra input across encoder capacities, and against scale-matched boundary controls.

Linear probing. Linear probes test whether a property is linearly decodable from frozen features [29]. They are common in language and vision representation studies but have rarely been used to check whether a proposed forensic feature is already present in a pretrained backbone.

3 A Statistical-Mechanical Model of Local Image Statistics

We model local intensity distributions as an idealized canonical ensemble. Identities (2)–(5) are exact within this discrete formulation; the model describes image statistics and does not imply that sensors are thermodynamic engines.

Pixels as canonical ensembles. Let $I(\mathbf{x}) \in [0,1]$ be the luminance at pixel $\mathbf{x}$, discretized into K states with centres $c_1, \dots, c_K$ and energies

$$\varepsilon_k(\mathbf{x}) = \tfrac{1}{2}\big(I(\mathbf{x}) - c_k\big)^2. \tag{1}$$

At temperature θ the state occupancy follows the Boltzmann distribution

$$\begin{aligned} a_k(\mathbf{x}) &= \frac{\exp\big(-\varepsilon_k(\mathbf{x})/\theta\big)}{Z(\mathbf{x})}, \\ Z(\mathbf{x}) &= \sum_{j=1}^{K} \exp\big(-\varepsilon_j(\mathbf{x})/\theta\big), \end{aligned} \tag{2}$$

which is the maximum-entropy distribution for a fixed mean energy [8]. With $S[a(\mathbf{x})] = -\sum_k a_k(\mathbf{x}) \ln a_k(\mathbf{x})$, the Gibbs–Helmholtz identity holds at each pixel [9]:

$$F_{\mathrm{px}}(\mathbf{x}) \equiv -\theta \ln Z(\mathbf{x}) = \langle \varepsilon \rangle_{a(\mathbf{x})} - \theta\, S[a(\mathbf{x})]. \tag{3}$$

As $\theta \to 0$ the distribution collapses to the nearest bin, recovering an ordinary histogram.

Patch entropy and entropy of mixing. For a neighbourhood $W(\mathbf{x})$ of $|W|$ pixels, the coarse-grained distribution is

$$\bar{a}_k(\mathbf{x}) = \frac{1}{|W|} \sum_{\mathbf{y} \in W(\mathbf{x})} a_k(\mathbf{y}). \tag{4}$$

By the concavity of entropy [30], its entropy splits into two non-negative terms [31, 32]:

$$\begin{aligned} S[\bar{a}] = &\underbrace{\frac{1}{|W|} \sum_{\mathbf{y} \in W} S[a(\mathbf{y})]}_{\text{mean pixel entropy}} \\ &+ \underbrace{\mathrm{JS}_W\big(a(\mathbf{y}_1), \dots, a(\mathbf{y}_{|W|})\big)}_{\text{entropy of mixing}}. \end{aligned} \tag{5}$$

The mixing term is the generalized Jensen–Shannon divergence of the pixel distributions in W; it vanishes only when all pixels share one distribution. A patch that straddles an authentic–tampered boundary mixes two populations. *Prediction 1:* mixing entropy peaks along splice boundaries.

Residual energy and noise temperature. Sensor noise and demosaicing residuals approximate a zero-mean random field with variance $s^2(\mathbf{x})$. For a high-pass filter h with $\sum_{\mathbf{u}} h(\mathbf{u}) = 0$

and $\|h\|_2 = 1$, the residual $r = h * I$ satisfies $\langle r^2(\mathbf{x})\rangle = s^2(\mathbf{x})$ away from edges. By analogy with equipartition we call $s^2(\mathbf{x})$ the local noise temperature; it is stationary within one acquisition chain. *Prediction 2:* noise temperature jumps across splice boundaries.

Free energy and phase interfaces. Combining internal energy U, patch entropy S and noise temperature T gives the Helmholtz free-energy field

$$F(\mathbf{x}) = U(\mathbf{x}) - T(\mathbf{x})\, S(\mathbf{x}). \tag{6}$$

Treating the predicted tampering map $M(\mathbf{x}) \in [0,1]$ as a phase field, we define an interface energy and an intra-phase equilibrium energy [24, 33]:

$$E_{\text{int}}[M] = \int_\Omega \gamma(\mathbf{x})\, |\nabla M(\mathbf{x})|\, d\mathbf{x}, \qquad \gamma(\mathbf{x}) = \exp\!\big(-|\nabla \hat{F}(\mathbf{x})|/\tau\big), \tag{7}$$

$$E_{\text{eq}}[M] = \text{Var}_M(F) + \text{Var}_{1-M}(F), \tag{8}$$

where $\hat{F}$ is F normalized to $[0,1]$ and $\text{Var}_M(F)$ is the variance of F weighted by M. *Prediction 3:* if the thermodynamic state carries information not accessible from RGB, adding it will improve localization on unseen acquisition pipelines; if encoders already decode it, the gain will be negligible.

4 Models and Architectures

4.1 Fixed thermodynamic descriptors

Images are split into non-overlapping 16×16 patches P. With h_k the count in bin k of a 32-bin histogram and $N = 256$,

$$S = -\sum_{k=1}^{32} \tilde{p}_k \ln \tilde{p}_k, \quad \tilde{p}_k = \frac{h_k/N + \epsilon}{\sum_j (h_j/N + \epsilon)}, \tag{9}$$

$$U = \frac{1}{|P|} \sum_{\mathbf{x}\in P} I(\mathbf{x})^2, \qquad r = I - b_3 * I, \qquad T = \text{Var}_P(r) + \epsilon, \qquad N_r = \text{Std}_P(r), \tag{10}$$

where b_3 is a 3×3 box filter and $F = U - TS$ by Eq. (6). The five maps are upsampled bilinearly and normalized with statistics from 200 training images. Two properties matter later: $N_r = (T - \epsilon)^{1/2}$ is a monotonic transform of T, and since $U \in [0,1]$ while $TS \lesssim 0.18$ ($T \le 0.052$, $S \le \ln 32$), F is dominated by U, which tracks patch brightness.

4.2 ThermoDiff-F and ThermoDiff++

ThermoDiff-F is a U-Net [34] with an ImageNet-pretrained ResNet34 [35] or HRNet-W32 [36] encoder; the fixed maps enter the bottleneck through four-head cross-attention. Variants add Perona–Malik diffusion channels [19, 22], a high-pass noise stream or a Siamese contrastive loss (24.4–27.6 M parameters for ResNet34).

ThermoDiff++ uses a SegFormer MiT-B2 encoder [37]. A learnable thermodynamic layer maps patches onto $K = 16$ microstates via Eq. (2), computes coarse-grained entropy over 7×7, 15×15 and 31×31 windows, and extracts internal energy with eight learnable zero-mean 5×5 filters,

$$U_j(\mathbf{x}) = \ln\!\Big(1 + \lambda\, \overline{(\tilde{h}_j * I)^2}(\mathbf{x})\Big), \qquad \lambda = 100, \tag{11}$$

while a three-layer MLP estimates T. The 32-channel embedding t enters the encoder at 1/4, 1/8 and 1/16 scale through gated residuals initialized with $W_p = 0$:

$$f \leftarrow f + \sigma\big(W_g[f;t]\big) \odot (W_p\, t). \tag{12}$$

An FPN decoder predicts the mask together with image-level and confidence outputs, trained with

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + 0.3\,\mathcal{L}_{\text{cls}} + 0.1\, w(e)\,\mathcal{L}_{\text{phys}} + 0.1\,\mathcal{L}_{\text{conf}}, \qquad \mathcal{L}_{\text{phys}} = E_{\text{int}} + 0.5\, E_{\text{eq}}, \tag{13}$$

where $w(e)$ ramps linearly over the first three epochs.

4.3 Controlled HRNet-W32 testbed

To isolate absorption from architectural differences we use one HRNet-W32 testbed (Fig. 2a). All five encoder outputs are projected to 96 channels, upsampled to half resolution, concatenated and classified by a 3×3 and a 1×1 convolution (31.5 M parameters). In the thermodynamic variant, the five maps of Section 4.1 pass through two

Table 1 Datasets. CASIA v2 is used for training (COCO only for ThermoDiff++ pretraining); all other sets are zero-shot tests.

Dataset	Images	Manipulation type
COCO 2017 [39]	24,000 / epoch	Synthetic splicing
CASIA v2 [40]	5,123 + 7,491 authentic	Splicing and copy-move
CASIA v1 [40]	920	Splicing and copy-move
Columbia [11]	180	Uncompressed splicing
NIST16 [41]	564	Splicing, copy-move, removal
COVERAGE [42]	100	Similar-object copy-move
MISD [43]	235	Multi-source splicing
IMD2020 [44]	2,010	In-the-wild manipulations
DEFACTO-CM [45]	1,000 subsample	Object copy-move

3×3 convolutions to 32 channels and are concatenated at every scale (+0.07 M parameters). Three encoder regimes are compared under the same recipe: ImageNet-pretrained on all data, pretrained on a 10% subset, and randomly initialized.

5 Experimental Protocol

All models are trained only on CASIA v2; every other dataset is a zero-shot test (Table 1). Columbia masks are converted to solid regions, MISD masks are merged per image, NIST16 masks are inverted, and COVERAGE masks are paired with their forged images. For ThermoDiff++ pretraining, COCO objects are pasted with random alpha-feathering, hard edges or Poisson blending [38], and 70% of the composites are recompressed. Mean forged-area fractions are 9% (CASIA v2), 26% (Columbia), 11% (COVERAGE), 18% (MISD), 8% (IMD2020) and 2% (DEFACTO-CM). Training settings are listed in Table 2.

Metrics and controls. The primary metric is pixel F_1 at threshold 0.5, computed per tampered image and averaged, as in IMDL-BenCo [18]; IoU and pixel AUC are also reported. An all-forged mask gives $F_1 = 2a/(1+a)$ for forged fraction a; on Columbia it reaches a measured 0.407, higher than a trained RGB network (0.378), so it is reported as a trivial baseline. Linear probes are ridge regressions ($\alpha = 10\,C$) from frozen encoder activations, pooled to a 24×24 grid, onto the thermodynamic maps (300 training and 100 validation images). For boundary analysis, positives are pixels within an 8-pixel band around the true seam and negatives are an equal number of interior pixels; we compute the AUC of $|\nabla X|$ for each map and for three scale-matched controls: Sobel pixel edges, patch-mean brightness and patch standard deviation (contrast) on the same 16×16 grid.

Table 2 Training settings of ThermoDiff++ and the controlled HRNet-W32 testbed.

Setting	ThermoDiff++	HRNet-W32 testbed
Backbone Initialization	SegFormer MiT-B2 ImageNet-1k	HRNet-W32 ImageNet-1k vs. Random
CASIA v2 Split	80 / 10 / 10 (all images)	80 / 10 / 10 (tampered only)
Input Dimension	384×384 (crops)	384×384
Pretraining	8 epochs COCO synthetic	None
Loss Function	Eq. (13)	BCE + Dice
Optimizer	AdamW ($\eta_e = 6 \times 10^{-5}$)	AdamW ($\eta = 10^{-4}$)
Schedule	Warm-up + Cosine (40 ep.)	Cosine (25 to 40 epochs)
Hardware Profile	NVIDIA P100	NVIDIA T4 (FP16 mixed)

6 Results

6.1 Benchmark performance

Under the IMDL-BenCo protocol [18] (CASIA v2 training, threshold 0.5, no test-time augmentation), ThermoDiff++ without its thermodynamic stream reaches $F_1 = 0.766$ on CASIA v1, 0.349 on NIST16 and 0.568 on IMD2020, with a five-dataset mean of 0.528, comparable to TruFor (0.530) (Table 3).

6.2 The prior alone

Used directly as localizers, the fixed maps are weak. Unsupervised deviation thresholds ($|X - \text{median}|/\text{MAD}$) do not beat the all-forged mask on five of six test sets, and supervised logistic readouts collapse to the all-forged output. Only noise temperature T clearly exceeds the trivial mask: on MISD ($F_1 = 0.354$ vs. 0.282), IMD2020 (0.173 vs. 0.131) and DEFACTO-CM (0.051 vs. 0.041). On Columbia the best map only ties the trivial mask (0.407), and on CASIA v2 and COVERAGE it exceeds it by less than 0.005 (Table 4, Fig. 3).

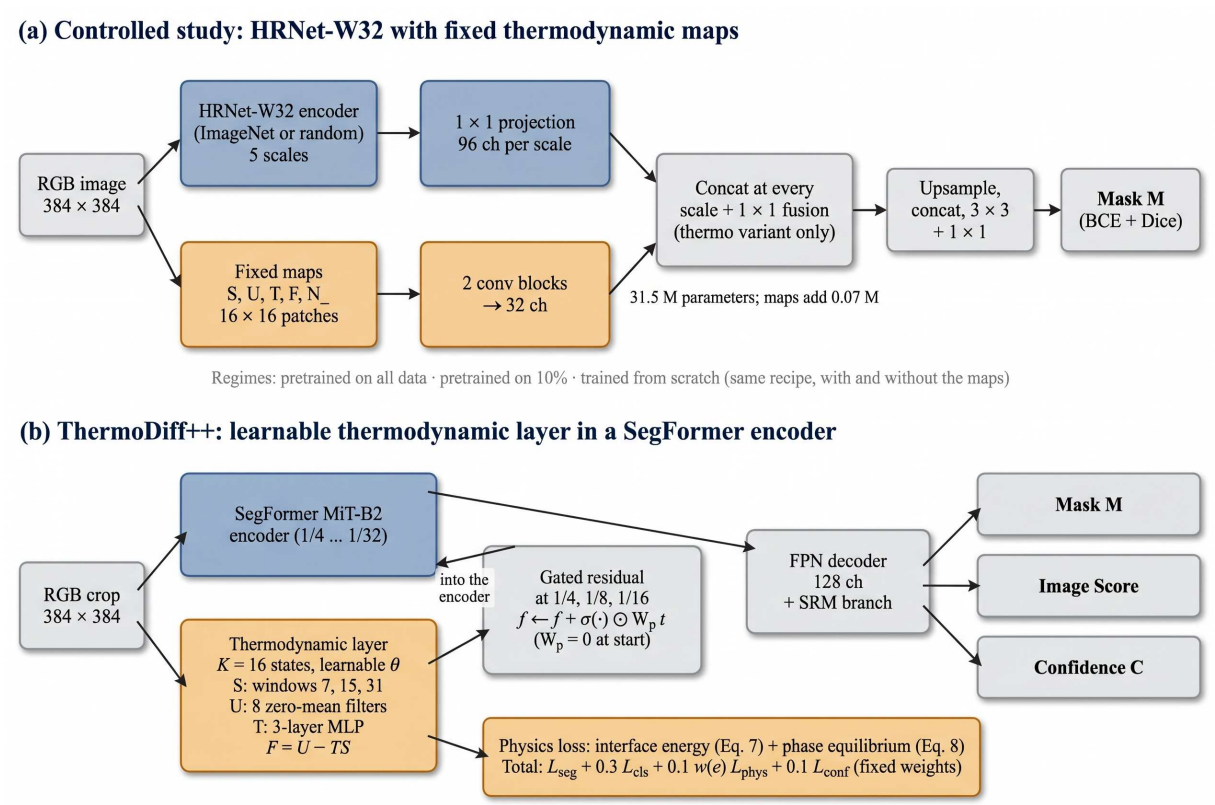

Fig. 2 Architectures. (a) The controlled HRNet-W32 testbed, in which five fixed thermodynamic maps are fused into every scale (+0.07 M parameters on 31.5 M). (b) ThermoDiff++, with a learnable thermodynamic layer inside a SegFormer MiT-B2 encoder.

Table 3 Pixel-level F1 localization performance at threshold 0.5 under the CASIA v2 benchmark protocol. Comparative values are obtained from the standardized IMDL-BenCo benchmark [18].

Method	CASIA v1	Columbia	NIST16	COVERAGE	IMD2020	Five-Dataset Mean
MVSS-Net [14]	0.534	0.386	0.246	0.259	0.279	0.341
CAT-Net [5]	0.581	0.584	0.269	0.296	0.273	0.401
ObjectFormer [15]	0.429	0.336	0.173	0.294	0.173	0.281
PSCC-Net [16]	0.378	0.604	0.214	0.231	0.235	0.333
TruFor [6]	0.721	0.865	0.324	0.419	0.322	0.530
IML-ViT [17]	0.721	0.780	0.331	0.435	0.327	0.519
Ours (No Thermo Layer)	0.766	0.555	0.349	0.403	0.568	0.528
Ours (+ Flip TTA)	0.762	0.573	0.340	0.380	0.567	0.524

6.3 Linear probing of pretrained encoders

A frozen ImageNet-pretrained HRNet-W32 linearly reconstructs the thermodynamic variables without any forensic training: $R^2 = 0.83$ for S, 0.80 for N_r and 0.71 for T on held-out images. A randomly initialized copy reaches only 0.56, 0.57 and 0.44. Both encoders reconstruct U and F almost perfectly ($R^2 \geq 0.91$), confirming that these maps are close to linear functions of brightness (Table 5, Fig. 4).

6.4 Gain across encoder capacity

In the controlled testbed, the maps help only capacity-limited encoders, and the gain approaches zero as the encoder strengthens (Table 6, Fig. 5). With 10% of the data, the mean zero-shot gain is $+0.007 \pm 0.044$ over three seeds

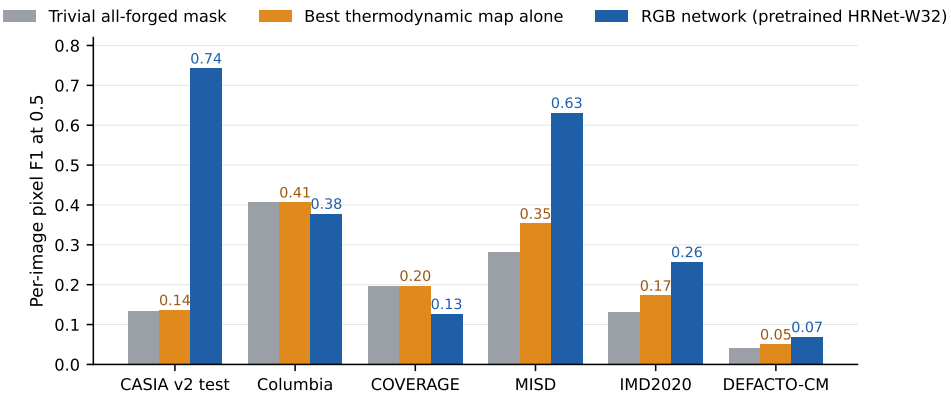

Fig. 3 Each thermodynamic map used alone as a localizer, compared with a trivial all-forged mask and the pretrained RGB HRNet-W32 network. Only the noise temperature T clearly beats the trivial mask, on MISD, IMD2020 and DEFACTO-CM.

Table 4 Each fixed map used alone as a localizer (pixel F_1), against the all-forged mask and the pretrained RGB HRNet-W32. The best map per dataset is given in parentheses.

Dataset	All-forged	Best map	RGB network
CASIA v2 Test	0.133	0.137 (F)	0.743
Columbia	0.407	0.407 (tie)	0.378
COVERAGE	0.196	0.198	0.127
MISD	0.282	0.354 (T)	0.631
IMD2020	0.131	0.173 (T)	0.257
DEFACTO-CM	0.041	0.051 (T)	0.069

Table 5 Linear probe decoding accuracy (R^2 on held-out data) from frozen encoder representations.

Encoder	S	T	N_r	U	F
Random init.	0.56	0.44	0.57	0.91	0.91
ImageNet	0.83	0.71	0.80	0.93	0.93

(+0.003, −0.035, +0.053). Columbia is the exception, with +0.056 ± 0.016. For the fully trained pretrained encoder the effect reverses (−0.015 on Columbia).

The transformer shows the same pattern. Removing the learnable thermodynamic layer from ThermoDiff++ changes CASIA v2 F_1 by only +0.002; on the eight test suites removal never lowers F_1 by more than 0.004 and raises it on several (Columbia +0.043, IMD2020 +0.032, DEFACTO-CM +0.018), while cutting training compute by 63% (Table 7).

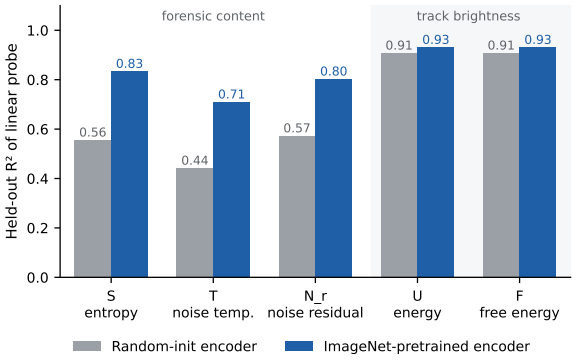

Fig. 4 Held-out R^2 of ridge probes that predict each thermodynamic map from frozen HRNet-W32 features. Pretraining raises R^2 for S, T and N_r by about 0.25, while U and F are decoded almost equally well by a random encoder because they track brightness.

6.5 Scale-matched boundary analysis

Gradients of the thermodynamic maps align with splice seams better than Sobel pixel edges, but the scale-matched controls explain most of this (Table 8). Patch brightness beats free energy F on five of six datasets, and patch entropy S is below both brightness and contrast everywhere. Only the noise-temperature maps (T, N_r) beat every control, and only on two datasets: MISD (+0.021, +0.023) and DEFACTO-CM (+0.015, +0.012) (Fig. 6).

6.6 Loss balancing in physics-informed training

With homoscedastic uncertainty weighting ($w_i = \exp(-s_i)$) [46], ThermoDiff++ training becomes

Table 6 Marginal change in pixel F1 (threshold 0.5) resulting from the addition of fixed thermodynamic maps. The 10% pretrained regime reports the sample mean and sample standard deviation across three random seeds (42, 43, 44).

Evaluation Benchmark	Scratch (25 ep.)	Scratch (40 ep.)	Pretrained (10% Data)	Pretrained (100% Data)
In-Domain CASIA v2 F1	0.219	0.282	0.496	0.743
Mean Zero-Shot F1	0.135	0.148	0.209	0.293
Δ CASIA v2 Test	+0.019	+0.010	+0.021 ± 0.016	+0.005
Δ Columbia	+0.036	+0.027	+0.056 ± 0.016	-0.015
Δ COVERAGE	-0.002	-0.040	-0.017 ± 0.035	-0.010
Δ MISD	+0.030	+0.014	+0.002 ± 0.104	+0.015
Δ IMD2020	0.000	+0.017	+0.003 ± 0.048	-0.013
Δ DEFACTO-CM	0.000	-0.005	-0.009 ± 0.022	0.000
Δ Zero-Shot Mean	+0.013	+0.003	+0.007 ± 0.044	-0.005

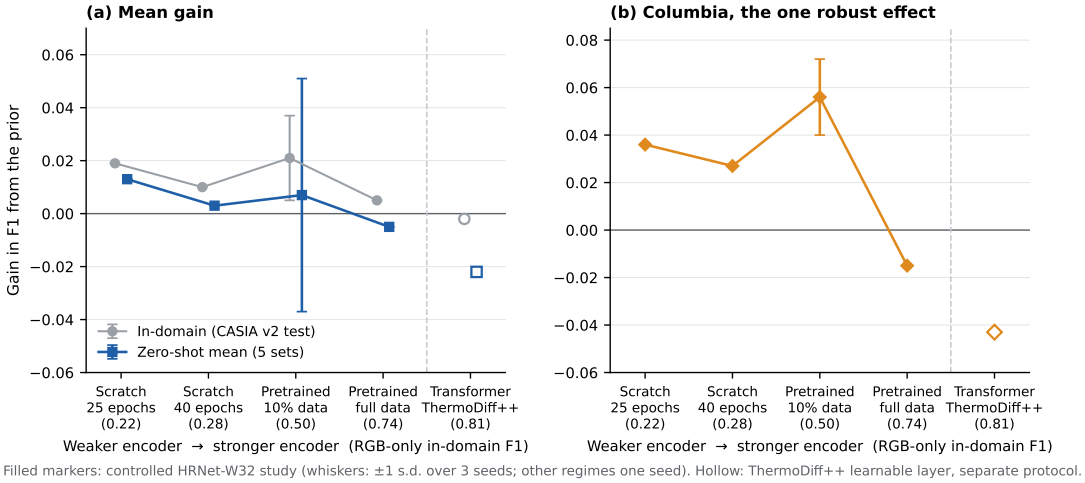

Fig. 5 Absorption curve. (a) Mean F1 gain from adding the thermodynamic maps, across encoder regimes; the 10% regime shows mean ± s.d. over seeds 42, 43 and 44. (b) Columbia, the one dataset where the gain is robust in weak regimes.

Table 7 Zero-shot localization ablation of the learnable thermodynamic layer in ThermoDiff++. Evaluations report per-image F1 at threshold 0.5 (with pixel AUC reported in parentheses).

Configuration	CASIA v2 Test	Columbia	MISD	IMD2020	DEFACTO-CM
Full Architecture	0.805 (0.975)	0.512 (0.795)	0.543 (0.861)	0.536 (0.807)	0.187 (0.708)
Without Thermo Layer	0.807 (0.975)	0.555 (0.822)	0.541 (0.869)	0.568 (0.814)	0.205 (0.715)

Table 8 Boundary localization AUC across an 8-pixel seam margin (500 images per evaluation set). Controls are evaluated over identical 16 × 16 spatial windows. Bold indicates the highest AUC per benchmark.

Dataset	Grad(S)	Grad(T)	Grad(F)	Grad(N_r)	Pixel Edges	Patch Brightness	Patch Contrast
CASIA v2 Test	0.615	0.664	0.681	0.669	0.628	**0.683**	0.655
Columbia	0.654	0.732	**0.740**	0.730	0.614	0.734	0.689
COVERAGE	0.598	0.637	0.650	0.633	0.623	**0.665**	0.614
MISD	0.625	0.694	0.645	**0.696**	0.600	0.655	0.673
IMD2020	0.616	0.690	0.686	0.681	0.652	**0.699**	0.653
DEFACTO-CM	0.594	**0.626**	0.574	0.624	0.610	0.604	0.611

unbalanced. The physics loss falls quickly to about 10^{-3}, so its weight reaches the upper bound of 7.39 by epoch 5; the classification weight reaches the bound at epoch 19, while the segmentation loss receives only 4.3. Fixed weights (1.0, 0.3, 0.1, 0.1 for segmentation, classification, physics and confidence) avoid this collapse and improve every

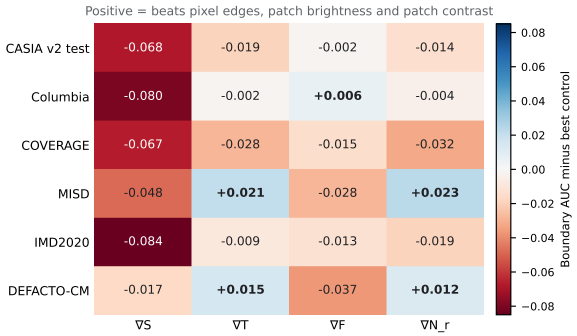

Fig. 6 Boundary AUC of each thermodynamic map minus the best scale-matched control (pixel edges, patch brightness, patch contrast). Positive values mean the map carries boundary evidence beyond these simple statistics.

metric, e.g. CASIA v2 F_1 from 0.740 to 0.805 and Columbia F_1 from 0.483 to 0.512 (Table 9).

7 Discussion

Representation absorption. Four findings point the same way. The maps alone barely beat a trivial mask (Section 6.2); a frozen ImageNet encoder already decodes them linearly (Section 6.3); their benefit shrinks with encoder capacity and data and disappears for pretrained CNNs and transformers (Section 6.4); and their boundary evidence is largely brightness and contrast (Section 6.5). Prediction 3 therefore holds in its null form: once an encoder can decode noise and entropy statistics from RGB, an explicit physical stream is redundant. Fig. 7 shows typical maps and predictions.

Why Columbia differs. Columbia is the only dataset where the maps help weak encoders. Its splices join uncompressed regions from different cameras without recompression or feathering, so jumps in sensor noise are large. In recompressed and post-processed sets such as CASIA v2 and IMD2020, processing erodes these discontinuities.

Implications for forensic studies. (i) The value of an auxiliary stream cannot be shown with single runs: our run-to-run spread (± 0.044 F_1) is as large as many reported gains. (ii) Before designing a fusion branch, authors should probe whether the backbone already encodes the signal. (iii) Physical maps offered as explanations should be compared with scale-matched brightness and contrast controls. (iv) Physics priors remain useful where strong pretraining is not available, such as very small edge models, non-standard sensors (infrared, satellite) or very small datasets.



Fig. 7 Qualitative examples: input image, ground-truth mask, the entropy S, noise temperature T (contrast-stretched) and free energy F maps, and predictions of the from-scratch RGB and RGB + thermodynamic HRNet-W32 models (40 epochs, seed 42).

8 Conclusion

We modelled local image statistics as a canonical ensemble and tested thermodynamic priors for splicing localization under controlled conditions. Fixed thermodynamic maps alone perform on par with an all-forged mask on five of six benchmarks; ImageNet pretraining makes patch entropy ($R^2 = 0.83$) and noise temperature ($R^2 = 0.71$) linearly decodable; the maps help capacity-limited or low-data encoders but not strong pretrained ones; their boundary evidence is mostly brightness and contrast; and physics losses need fixed weights, because uncertainty weighting lets them starve the segmentation loss. Trained only on CASIA v2, our baseline reaches zero-shot F_1 of 0.766 on CASIA v1, 0.349 on NIST16 and 0.568 on IMD2020, comparable to recent detectors. Future work will test whether the same absorption occurs for audio manipulation and generative-image detection.

Table 9 Quantitative impact of loss weighting mechanisms on ThermoDiff++ performance.

Weighting Formulation	CASIA v2 F1	IoU	Pixel AUC	Columbia F1	MISD F1	IMD2020 F1
Uncertainty [46]	0.740	0.678	0.928	0.483	0.505	0.523
Fixed Weights (Eq. 13)	0.805	0.753	0.975	0.512	0.543	0.536

Acknowledgements. Not applicable.

Statements and Declarations

- **Funding:** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
- **Competing interests:** The authors declare that they have no competing interests.
- **Ethics approval and consent to participate:** Not applicable. The study uses only publicly available image datasets.
- **Consent for publication:** Not applicable.
- **Data availability:** All evaluation datasets utilized in this study are openly accessible through their original institutional repositories [11, 39–45]. The source code, training logs, evaluation scripts and model checkpoints are available from the corresponding author upon reasonable request.
- **Materials availability:** Not applicable.
- **Code availability:** Available from the corresponding author upon reasonable request.
- **Author contributions:** Saleema K: conceptualization, methodology, software, data curation, formal analysis, investigation, validation, visualization, writing – original draft. Lajish V. L.: conceptualization, supervision, project administration, writing – review and editing.
- **Use of AI tools:** Generative AI tools were used only to assist with text formatting, language editing and manuscript organization. The authors reviewed all content, verified all derivations and results, and take full responsibility for the manuscript.

References

[1] L. Verdoliva, Media forensics and deepfakes: an overview, IEEE J. Sel. Top. Signal Process. 14 (2020) 910–932. https://doi.org/10.1109/JSTSP.2020.3002101.

[2] M. Zampoglou, S. Papadopoulos, Y. Kompatsiaris, Large-scale evaluation of splicing localization algorithms for web images, Multimed. Tools Appl. 76 (2017) 4801–4834. https://doi.org/10.1007/s11042-016-3795-2.

[3] P. Zhou, X. Han, V.I. Morariu, L.S. Davis, Learning rich features for image manipulation detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 1053–1061. https://doi.org/10.1109/CVPR.2018.00116.

[4] B. Bayar, M.C. Stamm, Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection, IEEE Trans. Inf. Forensics Secur. 13 (2018) 2691–2706. https://doi.org/10.1109/TIFS.2018.2825953.

[5] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, C. Kim, Learning JPEG compression artifacts for image manipulation detection and localization, Int. J. Comput. Vis. 130(8) (2022) 1875–1895. https://doi.org/10.1007/s11263-022-01617-5.

[6] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, L. Verdoliva, TruFor: leveraging all-round clues for trustworthy image forgery detection and localization, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 20606–20615. https://doi.org/10.1109/CVPR52729.2023.01974.

[7] H.A. Jalab, T. Subramaniam, R.W. Ibrahim, H. Kahtan, N.F.M. Noor, New texture descriptor based on modified fractional entropy for digital image splicing forgery detection, Entropy 21(4) (2019) 371. https://doi.org/10.3390/e21040371.

[8] E.T. Jaynes, Information theory and statistical mechanics, Phys. Rev. 106 (1957) 620–630. https://doi.org/10.1103/PhysRev.106.620.

[9] H.B. Callen, Thermodynamics and an Introduction to Thermostatistics, second ed., John Wiley & Sons, New York, 1985.

[10] B. Mahdian, S. Saic, Using noise inconsistencies for blind image forensics, Image Vis. Comput. 27 (2009) 1497–1503. https://doi.org/10.1016/j.imavis.2009.02.002.

[11] Y.-F. Hsu, S.-F. Chang, Detecting image splicing using geometry invariants and camera characteristics consistency, in: Proc. IEEE Int. Conf. Multimed. Expo (ICME), 2006, pp. 549–552. https://doi.org/10.1109/ICME.2006.262450.

[12] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Trans. Inf. Forensics Secur. 7 (2012) 868–882. https://doi.org/10.1109/TIFS.2012.2190402.

[13] Y. Wu, W. AbdAlmageed, P. Natarajan, ManTra-Net: manipulation tracing network for detection and localization of image forgeries with anomalous features, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 9543–9552. https://doi.org/10.1109/CVPR.2019.00977.

[14] X. Chen, C. Dong, J. Ji, J. Cao, X. Li, Image manipulation detection by multi-view multi-scale supervision, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 14185–14193. https://doi.org/10.1109/ICCV48922.2021.01392.

[15] J. Wang, Z. Wu, J. Chen, X. Han, A. Shrivastava, S.-N. Lim, Y.-G. Jiang, ObjectFormer for image manipulation detection and localization, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 2354–2363. https://doi.org/10.1109/CVPR52688.2022.00240.

[16] X. Liu, Y. Liu, J. Chen, X. Liu, PSCC-Net: progressive spatio-channel correlation network for image manipulation detection and localization, IEEE Trans. Circuits Syst. Video Technol. 32(11) (2022) 7505–7517. https://doi.org/10.1109/TCSVT.2022.3189545.

[17] X. Ma, B. Du, Z. Jiang, A.Y. Al Hammadi, J. Zhou, IML-ViT: benchmarking image manipulation localization by vision transformer, arXiv:2307.14863, 2023. https://doi.org/10.48550/arXiv.2307.14863.

[18] X. Ma, X. Zhu, L. Su, B. Du, Z. Jiang, B. Tong, Z. Lei, X. Yang, C.-M. Pun, J. Lv, J. Zhou, IMDL-BenCo: a comprehensive benchmark and codebase for image manipulation detection & localization, in: Adv. Neural Inf. Process. Syst. 37 (NeurIPS), Datasets and Benchmarks Track, 2024. https://doi.org/10.48550/arXiv.2406.10580.

[19] P. Perona, J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 629–639. https://doi.org/10.1109/34.56205.

[20] S. Kichenassamy, The Perona–Malik paradox, SIAM J. Appl. Math. 57 (1997) 1328–1342. https://doi.org/10.1137/S0036139995287637.

[21] F. Catté, P.-L. Lions, J.-M. Morel, T. Coll, Image selective smoothing and edge detection by nonlinear diffusion, SIAM J. Numer. Anal. 29 (1992) 182–193. https://doi.org/10.1137/0729012.

[22] M.J. Black, G. Sapiro, D.H. Marimont, D. Heeger, Robust anisotropic diffusion, IEEE Trans. Image Process. 7 (1998) 421–432. https://doi.org/10.1109/83.661192.

[23] J. Weickert, Anisotropic Diffusion in Image Processing, B.G. Teubner, Stuttgart, 1998.

[24] L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D 60 (1992) 259–268. https://doi.org/10.1016/0167-2789(92)90242-F90242-F).

[25] D. Cozzolino, L. Verdoliva, Noiseprint: a CNN-based camera model fingerprint, IEEE Trans. Inf. Forensics Secur. 15 (2020) 144–159. https://doi.org/10.1109/TIFS.2019.2916364.

[26] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 721–741. https://doi.org/10.1109/TPAMI.1984.4767596.

[27] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. 378 (2019) 686–707. https://doi.org/10.1016/j.jcp.2018.10.045.

[28] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Physics-informed machine learning, Nat. Rev. Phys. 3 (2021) 422–440. https://doi.org/10.1038/s42254-021-00314-5.

[29] G. Alain, Y. Bengio, Understanding intermediate layers using linear classifier probes, in: Int. Conf. Learn. Represent. (ICLR), Workshop Track, 2017. https://doi.org/10.48550/arXiv.1610.01644.

[30] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[31] T.M. Cover, J.A. Thomas, Elements of Information Theory, second ed., John Wiley & Sons, Hoboken, 2006.

[32] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theory 37 (1991) 145–151. https://doi.org/10.1109/18.61115.

[33] J.W. Cahn, J.E. Hilliard, Free energy of a nonuniform system. I. Interfacial free energy, J. Chem. Phys. 28 (1958) 258–267. https://doi.org/10.1063/1.1744102.

[34] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: Med. Image Comput. Comput.-Assist. Interv. (MICCAI), Lect. Notes Comput. Sci. 9351, 2015, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[36] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686.

[37] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, in: Adv. Neural Inf. Process. Syst. 34 (NeurIPS), 2021, pp. 12077–12090.

[38] P. Pérez, M. Gangnet, A. Blake, Poisson image editing, ACM Trans. Graph. 22 (2003) 313–318. https://doi.org/10.1145/882262.882269.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: Proc. Eur. Conf. Comput. Vis. (ECCV), Lect. Notes Comput. Sci. 8693, 2014, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

[40] J. Dong, W. Wang, T. Tan, CASIA image tampering detection evaluation database, in: Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP), 2013, pp. 422–426. https://doi.org/10.1109/ChinaSIP.2013.6625374.

[41] H. Guan, M. Kozak, E. Robertson, Y. Lee, A.N. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, J. Fiscus, MFC datasets: large-scale benchmark datasets for media forensic challenge evaluation, in: Proc. IEEE Winter Conf. Appl. Comput. Vis. Workshops (WACVW), 2019, pp. 63–72. https://doi.org/10.1109/WACVW.2019.00018.

[42] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, S. Winkler, COVERAGE – a novel database for copy-move forgery detection, in: Proc. IEEE Int. Conf. Image Process. (ICIP), 2016, pp. 161–165. https://doi.org/10.1109/ICIP.2016.7532339.

[43] K.D. Kadam, S. Ahirrao, K. Kotecha, Multiple image splicing dataset (MISD): a dataset for multiple splicing, Data 6 (2021) 102. https://doi.org/10.3390/data6100102.

[44] A. Novozamsky, B. Mahdian, S. Saic, IMD2020: a large-scale annotated dataset tailored for detecting manipulated images, in: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW), 2020, pp. 71–80. https://doi.org/10.1109/WACVW50321.2020.9096940.

[45] G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J.L. Dugelay, M. Pic, DEFACTO: image and face manipulation dataset, in: Proc. Eur. Signal Process. Conf. (EUSIPCO), 2019, pp. 1–5. https://doi.org/10.23919/EUSIPCO.2019.8903181.

[46] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7482–7491. https://doi.org/10.1109/CVPR.2018.00781.