Can We Build a Monolithic Model for Fake Image Detection? SICA: Semantic-Induced Constrained Adaptation for Unified-Yet-Discriminative Artifact Feature Space Reconstruction

Bo Du, Xiaochen Ma, Xuekang Zhu, Zhe Yang, Chaoqun Niu, Jian Liu, Ji-Zhe Zhou

Sichuan University • Guangjian Team, Ant Group

ICML International Conference On Machine Learning

Background & Motivation

Fake Image Detection (FID) Challenges:

- FID, aiming at unified detection across four image forensic subdomains, is critical in real-world forensic scenarios.
- Compared with ensemble approaches, monolithic FID models are theoretically more promising, but consistently yield inferior performance in practice.

Table 1. **Overview of dominant artifacts across different domains.** The distinct **Dependencies** indicate that artifact strategies tailored for one domain are often non-transferable to others, highlighting the heterogeneous nature of artifacts. △ indicates artifacts that are conceptually transferable but barely work in other subdomains, while ✗ denotes conceptually non-transferable artifacts.

Domain	Dominant Artifacts	Dependency (Constraint)	Transfer?	Rep. Method
Deepfake	Blending Boundary	Face Swapping Mask	✗	Face X-ray (Li et al., 2020b)
	Frequency	Facial Region	△	F3-Net (Qian et al., 2020)
	Physiology	Human Pulse (rPPG)	✗	FakeCatcher (Ciftci et al., 2020)
	Landmark	Facial Structure	✗	CSH (Hu et al., 2021)
	Semantic Motion	Lip-Voice Sync	✗	LipForensics (Haliassos et al., 2021)
AIGC	Global Spectrum	Checkerboard Pattern	✗	CNN-Gen (Wang et al., 2020)
	Fingerprint	GAN/Gen. Architecture	✗	Freq-Analysis (Frank et al., 2020)
	Reconstruction	Diffusion Prior	✗	DIRE (Wang et al., 2023)
IMDL	Noise Residual	Camera Sensor (PRNU)	△	ManTra-Net (Wu et al., 2019)
	Edge	Boundary Inconsistency	✗	MVSS-Net (Chen et al., 2021)
Doc	Text Morphology	Font/Glyph Rendering	✗	DocTamper (Qu et al., 2023)

The Root Cause — Feature Space Collapse:

- **"Ji-Zhe phenomenon"**: We identify the intrinsic distinctness of artifacts across subdomains—a critical barrier we term the "Ji-Zhe phenomenon".
- Artifacts in FID are highly distinct across subdomains, projecting high-dimensional artifact features into a unified low-dimensional representation space leads to the collapse of the artifact feature space.
- The core challenge boils down to the "unified-yet-discriminative" reconstruction of the artifact feature space.

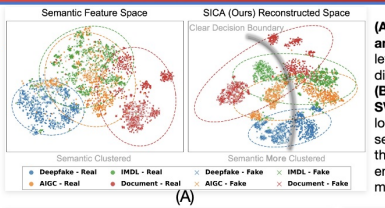
SICA Paradigm

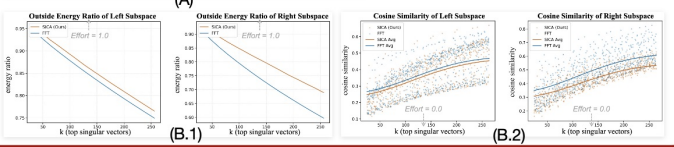
The Semantic Structural Hypothesis:

- We hypothesize that high-level semantics can serve as a structural prior for the reconstruction to address the paradoxical challenge.

Semantic-Induced Constrained Adaptation (SICA):

- **Semantic-Induced:** Employing a robust semantic manifold where inputs are naturally clustered by subdomain.
- **Constrained Adaptation**: Utilizing low-rank adaptation to selectively bridge the semantic distribution gap between the reference manifold and training data, while preserving the reference manifold.





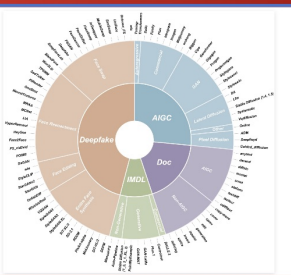
(A) t-SNE visualization of semantic feature space and SICA (ours) reconstructed space. SICA leverages semantics to reconstruct unified-yet-discriminative artifact feature space.
(B) Spectral analysis of left and right subspace via SVD. SICA exhibits higher outside energy ratio and lower cosine similarity with respect to the dominant semantic subspace, proving to learn less semantics, thereby reducing the risk of semantic overfitting and enabling better artifact learning. Please refer to the main text for a more detailed description.

OpenMMSec Benchmark

Constructing a New *OpenMMSec* Benchmark (Open-Sourced):

To conduct a systematic and fair study of FID, we construct ***OpenMMSec*** (***Open M***ulti-***M***edia ***Sec***urity), the first large-scale comprehensive FID dataset.

- **Scale:** Contains over 330K images.
- **Diversity:** Covers all 4 subdomains (Deepfake, AIGC, IMDL, Doc) with 15 primary faking types and 98 fine-grained image faking types.
- **Rich Sources**: Integrates 19 subdomain datasets and spans over 10 real-world datasets.



Experimental Results

Generalization Comparison:

SICA outperforms 15 state-of-the-art methods across all 4 evaluation metrics on *OpenMMSec*.

Table 2. **Accuracy results on *OpenMMSec*.** The overall average is the macro-average of the domain averages. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Deepfake					AIGC						IMDL				Doc			Avg
	EFS	FE	FR	FS	Avg	GAN	Lat-Diff	Pix-Diff	AR	Comm	Other	Avg	Gen	Non-Gen	Avg	AIGC	Non-AIGC	Avg	
Resnet (He et al., 2016)	71.3	65.4	70.0	64.7	67.8	77.4	65.2	75.7	78.5	72.7	75.5	74.2	80.3	74.2	77.3	61.9	81.5	71.7	72.7
EfficientNet (Tan & Le, 2019)	45.0	29.0	50.8	48.5	43.3	65.2	58.9	63.1	65.1	62.1	63.7	63.0	51.7	51.1	51.4	47.0	76.8	61.9	54.9
CapsuleNet (Nguyen et al., 2019)	63.4	58.7	66.5	59.6	62.0	79.2	63.1	77.5	81.3	73.2	77.8	75.4	79.2	75.3	77.2	59.8	84.8	72.3	71.7
SegFormer (Xie et al., 2021)	83.8	80.5	83.5	75.1	80.7	87.4	80.9	87.1	89.7	87.8	82.3	85.9	83.2	80.3	81.7	66.8	80.1	73.4	80.4
Swin (Liu et al., 2021b)	83.6	80.5	80.6	71.5	79.0	86.9	83.8	85.8	87.6	86.8	81.3	85.4	85.0	80.8	82.9	66.9	77.1	72.0	79.8
ConvNeXt (Liu et al., 2022)	80.5	84.0	83.9	72.4	80.2	89.0	79.8	87.9	92.0	89.6	83.4	87.0	83.6	80.2	81.9	65.0	75.3	70.1	79.8
Recce (Cao et al., 2022)	59.0	50.8	50.7	53.7	53.5	90.4	66.1	90.8	97.1	85.3	87.8	86.3	63.1	77.1	70.1	78.6	52.8	65.7	68.9
Sia (Sun et al., 2022)	64.0	63.5	73.1	63.7	66.1	81.6	68.7	80.2	84.1	79.2	80.4	79.0	71.9	63.8	67.9	64.4	83.4	**73.9**	71.7
IML-ViT (Ma et al., 2023)	66.6	64.9	74.6	66.0	68.0	83.3	70.6	81.8	86.1	81.5	80.9	80.7	81.5	78.9	80.2	68.9	78.5	73.7	75.7
Trufor (Guillaro et al., 2023)	73.7	72.7	77.1	65.4	72.2	83.7	77.9	83.0	86.0	84.0	80.3	82.5	82.7	78.3	80.5	58.1	86.6	72.3	76.9
UniFD (Ojha et al., 2023)	57.0	44.8	57.6	57.4	54.2	84.1	68.3	81.7	84.9	76.9	82.9	79.8	65.7	76.2	70.9	39.0	86.3	62.7	66.9
FFTIN (Chen et al., 2024)	73.7	75.5	81.2	69.5	75.0	93.6	89.2	93.9	96.1	94.5	86.8	92.4	84.7	80.6	82.6	64.8	75.3	70.0	80.0
Effort (Yan et al., 2024a)	87.1	87.2	80.7	85.0	85.0	86.6	71.3	84.6	86.3	78.8	83.9	81.9	81.7	85.7	83.7	61.1	78.1	69.6	80.1
Mesorch (Zhu et al., 2025a)	76.7	72.6	79.9	73.4	75.7	83.0	73.8	82.1	85.5	83.3	81.0	81.4	81.2	78.7	79.9	63.6	81.8	72.7	77.4
CO-SPY (Cheng et al., 2025)	75.1	71.3	77.0	65.5	72.2	84.7	80.8	84.2	86.2	85.7	78.5	83.3	77.8	74.2	76.0	61.9	78.0	70.0	75.4
CLIP-FFT	77.4	78.2	80.4	68.0	76.0	92.2	82.0	91.0	95.1	91.7	85.4	89.6	83.5	79.3	81.4	66.6	81.2	**73.9**	80.2
SICA (Ours)	91.3	89.4	86.5	86.6	**88.4**	96.3	94.5	94.7	96.7	94.8	86.9	**94.0**	85.4	85.1	**85.3**	69.2	78.4	73.8	**85.4**

Feature Space Reconstruction Analysis:

SICA reconstructs the target unified-yet-discriminative artifact feature space in a near-orthogonal manner, thus firmly validating our hypothesis.

Table 3. Cross-domain performance comparison under different training strategies. Rows indicate training domains and columns indicate test domains. The last row reports unified training results.

Train	Deepfake	AIGC	IMDL	Doc
Deepfake	**0.8183**	0.7639	0.4605	0.3488
AIGC	0.5704	**0.9291**	0.4161	0.5263
IMDL	0.5862	0.4706	**0.8535**	0.7870
Doc	0.5604	0.4474	0.4843	**0.8657**
Unified	0.7900	0.8987	0.8223	0.8080

(a) FFT-SD

Train	Deepfake	AIGC	IMDL	Doc
Deepfake	**0.9235**	0.6497	0.5612	0.4455
AIGC	0.5496	**0.9623**	0.4205	0.6425
IMDL	0.8093	0.5869	**0.8817**	0.6980
Doc	0.4773	0.5032	0.5467	**0.8893**
Unified	0.9234	0.9572	0.8804	0.8646

(b) SICA-SD

Train Wo.	Deepfake	AIGC	IMDL	Doc
Deepfake	**0.7660**	0.9642	0.8662	0.8807
AIGC	0.9191	**0.6491**	0.8730	0.8622
IMDL	0.9096	0.9541	**0.4674**	0.8575
Doc	0.9221	0.9532	0.8812	**0.7081**
Unified	0.9234	0.9572	0.8804	0.8646

(c) SICA-LODO

Ablation study:

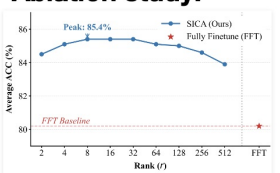


Figure 8. **Ablation study of rank.**

Table 4. **Ablation of pretrain backbones.** The evaluation metric used is accuracy (ACC).

Backbone	Pretrained	Deepfake	AIGC	IMDL	Doc	Avg
Resnet-50	ImageNet-1K	67.8	74.2	77.3	71.7	72.7
Resnet-50	CLIP	73.6	88.1	79.7	71.8	78.3
ViT-L/14	DINOv2	86.8	86.4	83.8	74.7	82.9
ViT-L/16	SigLIP	81.7	88.8	83.9	74.4	82.2
ViT-L/14	CLIP	88.4	94.0	85.3	73.8	85.4

Attention Map Visualization:

Figure 11. **Attention map of fake images across four domains.**

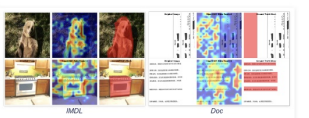


Figure 12. **Attention map of fake images in IMDL and Doc with corresponding regions.**

Robustness Experiment:

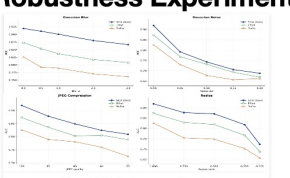


Figure 10. **Robustness experiment.**