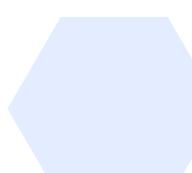
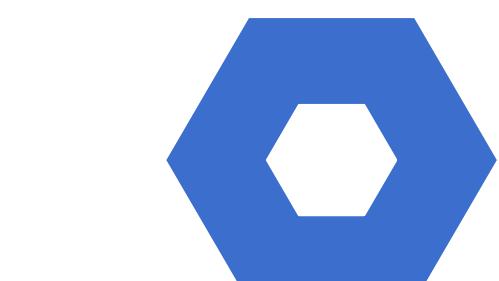
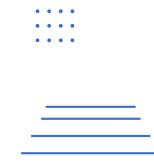
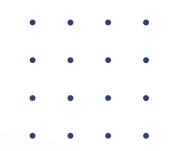
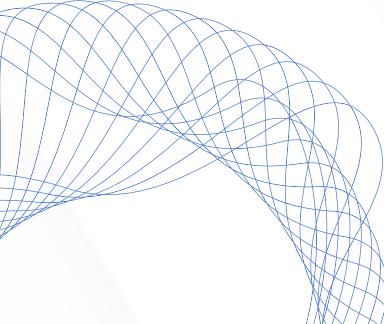


RESEARCH PROJECT

Kanokporn Pakdeenual

675020069-9



4. Result

4.1

Datasets and Metrics

4.2

Ablation Study

4.3

SoTA Comparison

4.4

Limitations

4.1 Datasets and Metrics

- **Datasets**

The **29,000 synthetic images** were generated using 19 different generative models, including



4.1 Datasets and Metrics

- **Datasets**

The **6,000 real images** used to evaluate the model were collected from high-quality image datasets, including the following sources:

LSUN

FFHQ

ImageNet

COCO

LAION

RAISE

4.1 Datasets and Metrics

Real images are carefully matched in semantic content and compression quality to synthetic ones

- **Datasets**

real images

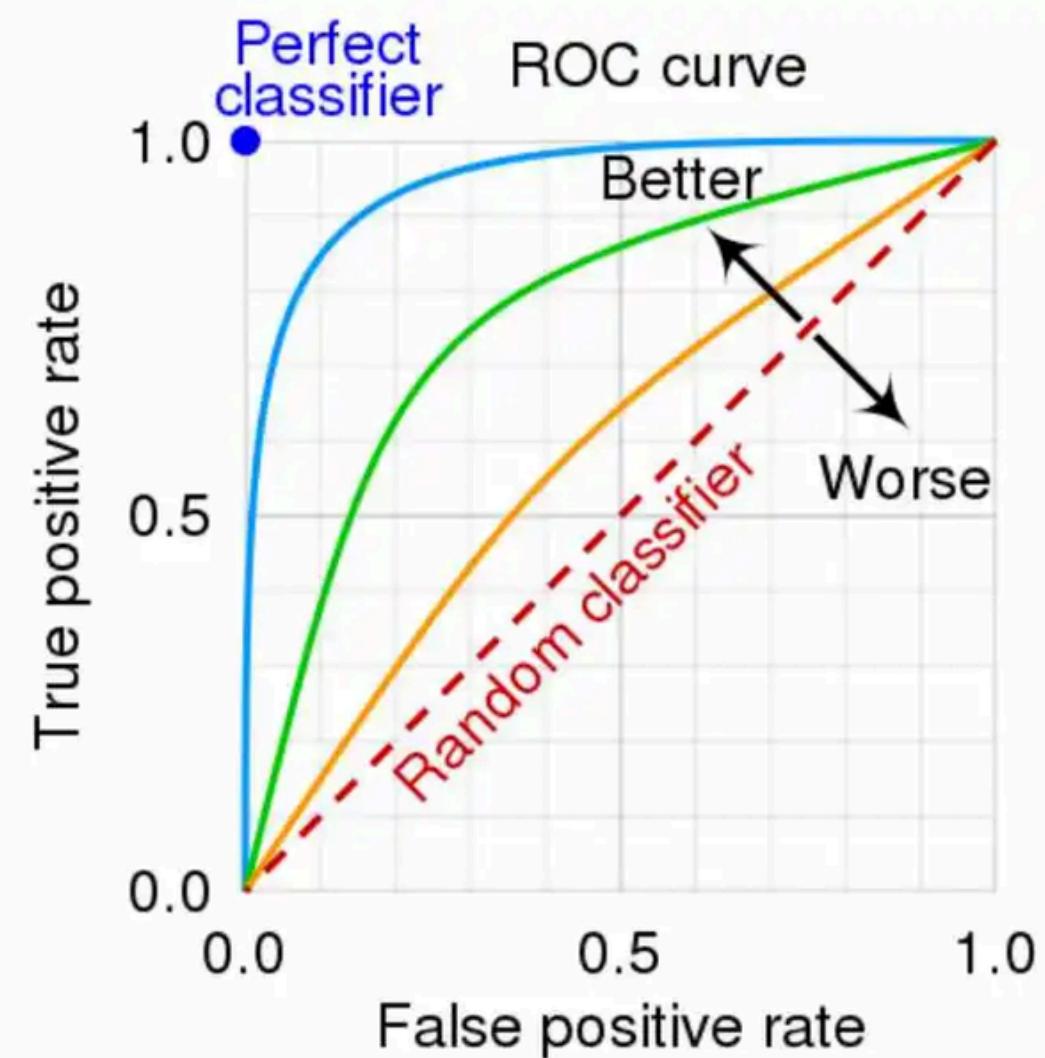


Fig. 4: Examples of real and AI-generated images of different categories used in our experiments. Top: real images from LSUN, FFHQ, ImageNET and COCO. Bottom: generated images from DiffusionGAN, StyleGAN2, DiT and SDXL.

4.1 Datasets and Metrics

- **Evaluation Metrics**

- **AUC (Area Under ROC Curve):** Measures the model's ability to distinguish between real and AI-generated images. The closer to 100%, the better the model's performance. Commonly used in binary classification task



4.1 Datasets and Metrics

• Evaluation Metrics

- **Balanced Accuracy (bACC):** Used when the number of real and fake images is imbalanced. Calculated as the average of true positive rates for each class.
→ Ensures that both real and fake image detection are fairly evaluated.

4.1 Datasets and Metrics

• Evaluation Metrics

- **Threshold Analysis:** The study also analyzes how different decision thresholds affect performance. Threshold = the value used to decide whether an image is “real” or “fake”. This helps assess the model’s robustness and stability in real-world scenarios

4.2 Ablation Study

ในงานวิจัยนี้ ค่า **Entropy (H)** และ **NLL (Negative Log-Likelihood)** ถูกใช้เพื่อวัดความยากในการเข้ารหัสภาพผ่านโมเดลบีบอัด โดยมีเป้าหมายเพื่อแยกแยะระหว่างภาพจริงและภาพปลอมผ่านความซับซ้อนและความสามารถในการคาดเดาของภาพแต่ละภาพ

Entropy (H) – เอนโทรปี

เป็นการวัดว่า "ข้อมูลนี้มีความคาดเดาได้ยากแค่ไหน" ถ้าภาพมีความซับซ้อนมาก (รายละเอียดเยอะคาดเดายาก) → ค่า Entropy จะ สูง ถ้าภาพเรียบง่าย (เช่น พื้นหลังเรียบ ๆ) → ค่า Entropy จะ ต่ำ เช่น ภาพที่เต็มไปด้วยรายละเอียด เช่น ภาพถนนในเมืองตอนกลางคืน จะมี Entropy สูงกว่าภาพพื้นหลังขาว ๆ

NLL (Negative Log-Likelihood)

ต้นทุนการเข้ารหัส

เป็นการวัดว่า "ถ้าโมเดลพยายามคาดเดาค่าของแต่ละ pixel ในภาพนี้ มันจะยากแค่ไหน" ถ้า โมเดลสามารถคาดเดาได้ดี (ภาพดูเป็นธรรมชาติ) → ค่า NLL จะ ต่ำ ถ้าโมเดลคาดเดาไม่เก่ง (ภาพเปลก ๆ ปลอม ๆ) → ค่า NLL จะ สูง พูดง่าย ๆ ก็คือ NLL คือค่าความยากในการบีบอัดหรือคาดเดาภาพ

4.2 Ablation Study

เพื่อวิเคราะห์ว่า **ตัวชี้วัด (features)** ใดมีผลมากที่สุดต่อความสามารถของโมเดลในการแยกแยะภาพจริงกับภาพที่สร้างโดย AI ซึ่งมีตัวชี้วัดดังต่อไปนี้

- $D^{(0)}$ (the 0-level coding cost gap),
- its slope $\Delta^{01} = D^{(0)} - D^{(1)}$,
- and their absolute values.

4.2 Ablation Study

1. $D^{(0)}$ = O-level coding cost gap

คือ ความแตกต่างของต้นทุนการเข้ารหัส ระหว่างภาพจริงกับภาพปลอม ที่ระดับ 0 (ภาพที่ไม่มีการบีบอัด) เช่น ถ้าโมเดลใช้ NLL กับภาพจริงได้ค่าเท่ากับ 2.0 แต่กับภาพปลอมได้ 3.5 $\rightarrow D^{(0)} = 3.5 - 2.0 = 1.5$ ถ้า $D^{(0)}$ สูง \rightarrow แสดงว่าโมเดล “รับรู้” ความแตกต่างระหว่างภาพจริงกับปลอมได้ชัด ถ้า $D^{(0)}$ ใกล้ 0 \rightarrow ภาพจริงกับปลอมแยกกันไม่ออกร

4.2 Ablation Study

2. $\Delta^{01} = D^{(0)} - D^{(1)}$

คือ ความชันของช่องว่างต้นทุนการเข้ารหัส ระหว่างระดับ 0 กับระดับ 1 หมายถึง ความต่างระหว่าง
ภาพจริง-ปلوم ลดลงหรือเพิ่มขึ้น เมื่อความละเอียดลดลง

ตัวอย่าง:

- $D^{(0)} = 1.5$
- $D^{(1)} = 0.5$
- $\Delta^{01} = 1.5 - 0.5 = 1.0 \rightarrow$ แสดงว่าระดับ 0 แยกແຍະได้ดีกว่าชัดเจน

4.2 Ablation Study

3. Absolute values

การนำเอาค่าที่ได้จาก $D^{(0)}$ หรือ Δ^{01} มาทำให้เป็นบวกหมด เช่น

- ถ้า $D^{(0)} = -2.3 \rightarrow \text{Absolute} = 2.3$
- ใช้เพื่อถูกว่า "มาน้อยแค่ไหน" โดยไม่สนใจว่าบวกหรือลบ

4.2 Ablation Study

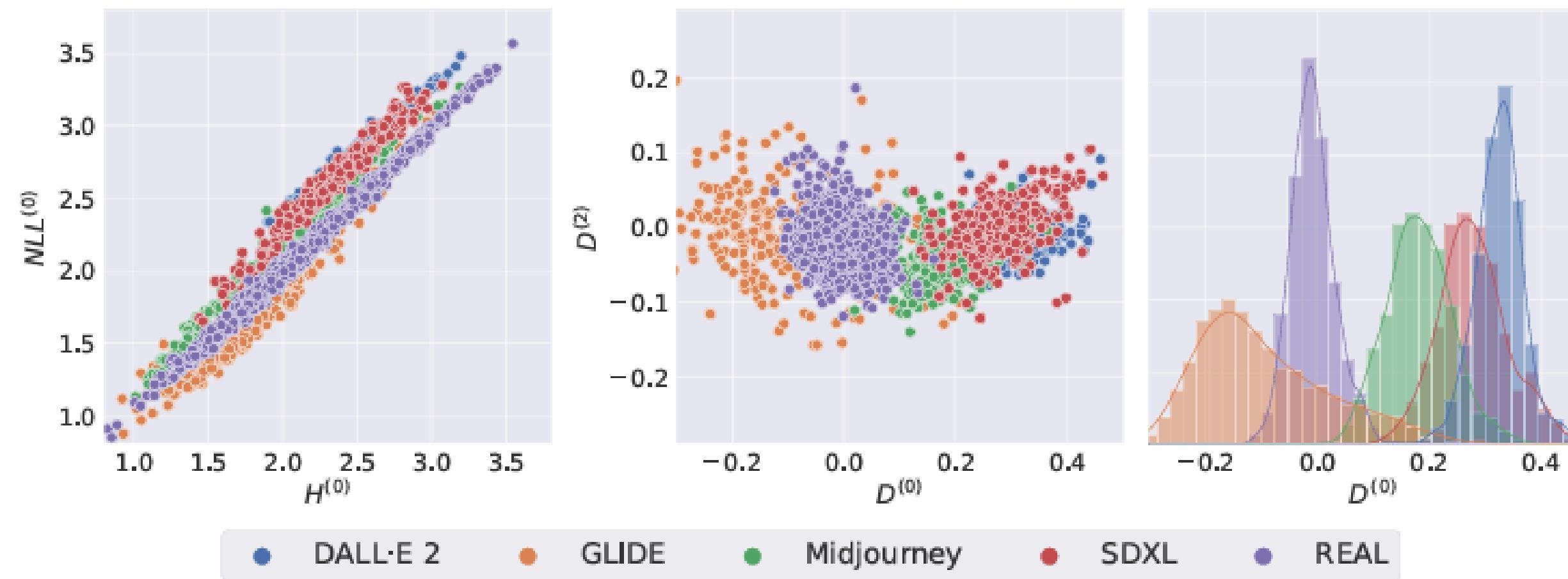


Fig. 5: Decision statistics. NLL and entropy by themselves are not discriminant (left). Their difference (center) is much more useful for detection, but only at high resolution, $D^{(0)}$, while $D^{(1)}$ is less discriminant and $D^{(2)}$ basically useless. Right box shows histograms of $D^{(0)}$ for real and synthetic images. Note that for GLIDE, $D^{(0)}$ is negative, on the average. Good discrimination is still possible based on the absolute value.

4.2 Ablation Study

Influence of the real class

เพื่อศึกษาว่าการเลือกชุดข้อมูลภาพจริง (real images) ที่ใช้ฝึกโมเดล lossless encoder มีผลต่อประสิทธิภาพของการตรวจจับภาพปลอมจาก AI หรือไม่

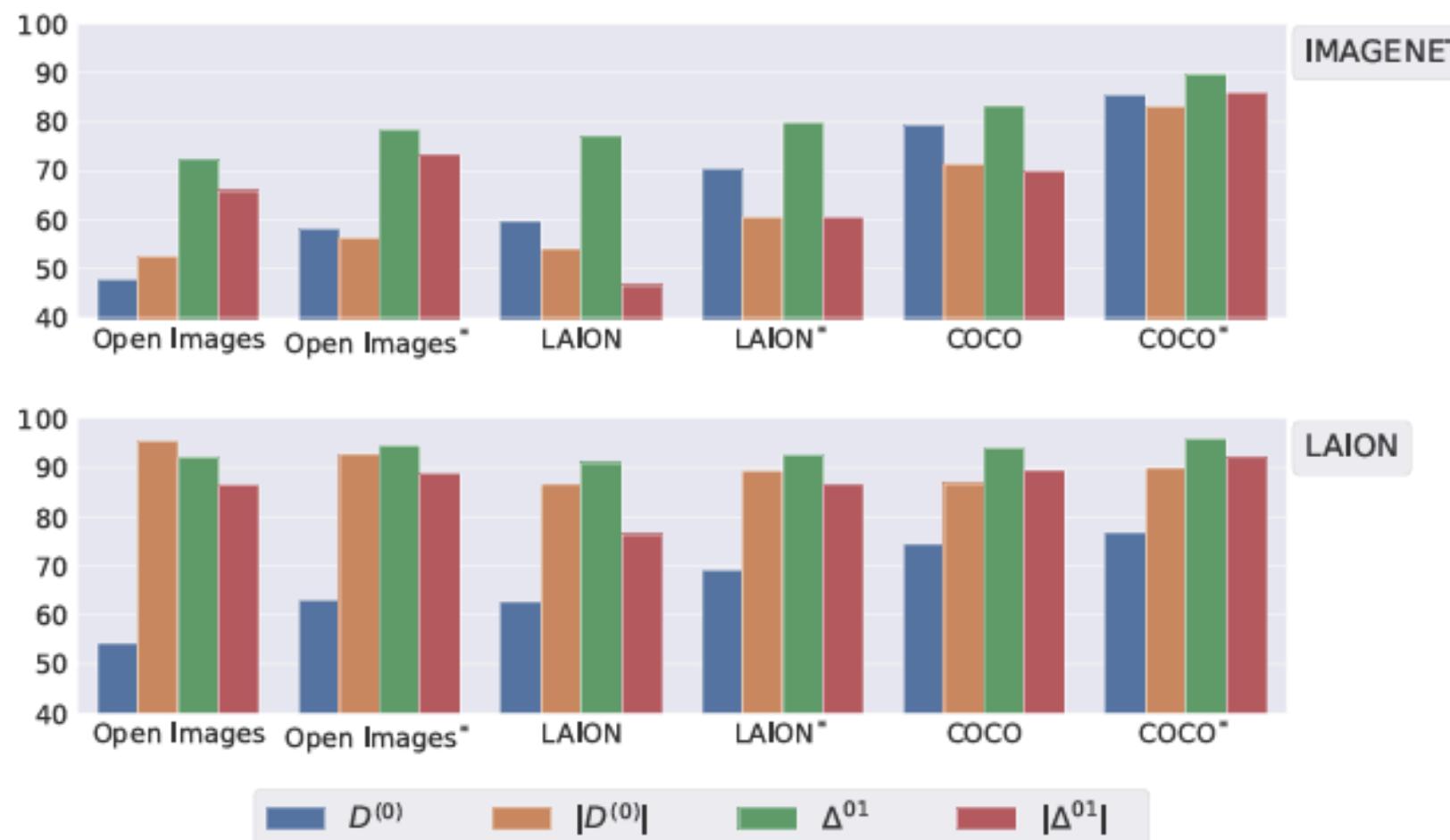


Fig. 6: AUC of proposed method as a function of decision statistic (see Section 3.4) and dataset of real images used to train the lossless encoder: Open Images, LAION, COCO, and their augmented versions (*). Synthetic test images are selected to match the corresponding real test images: ImageNet (top), and LAION (bottom).

4.2 Ablation Study

ตารางแสดงรายละเอียดของวิธีการตรวจจับภาพปลองจากการวิจัยก่อนหน้า โดยเปรียบเทียบในด้านแนวคิด วิธีท伦 (ภาพจริง/ภาพปลองที่ใช้), ขนาดข้อมูล, การใช้การเพิ่มข้อมูล (augmentation) และกลยุทธ์ในการทดสอบ ซึ่งจะใช้เป็นพื้นฐานในการเปรียบเทียบกับวิธีใหม่ของผู้วิจัย เพื่อประเมินว่าแนวทางที่นำเสนอสามารถให้ผลลัพธ์ที่ดีกว่าเพียงใดในบริบทที่หลากหลาย

Table 1: Reference methods. For each one we indicate the key idea, the datasets of real and synthetic images used for training with their sizes, whether or not augmentation is used, the test strategy.

Acronym [ref]	Idea/Approach	Training	Real/Fake	Size(K)	Augment.	Test Strategy
Wang2020 [74]	High diversity	LSUN/ProGAN	360/360	✓	global pooling	
PatchFor. [7]	Patch-based	CelebA,FF/various	84/272		resizing	
Liu2022 [43]	Noise-based	LSUN/ProGAN	360/360	✓	global pooling	
Corvi2023 [10]	No-downsampling	COCO,LSUN/Latent	180/180	✓	global pooling	
LGrad [72]	Gradient-based	LSUN/ProGAN	72/72	✓	resizing	
DIRE [75]	Inversion	LSUN-Bed/ADM	40/40		resizing	
DE-FAKE [67]	Prompt-based	LSUN/Stable Diff.	20/20		resizing	
Ojha2023 [51]	CLIP	LSUN/ProGAN	360/360	✓	cropping	
NPR [71]	Residual	LSUN/ProGAN	72/72		resizing	
AEROBLADE [60]	AE rec. error	- / -	- / -		global distance	

real images, where this latter class comes from ImageNet [15] (Fig.6, top) or LAION [66] (Fig.6, bottom). We can observe that the best and more uniform results across the four decision statistics are obtained using COCO*, while training on Open Images guarantees good performance if the real class is LAION, but bad performance if it is ImageNet. Additional results are included in the supplementary material.

4.3 SoTA Comparison

เปรียบเทียบวิธีใหม่กับ SoTA “State of the Art” ที่มีอยู่ในงานวิจัยก่อนหน้าทั้งหมด 10 วิธี วิธีใหม่ของผู้วิจัยให้ผล AUC เฉลี่ยเกิน 80% ทุกกรณี โดยสูงสุดถึง 90% แสดงให้เห็นว่าวิธีที่นำเสนอมานี้มีประสิทธิภาพมากกว่าวิธีอื่น ๆ โดยเฉพาะเมื่อต้องรับมือกับข้อมูลที่เปลี่ยนไป

Table 2: AUC for reference and proposed methods. Best score in bold with a 0.5% margin. S = LSUN, F = FFHQ, I = ImageNet, C = COCO, L = LAION, R = RAISE.

	Real data	Wang2020	PatchFor.	Liu2022	Corvi2023	LG Grad	DIRE	DEFAKE	Ojha2023	NPR	AEROBLADE	Ours $D^{(0)}$	Ours $ D^{(0)} $	Ours $\Delta^{(1)}$	Ours $ \Delta^{(1)} $
GauGAN	C	98.9	80.8	99.7	83.8	81.6	99.9	43.8	100.	89.1	55.1	99.8	99.8	99.9	99.7
BigGAN	I	92.7	85.5	94.7	83.4	77.2	99.8	59.0	99.6	86.8	51.9	92.3	88.6	95.9	92.6
StarGAN	F	94.7	100.	99.9	95.9	73.9	40.4	45.9	99.7	81.5	84.0	100.	100.	100.	100.
StyleGAN2	S	98.1	83.8	99.7	89.1	99.8	58.3	39.1	96.7	100.	30.0	96.6	96.1	96.7	96.5
	F	94.9	85.1	99.9	58.4	82.7	55.5	47.6	91.0	71.3	60.1	43.1	87.7	41.1	88.7
GigaGAN	I	73.7	61.0	97.3	50.5	76.4	99.9	64.3	94.6	82.4	47.5	72.4	68.1	72.4	68.1
	C	79.5	84.0	99.6	90.9	76.7	99.9	87.9	97.6	95.5	80.6	96.5	94.0	96.7	93.4
Diff.GAN	S	89.8	92.6	99.5	96.6	99.5	49.8	44.8	97.4	100.	43.9	99.4	99.4	99.5	99.5
GALIP	C	89.7	98.2	94.3	87.7	56.7	100.	75.6	98.6	90.7	65.0	98.4	96.3	99.7	99.4
DALL-E	L	66.4	71.7	95.0	98.3	95.2	99.8	55.9	97.3	99.5	24.1	99.2	95.8	98.2	95.4
DDPM	F	31.6	98.4	22.8	100.	9.8	23.1	50.5	77.7	92.4	81.7	76.6	25.2	93.8	79.6
	S	67.6	67.6	70.6	80.3	81.1	52.0	37.4	88.2	94.1	53.1	49.5	53.5	69.4	71.0
ADM	I	61.0	81.9	94.4	81.1	72.7	99.5	69.1	85.3	78.5	80.3	87.8	90.5	95.3	92.1
	C	64.8	97.4	96.3	97.2	81.5	99.9	92.4	88.8	95.4	98.0	47.8	88.5	91.1	91.1
GLIDE	R	32.2	95.0	56.6	86.5	50.6	42.9	92.2	72.8	63.3	87.7	23.2	89.4	51.1	65.1
	L	72.6	74.1	90.8	86.9	90.3	100.	60.2	95.3	99.8	68.7	54.5	84.2	93.8	88.5
DiT	I	58.6	83.1	88.0	100.	56.2	99.6	87.4	77.8	78.4	99.8	89.4	84.3	94.9	91.0
Stable D. 1.4	C	68.2	86.1	95.3	100.	54.7	99.9	93.3	97.9	76.5	99.8	48.4	74.8	54.6	71.4
	R	37.9	61.8	73.4	100.	50.0	37.6	88.0	87.7	43.0	96.9	99.4	98.7	97.0	97.2
Stable D. 2	C	56.5	78.6	94.2	100.	62.8	99.3	97.9	82.3	89.3	99.9	83.0	90.3	84.5	89.1
	R	50.2	38.7	34.8	100.	41.4	35.5	80.7	89.5	44.0	97.4	98.5	96.8	95.8	95.9
SDXL	C	83.8	60.8	89.3	100.	89.3	99.5	94.0	80.0	99.3	87.9	99.9	99.9	99.9	99.8
	R	54.3	68.4	31.1	100.	57.2	47.1	84.4	85.1	76.7	69.7	100.	100.	99.1	99.2
Deep.-IF	C	78.0	62.7	72.2	99.9	68.8	98.9	96.9	92.9	91.6	81.9	91.7	82.3	88.4	79.4
DALL-E 2	C	88.5	52.4	98.9	88.2	78.6	99.9	80.6	97.1	90.0	59.3	100.	100.	100.	99.9
	R	64.8	41.9	70.4	69.4	58.6	44.7	70.9	95.2	39.5	32.8	100.	100.	100.	100.
DALL-E 3	C	65.0	47.3	99.5	100.	88.4	99.9	96.2	86.4	97.7	99.7	99.7	99.5	98.3	98.2
	R	10.9	52.7	0.2	60.8	37.9	47.6	92.4	36.4	48.7	48.3	79.1	66.7	78.0	78.1
Midjourney	R	40.2	57.8	40.7	100.	56.3	51.0	78.1	66.2	77.0	99.0	99.7	99.3	98.5	98.5
Adobe Firefly	R	84.8	49.4	11.8	98.0	40.6	57.4	81.4	97.5	32.1	52.8	73.6	41.2	80.8	80.4
AVG		68.3	73.3	77.0	89.4	68.2	74.6	72.9	88.4	80.1	71.2	83.3	86.4	88.8	90.0

4.3 SoTA Comparison

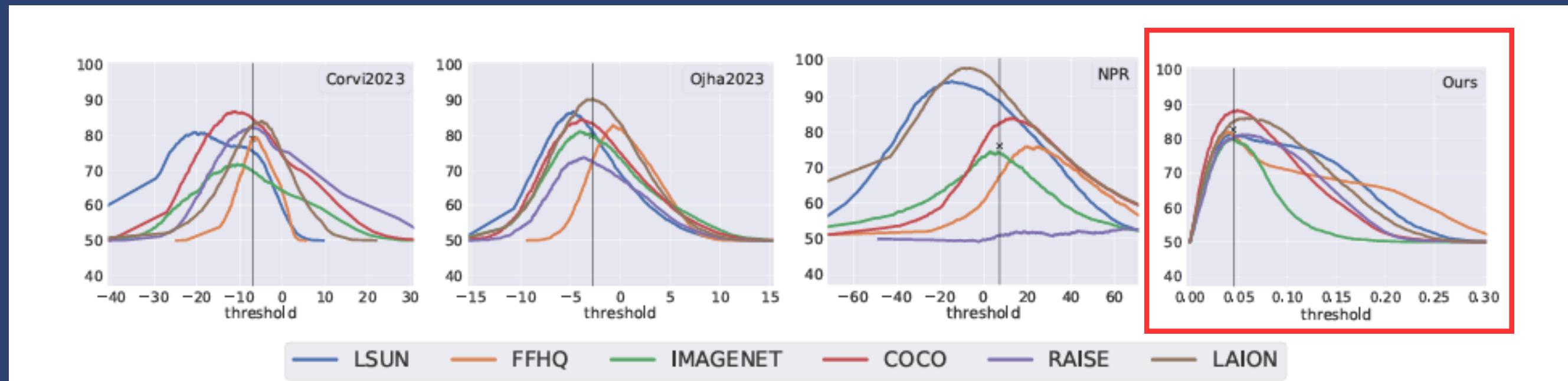


Fig. 7: Balanced accuracy as a function of the detection threshold. For each dataset of real images, we average accuracy over all associated synthetic generators. The dotted vertical line indicates the global optimal threshold and the \times symbol the corresponding accuracy. Note that only for the proposed method all peaks are very close, indicating the presence of a single threshold. Charts for other methods are reported in the Suppl.

กราฟแสดง Balanced Accuracy เมื่อเปลี่ยนค่าการตัดสิน (threshold) ของแต่ละวิธี โดยวิธีที่นำเสนอ มีจุดสูงสุด ของแต่ละเส้นอยู่ใกล้กันมากที่สุด แสดงถึงความเสถียรของวิธีที่ไม่ขึ้นกับชุดข้อมูลจริงที่ใช้แทน สามารถเลือก threshold ค่าเดียวเพื่อใช้กับทุกชุดข้อมูลได้ ตรงข้ามกับวิธีอื่นที่ต้องปรับ threshold ใหม่ในแต่ละกรณีเพื่อให้ได้ผลดีเหมาะสมสำหรับการใช้งานจริง เพราะลดขั้นตอนการปรับแต่งก่อน deployment

4.4 Limitations

วิธีที่นำเสนอถูกออกแบบมาเพื่อแยกภาพที่สร้างขึ้นทั้งหมด ไม่ใช่เพื่อตรวจจับการดัดแปลงเฉพาะจุด สามารถขยายผลเพื่อตรวจจับเฉพาะจุดได้ เนื่องจากมีการคำนวณค่าสถิติรายพิกเซลอยู่แล้ว อิงกับโมเดลของภาพจริงที่เรียนรู้โดย encoder หากภาพจริงไม่ตรงตามโมเดลนี้ ผลลัพธ์อาจคลาดเคลื่อน ภาพที่ถูกบีบอัดหรือปรับขนาดมากเกินไปอาจทำให้การวิเคราะห์สถิติผิดพลาดได้ โดยเฉพาะภาพจากเว็บซึ่งมักมี artifact หรือคุณภาพไม่สม่ำเสมอ

THANK YOU