VRIJE UNIVERSITEIT BRUSSEL

ICIP abu dhabi 2024

ON THE DETECTION OF IMAGES GENERATED FROM TEXT

2024.10.29

Authors: Yuqing Yang
Charuka Moremada
Nikolaos Deligiannis
Presenter: Yuqing Yang
(yuqing.yang @vub.be)

CONTENT

- **Motivation & Related Work**
- **Proposed Method**
- **Experiments & Results**
- **Conclusion**

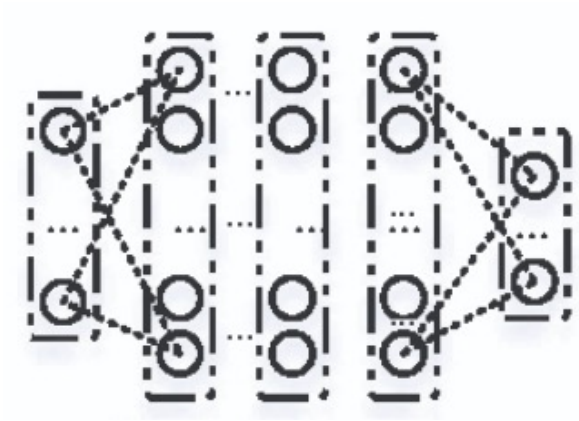
MOTIVATION

- **Advanced Generative Model**
  - Text-to-Image generation models
  - Stable Diffusion (SD), Latent Diffusion (LD), GLIDE, and DALL·E
- **Large Scale Dataset**
  - Images > corresponding prompts
  - Extensive datasets

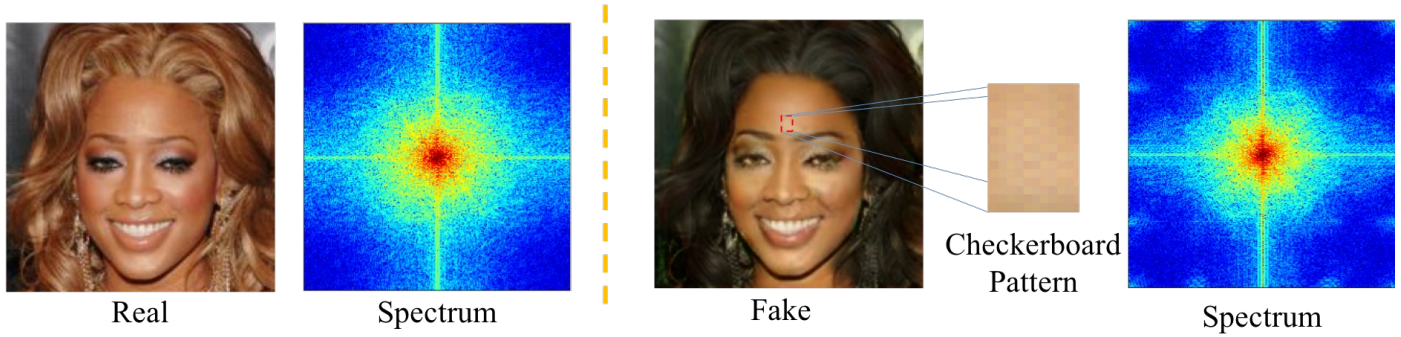
'A street sign that reads "Latent Diffusion" '

RELATED WORK

- **Synthetic Image Detection Method**
  - Deep Feature Methods
  - Frequency Features Methods



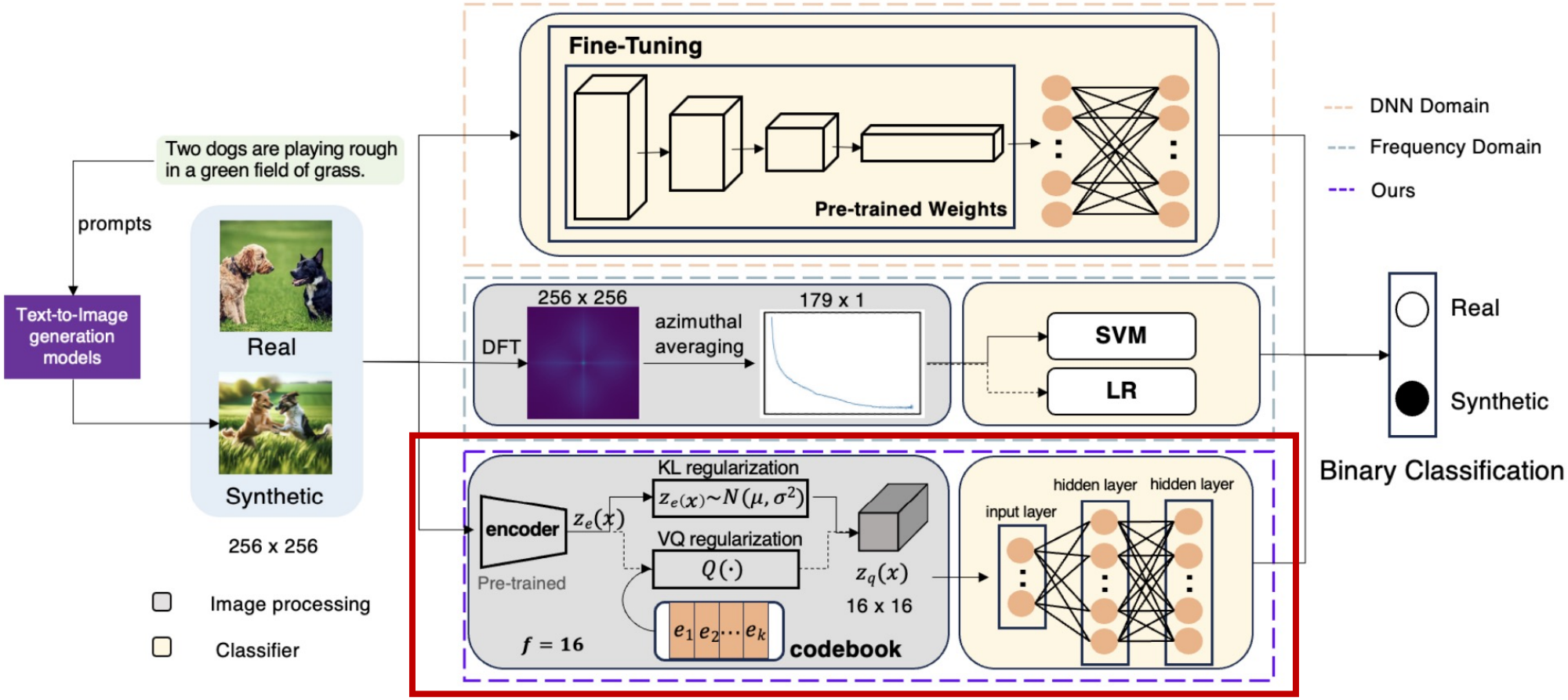
DNN Classifier



Frequency Analysis

PROPOSED METHOD

- **Latent Space Features**



EXPERIMENTAL SETTINGS

- **Used Datasets**

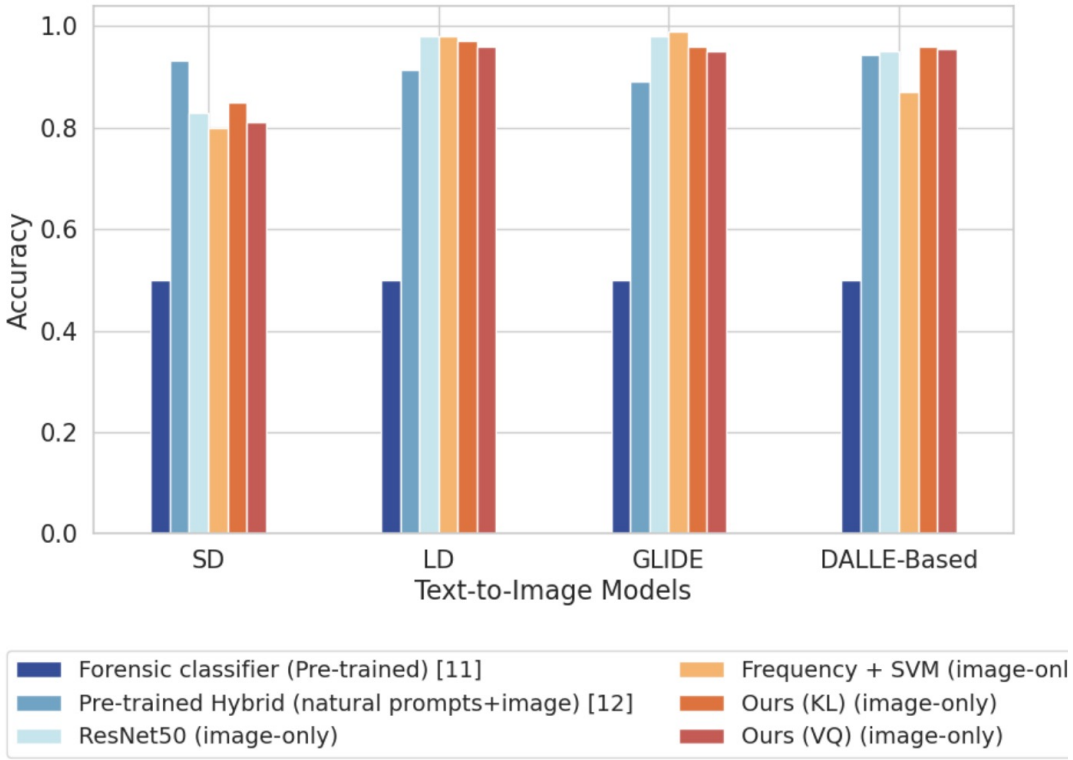
Real dataset	# of samples	Generative model	# of samples	Generated image size
MSCOCO	8000	SD [6]	2000	512 × 512
		LD [6]	2000	256 × 256
		GLIDE [5]	2000	256 × 256
		DM [7]	2000	256 × 256
Flickr30k	8000	SD [6]	2000	512 × 512
		LD [6]	2000	256 × 256
		GLIDE [5]	2000	256 × 256
		DM [7]	2000	256 × 256
Total	Real samples: 16000		Synthetic samples:16000	

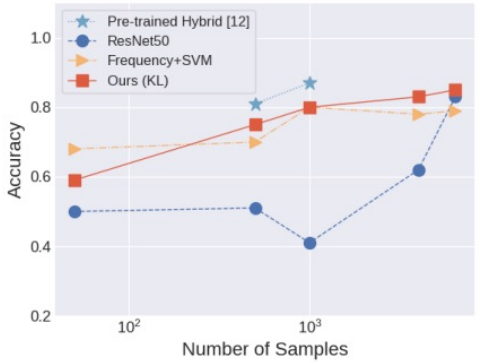
- **Used Classifiers**

Classifier	input size	# param.	MACs
Logistic Regression	179×1	180	180
SVM	179×1	$N_{sv}+1$	$N_{sv}\times 179$
ResNet50	256×256	25.56M	5.40G
VGG16	256×256	138.36M	20.24G
Hybrid [12]	512×1	0.59M	0.60M
Ours	16×16	0.09M	1.13M

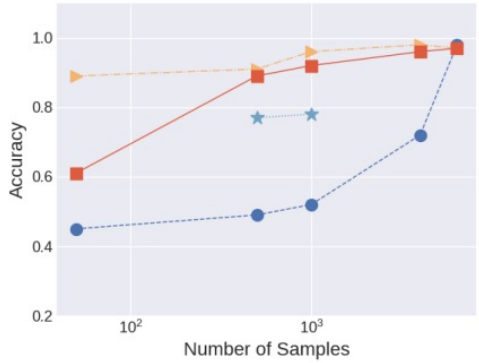
RESULTS

- **Detection Performance**

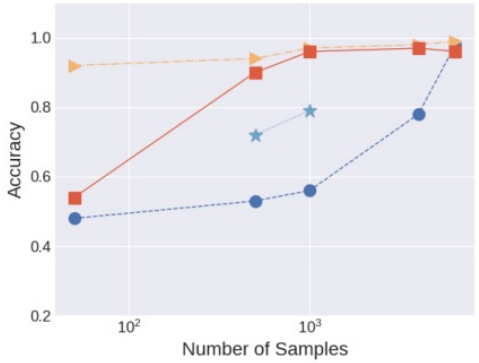




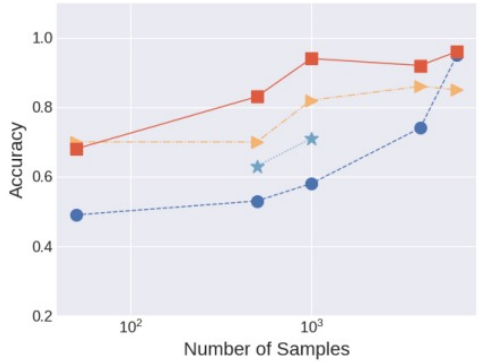
(a) Stable Diffusion (SD)



(b) Latent Diffusion (LD)



(c) GLIDE



(d) DALL·E-MINI (DM)

RESULTS

• Perturbation Test

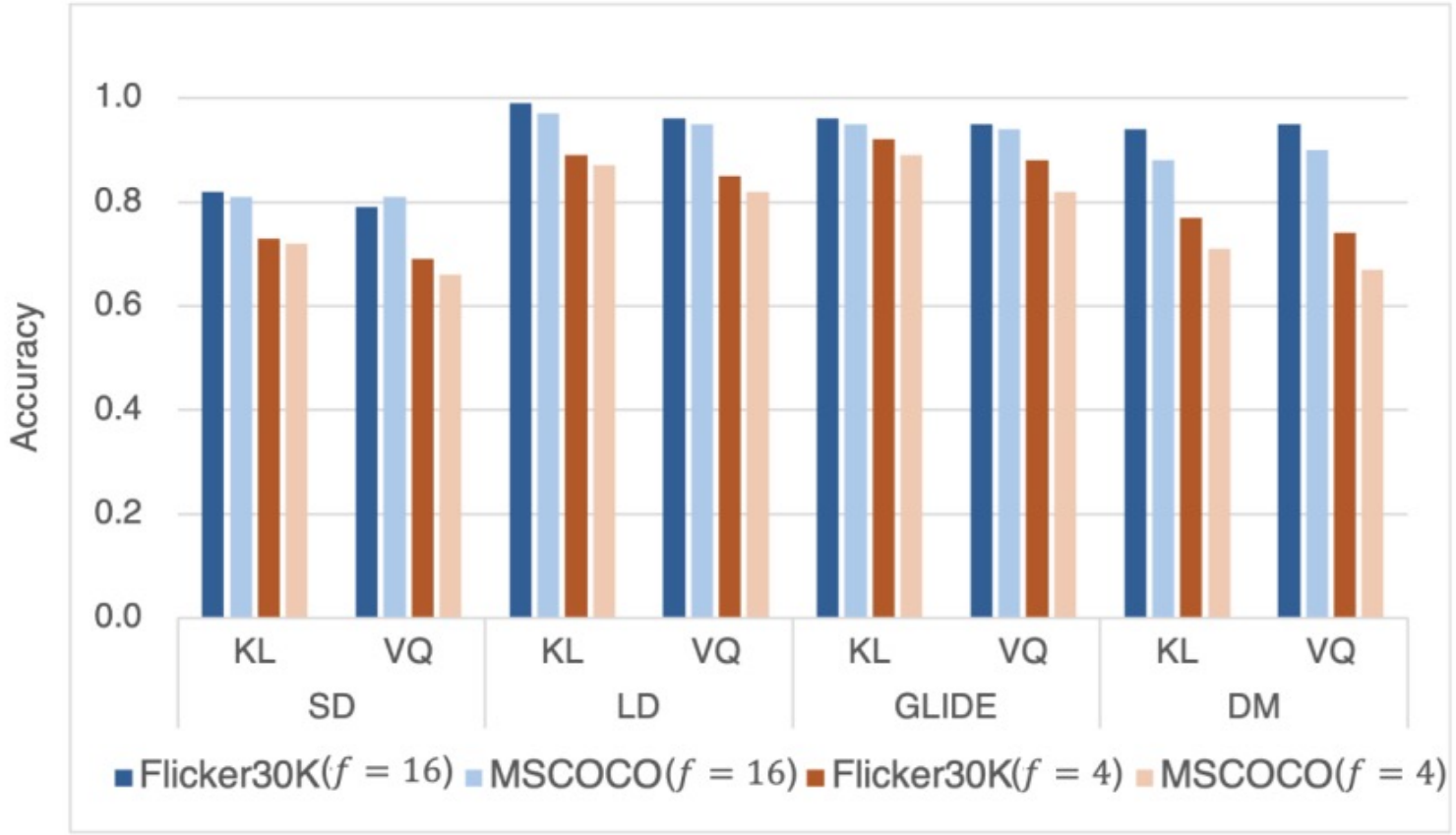
Training Dataset	Deep Features Method					Frequency Features Method					Latent Space Features Method				
	Classifier	Pertubation Methods				Classifier	Pertubation Methods				Regularization	Pertubation Methods			
		GB	GN	MB	SnP		GB	GN	MB	SnP		GB	GN	MB	SnP
SD	ResNet50	7.5	6.3	9.0	7.4	LR	36.5	34.9	22.2	28.1	KL	13.4	39.4	8.1	11.1
	VGG16	3.9	5.1	6.3	5.0	SVM	23.8	33.5	32.4	16.8	VQ	**2.2**	**3.4**	**2.0**	**0.8**
LD	ResNet50	16.5	16.7	15.9	14.6	LR	53.8	48.1	46.3	31.8	KL	6.7	17.6	8.2	19.3
	VGG16	18.6	17.2	22.2	**12.6**	SVM	47.7	42.6	16.2	37.3	VQ	**4.4**	**10.0**	**4.9**	16.25
GLIDE	ResNet50	21.8	20.7	19.5	16.4	LR	44.5	49.2	14.4	47.6	KL	36.8	46.1	34.5	40.6
	VGG16	21.5	**18.9**	21.4	**14.9**	SVM	43.0	49.2	**12.0**	46.8	VQ	**16.8**	39.8	21.8	30.5
DM	ResNet50	18.1	17.2	19.6	9.4	LR	72.2	41.3	39.2	32.5	KL	6.4	1.9	12.4	11.9
	VGG16	11.3	12.2	16.1	**2.1**	SVM	34.3	38.4	42.9	30.3	VQ	**3.3**	**0.3**	**7.1**	9.2

• Generalization Performance

Training Dataset	Deep Features Method					Frequency Features Method					Latent Space Features Method				
	Classifier	Test Datasets				Classifier	Test Datasets				Regularization	Test Datasets			
		SD	LD	GLIDE	DM		SD	LD	GLIDE	DM		SD	LD	GLIDE	DM
SD	ResNet50	-	87.5	79.4	77.4	LR	-	87.9	78.4	78.1	KL	-	89.4	77.8	76.4
	VGG16	-	90.3	73.9	75.8	SVM	-	73.8	89.7	79.2	VQ	-	74.5	82.4	77.5
LD	ResNet50	53.4	-	92.4	85.2	LR	58.2	-	95.4	63.8	KL	57.4	-	85.4	65.4
	VGG16	62.8	-	84.5	71.5	SVM	54.8	-	99.2	57.0	VQ	60.8	-	74.2	63.5
GLIDE	ResNet50	51.5	86.6	-	80.0	LR	49.8	86.8	-	54.8	KL	54.8	91.4	-	62.8
	VGG16	53.7	84.8	-	66.5	SVM	48.6	86.9	-	54.1	VQ	55.6	82.4	-	59.2
DM	ResNet50	58.4	96.4	95.4	-	LR	76.6	88.7	79.8	-	KL	56.4	94.3		-
	VGG16	65.3	93.8	91.0	-	SVM	76.8	91.7	82.8	-	VQ	59.2	72.4	64.2	-

RESULTS

- **Latent Representation Size**



CONCLUSION

- Compared the performance of several detectors in distinguishing synthetic images generated by various text-to-image generation models from real images.
- A lightweight detector based on latent space features, specifically designed for identifying synthetic images generated using text-to-image generation models.
- When the proposed model is trained on synthetic images generated by the *SD method*, it demonstrates good generalization properties in terms of detecting fake images produced by various text-to-image generation models.