ForgeLens: Data-Efficient Forgery Focus for Generalizable Forgery Image Detection

Yingjian Chen, Lei Zhang, Yakun Niu*.

Henan Key Laboratory of Big Data Analysis and Processing,

School of Computer and Information Engineering, Henan University

Introduction

Background:

- *AIGC (GANs & Diffusion) → produce realistic fake images*
- *"Seeing is believing" no longer holds → societal risks*
- *Detecting synthetic images is urgent*

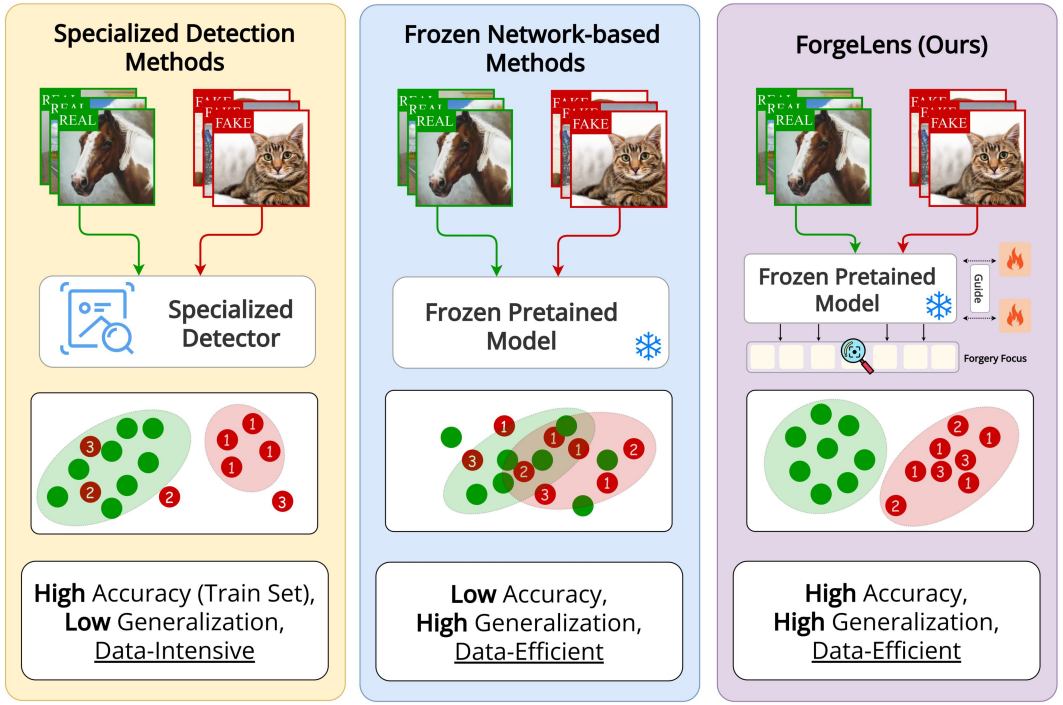
A key challenge:

- ***Specialized detectors**: High accuracy, poor generalization, data-hungry*
- ***Frozen pre-trained models**: Good generalization, low accuracy*
- ***Need**: High accuracy + High generalization + Data efficiency*

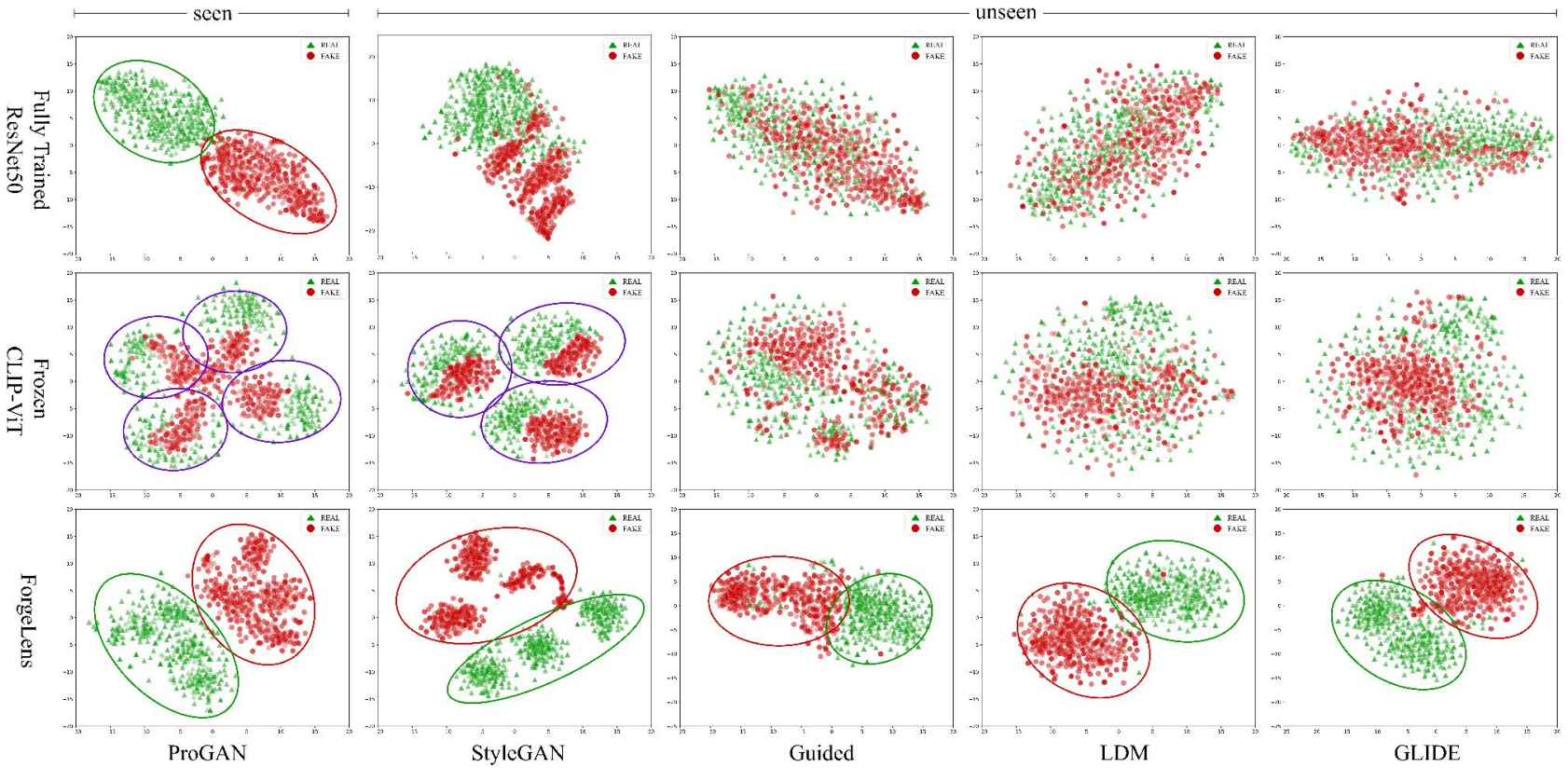
Solution: ForgeLens

- *Build on frozen CLIP-ViT*

*→ Guide the model to focus on **forgery-relevant features** with limited training data.*

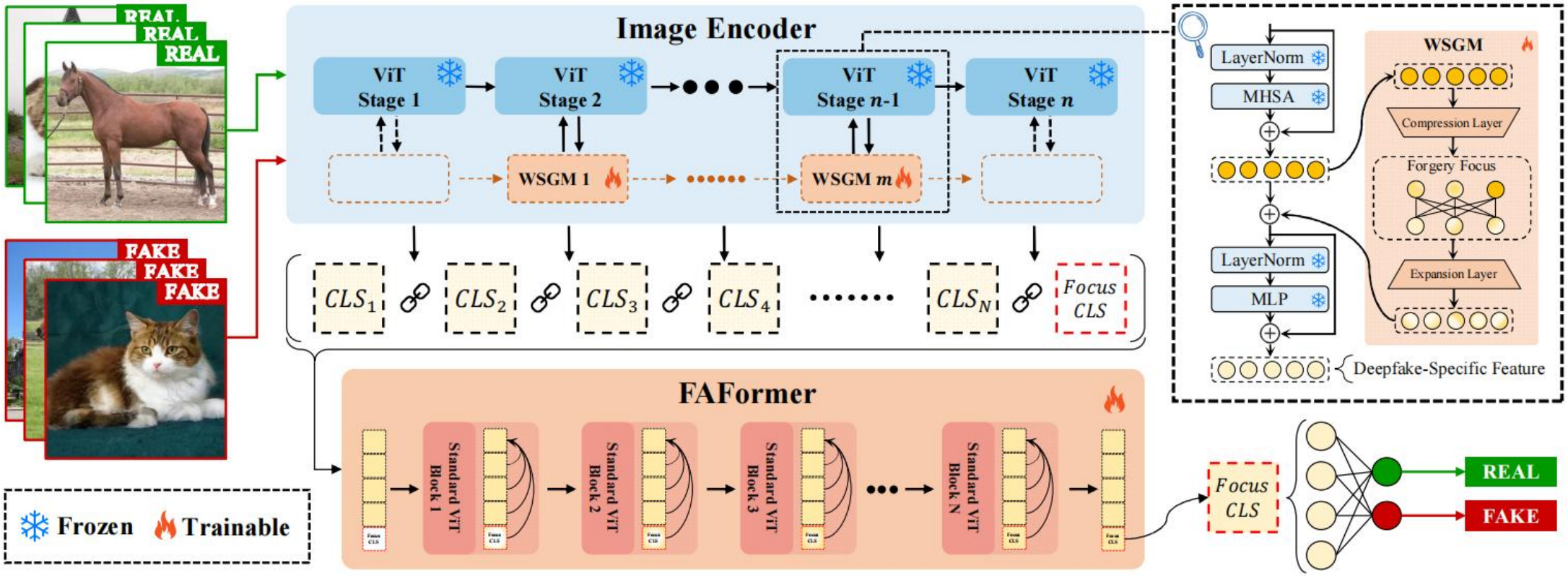


Visualization of extracted image features



- **ResNet50 (specialized model)**: *clear separation on seen data (e.g., ProGAN) but fails on unseen data*
- **CLIP-ViT (frozen model)**: *features cluster by image category rather than real vs. fake → general-purpose and not optimized for forgery detection*
- *General-purpose features (e.g., frozen CLIP-ViT) contain much forgery-irrelevant info → poor fake/real separation*

ForgeLens



WSGM

Guide frozen image encoder to focus on the forgery-specific information within the general-purpose image features it extracts during training

FAFormer

Refines forgery information from the multi-stage image features (low-level and high-level)

Datasets

Training Dataset

- *ForenSynths*
- *Subsampling: 1%, 4%, 20%, 50% of original training set*

Evaluation Dataset

- *UniversalFakeDetect*
- *19 generative models (GANs & diffusion)*
- *Includes real & fake images, some subsets by class*

Volume	Total Size	Classes	Neg%
1%	1,600	car, cat, chair, horse	50%
4%	6,400	car, cat, chair, horse	50%
20%	28,800	car, cat, chair, horse	50%
50%	72,000	car, cat, chair, horse	50%
100%	144,024	car, cat, chair, horse	50%

Table 6. Statistics of Training Dataset. We report the number of images in each data split, the class distribution (all classes share the total data volume), and the proportion of negative samples.

Generative Models	Size	Class Count	Neg%
ProGAN	8,000	20	50%
CycleGAN	2,642	6	50%
BigGAN	4,000	N/A	50%
StyleGAN	11,982	3	50%
GauGAN	10,000	N/A	50%
StarGAN	3,998	N/A	50%
Deepfakes	5,405	N/A	49.9%
SITD	360	N/A	50%
SAN	438	N/A	50%
CRN	12,764	N/A	50%
IMLE	12,764	N/A	50%
Guided	2,000	N/A	50%
LDM 200 steps	3,000	N/A	33.3%
LDM 200 w/CFG	3,000	N/A	33.3%
LDM 100 steps	3,000	N/A	33.3%
Glide-100-27	3,000	N/A	33.3%
Glide-50-27	3,000	N/A	33.3%
Glide-100-10	3,000	N/A	33.3%
DALL-E	3,000	N/A	33.3%

Table 7. Statistics of the UniversalFakeDetect Dataset. We report the size of each subset, the number of classes (N/A indicates no class split), and the proportion of negative samples.

Experimental Results

Method	Generative Adversarial Networks						Deep fakes	Low level vision		Perceptual loss		Guided	LDM			Glide			DALL-E	Avg. Acc
	Pro GAN	Cycle GAN	Big GAN	Style GAN	Gau GAN	Star GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/CFG	100 steps	100 27	50 27	100 10		
Specialized																				
Patchfor [5]	75.03	68.97	68.47	79.16	64.23	63.94	75.54	75.14	75.28	72.33	55.30	67.41	76.50	76.10	75.77	74.81	73.28	68.52	67.91	71.24
F3Net [42]	99.38	76.38	65.33	92.56	58.10	100.0	63.48	54.17	47.26	51.47	51.47	96.20	68.15	75.35	68.80	81.65	83.25	83.05	66.30	71.33
FreqNet [49]	97.90	95.84	90.45	97.55	90.24	93.41	97.40	88.92	59.04	71.92	67.35	86.70	84.55	99.58	65.56	85.69	97.40	88.15	59.06	85.09
Preprocessing-based																				
CNN-Spot [54]	99.99	85.20	70.20	85.70	78.95	91.70	53.47	66.67	48.69	86.31	86.26	60.07	54.03	54.96	54.14	60.78	63.80	65.66	55.58	69.58
LGrad [47]	99.84	85.39	82.88	94.83	72.45	99.62	58.00	62.50	50.00	50.74	50.78	77.50	94.20	95.85	94.80	87.40	90.70	89.55	88.35	80.28
NPR [50]	99.84	95.00	87.55	96.23	86.57	99.75	76.89	66.94	98.63	50.00	50.00	84.55	97.65	98.00	98.20	96.25	97.15	97.35	87.15	87.56
Frozen Model-based																				
UniFD [39]	100.0	98.50	94.50	82.00	99.50	97.00	66.60	63.00	57.50	59.50	72.00	70.03	94.19	73.76	94.36	79.07	79.85	78.14	86.78	81.38
FatFormer [32]	99.89	99.32	99.50	97.15	99.41	99.75	93.23	81.11	68.04	69.45	69.45	76.00	98.60	94.90	98.65	94.35	94.65	94.20	98.75	90.86
RINE [29]	100.0	99.30	99.60	88.90	99.80	99.50	80.60	90.60	68.30	89.20	90.60	76.10	98.30	88.20	98.60	88.90	92.60	90.70	95.00	91.31
C2P-CLIP [48]	99.98	97.31	99.12	96.44	99.17	99.60	93.77	95.56	64.38	93.29	93.29	69.10	99.25	97.25	99.30	95.25	95.25	96.10	98.55	<u>93.79</u>
Ours																				
FreLens	99.95	99.24	97.67	96.64	98.84	95.24	88.97	85.83	93.75	97.23	97.55	73.34	98.72	96.98	98.86	96.07	96.17	95.43	98.29	**94.99**

Table 1. Accuracy (Acc) results of forgery detection methods on UniversalFakeDetect, covering both GANs and diffusion models. Methods are categorized into ***Specialized*** methods | ***Preprocessing-based*** methods | ***Frozen Model-based*** methods | ***Ours***. **Bold** and <u>underline</u> represent the best and second-best performance, respectively.

Method	Generative Adversarial Networks						Deep fakes	Low level vision		Perceptual loss		Guided	LDM			Glide			DALL-E	Avg. Acc
	Pro GAN	Cycle GAN	Big GAN	Style GAN	Gau GAN	Star GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/CFG	100 steps	100 27	50 27	100 10		
Specialized																				
Patchfor	80.88	72.84	71.66	85.75	65.99	69.25	76.55	76.19	76.34	74.52	68.52	75.03	87.10	86.72	86.40	85.37	83.73	78.38	75.67	77.73
F3Net	99.96	84.32	69.90	99.72	56.71	100.0	78.82	52.89	46.70	63.39	64.37	70.53	73.76	81.66	74.62	89.81	91.04	90.86	71.84	76.89
FreqNet	99.92	99.63	96.05	99.89	99.71	98.63	99.92	94.42	74.59	80.10	75.70	96.27	96.06	100.0	62.34	99.80	99.78	96.39	77.78	91.95
Preprocessing-based																				
CNN-Spot	100.0	93.47	84.50	99.54	89.49	98.15	89.02	73.75	59.47	98.24	98.40	73.72	70.62	71.00	70.54	80.65	84.91	82.07	70.59	83.58
LGrad	100.0	93.98	90.69	99.86	79.36	99.98	67.91	59.42	51.42	63.52	69.61	87.06	99.03	99.16	99.18	93.23	95.10	94.93	97.23	86.35
NPR	100.0	99.53	94.53	99.94	88.82	100.0	84.41	97.95	99.99	50.16	50.16	98.26	99.92	99.91	99.92	99.87	99.89	99.92	99.26	92.76
Frozen Model-based																				
UniFD	100.0	98.13	94.46	86.66	99.25	99.53	91.67	78.54	67.54	83.12	91.06	79.24	95.81	79.77	95.93	93.93	95.12	94.59	88.45	90.14
FatFormer	100.0	100.0	99.98	99.75	100.0	100.0	97.99	97.94	81.21	99.84	99.93	91.99	99.81	99.09	99.87	99.13	99.41	99.20	99.82	98.16
RINE	100.0	100.0	99.90	99.40	100.0	100.0	97.90	97.20	94.90	97.30	99.70	96.40	99.80	98.30	99.90	98.80	99.30	98.90	99.30	<u>98.78</u>
C2P-CLIP	100.0	100.0	99.96	99.50	100.0	100.0	98.59	98.92	84.56	99.86	99.95	94.13	99.99	99.83	99.98	99.72	99.79	99.83	99.91	98.66
Ours																				
FreLens	100.0	100.0	99.83	99.82	99.98	100.0	95.44	94.20	98.69	99.94	99.99	92.92	99.92	99.49	99.87	99.10	99.45	99.33	99.80	**98.83**

Table 2. Average Precision (AP) results of forgery detection methods on UniversalFakeDetect.

ForgeLens

Outperforms:

- *Specialized detector: +9.90% Avg.Acc, +6.88% Avg.AP*
- *Preprocessing-based detector: +7.43% Avg.Acc, +6.07% Avg.AP*
- *Frozen model-based detector: +1.20% Avg.Acc, +0.17% Avg.AP*

Compared to the base model UniFD:

- *Improves the Avg.Acc by 13.61% and Avg.AP by 8.69% (introducing minimal training parameters)*

Ablation Study and Robustness Evaluation

Effectiveness Analysis of Forgery Focus

w/ WSGM	W/ FAFormer	Avg.Acc.(%)	Avg.AP.(%)
✗	✗	81.38	90.14
✗	✓	87.89	92.26
✓	✗	94.52	98.12
✓	✓	**94.99**	**98.83**

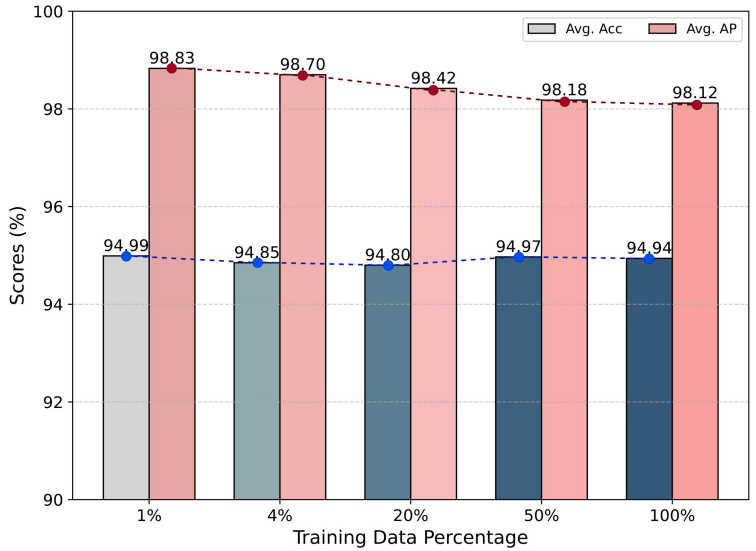
Table 3. Average accuracy and average precision (Acc/AP) in the Ablation Study of WSGM and FAFormer on UniversalFakeDetect. Results without each module are denoted as 'w/o'. The top-performing results are highlighted in bold.

Comparative Analysis of WSGM and Fine-Tuning Methods.

Method	Backbone	Avg.Acc.(%)	Avg.AP.(%)
✗	CLIP-ViT	81.38	90.14
Adapter	CLIP-ViT	85.49 (4.11↑)	94.29 (4.15↑)
LoRA	CLIP-ViT	86.96 (5.58↑)	96.41 (6.27↑)
WSGM (Ours)	CLIP-ViT	**94.52 (12.86↑)**	**98.12 (7.98↑)**

Table 4. Average Accuracy (Avg.Acc) and Average Precision (Avg.Ap) evaluated on UniversalFakeDetect. We compare our proposed WSGM with the previous fine-tuning method Adapter and LoRA. The top-performing results are highlighted in bold.

Impact of Different Training Data Size

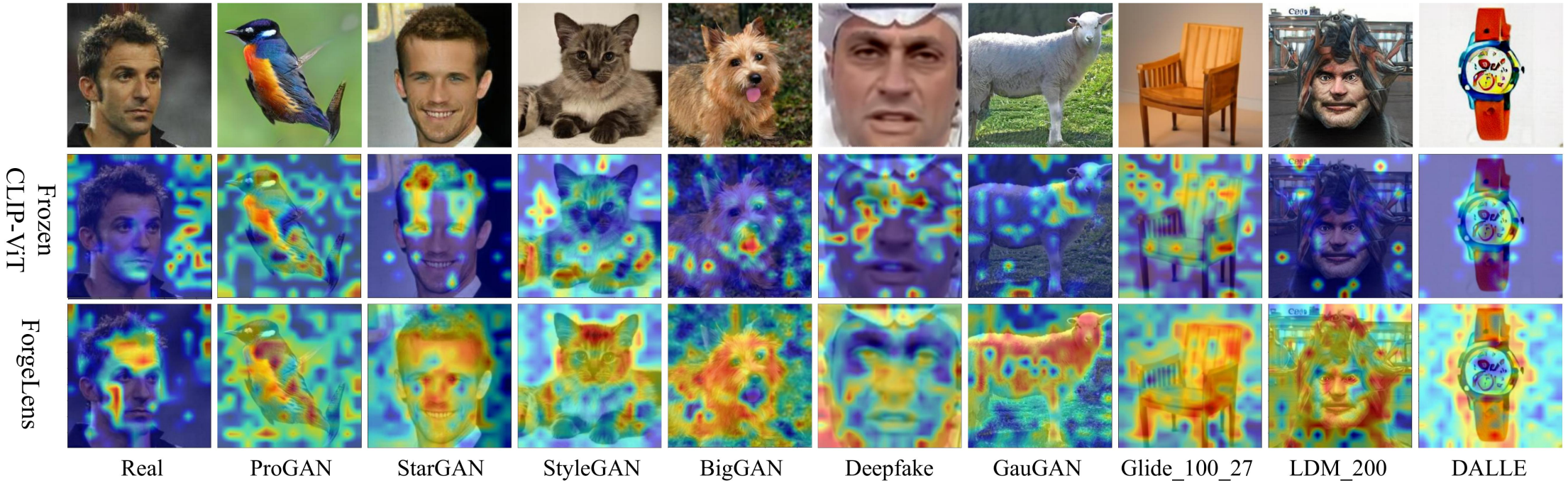


Robustness Evaluation

w/Blur	w/Cropping	w/JPEG	w/Noise	Avg.Acc.(%)	Avg.AP.(%)
✓	✗	✗	✗	81.98 (13.01↓)	92.10 (6.73↓)
✗	✓	✗	✗	91.17 (3.28↓)	96.97 (1.86↓)
✗	✗	✓	✗	91.62 (3.37↓)	97.74 (1.09↓)
✗	✗	✗	✓	82.64 (12.35↓)	93.41 (5.42↓)
✓	✓	✓	✓	86.54 (8.45↓)	94.87 (3.96↓)

Table 5. Average Accuracy (Avg.Acc) and Average Precision (Avg.Ap) evaluated on UniversalFakeDetect under various perturbations.

Visualization



CLIP-ViT (frozen base model)

- *Captures general-purpose features*
- *Attention dispersed, fails to localize forgery artifacts*
- *Ineffective for forgery detection*

ForgeLens (ours)

- *Focuses on forgery-specific regions (e.g., face swap artifacts, hair textures)*
- *Provides targeted and interpretable responses*

Conclusion

Key contributions

- *Proposed ForgeLens, a data-efficient & feature-guided framework for generalizable forgery detection.*
- *SOTA performance on 19 generative models (GANs & diffusion)*
- *Requires only minimal training data*

Limitations & Future Work

- *Two-stage training adds complexity*
- *FAFormer sensitive to hyperparameters*
- *Improve training stability and effectiveness. Enable learning from limited data to distinguish real vs. fake, including latest generative techniques.*

Thanks for listening!

Email: yingjianchen@henu.edu.cn