DIFFUSION EPISTEMIC UNCERTAINTY WITH ASYMMETRIC LEARNING FOR DIFFUSION-GENERATED IMAGE DETECTION

Yingsong Huang[1*], Hui Guo[1*], Jing Huang[2*], Bing Bai[3†] , Qi Xiong[1†]

[1] Tencent Inc.

[2] Hikvision

[3] Microsoft MAI

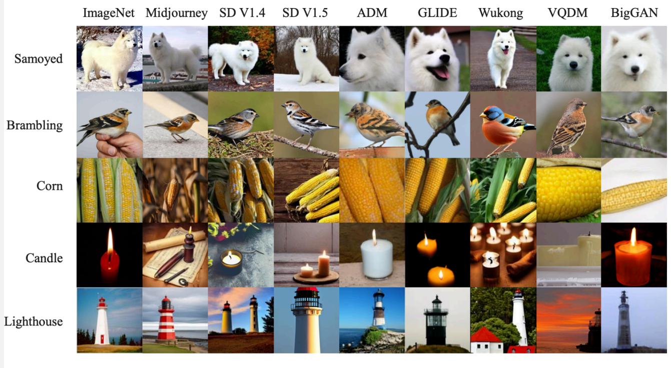
[*] Equal contributions from both authors.

OUTLINE

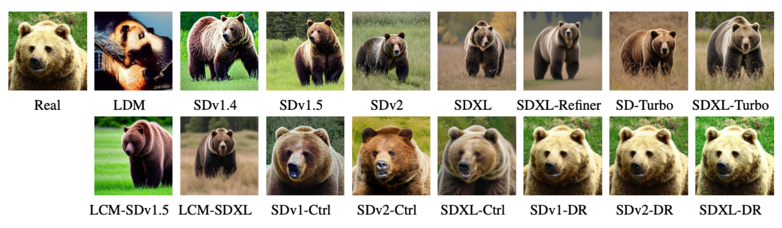
- **Background: Diffusion-generated Images Detection and Uncertainty Estimation**
- Perspective: Uncertainty Matters in Diffusion-Generated Image Detection
  - Predictive ambiguity in Reconstruction error
  - Rethinking Uncertainty in the Diffusion Process
- Method
  - Diffusion Epistemic Uncertainty Feature Estimation
  - Asymmetric Learning method
- Experiments
- Conclusion

BACKGROUND

- Diffusion-generated images detection
  - Effective methods to detect diffusion-generated images are in increasing demand as diffusion-based models can produce highly realistic images.
  - Recent studies on fake image detection fall into two categories.
    - Methods based on obvious artifacts have largely disappeared as generators improve.
    - Methods based on Statistical features such as reconstruction error by diffusion model are gaining prominence.
  - Even when reconstruction error-based method works on in distribution data, it struggles to generalize to data with unfamiliar pattern.



Visualization of images on GenImage dataset. SD is short for Stable Diffusion



Samples from our proposed DRCT-2M dataset.

BACKGROUND

- A brief introduction to Diffusion Model
  - A diffusion model typically involves two processes.
    - The forward process is defined as below, where x_t is the noisy image at the t-th step, α_t is a predefined noise schedule, and T is the total steps.

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\mathbf{x}_{t-1}, (1 - \frac{\alpha_t}{\alpha_{t-1}}\mathbf{I}))\,, \qquad (1)$$

      - According to the property of Markov chain, we can get x_t from x_0 via:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1 - \alpha_t)\mathbf{I})\,. \qquad (2)$$

    - The reverse process is also formulated as a Markov chain in DDPM, using a network $p_\theta(x_{t-1}$ |xt) to fit the real distribution q(x_{t-1} | x_t):

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))\,. \qquad (3)$$

      - DDIM proposes a new deterministic method without the Markov hypothesis. The new reverse process is formulated as:

$$\begin{aligned}\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}(&\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\alpha_t}})+ \\ &\sqrt{1 - \alpha_{t-1} - \sigma_t^2}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) + \sigma_t\boldsymbol{\epsilon}_t\,.\end{aligned} \qquad (4)$$

BACKGROUND

- Uncertainty in the Diffusion Process
  - Uncertainty can have a significant impact on statistical feature, such as the commonly used reconstruction error, in diffusion-generated image detection.
  - Types of uncertainty:
    - Epistemic. Uncertainty in model structure and parameters, which captures the model's lack of knowledge about unfamiliar patterns
    - Aleatoric. Uncertainty inherent in the observation data, which reflects data noise and randomness.
  - For detecting generated images, it is essential to disentangle the overall uncertainty in diffusion measurements and focus on more effective features.

BACKGROUND

- A brief introduction to Epistemic uncertainty
  - Epistemic uncertainty can be modeled by placing an isotropic Gaussian prior $p(\theta)$ over the model's parameters $\epsilon_\theta(x_t, t)$, in Bayesian Neural Networks (BNNs)
    - Let $\mathcal{D}$ be the training dataset, the target of BNNs is to infer the posterior $p(\theta|\mathcal{D})$ and predict distribution of noise for noise-corrupted data x_t^* with

$$p(\epsilon_t^*|x_t^*, \mathcal{D}) \; = \int p(\epsilon|\epsilon_\theta(x_t^*, t))\, p(\theta|\mathcal{D}) \; d\theta$$

    - Since $p(\theta|\mathcal{D})$ cannot be computed directly, Laplace Approximation (LA) approximates it with

$$q(\theta) = \mathcal{N}(\theta; \theta_{\text{MAP}}, \Sigma)$$

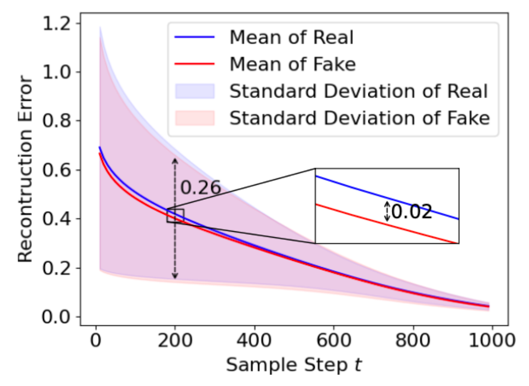
      - Where θ_{MAP} is the maximum a posteriori (MAP) estimate, and $\Sigma \; = \; \left[-\nabla_\theta^2(\, \log p(\mathcal{D}|\theta) + \; \log p(\theta))|_\theta = \theta_{MAP}\right]^{-1}$. Two techniques have been proposed to simplify the estimation of Σ
        - Hessian approximations with factorization, and the most lightweight case is a diagonal factorization which ignores off-diagonal elements
        - The subnetwork LA, and the last-layer LA (LLLA) is its special case which only treats the parameters of the last probabilistically

OUTLINE

- Background: Diffusion-generated Images Detection and Uncertainty Estimation
- **Perspective: Uncertainty Matters in Diffusion-Generated Image Detection**
  - Predictive ambiguity in Reconstruction error
  - Rethinking Uncertainty in the Diffusion Process
- Method
  - Diffusion Epistemic Uncertainty Feature Estimation
  - Asymmetric Learning method
- Experiments
- Conclusion

PREDICTIVE AMBIGUITY IN RECONSTRUCTION ERROR

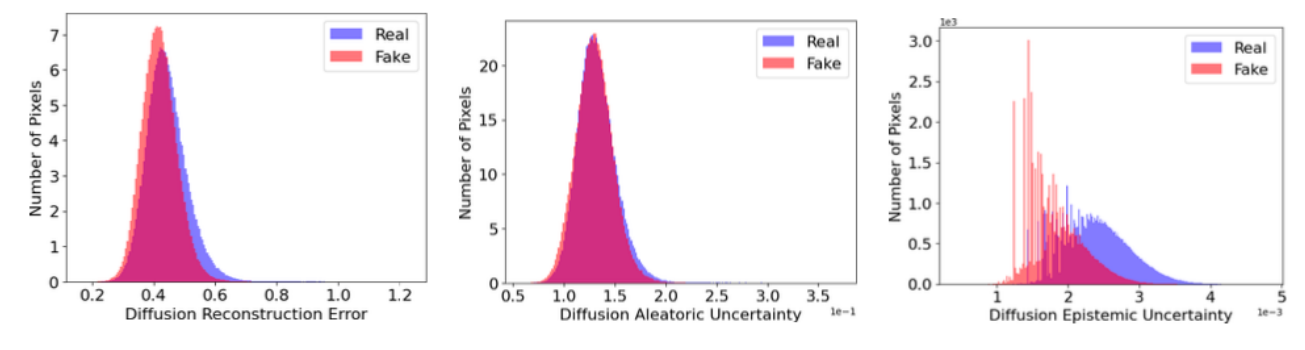
- Relying on diffusion-based measurements (e.g., reconstruction error) to identify generated images can become misleading when aleatoric uncertainty dominates, as it inflates prediction ambiguity.
  - Aleatoric uncertainty arises from inherent noise in the observed data and reflects aspects of the task that are intrinsically difficult. It does not increase for out-of-distribution (OOD) samples, making it unsuitable for detecting such anomalies
    - The difference in reconstruction error becomes less significant when compared to the standard deviation caused by aleatoric uncertainty.



Diffusion Reconstruction error and predictive ambiguity due to aleatoric uncertainty

RETHINKING UNCERTAINTY IN THE DIFFUSION PROCESS

- Both aleatoric and epistemic uncertainty, inherent in the diffusion, contribute to diffusion reconstruction error.
  - Relying solely on reconstruction error is less effective in distinguishing real from fake images when aleatoric uncertainty is high.
  - Epistemic uncertainty increases significantly for samples that fall outside the domain of the training data and consequently provides a more reliable basis for anomaly detection.
    - Epistemic uncertainty more accurately distinguishes real samples from fake ones.



Distribution of diffusion reconstruction error (real samples overlapped with that of fake samples due to the presence of aleatoric uncertainty), aleatoric uncertainty (nearly indistinguishable), and
epistemic uncertainty in real and generated samples

OUTLINE

- Background: Diffusion-generated Images Detection and Uncertainty Estimation
- Perspective: Uncertainty Matters in Diffusion-Generated Image Detection
  - Predictive ambiguity in Reconstruction error
  - Rethinking Uncertainty in the Diffusion Process
- **Methods (DEUA)**
  - Diffusion Epistemic Uncertainty Feature Estimation
  - Asymmetric Learning method
- Experiments
- Conclusion

DIFFUSION EPISTEMIC UNCERTAINTY FEATURE ESTIMATION

- Epistemic uncertainty for generated image detection
  - To estimate diffusion epistemic uncertainty, the forward and reverse diffusion process are performed in the latent space $\mathcal{X} = \{\mathbf{x}^{(\mathbf{i})}\}$ encoded by a pre-trained VAE model.
  - The epistemic uncertainty in the diffusion model $\epsilon_\theta(x_t, t)$ is captured by

$$Var_{q(\,x_{t-1}|x,\,t\,)}(x_{t-1}) \propto Var_\theta(\mathbb{E}_\epsilon(\mu_\theta(\sqrt{\alpha_\mathrm{t}}\mathrm{x} + (1-\alpha_\mathrm{t})\epsilon, \mathrm{t})))$$

    - A Bayesian diffusion model is built by placing a prior distribution $p(\theta)$ over the parameters of a pre-trained diffusion model $\epsilon_\theta(x_t, t)$ .With the inferred posterior $p(\theta|\mathcal{D})$, we cat get the predict mean via:

$$\mathbb{E}_\theta(\mathrm{x}_{\mathrm{t}-1}) = \int p(\theta|\mathcal{D})\mathbb{E}_\epsilon(\mu_\theta(\sqrt{\alpha_\mathrm{t}}\mathrm{x} + (1-\alpha_\mathrm{t})\epsilon, \mathrm{t}))\mathrm{d}\theta$$

      - We leverage the LLLA with diagonal factorization to approximate $p(\theta|\mathcal{D})$ as

$$q(\theta) = \mathcal{N}(\theta; \theta_{\mathrm{MAP}}, \Sigma)$$

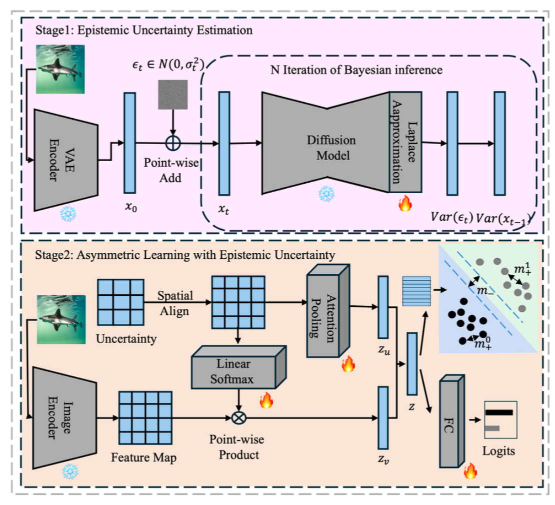
      - By the Monte Carlo approximation, we sample model parameters $\theta_\mathrm{i} \sim q(\theta) = \mathcal{N}(\theta; \theta_{\mathrm{MAP}}, \Sigma)$ and noise $\epsilon_\mathrm{j} \sim \mathcal{N}(0, \mathrm{I})$, where i= 1, .., M and j = 1, ..., N. Then epistemic uncertainty is computed as

$$U(\,x_{t-1} \mid x, t\,) = Var_i\left(\mathbb{E}_j\left[\mu_{\theta_i}\left(\sqrt{\alpha_t x} + (1-\alpha_t)\epsilon_j, t\right)\right]\right)$$

DIFFUSION EPISTEMIC UNCERTAINTY FEATURE ESTIMATION

- Epistemic uncertainty for generated image detection
  - Let u be the spatially aligned epistemic uncertainty and v be the visual feature map. $\bar{u}$ be the mean epistemic uncertainty feature, which is the average of u. The global epistemic uncertainty feature be z_u, which is computed using an attention pooling on u. And the refined visual feature is computed with a multi-head attention module:

$$z_v = MHA(\overline{u}, u, v)$$



Workflow of our method. In the first stage, we utilize the Laplace approximation to estimate diffusion epistemic uncertainty. In the second stage, we exploit diffusion epistemic uncertainty to train a binary classifier with asymmetric learning.

ASYMMETRIC LEARNING METHOD

- To learn an even decision boundary with larger margins while preserving the information content of the representation, ASL is proposed.
  - Using an asymmetric contrastive loss which maximizes the distance of negative pairs and minimizes the distance of positive pairs by class-specific margins.The loss is as:

$$\ell_m = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[y_i^{(1)} = y_i^{(2)} = c\right] \cdot \max\left(0, m^c - s_W(i)\right) + \mathbb{I}\left[y_i^{(1)} \neq y_i^{(2)}\right] \cdot \max\left(0, s_W(i)\right)$$

    - where N is the total number of sample pairs, yi is the bi- nary label for each sample (0 denotes the real class), sW (i) is the Cosine similarity between the samples in each pair, mc is the margin specific to the class c.
    - Consider the wide range of features exhibited by the real class, we propose the implementation of a smaller similarity margin specifically for the real class.
  - The overall loss function is a weighted average of cross-entropy loss, asymmetric contrastive loss:

$$\ell(W) = \ell_c(W) + \lambda\ell_m(W)$$

    - where l_c is the cross-entropy loss, λ is the hyper-parameter controlling the importance of l_m .

OUTLINE

- Background: Diffusion-generated Images Detection and Uncertainty Estimation
- Perspective: Uncertainty Matters in Diffusion-Generated Image Detection
  - Predictive ambiguity in Reconstruction error
  - Rethinking Uncertainty in the Diffusion Process
- Method
  - Diffusion Epistemic Uncertainty Feature Estimation
  - Asymmetric Learning method
- **Experiments**
- Conclusion

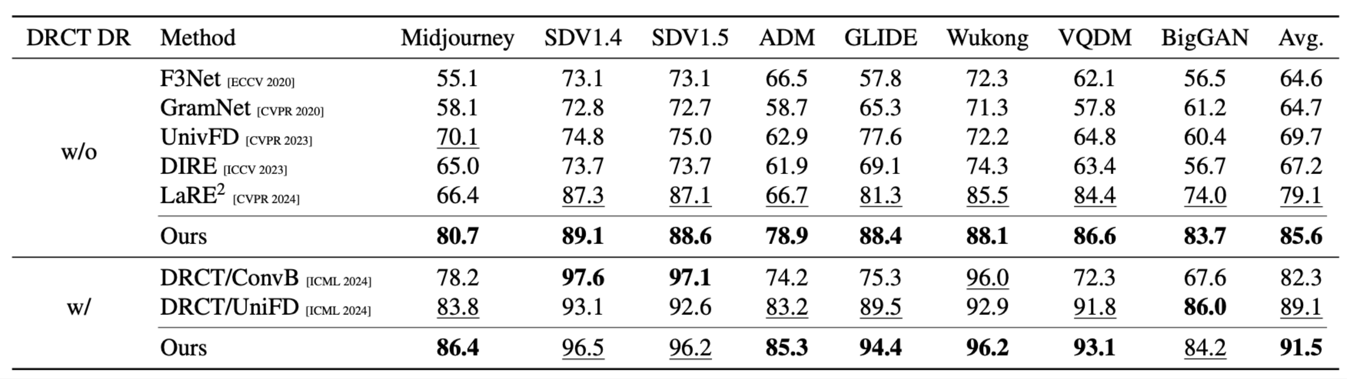
EXPERIMENTAL SETUP

- Datasets
  - We conduct evaluations on two large-scale datasets: GenImage and DRCT-2M
    - GenImage comprises 2,681,167 images, segregated into 1,331,167 real images from ImageNet and 1,350,000 fake images generated from eight generative model.
    - DRCT-2M consists of two parts: 1,920,000 images generated by various diffusion-based generative models and the real images from MSCOCO

- Experimental settings
    - We compared our method with several state-of-the-art image generation detection approaches. All experimental setups followed the guidelines established by the GenImage and DRCT-2M benchmarks. Each trained model was then evaluated on all eight test subsets.
    - We conducted generalizability comparisons across generators on GenImage and DRCT-2M, respectively.
    - We conducted cross-dataset experiments following DRCT

EXPERIMENTAL RESULTS

- 1 Performance in GenImage
  - In the scenario where DRCT DR was not applied, our method outperformed the state-of-the-art by 6.5% in average accuracy (ACC).
  - When integrated with DRCT DR, our method shown an additional gain, outperforming DRCT by 2.4% in average ACC.

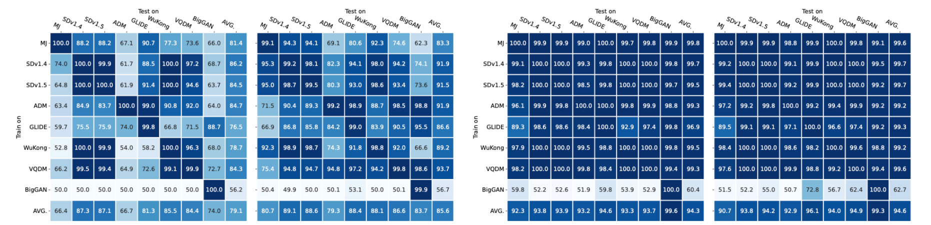


DRCT DR	Method	Midjourney	SDV1.4	SDV1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
w/o	F3Net [ECCV 2020]	55.1	73.1	73.1	66.5	57.8	72.3	62.1	56.5	64.6
	GramNet [CVPR 2020]	58.1	72.8	72.7	58.7	65.3	71.3	57.8	61.2	64.7
	UnivFD [CVPR 2023]	70.1	74.8	75.0	62.9	77.6	72.2	64.8	60.4	69.7
	DIRE [ICCV 2023]	65.0	73.7	73.7	61.9	69.1	74.3	63.4	56.7	67.2
	LaRE^2 [CVPR 2024]	66.4	87.3	87.1	66.7	81.3	85.5	84.4	74.0	79.1
	Ours	**80.7**	**89.1**	**88.6**	**78.9**	**88.4**	**88.1**	**86.6**	**83.7**	**85.6**
w/	DRCT/ConvB [ICML 2024]	78.2	**97.6**	**97.1**	74.2	75.3	96.0	72.3	67.6	82.3
	DRCT/UniFD [ICML 2024]	83.8	93.1	92.6	83.2	89.5	92.9	91.8	**86.0**	89.1
	Ours	**86.4**	96.5	96.2	**85.3**	**94.4**	**96.2**	**93.1**	84.2	**91.5**

Accuracy (ACC, %) comparisons on GenImage test subsets. Eight models are trained on eight generators respectively. All the eight models are tested on the specified test subsets, and averaging the accuracy scores yields the final results

EXPERIMENTAL RESULTS

- 2 Generalizability Across Generators
  - Comparison on GenImage. Our method experienced less performance degradation across subsets with different generators, achieving an overall average ACC improvement from 79.1% to 85.6%



(a) Comparison on ACC. Left: baseline, right: ours.

Train on \ Test on	MJ	SDv1.4	SDv1.5	ADM	GLIDE	WuKong	VQDM	BigGAN	AVG.
MJ	100.0	88.2	88.2	67.1	90.7	77.3	73.6	66.0	81.4
SDv1.4	74.0	100.0	99.9	61.7	88.5	100.0	97.2	68.7	86.2
SDv1.5	64.8	100.0	100.0	61.9	91.4	100.0	94.6	63.7	84.5
ADM	63.4	84.9	83.7	100.0	99.0	90.8	92.0	64.0	84.7
GLIDE	59.7	75.5	75.9	74.0	99.8	66.8	71.5	88.7	76.5
WuKong	52.8	100.0	99.9	54.0	58.2	100.0	96.3	68.0	78.7
VQDM	66.2	99.5	99.4	64.9	72.6	99.1	99.9	72.7	84.3
BigGAN	50.0	50.0	50.0	50.0	50.0	50.0	50.0	100.0	56.2
AVG.	66.4	87.3	87.1	66.7	81.3	85.5	84.4	74.0	79.1

Train on \ Test on	MJ	SDv1.4	SDv1.5	ADM	GLIDE	WuKong	VQDM	BigGAN	AVG.
MJ	99.1	94.3	94.1	69.1	80.6	92.3	74.6	62.3	83.3
SDv1.4	95.3	99.2	98.1	82.3	94.1	98.0	94.2	74.1	91.9
SDv1.5	95.0	98.7	99.5	80.3	93.0	98.6	93.4	73.6	91.5
ADM	71.5	90.4	89.3	99.2	98.9	88.7	98.5	98.8	91.9
GLIDE	66.9	86.8	85.8	84.2	99.0	83.9	90.5	95.5	86.6
WuKong	92.3	98.9	98.7	74.3	91.8	98.8	92.0	66.6	89.2
VQDM	75.4	94.8	94.7	94.8	97.2	94.2	99.8	98.6	93.7
BigGAN	50.4	49.9	50.0	50.1	53.1	50.0	50.1	99.9	56.7
AVG.	80.7	89.1	88.6	79.3	88.4	88.1	86.6	83.7	85.6

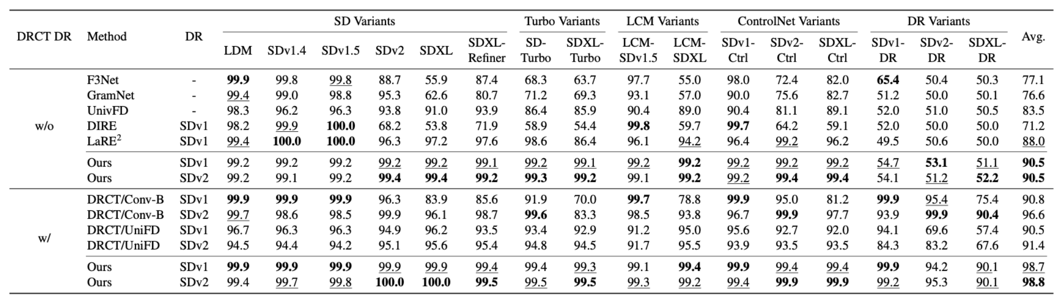
(b) Comparison on AP. Left: baseline, right: ours.

Train on \ Test on	MJ	SDv1.4	SDv1.5	ADM	GLIDE	WuKong	VQDM	BigGAN	AVG.
MJ	100.0	99.9	99.9	99.0	100.0	99.7	99.8	99.9	99.8
SDv1.4	99.1	100.0	100.0	99.3	99.8	100.0	100.0	99.8	99.7
SDv1.5	98.2	100.0	100.0	98.5	99.8	100.0	100.0	99.7	99.5
ADM	96.1	99.9	99.8	100.0	100.0	99.8	99.9	98.8	99.3
GLIDE	89.3	98.6	98.6	98.4	100.0	92.9	97.4	99.8	96.9
WuKong	97.9	100.0	100.0	99.5	98.8	100.0	100.0	99.8	99.5
VQDM	98.2	100.0	100.0	99.8	98.4	100.0	100.0	99.4	99.3
BigGAN	59.8	52.2	52.6	51.9	59.8	53.9	52.9	100.0	60.4
AVG.	92.3	93.8	93.9	93.2	94.6	93.3	93.7	99.6	94.3

Train on \ Test on	MJ	SDv1.4	SDv1.5	ADM	GLIDE	WuKong	VQDM	BigGAN	AVG.
MJ	100.0	99.9	99.9	98.8	99.9	100.0	99.8	99.1	99.6
SDv1.4	99.2	100.0	100.0	99.1	99.9	100.0	100.0	99.5	99.7
SDv1.5	99.4	100.0	100.0	99.2	99.9	100.0	100.0	99.2	99.7
ADM	97.2	99.2	99.8	100.0	99.2	99.4	99.9	99.2	99.2
GLIDE	89.5	99.1	99.1	97.1	100.0	96.6	97.4	99.2	99.3
WuKong	98.4	100.0	100.0	98.6	98.2	100.0	99.6	98.8	99.2
VQDM	97.6	100.0	100.0	99.9	98.8	99.2	100.0	99.4	99.6
BigGAN	51.5	52.2	55.0	50.7	72.8	56.7	62.4	100.0	62.7
AVG.	90.7	93.8	94.2	92.9	96.1	94.0	94.9	99.3	94.6

Results of cross-validation on GenImage. We train eight models on eight subsets respectively, each corresponding to a different generator.

  - Comparison on DRCT-2M. Our method is robust in previously unseen diffusion models, such as SDXL, SDXL-Turbo, and LCM-SDXL especially, results in an overall improvement in average accuracy from 88.0% to 90.5%.

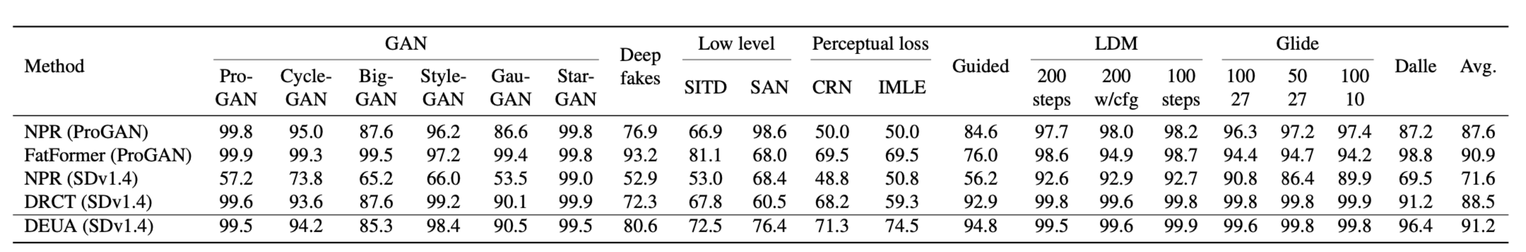


DRCT DR	Method	DR	SD Variants						Turbo Variants		LCM Variants		ControlNet Variants			DR Variants			Avg.
			LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	
w/o	F3Net	-	**99.9**	99.8	99.8	88.7	55.9	87.4	68.3	63.7	97.7	55.0	98.0	72.4	82.0	**65.4**	50.4	50.3	77.1
	GramNet	-	99.4	99.0	98.8	95.3	62.6	80.7	71.2	69.3	93.1	57.0	90.0	75.6	82.7	51.2	50.0	50.1	76.6
	UnivFD	-	98.3	96.2	96.3	93.8	91.0	93.9	86.4	85.9	90.4	89.0	90.4	81.1	89.1	52.0	51.0	50.5	83.5
	DIRE	SDv1	98.2	99.9	**100.0**	68.2	53.8	71.9	58.9	54.4	**99.8**	59.7	**99.7**	64.2	59.1	52.0	50.0	50.0	71.2
	LaRE^2	SDv1	99.4	**100.0**	**100.0**	96.3	97.2	97.6	98.6	86.4	96.1	94.2	96.4	99.2	96.2	49.5	50.6	50.0	88.0
	Ours	SDv1	99.2	99.2	99.2	99.2	99.2	99.1	99.2	99.1	99.2	**99.2**	99.2	99.2	99.2	54.7	**53.1**	51.1	**90.5**
	Ours	SDv2	99.2	99.1	99.2	**99.4**	**99.4**	**99.2**	**99.3**	**99.2**	99.1	**99.2**	99.2	**99.4**	**99.4**	54.1	51.2	**52.2**	**90.5**
w/	DRCT/Conv-B	SDv1	**99.9**	**99.9**	**99.9**	96.3	83.9	85.6	91.9	70.0	**99.7**	78.8	**99.9**	95.0	81.2	**99.9**	95.4	75.4	90.8
	DRCT/Conv-B	SDv2	99.7	98.6	98.5	99.9	96.1	98.7	**99.6**	83.3	98.5	93.8	96.7	**99.9**	97.7	93.9	**99.9**	**90.4**	96.6
	DRCT/UniFD	SDv1	96.7	96.3	96.3	94.9	96.2	93.5	93.4	92.9	91.2	95.0	95.6	92.7	92.0	94.1	69.6	57.4	90.5
	DRCT/UniFD	SDv2	94.5	94.4	94.2	95.1	95.6	95.4	94.8	94.5	91.7	95.5	93.9	93.5	93.5	84.3	83.2	67.6	91.4
	Ours	SDv1	**99.9**	**99.9**	**99.9**	99.9	99.9	99.4	99.4	99.3	99.1	**99.4**	**99.9**	99.4	99.4	**99.9**	94.2	90.1	98.7
	Ours	SDv2	99.4	99.7	99.8	**100.0**	**100.0**	**99.5**	99.5	**99.5**	99.3	99.2	99.4	**99.9**	**99.9**	99.2	95.3	90.1	**98.8**

Accuracy (ACC, %) comparisons on DRCT-2M. All methods are only trained on SDv1.4 and evaluated on different test subsets on DRCT-2M.

EXPERIMENTAL RESULTS

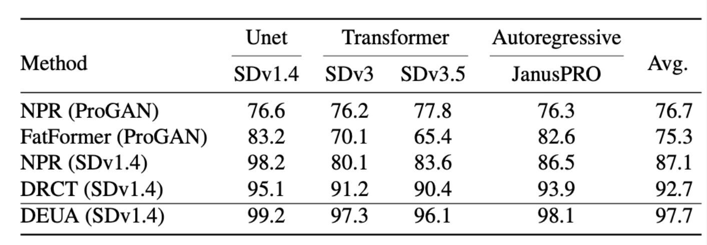
- 2 Generalizability Across more Generators
  - Comparison on UniversalFakeDetect dataset. DEUA achieved competitive or even superior performance compared to existing unified detectors when trained on diffusion models



Method	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	Avg.
	ProGAN	CycleGAN	BigGAN	StyleGAN	GauGAN	StarGAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
NPR (ProGAN)	99.8	95.0	87.6	96.2	86.6	99.8	76.9	66.9	98.6	50.0	50.0	84.6	97.7	98.0	98.2	96.3	97.2	97.4	87.2	87.6
FatFormer (ProGAN)	99.9	99.3	99.5	97.2	99.4	99.8	93.2	81.1	68.0	69.5	69.5	76.0	98.6	94.9	98.7	94.4	94.7	94.2	98.8	90.9
NPR (SDv1.4)	57.2	73.8	65.2	66.0	53.5	99.0	52.9	53.0	68.4	48.8	50.8	56.2	92.6	92.9	92.7	90.8	86.4	89.9	69.5	71.6
DRCT (SDv1.4)	99.6	93.6	87.6	99.2	90.1	99.9	72.3	67.8	60.5	68.2	59.3	92.9	99.8	99.6	99.8	99.8	99.8	99.9	91.2	88.5
DEUA (SDv1.4)	99.5	94.2	85.3	98.4	90.5	99.5	80.6	72.5	76.4	71.3	74.5	94.8	99.5	99.6	99.9	99.6	99.8	99.8	96.4	91.2

ACC comparisons on the UniversalFakeDetect Dataset. Results of NPR, FatFormer and C2P-CLIP trained on ProGAN are from paper C2P-CLIP. Results of NPR and DRCT trained on GenImage SDv1.4 are obtained using their official checkpoints.

  - Comparison on novel generative paradigms, including diffusion transformers and autoregressive models. DEUA consistently maintains high detection accuracy on these emerging architectures, further confirming its adaptability and effectiveness

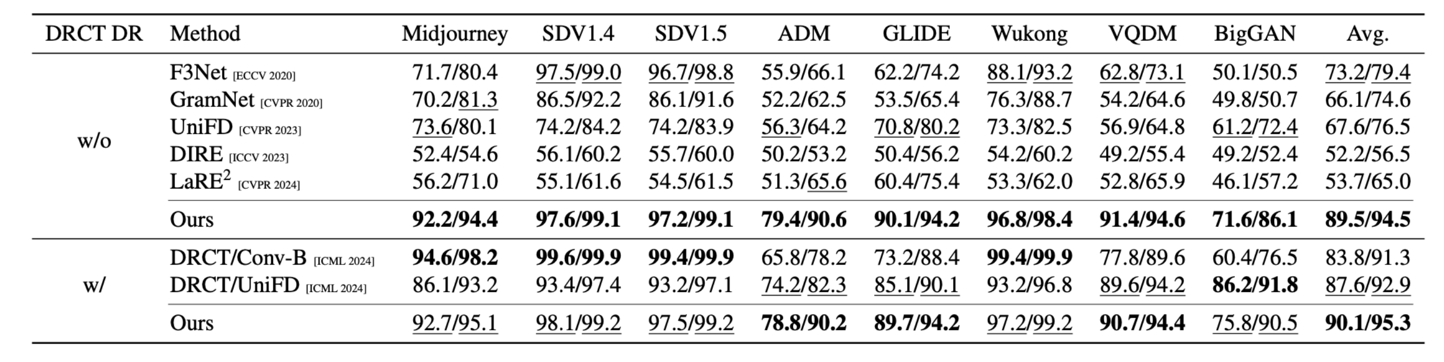


Method	Unet	Transformer		Autoregressive	Avg.
	SDv1.4	SDv3	SDv3.5	JanusPRO	
NPR (ProGAN)	76.6	76.2	77.8	76.3	76.7
FatFormer (ProGAN)	83.2	70.1	65.4	82.6	75.3
NPR (SDv1.4)	98.2	80.1	83.6	86.5	87.1
DRCT (SDv1.4)	95.1	91.2	90.4	93.9	92.7
DEUA (SDv1.4)	99.2	97.3	96.1	98.1	97.7

ACC comparison on new generators. SDv3, sdv3.5 and JanusPRO are collected following GenImage. Results of NPR, FatFormer and DRCT are obtained using their official checkpoints

EXPERIMENTAL RESULTS

- 3 Generalizability Across Datasets
  - Our method demonstrated stronger generalization, with smaller declines in both ACC and AP metrics under identical cross-dataset conditions, which offers improved robustness across varying image content.
    - Methods based on diffusion reconstruction errors, such as DIRE and LaRE2 , encountered substantial challenges in maintaining performance across different datasets.
    - Additionally, despite its diffusion focus and orthogonality to unified schemes, DEUA stays competitive and is an effective module in larger frameworks

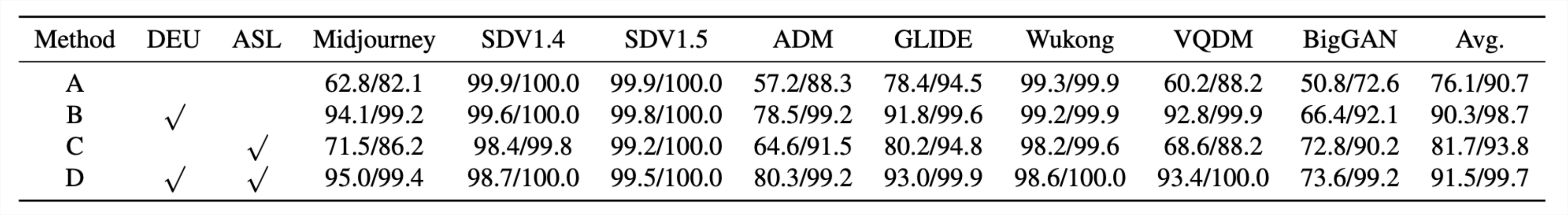


DRCT DR	Method	Midjourney	SDV1.4	SDV1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
w/o	F3Net [ECCV 2020]	71.7/80.4	97.5/99.0	96.7/98.8	55.9/66.1	62.2/74.2	88.1/93.2	62.8/73.1	50.1/50.5	73.2/79.4
	GramNet [CVPR 2020]	70.2/81.3	86.5/92.2	86.1/91.6	52.2/62.5	53.5/65.4	76.3/88.7	54.2/64.6	49.8/50.7	66.1/74.6
	UniFD [CVPR 2023]	73.6/80.1	74.2/84.2	74.2/83.9	56.3/64.2	70.8/80.2	73.3/82.5	56.9/64.8	61.2/72.4	67.6/76.5
	DIRE [ICCV 2023]	52.4/54.6	56.1/60.2	55.7/60.0	50.2/53.2	50.4/56.2	54.2/60.2	49.2/55.4	49.2/52.4	52.2/56.5
	LaRE^2 [CVPR 2024]	56.2/71.0	55.1/61.6	54.5/61.5	51.3/65.6	60.4/75.4	53.3/62.0	52.8/65.9	46.1/57.2	53.7/65.0
	Ours	**92.2/94.4**	**97.6/99.1**	**97.2/99.1**	**79.4/90.6**	**90.1/94.2**	**96.8/98.4**	**91.4/94.6**	**71.6/86.1**	**89.5/94.5**
w/	DRCT/Conv-B [ICML 2024]	**94.6/98.2**	**99.6/99.9**	**99.4/99.9**	65.8/78.2	73.2/88.4	**99.4/99.9**	77.8/89.6	60.4/76.5	83.8/91.3
	DRCT/UniFD [ICML 2024]	86.1/93.2	93.4/97.4	93.2/97.1	74.2/82.3	85.1/90.1	93.2/96.8	89.6/94.2	**86.2/91.8**	87.6/92.9
	Ours	92.7/95.1	98.1/99.2	97.5/99.2	**78.8/90.2**	**89.7/94.2**	97.2/99.2	**90.7/94.4**	75.8/90.5	**90.1/95.3**

Accuracy (ACC, %) / average precision (AP, %) comparisons of generalizability across datasets. All methods are trained on DRCT-2M/SDv1.4 using SDv1 as the diffusion reconstruction model and evaluated on different testing subsets of GenImage.

EXPERIMENTAL RESULTS

- 4 Ablation Experiments
  - Each individual element provides an independent performance improvement.
  - Influence of Diffusion Epistemic Uncertainty
    - Integrating diffusion epistemic uncertainty into our model led to the most substantial improvement of 14.2%/8.0% ACC/AP compared to the baseline
  - Influence of Asymmetric Learning
    - led to substantial performance gains on the Big- GAN subset, achieving increases of 22.0% in ACC and 17.6% in AP over the baseline



Method	DEU	ASL	Midjourney	SDV1.4	SDV1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
A			62.8/82.1	99.9/100.0	99.9/100.0	57.2/88.3	78.4/94.5	99.3/99.9	60.2/88.2	50.8/72.6	76.1/90.7
B	√		94.1/99.2	99.6/100.0	99.8/100.0	78.5/99.2	91.8/99.6	99.2/99.9	92.8/99.9	66.4/92.1	90.3/98.7
C		√	71.5/86.2	98.4/99.8	99.2/100.0	64.6/91.5	80.2/94.8	98.2/99.6	68.6/88.2	72.8/90.2	81.7/93.8
D	√	√	95.0/99.4	98.7/100.0	99.5/100.0	80.3/99.2	93.0/99.9	98.6/100.0	93.4/100.0	73.6/99.2	91.5/99.7

Ablative study results on GenImage test subsets

OUTLINE

- Background: Diffusion-generated Images Detection and Uncertainty Estimation
- Perspective: Uncertainty Matters in Diffusion-Generated Image Detection
  - Predictive ambiguity in Reconstruction error
  - Rethinking Uncertainty in the Diffusion Process
- Method
  - Diffusion Epistemic Uncertainty Feature Estimation
  - Asymmetric Learning method
- Experiments
- **Conclusion**

CONCLUSION

- In this work, we propose a novel framework for detecting diffusion-generated images with DEUA, Diffusion Epistemic Uncertainty and Asymmetric Learning method

- We observe that recent approaches relying on diffusion reconstruction error as a feature exhibit limited generalizability due to the influence of aleatoric uncertainty and introduce a new feature, diffusion epistemic uncertainty, which quantifies the deviation of an image from the manifold of diffusion-generated images. And Asymmetric Learning further boosts generalization.

- Experimental results demonstrate that our method achieves state-of-the- art generalizability across a variety of generation methods and datasets.

THANKS FOR YOUR TIME!