Introduction and On-Going Challenges with Machine Unlearning and Deepfake Detection

CSIRO's Data61 Cybersecurity Seminar

April 28, 2025

Associate Prof. Simon S. Woo

Topics for Today’s Talk

- Machine Unlearning

  - Fast Machine Unlearning Algorithm

  - Future Research Direction

- Deepfake Detection

  - Generalized Deepfake Detection

  - Proactive Defense: Generation Suppression/Concept Erasing

Brief introduction about me

Personal Background

Born & Lived in S. Korea (for 18 years) | Immigrated to USA (for 20 years) | Since 2017, living in S. Korea

Educational Background

BS EE | MS ECE | MS CS | PhD CS

Work & Research Experiment

Intel | NASA's Jet Propulsion Lab | USC/ISI | Verisign | SUNY/Korea | SKKU

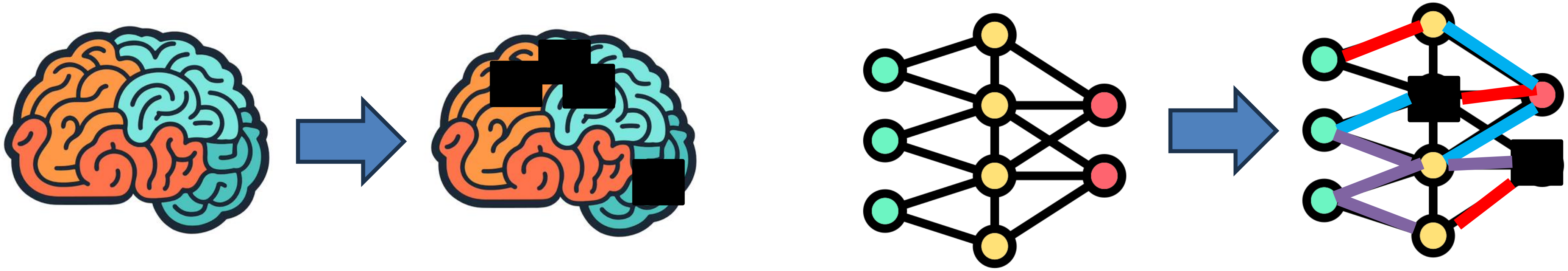
My Main Research Areas

- AI Security & Privacy
  - Multimedia Forensics (Deepfakes)
  - Machine Unlearning
- Other Topics
  - Anomaly Detection (vision, time-series)
  - Medical Anomaly Detection
  - Satellite Object Detection

Machine Unlearning

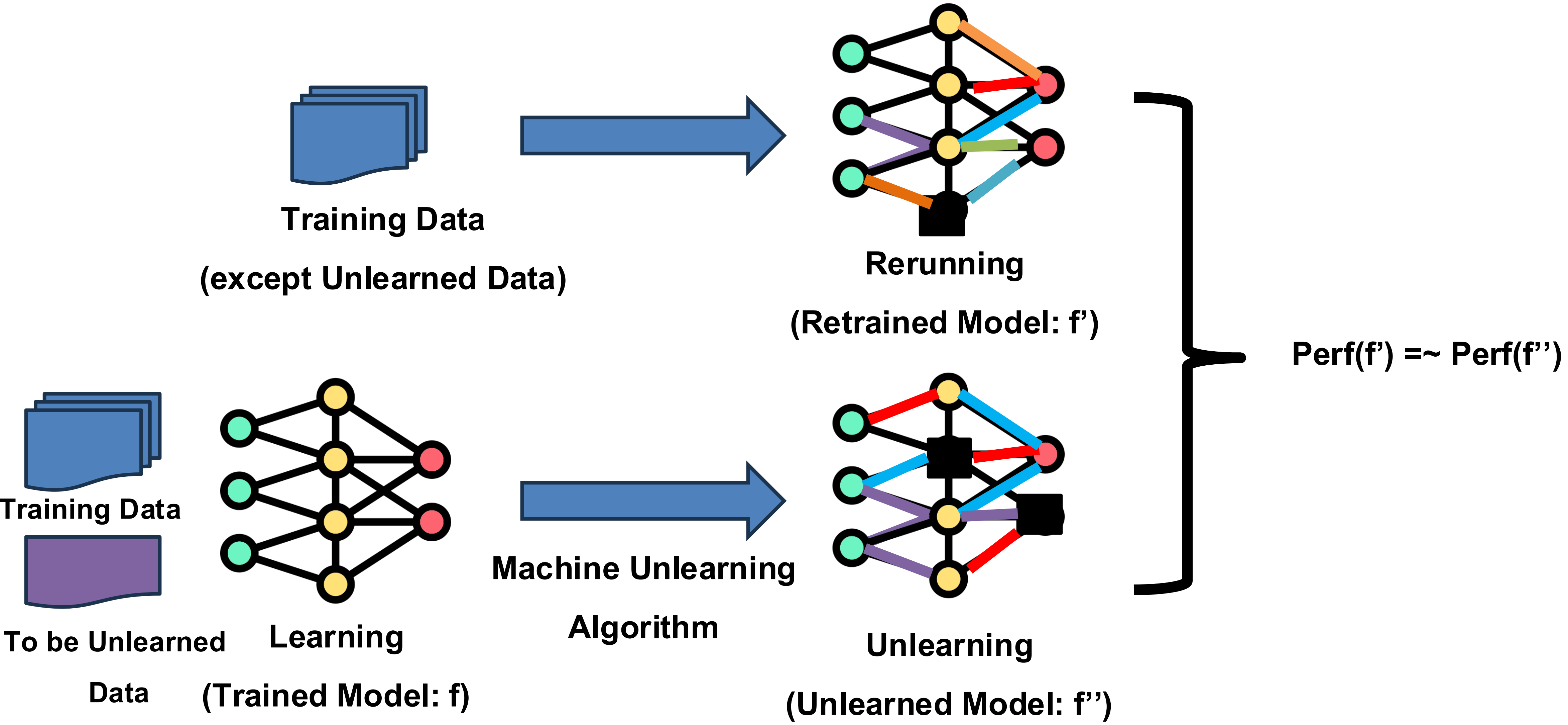
What is Machine Unlearning?

- Make a machine to forget what it leaned
  (specific information, image, class, instances, etc)
- Make a machine to erase some parts of its memory



Learn Unlearn Learn Unlearn

What is Machine Unlearning?



Layer Attack Unlearning: Fast and Accurate Machine Unlearning via Layer Level Attack and Knowledge Distillation

Hyunjune Kim, Sangyong Lee, Simon S. Woo*

Sungkyunkwan University, Suwon, South Korea

The 38th Annual AAAI Conference on Artificial Intelligence

Motivations

- Companies handling such personal data should delete the information from their ML models in response to the **user's request for forgetting**: GDPR, Privacy, Copyright issues
- Simply retraining models to exclude information subject to user's request for forgetting requires **significant costs and time**.
- Machine unlearning offers a solution by **selectively forgetting specific data without retraining in ML models.**

Related work

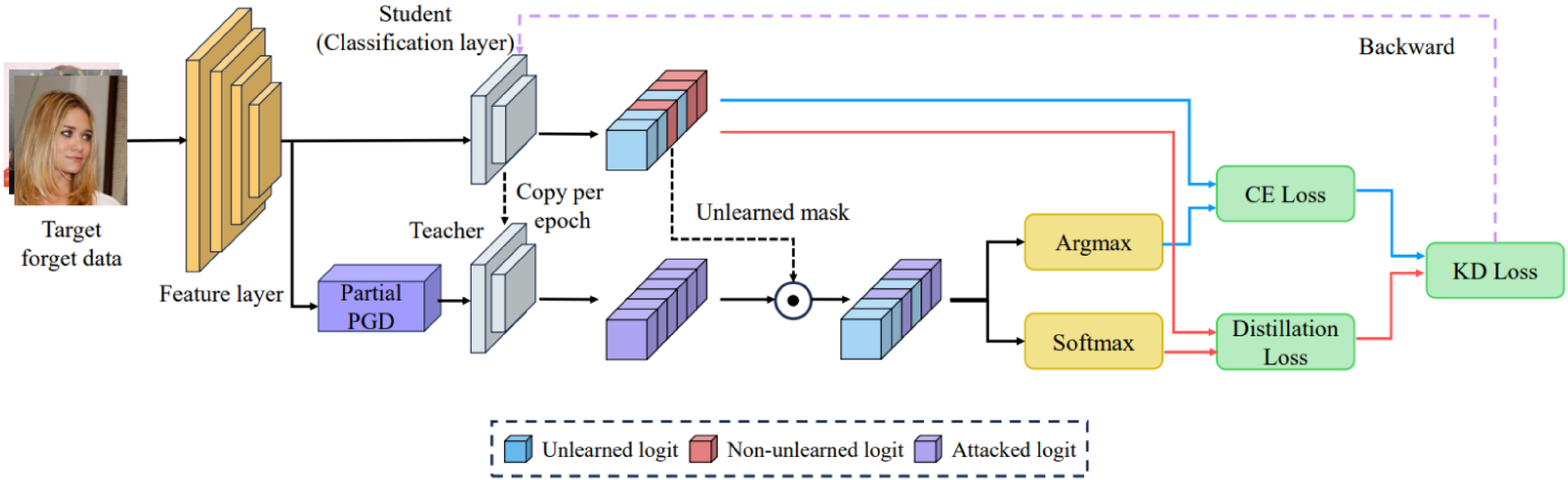
- There are several approaches to solve the problem of machine unlearning
  - Data-driven
    - This strategy involves effectively managing data by partitioning or augmenting to make unlearning model.
  - Model-agnostic
    - This strategy is a methodology by adjusting the model's learning parameters for forgetting.
- Our approach we will introduce among these is model-agnostic to solve class-wise unlearning problem.

Our contributions

- To efficiently perform unlearning task, we propose layer-level unlearning and Partial PGD instead of unlearning the entire model.
- By utilizing knowledge distillation (KD), we preserved the model's utility after the unlearning task.
- Our approach achieves good results in terms of time and accuracy through experiments in diverse environments.

Overall Architecture and Procedures

- We focus on only <u>modifying the parameters of the classification layer</u> tied to classification instead of the entire layers for unlearning
- This approach uses of classification layer as Student and Teacher at each epoch for KD
- The role of Partial PGD is to find target information in the vicinity of the forget data samples, which is then distilled into knowledge for Student



Partial-Projected Gradient Descent (PGD)

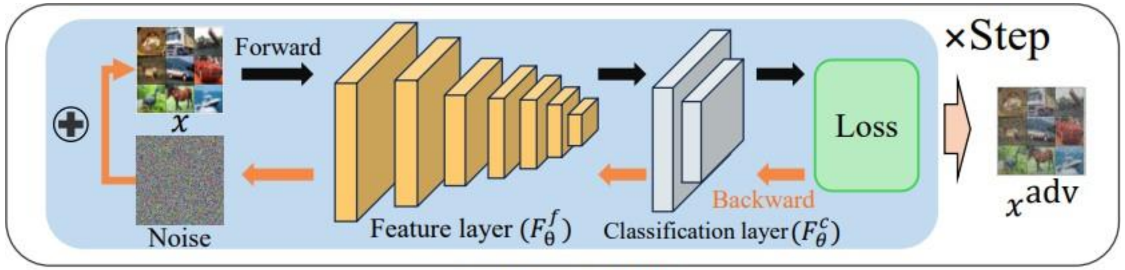
Adversarial examples in unlearning:

- Random or irrelevant class assignments significantly impair task performance
- Enhance the search for appropriate neighboring spaces for forgetting data assignment

Differentiation in adversarial approach:

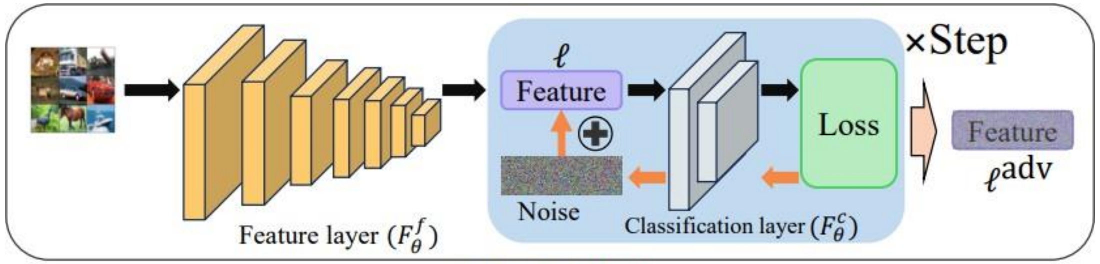
- Clarifies the unique role of adversarial examples in the study, unlike previous methods
- Original PGD approach may introduce slow calculation
- No requirement for full model gradient calculation, optimizing the adversarial creation process

$$x^{t+1} = \Pi(x^t + (\epsilon \cdot sign(\nabla_x \mathcal{L}(x, y, \theta)))$$



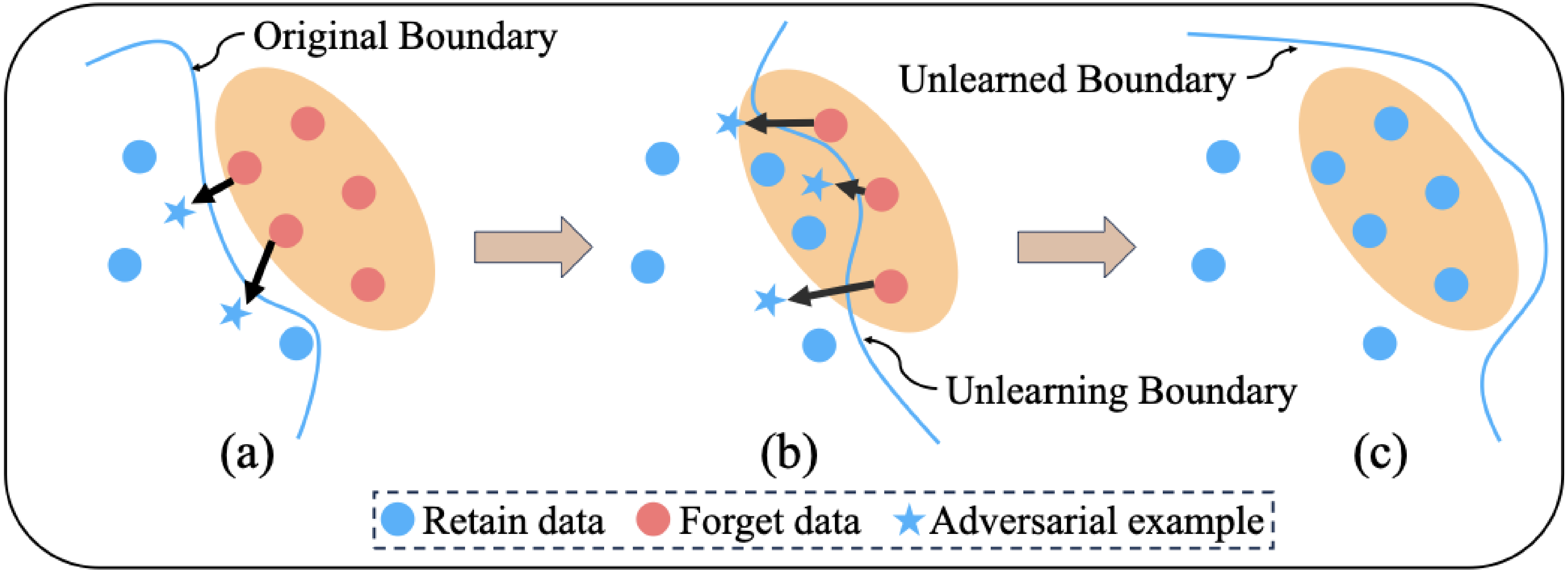
(a) Original PGD

VS.



(b) Partial-PGD

$$\ell^{t+1} = \Pi(\ell^t + (\epsilon \cdot sign(\nabla_\ell \mathcal{L}(\ell, y, \theta)))$$



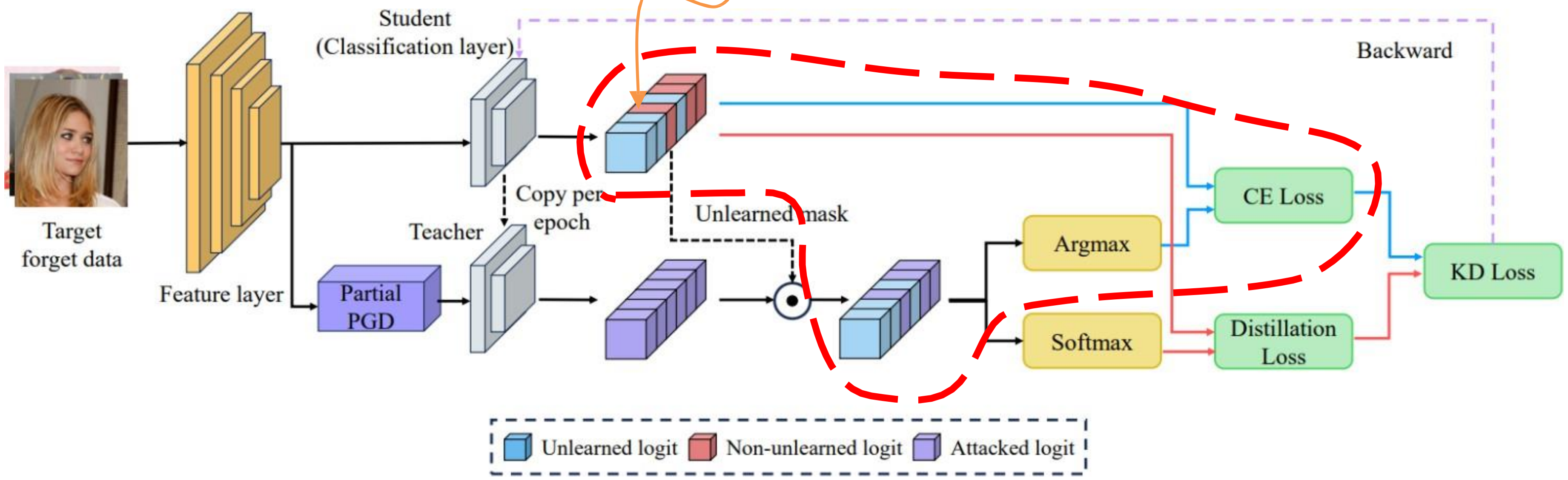
Boundary evolution in the unlearning process. As shown in (a), the original model receives the initial knowledge about the boundary. As the epoch progresses, the boundary information updates as depicted in (b) and (c) from the distilled knowledge

End-to-End Unlearning Process

Cross-Entropy (CE) Loss:

- Application of Partial-PGD on the teacher model to generate adversarial examples to find the space near the forget data
- In this CS Loss, the step involves injecting the Teacher's hard label information into the Student
- Then non-unlearned logits replace adversarial ones for loss computation to make the unlearned mask
- The replaced logtis are represented by the argmax results, indicated as y_f^{adv} and y_{s_θ}

$$\mathcal{L}_{CE} = \begin{cases} \mathrm{CE}(S_\theta(\ell_f), y_f^{\mathrm{adv}}) & \text{if } y_{S_\theta} = y_f \\ \mathrm{CE}(S_\theta(\ell_f), y_{S_\theta}) & \text{otherwise,} \end{cases}$$

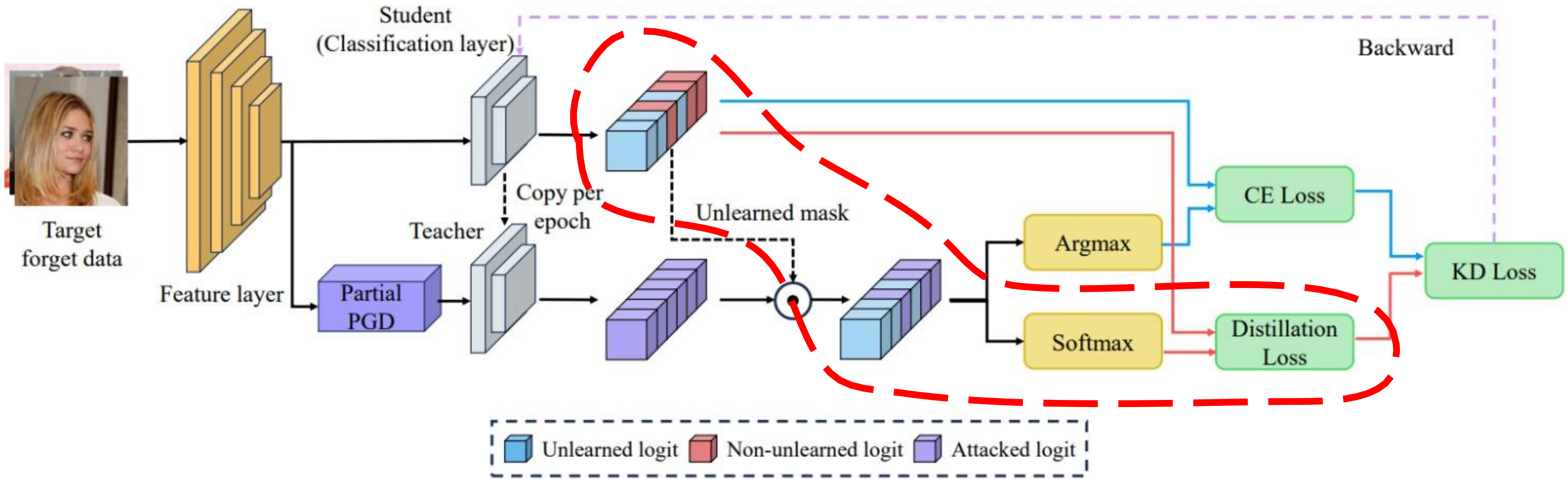


End-to-End Unlearning Process

Distillation Loss:

- In this Loss, the step involves injecting the Teacher's soft label information from Partial-PGD into Student
- Z represents the double Softmax representation, σ represents the softmax function
- Use of double Softmax to adjust probability distribution from Teacher to convey soft label information to Student
- $\mathcal{L}_{DI}$ focuses on creating a similar boundary to the teacher model, ensuring performance while removing D_f information

$$Z = \begin{cases} \sigma(\mathcal{T}_\theta(\ell_f^{\text{adv}})) & \text{if } y_{\mathcal{S}_\theta} = y_f \\ \sigma(\mathcal{S}_\theta(\ell_f)) & \text{otherwise,} \end{cases} \qquad \mathcal{L}_{DI} = \mathrm{KL}\left(\sigma(\frac{\mathcal{S}_\theta(\ell_f)}{T}), \sigma(\frac{Z}{T})\right)$$



End-to-End Unlearning Process

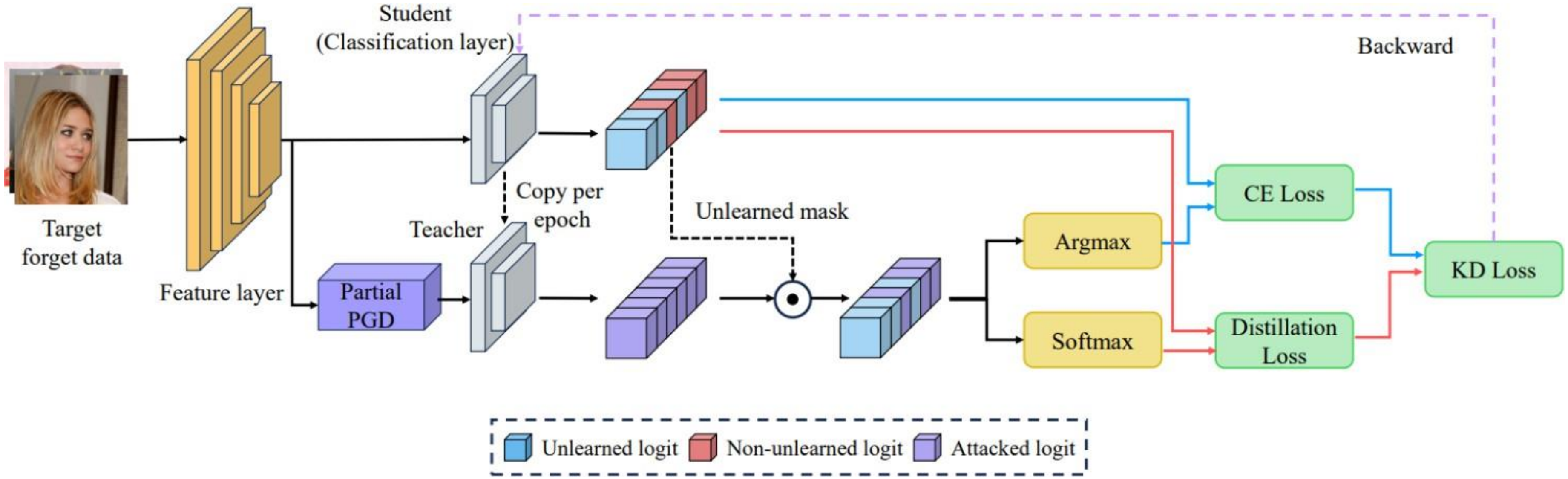
Final Loss Function Composition:

- Combines $\mathcal{L}_{CE}$ and $\mathcal{L}_{DI}$ for the ultimate loss function

$$\mathcal{L} = (1-\alpha)\cdot\mathcal{L}_{CE} + \alpha\cdot T^2\cdot\mathcal{L}_{DI}$$

- Formation of the unlearning model $\mathcal{M}_{\theta^*}$ by merging the feature layer $\mathcal{F}_{\theta}^{f}$ with the classification layer $\mathcal{F}_{\theta^*}^{c}$

$$\mathcal{M}_{\theta*} = \mathcal{F}_{\theta*}^{c} \circ \mathcal{F}_{\theta}^{f}$$



Setup

Datasets: CIFAR-10, Fashion-MNIST, and VGGFace2

Models: VGG16, ResNet18, ResNet50, and ViT

Baselines: Negative Gradient, Fine-tune, Random Label, Fisher Forgetting, Boundary Shrink, and IWU

Metrics

Accuracy (ACC) : Accuracy of a model $\mathcal{M}_\theta$ tested on D_{train} or D_{test}, δ is a the Kronecker delta function.

$$\mathrm{ACC} = 100 \cdot \frac{\sum_{i=1}^{N} \delta(\sigma(\mathcal{M}_\theta(x_i)), y_i)}{\mathbf{N}}$$

Unlearning Score (US) : We calculate US through the retain data accuracy(acc_r) and forget data accuracy(acc_f). A score closer to 1 indicates a higher quality of unlearning results.

$$\mathrm{US}(\mathrm{acc}_r, \mathrm{acc}_f) = \frac{\exp(\frac{\mathrm{acc}_r}{100}) + \exp(1 - \frac{\mathrm{acc}_f}{100}) - 2}{2 \cdot (\exp(1) - 1)}$$

Results **Utility Performance:** Accuracy and Unlearning Score (US)

Dataset	Model	VGG16					ResNet18					ResNet50					ViT				
	Metrics	$D_r \uparrow$	$D_f \downarrow$	$D_{tr} \uparrow$	$D_{tf} \downarrow$	US	$D_r \uparrow$	$D_f \downarrow$	$D_{tr} \uparrow$	$D_{tf} \downarrow$	US	$D_r \uparrow$	$D_f \downarrow$	$D_{tr} \uparrow$	$D_{tf} \downarrow$	US	$D_r \uparrow$	$D_f \downarrow$	$D_{tr} \uparrow$	$D_{tf} \downarrow$	US
CIFAR-10	Original	99.98	100	92.07	96.70	0.4494	99.98	100	93.13	96.60	0.4575	99.94	99.96	93.44	95.0	0.4646	88.06	93.52	81.48	88.40	0.4020
	Retrain (Optimal)	99.89	0	91.98	0	0.9390	99.79	0	92.50	0	0.9428	99.77	0	92.48	0	0.9426	95.0	0	81.0	0	0.8631
	Negative Gradient	88.53	16.96	79.86	17.0	0.7320	93.85	28.38	86.30	25.54	0.7204	88.75	24.77	82.52	23.30	0.7087	85.264	18.69	79.74	16.7	0.7332
	Fine-tune	99.63	0	90.09	0	0.9253	99.63	0	91.25	0	0.9337	99.45	0	90.79	0	0.9304	90.96	1.77	82.43	1.62	0.8598
	Random Label	80.99	3.56	72.40	3.69	0.7805	91.38	11.09	84.00	10.98	0.8007	81.30	12.91	76.62	11.84	0.7467	77.58	15.10	73.42	14.38	0.7094
	Fisher Forgetting	46.78	55.24	44.61	52.30	0.3414	59.0	52.34	55.57	52.2	0.3945	58.17	58.06	55.95	56.20	0.3781	42.68	66.34	43.34	62.30	0.2911
	Boundary Shrink	90.73	10.16	81.53	9.58	0.7943	95.88	9.75	87.91	10.24	0.8329	86.03	3.94	80.09	3.46	0.8303	85.22	0.61	79.29	0.28	0.8498
	IWU	90.81	0	82.35	0.10	0.8712	89.41	0	82.55	0	0.8733	86.11	0	79.98	0	0.8564	82.48	3.92	77.01	2.58	0.8173
	Ours	99.97	0	92.18	0	**0.9405**	99.97	0	93.53	0	**0.9504**	99.92	0	93.52	0	**0.9503**	87.51	0	81.14	0	**0.8640**
Fashion-MNIST	Original	99.83	100	94.38	99.60	0.4579	98.45	99.96	94.71	99.70	0.4601	98.49	99.98	94.68	99.6	0.4601	91.27	98.71	88.28	97.10	0.4210
	Retrain (Optimal)	100	0	93.40	0	0.9494	100	0	93.38	0	0.9493	100	0	93.28	0	0.9485	89.44	0	86.76	0	0.9019
	Negative Gradient	97.77	0	92.63	0	0.9438	92.57	1.39	90.04	0.84	0.9183	84.44	12.63	81.42	10.22	0.7890	71.77	0.10	70.38	0.10	0.7964
	Fine-tune	99.67	0	93.07	0	0.9470	97.23	0	91.93	0	0.9386	98.83	0	92.85	0	0.9454	96.08	0.01	88.72	0.10	**0.9148**
	Random Label	98.17	8.34	92.43	23.55	0.7763	76.80	11.47	74.80	11.54	0.7375	75.99	10.77	73.73	10.72	0.7368	84.18	11.36	82.10	13.04	0.7736
	Fisher Forgetting	62.33	28.81	60.32	28.10	0.5471	72.78	57.65	71.03	54.10	0.4705	60.59	84.01	60.25	82.60	0.2958	43.42	88.01	42.60	86.3	0.1972
	Boundary Shrink	86.88	1.47	81.66	1.12	0.8586	95.78	34.54	92.31	32.40	0.7225	83.50	30.23	80.60	27.08	0.6728	70.31	2.04	68.74	2.70	0.7665
	IWU	99.09	0	93.68	0	0.9515	93.82	0	90.80	0	0.9304	80.17	0	77.94	0	0.8434	82.85	0	81.21	0	0.8645
	Ours	99.51	0	93.89	0	**0.9531**	97.98	0	94.54	0	**0.9579**	98.14	0	94.48	0	**0.9575**	90.11	0	87.44	0	0.9066
VGGFace2	Original	100	100	96.67	98.41	0.4787	100	100	95.88	98.41	0.4727	99.12	98.43	93.67	100	0.4514	94.71	96.86	95.43	93.82	0.4832
	Retrain (Optimal)	99.98	0	96.67	0	0.9740	100	0	96.20	0	0.9705	99.10	0	94.77	0	0.9596	92.63	0	93.32	0	0.9488
	Negative Gradient	96.85	15.67	90.50	4.76	0.8915	97.32	9.75	89.55	12.69	0.8272	86.80	4.73	78.79	3.17	0.8241	91.16	1.63	92.34	0	0.9416
	Fine-tune	97.86	0	89.87	0	0.9416	91.42	0	85.91	0	0.8960	95.18	0	90.03	0	0.9249	96.91	1.63	84.85	3.70	0.8600
	Random Label	90.32	1.74	79.11	1.58	0.8384	96.76	6.44	87.34	0	0.9059	88.24	13.19	82.43	9.52	0.8007	92.06	9.68	91.04	8.64	0.8667
	Fisher Forgetting	46.24	31.01	42.72	50.79	0.3400	72.78	57.65	71.03	54.10	0.4705	76.28	4.52	71.83	7.93	0.7455	60.80	71.07	53.58	60.49	0.3472
	Boundary Shrink	99.48	17.25	93.04	5.36	0.9055	94.02	5.40	86.08	5.36	0.8559	93.85	5.36	85.78	5.0	0.8565	86.92	6.46	86.81	4.25	0.8693
	IWU	99.21	10.80	94.46	4.76	0.8650	75.23	0.17	69.77	0	0.7936	78.62	0	69.14	0	0.7899	76.25	0.27	78.66	0	0.8479
	Ours	99.70	0	96.70	0	**0.9743**	99.79	0	95.34	0	**0.9639**	97.46	0	93.28	0	**0.9485**	95.18	0	95.50	0	**0.9651**

Results

Efficiency: Extra data used & Time consumption

		Retrain	Fisher Forgetting	Fine-tune	NG	Random Label	Boundary Shrink	IWU	**Ours**
CIFAR-10	**Total Extra Data Used**	45,000	45,000	45,000	5,000	5,000	5,000	5,000	5,000
	Time w/ VGG16	3,683	9,710	433	73	24	116	1351	**3.76**
	Time w/ ResNet18	2,871	12,526	546	153	30	191	362	**4.37**
	Time w/ ResNet50	4,705	19,850	1,061	174	57	471	1513	**7.76**
	Time w/ ViT	4,441	13,238	479	78	23	163	1563	**25.93**
Fshion-MNIST	**Total Extra Data Used**	54,000	54,000	54,000	6,000	6,000	6,000	6,000	6,000
	Time w/ VGG16	2,309	8,526	430	85	23	214	1072	**8.75**
	Time w/ ResNet18	2,768	12,116	582	103	30	715	223	**5.19**
	Time w/ ResNet50	5,758	22,013	1,229	206	76	929	967	**9.14**
	Time w/ ViT	2,155	8,377	487	80	25	282	546	**13.39**
VGGFace2	**Total Extra Data Used**	5,726	5,726	5,726	574	574	574	574	574
	Time w/ VGG16	1,840	1,295	468	400	17	338	548	**5.6**
	Time w/ ResNet18	1,861	1,354	670	140	27	473	1258	**6.51**
	Time w/ ResNet50	3,721	2,597	3,291	484	157	503	1837	**17.77**
	Time w/ ViT	2,155	1,428	665	84	27	187	783	**6.74**

Ablation Study

Data Usage Ratio:

- The class-specific D_f dataset for one class in CIFAR-10 contains 5,000 samples. As shown in Table 5, we reduced the dataset size to 50% (2,500) and 10% (500) for each model to perform the unlearning task.

	Model	VGG16		ResNet18		ResNet50		ViT	
	Total Extra Data Used	2,500	500	2,500	500	2,500	500	2,500	500
Metrics	D_{tr}	92.42	92.38	93.51	93.38	93.63	93.37	81.14	81.6
	D_{tf}	0	0	0	0	0	0	0	0.1
	US	0.9422	0.9420	0.9503	0.9493	0.9512	0.9493	0.8640	0.8662
	Time	1.91	1.21	2.28	1.45	3.81	1.62	25.63	14.55

Original PGD vs. Partial-PGD:

- Compares unlearning performance when applying the original PGD vs. Partial-PGD within our method

	Original PGD			Partial PGD		
	D_{tr}	D_{tf}	Time (s)	D_{tr}	D_{tf}	Time (s)
VGG16	92.03	0	14.18	92.18	0	3.76
ResNet18	92.97	0	18.19	93.53	0	4.37
ResNet50	91.84	0	44.15	93.52	0	7.76
ViT	78.07	0	237.36	81.14	0	25.93

Double Softmax:

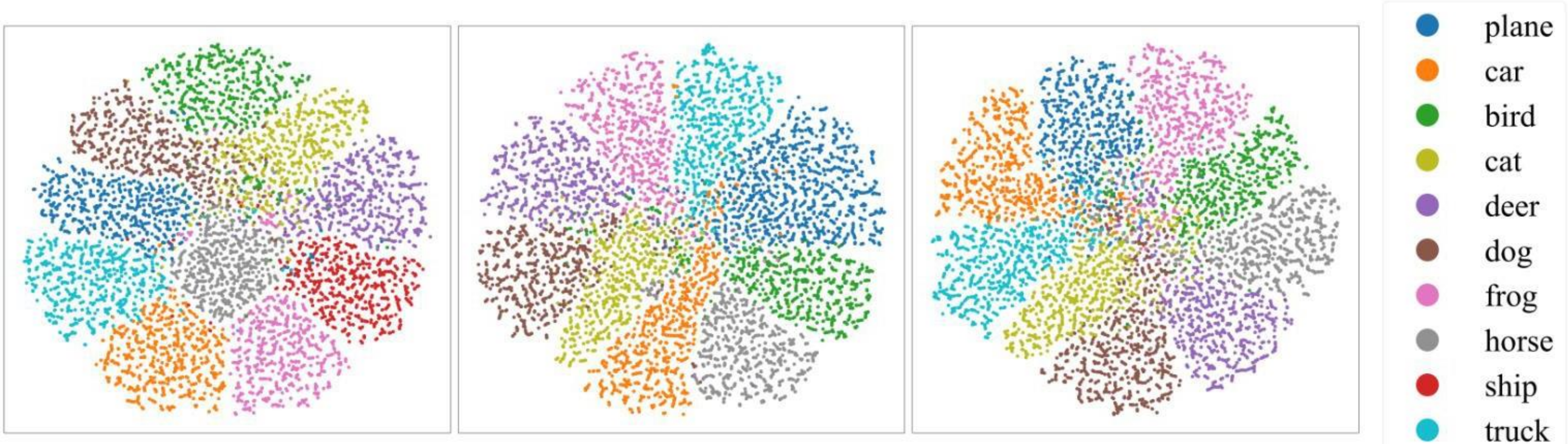
- Unlearning performance with and without double Softmax in our methods in Fashion-MNIST

	w/o Double Softmax			w/ Double Softmax		
	D_{tr}	D_{tf}	Time (s)	D_{tr}	D_{tf}	Time (s)
VGG16	84.74	0	10.9	93.89	0	8.75
ResNet18	91.42	0.1	25.87	94.54	0	5.19
ResNet50	80.91	0	93.49	94.48	0	9.13
ViT	87.01	0	61.37	87.44	0	13.39

Ablation Study

Visualization on Decision Boundary:

- The Figure presents the Original, Retrain, and Ours using tSNE on the CIFAR-10 dataset. The red dots represent samples of ship images, indicated as D_f .



(a) Original (b) Retrain (c) Ours

Conclusion

We proposed a novel machine unlearning algorithm **Layer Attack Unlearning**:

- Presents Partial-PGD as a layer unlearning method
- Proposes an end-to-end KD framework for enhancing accuracy and eliminating the forgetting dataset

Our experiments demonstrated success through extensive experiments:

- Modifying specific layers' learning objectives leads to effective unlearning
- Reduces parameters and computational cost, minimizing overall unlearning time

Layer Attack Unlearning offers a promising path for future research:

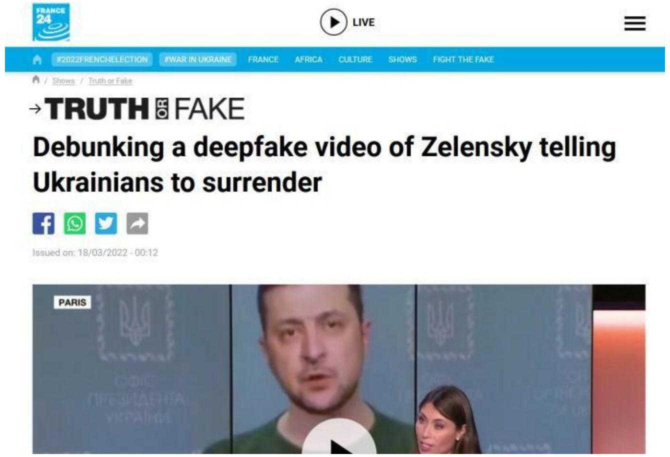
- Addresses diverse unlearning challenges effectively

Still Many Challenges on Machine Unlearning (MU) Research

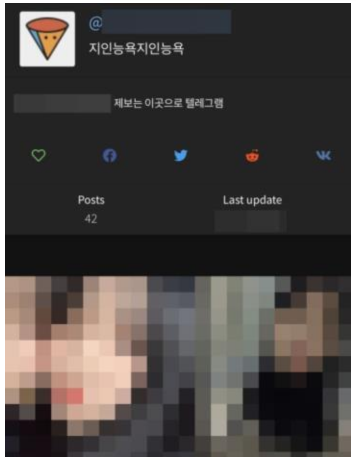
- Developing Practical Machine Unlearning Methods (vs. theoretical)
- Applying MU to Real World Datasets/Approaches
- Exploring LLM unlearning

Deepfake Abuses are Prevalent and Increasing!!!

Malicious use of Generative AI ⇒ Creating serious social problems



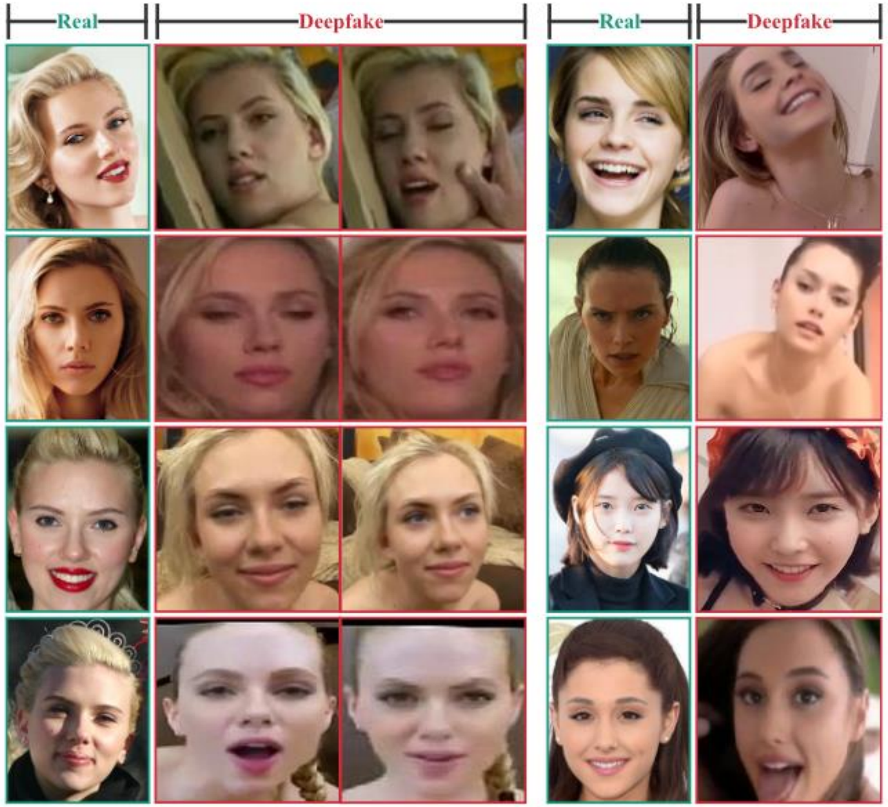
Fake news generation and propagation



Used for general individuals

Deep voices are used for new level of digital crimes (phishing, scamming, etc.)

More Serious Issues



The Korea Times

South Korea > Society

Deepfakes emerge as threat to presidential election

Listen

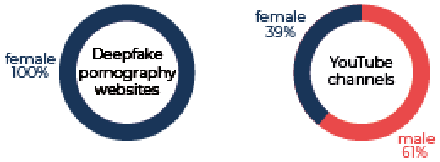
https://www.koreatimes.co.kr/southkorea/society/20250415/deepfakes-emerge-as-threat-to-presidential-election

Next S. Korea Presidential Election on June 3, 2025

Serious Social Problems

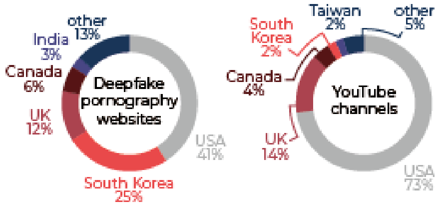
Demographic breakdown of deepfake videos from top five deepfake pornography websites and top 14 deepfake YouTube channels

We analyzed the gender, nationality, and profession of subjects in deepfake videos from the top 5 deepfake pornography websites, as well as the top 14 deepfake YouTube channels that host non-pornographic deepfake videos.



Gender

Deepfake pornography is a phenomenon that exclusively targets and harms women. In contrast, the non-pornographic deepfake videos we analyzed on YouTube contained a majority of male subjects.

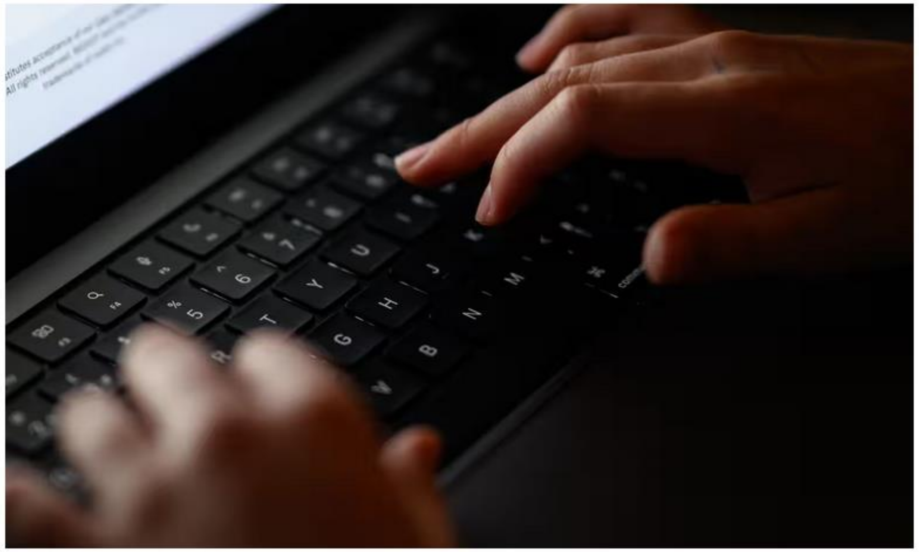


Nationality

We found that over 90% of deepfake videos on YouTube featured Western subjects. However, non-Western subjects featured in almost a third of videos on deepfake pornography websites, with South Korean K-pop singers making up a quarter of the subjects targeted. This indicates that deepfake pornography is an increasingly global phenomenon.

Sharing deepfake intimate images to be criminalised in England and Wales

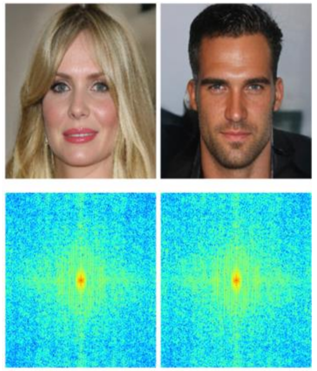
Under online safety bill, maximum sentence where intent to cause distress is proved will be two years



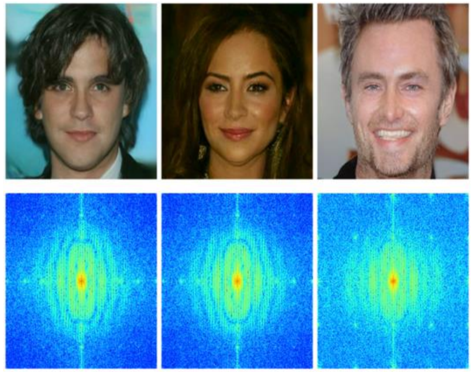
Offenders found guilty of sharing faked images for sexual gratification could be placed on the sex offender register. Photograph: Leon Neal/Getty Images

Sharing deepfake intimate images is to be criminalised in England and Wales. Amendments to the online safety bill will make it illegal to share explicit images or videos that have been digitally manipulated to look like someone else without their consent.

GAN | **Faceswap** | **Reennacment** | **Diffusion**



ProGAN VQGAN



DDIM PNDM LDM

GAN or DM? In-depth Analysis and Evaluation of AI-generated Face Data for Generalizable Deepfake Detection

SoK: Systematization and Benchmarking of Deepfake Detectors in a Unified Framework

Binh M. Le
Sungkyunkwan University, S. Korea
bmle@g.skku.edu

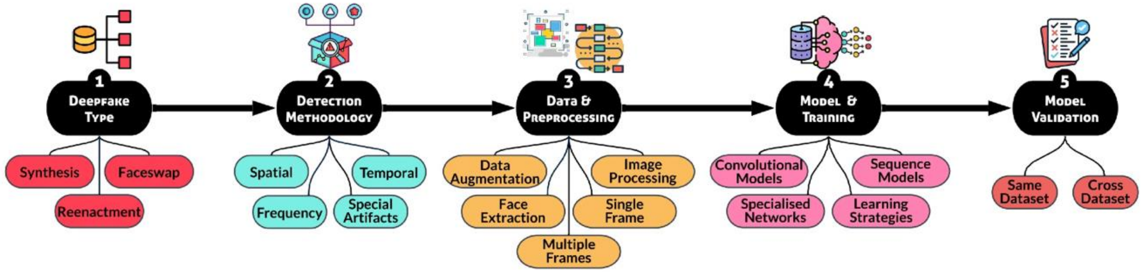
Jiwon Kim
Sungkyunkwan University, S. Korea
merwl0@g.skku.edu

Simon S. Woo*
Sungkyunkwan University, S. Korea
swoo@g.skku.edu

Kristen Moore
CSIRO's Data61, Australia
kristen.moore@data61.csiro.au

Alsharif Abuadbba
CSIRO's Data61, Australia
sharif.abuadbba@data61.csiro.au

Shahroz Tariq
CSIRO's Data61, Australia
shahroz.tariq@data61.csiro.au

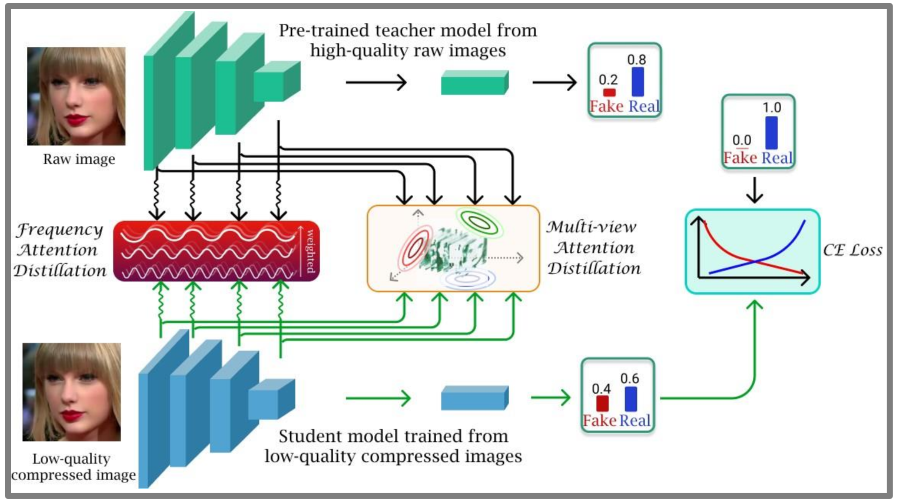


https://arxiv.org/pdf/2401.04364

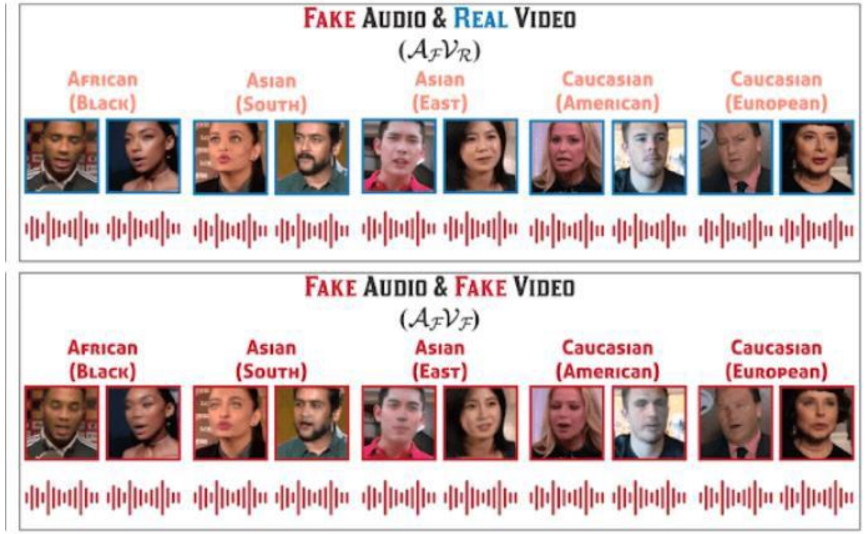
Table 3: **Systematic Classification of Deepfake Detectors.** In conceptual framework representations, white nodes indicate no papers fitting the category, half-colored nodes represent partial category representation, and fully colored nodes signify complete representation within the category (see Appx. Table 7 for details on detectors). The "FF++ Score" column displays each detector's performance on the FF++ dataset. Detectors marked with † were selected for further evaluations in Sec. 4.

	Focus of Methodology	Distinct Technique of Detector Architecture	Conceptual Framework Representation	Venue	Year	Detector Name	FF++ Score
Spatial Artifacts	**CF #1.** *ConvNet Models*	*Capsule Network*		ICASSP	'19	Cap.Forensics† [85]	96.60 (AUC)
		Depthwise Convolutions		ICCV	'19	XceptionNet† [98]	99.26 (ACC)
		Face X-ray Clues		CVPR	'20	Face X-ray [71]	98.52 (AUC)
		Unified Methodology		CVPR	'20	FFD [106]	-
		Bipartite Graphs		CVPR	'22	RECCE [6]	99.32 (AUC)
		Consistency Loss		CVPRW	'22	CORE [87]	99.94 (AUC)
		Face Implicit Identities		CVPR	'23	IID [49]	99.32 (AUC)
		Multiple Color Spaces		WACVW	'23	MCX-API† [127]	99.68 (AUC)
	CF #2. *Specialized Networks*	*Siamese Training*		ICPR	'20	EffB4Att† [3]	94.44 (AUC)
		Intra-class Compact Loss		AAAI	'21	LTW [107]	99.17 (AUC)
		Multi-attention losses		CVPR	'21	MAT† [132]	99.27 (AUC)
		Intra-instance CL		AAAI	'22	DCL [108]	99.30 (AUC)
		Self-blend Image		CVPR	'22	SBIs† [102]	99.64 (AUC)
	CF #3. *ConvNet Models with Learning Strategies*	*Adversarial Learning*		ACMMM	'21	MLAC [7]	88.29 (AUC)
		High Frequency Pattern		CVPR	'21	FRDM [77]	-
		Meta-learning		NeurIPS	'22	OST [8]	98.20 (AUC)
	CF #4. *ConvNet with Specialized Networks*	*Identity Representation*		CVPR	'23	CADDM† [27]	99.70 (AUC)
		Collaborative Learning		ICCV	'23	QAD [63]	95.60 (AUC)
	CF #5. *Sequence Models*	*Facial & Other Inconsistency*		CVPR	'22	ICT† [28]	98.56 (AUC)
		Unsupervised Inconsistency		ECCV	'22	UIA-ViT [137]	99.33 (AUC)
		Action Units		CVPR	'23	AUNet [2]	99.89 (AUC)
SpatioTemporal Artifacts	**CF #6.** *ConvNet Models*	*Facial Attentive Mask*		ACMMM	'20	ADDNet-3d [138]	86.69 (ACC)
		Anomaly Heartbeat Rhythm		ACMMM	'20	DeepRhythm [90]	98.50 (ACC)
		Multi-instance Learning		ACMMM	'20	S-IML-T [72]	98.39 (ACC)
		Time Discrepancy Modeling		IJCAI	'21	TD-3DCNN [130]	72.22 (AUC)
		Global-Local frame learning		IJCAI	'21	DIA [48]	98.80 (AUC)
		Local Dynamic Sync		AAAI	'22	DIL [41]	98.93 (ACC)
		Faces Predictive Learning		AAAI	'22	FInfer [47]	95.67 (AUC)
		Contrastive Learning		ECCV	'22	HCIL [42]	99.01 (ACC)
		Alternate Modules Freezing		CVPR	'23	AltFreezing† [126]	98.60 (AUC)
	CF #7. *ConvNet with Specialized Networks & Learning Strategies*	*SpatioTemporal Inconsistency*		ACMMM	'21	STIL [40]	98.57 (ACC)
		Reading Mouth Movements		CVPR	'21	LipForensics† [44]	97.10 (AUC)
		Temporal Transformer		ICCV	'21	FTCN† [134]	-
	CF #8. *Sequence Models*	*Combine ViT and CNN*		ICIAP	'21	CCViT† [14]	80.00 (ACC)
		Spatial-temporal Modules		WWW	'21	CLRNet† [116]	99.35 (F_1)
		Unsupervised Learning		NeurIPS	'22	LTTD [43]	97.72 (AUC)
Frequency Artifacts	**CF #9.** *ConvNet Models*	*Frequency Learning*		ECCV	'20	F3-Net [91]	98.10 (AUC)
		Single-center Loss		CVPR	'21	FDFL [69]	97.13 (AUC)
		Phase Spectrum Learning		CVPR	'21	SPSL [75]	95.32 (AUC)
		Spatial & Frequency Learning		AAAI	'23	LRL [11]	99.46 (AUC)
	CF #10. *ConvNet with Sequence Model & Learning Strategies*	*SpatioTemporal Frequency*		ECCV	'20	TRN [79]	99.12 (AUC)
		Knowledge Distillation		AAAI	'22	ADD† [64]	95.46 (ACC)
	CF #11. *Specialized Network & Learning Strategies*	*Intra-Sync with Frequency*		ECCV	'22	CD-Net [105]	98.50 (AUC)
		Collaborative Learning		CVPR	'23	SFDG [125]	95.98 (AUC)
Special Artifacts	**CF #12.** *ConvNet Models*	*Region Tracking*		CVPR	'21	RFM [122]	99.97 (AUC)
		Facial Features Modeling		CVPR	'21	FD2Net [136]	99.68 (AUC)
		2nd Order Anomaly		CVPR	'22	SOLA [37]	98.10 (AUC)
		Audio-video Anomaly		CVPR	'23	AVAD [38]	-
		Grad Pattern Learning		CVPR	'23	LGrad† [110]	66.70 (ACC)
	CF #13. *Sequence Model with Learning Strategies*	*Temporal Landmark Learning*		CVPR	'21	LRNet† [109]	99.90 (AUC)
		Noise Pattern Learning		AAAI	'23	NoiseDF [123]	93.99 (AUC)

- **Our group have researched deepfake detection and generation since 2017.**
- **We have published several top conference papers (AAAI, NeurIPS, WWW, ICML, ICCV) on deepfake detection, more than 25 publications in this area.**
- **Also, we have created and released deepfake video-audio dataset, “FakeVCeleb“.**
- **In addition, we have 2 international patents and transferred the deepfake detection technology**
- **We organized the workshop on deepfake and cheapfake (WDC) workshop 4 years in a row .**



RFP 주1) Woo et al., (2022). ADD: Frequency Attention and Multi-View Based Knowledge Distillation to Detect Low-Quality Compressed Deepfake Images. in Proc. of AAAI 2022 (pp. 122-130)



https://sites.google.com/view/fakeavcelebdash-lab/

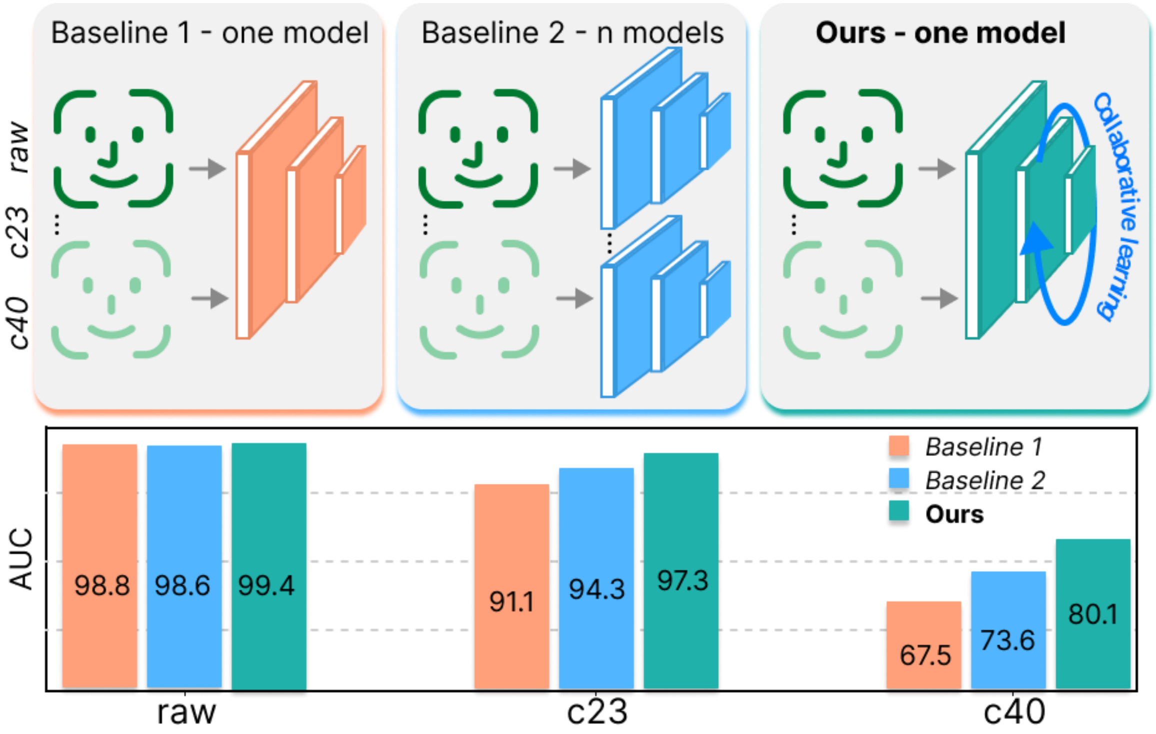
Still Many Challenges

- New generation methods (Attack and Defense)
  - How to handle new attacks and generation methods?
  - Is there a way to leverage existing architectures or pre-trained models?
- Challenges to generate new training dataset
  - Lack of training dataset?
  - Leverage existing dataset?
- Generalization & Explainability
- Low Quality Deepfakes
- International Synthetic Media Mitigation Efforts

Generalization

- Detection methods mainly assign various models to each quality of deepfake (ADD, BZNet), causing prohibitive overhead.

- In this work, we develop a unified model that can detect deepfake from various quality, called quality-agnostic deepfake detection (QAD), and improve overall performance.



Quality-Agnostic Deepfake Detection with Intra-model Collaborative Learning

Binh M. Le [1] and Simon S. Woo* [1]

Sungkyunkwan University, Suwon, South Korea [1]

IEEE/CVF International Conference on Computer Vision 2023

ICCV23 PARIS

성균관대학교 SUNGKYUNKWAN UNIVERSITY(SKKU)

DASH_LAB Data-driven AI Security HCI SUNGKYUNKWAN UNIVERSITY

email: bmle[AT]g.skku.edu

Notations

- A *raw sample* and its *compressed version* at quantile c are expressed as:

$$x_c = x_r - c$$

- Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} \subset \mathcal{X} \times \mathcal{Y}$
- Learning network $f: \mathcal{X} \to \mathbb{R}^2$ (*binary prediction*)
- Loss function $\mathcal{L}: \mathbb{R}^2 \times \mathcal{Y} \to \mathbb{R}$
- We consider loss as $\mathcal{L}(f(x), y) = 1 - \sigma_T(f(x), y)$ (σ_T is soft-max function)

Theorem and Our Motivation

For any network f, and with probability $1-\delta$ over the draw of $\mathcal{D}$:

$$\mathbb{E}[\mathbb{I}\{\hat{y}(x_c) \neq y\}] \leq 2\mathbb{E}_{\mathcal{D}}\mathcal{L}(f(x_c), y) + \frac{8}{T}\mathbb{E}_{\mathcal{D}}\mathcal{L}_{i-col}\big(f(x_c), f(x_r)\big) + 4\mathfrak{R}_{\mathcal{D}}(\Phi_{\mathcal{W}}) + \frac{16}{n} + \mathcal{O}\left(\sqrt{\frac{\log 2/\delta}{2n}}\right)$$

Where $\mathfrak{R}_{\mathcal{D}}$ is Rademacher complexity, $\Phi_{\mathcal{W}} = \{\mathcal{L}(f(x_r), y)\}$, and

$$\mathcal{L}_{i-col}\big(f(x_c), f(x_r)\big) = \|f(x_r) - f(x_c)\|$$

Theorem and Our Motivation

Minimizing *expectation error* by minimizing *classification loss* and *collaborative loss*

Expectation error

Classification loss

Apply Adversarial Weight Perturbation (***AWP***) to robust the model to corruptions

$$\mathbb{E}[\mathbb{I}\{\hat{y}(x_c) \neq y\}] \leq 2\mathbb{E}_{\mathcal{D}}\mathcal{L}(f(x_c), y)$$

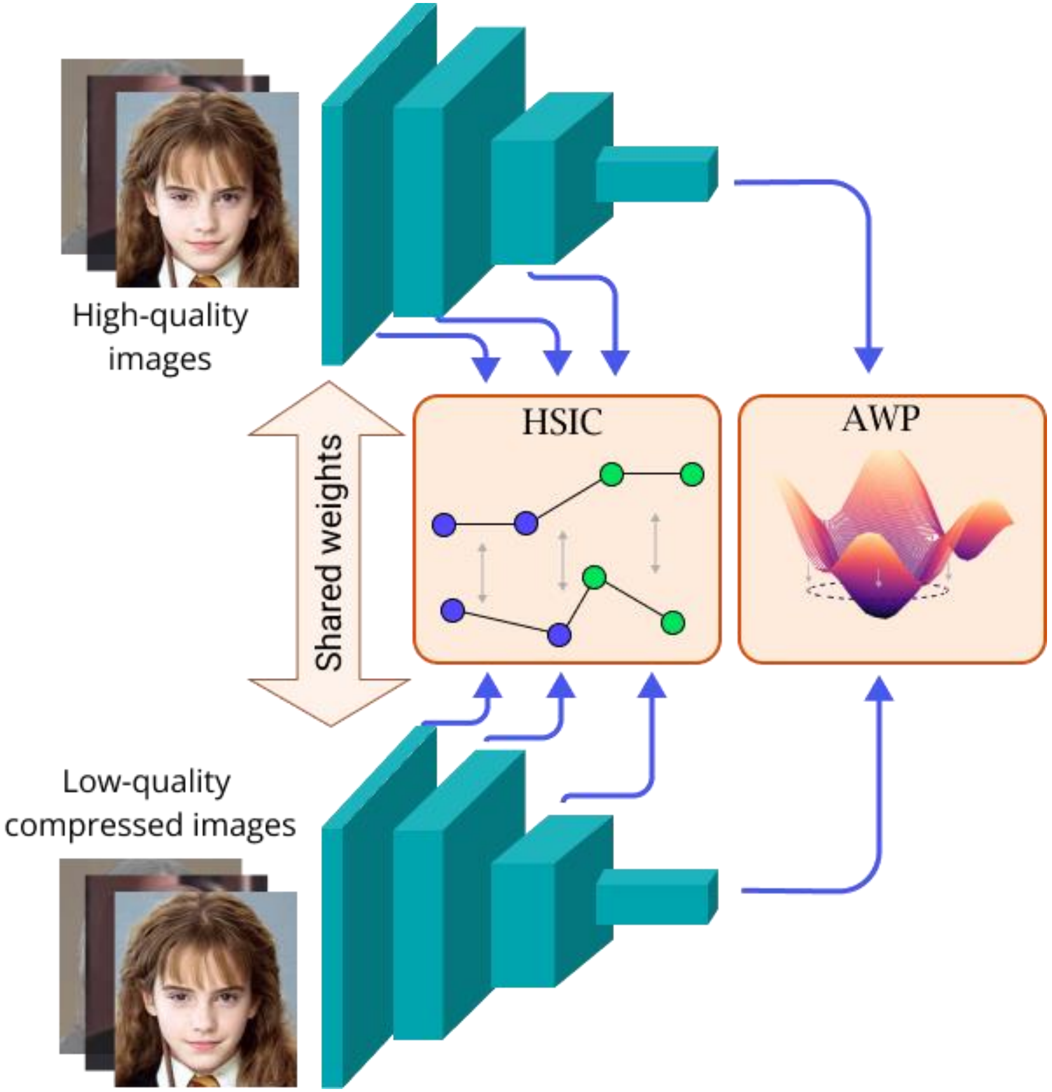
Instance-based collaborative loss

$$+\frac{8}{T}\mathbb{E}_{\mathcal{D}}\mathcal{L}_{i-col}\big(f(x_c), f(x_r)\big)$$

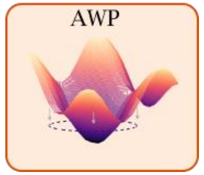
Apply Hilbert-Schmidt Independence Criterion (***HSIC***) to encourage geometrical similarity of raw data and the compressed data

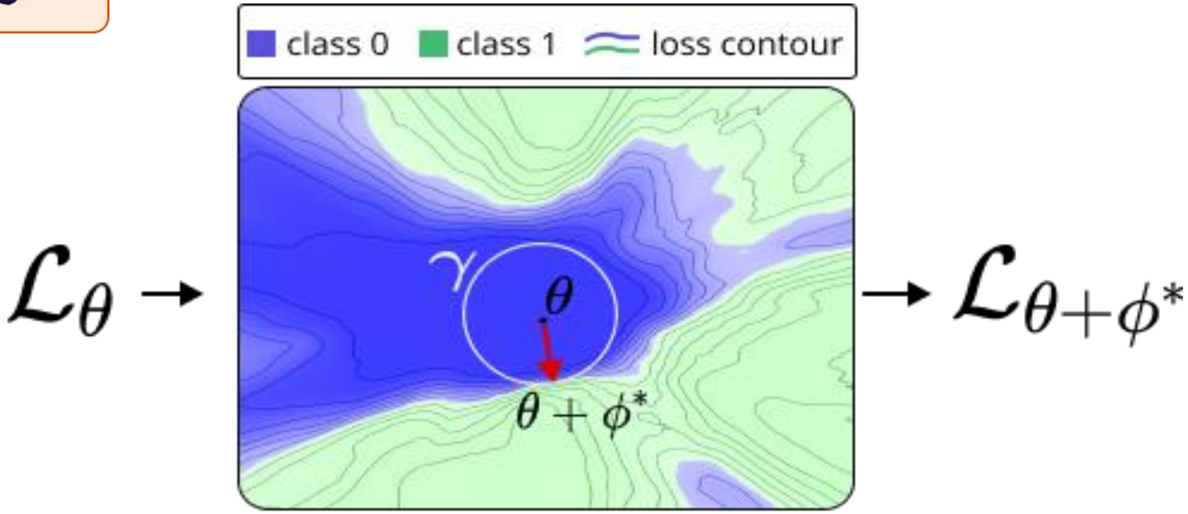
$$+ 4\mathfrak{R}_{\mathcal{D}}(\Phi_{\mathcal{W}}) + \frac{16}{n} + \mathcal{O}\left(\sqrt{\frac{\log 2/\delta}{2n}}\right)$$

Training scheme



Training scheme



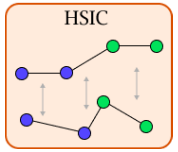


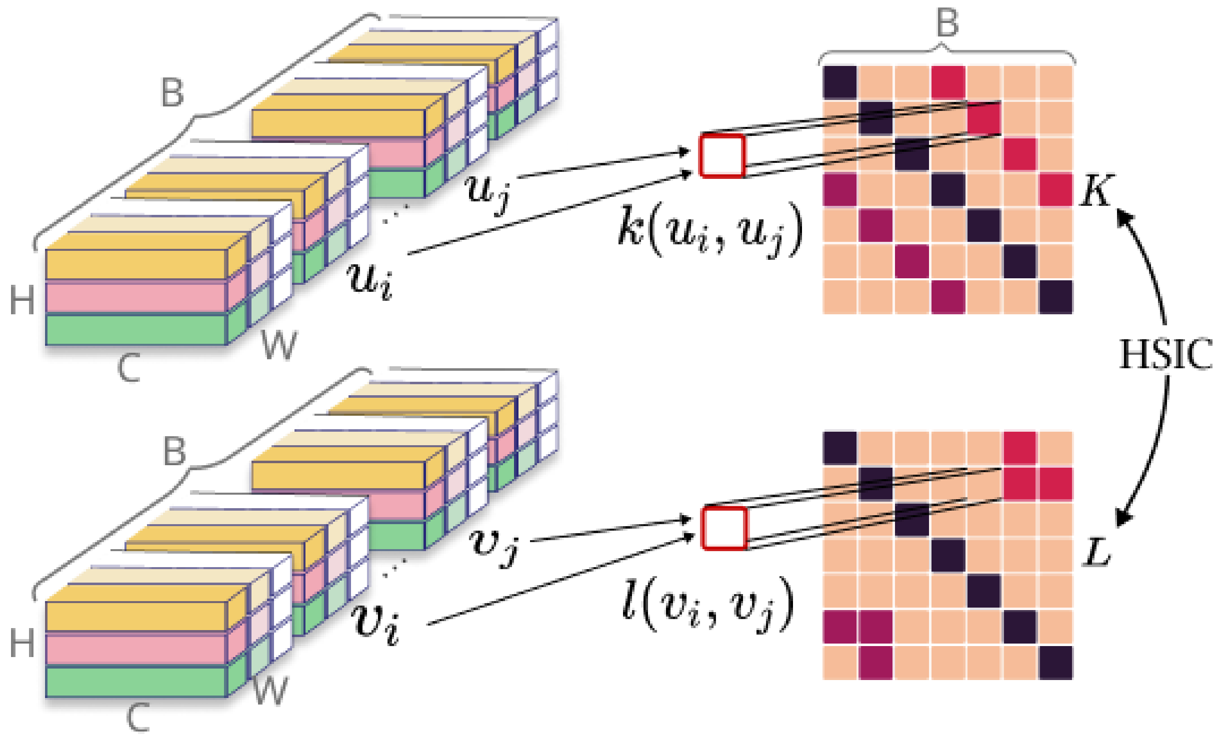
AWP in Weight loss landscape

- ϕ^*: worst-case perturbations of model weights *(significantly increase the loss).*
- ϕ^* are generated by adversarial methods:

$$\phi^* = \arg \max_{\phi \in \mathcal{B}(\theta,\gamma)} \mathcal{L}\big(f_{\theta+\phi}(x), y\big)$$

Training scheme





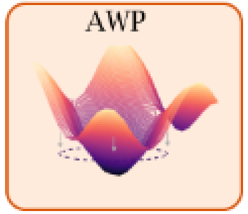
- HSIC measures the dependency between two random variables U and V via kernel k.
- A mini-batch included two quality τ and ρ at layer l^{th} are Z_l^{τ} and Z_l^{ρ}, collaborative loss:

$$\mathcal{L}_{col}(\tau, \rho) = -\sum_{l} \widehat{HSIC}\left(Z_l^{\tau}, Z_l^{\rho}\right)$$

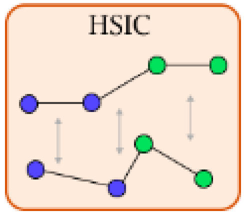
Overall training scheme

A mini-batch B include M quality modalities $\mathrm{T} = \{\mathrm{r}, \mathrm{c}_1, \dots, \mathrm{c}_{\mathrm{M}-1}\}$, end-to-end training objective:

$$\mathcal{L}_{QAD} = \frac{1}{\mathrm{M}B} \sum_{\tau \in T, i \in B} \mathcal{L}_{\phi^*}(x_{\tau,i}, y)$$



$$+\alpha \sum_{\tau,\rho \in T}^{\tau \neq \rho} \mathcal{L}_{col}(\tau, \rho)$$



Experimental results

Datasets (7): NeuralTextures (*NT*), Deepfakes (*DF*), Face2Face (*F2F*), FaceSwap (*FS*), FaceShifter (*FSH*), CelebDFV2 (*CDFv2*), and Face Forensics in the Wild (*FFIW10K*).

Compression (1+2): H.264 with quantile rate of 23 and 40: raw, c23 and c40.

Backbone (2): ResNet50 (QAD-R), and EfficientNet-B1 (QAD-E).

Baselines (8): MesoNet, XceptionNet, F3Net, Fan&Lin, SBIs, MAT, ADD, BZNet

Experimental results

Quality-agnostic: baselines models don't know input quality

Model	Test Set AUC (%)							
	NT	DF	F2F	FS	FSH	CDFv2	FFIW10K	*Avg*
Video Compression (raw + c23 + c40 of test set)								
MesoNet [1]◇	70.24	93.72	94.15	85.17	96.00	80.52	94.56	*87.77*
Rössler *et al.* [48] ◇	89.64	99.05	97.89	98.83	98.50	97.49	**99.17**	*97.22*
F^3Net [43] ◇	86.79	98.73	96.32	97.82	97.45	95.06	97.94	*95.73*
MAT [67]◇	86.79	98.73	96.32	97.82	97.45	95.06	97.94	*95.73*
Fang & Lin [11]	89.30	98.98	97.33	98.43	98.66	96.58	98.94	*96.89*
SBIs [51]†	78.33	95.19	79.74	80.37	80.48	-	-	*82.82*
BZNet [32]†	80.12	98.81	94.10	97.71	-	-	-	91.01
ADD [31]†	86.26	96.23	90.62	95.57	95.94	-	-	*92.92*
QAD-R (*ours*)	91.25	**99.54**	98.34	99.01	99.12	98.36	99.10	*97.82*
QAD-E (*ours*)	**94.92**	99.53	**98.94**	**99.27**	**99.12**	**98.38**	99.16	***98.47***

Video compression

Model	Test Set AUC (%)							
	NT	DF	F2F	FS	FSH	CDFv2	FFIW10K	*Avg*
Random Image Compression (JPEG on raw of test set)								
MesoNet [1]◇	70.23	92.02	88.32	82.60	91.84	81.12	91.87	*85.43*
Rössler *et al.* [48]◇	69.89	98.62	94.97	96.66	96.76	96.98	98.81	*93.24*
F^3Net [43]◇	70.95	97.89	92.83	96.34	94.72	95.44	97.19	*92.19*
MAT [67] ◇	69.53	98.96	**95.53**	97.99	96.97	98.21	98.91	*93.73*
Fang & Lin [11]	75.49	98.32	94.63	97.64	97.28	96.67	98.39	*94.06*
SBIs [51]†	77.75	97.83	82.05	86.10	85.42	-	-	*85.83*
BZNet [32]†	**79.00**	98.77	95.23	97.92	-	-	-	92.73
ADD [31]†	75.84	96.83	92.23	95.24	96.00	-	-	*91.23*
QAD-R (*ours*)	75.18	98.86	93.72	98.52	98.18	98.51	**98.96**	*94.56*
QAD-E (*ours*)	76.27	**99.20**	94.44	**98.69**	**98.60**	**98.52**	98.86	***94.94***

Random JPEG compression

Experimental results

Quality-aware: baselines models know input quality, except for our QAD

Method	w/ prior infor.	#params	Test Set AUC (%)							
			NT	DF	F2F	FS	FSH	CDFv2	FFIW10K	*Avg*
BZNet [32]† [×3]	Y	22M × 3	91.01	99.30	96.90	98.82	-	-	-	96.51
ADD [31]† [×3]	Y	23.5M × 3	89.08	99.25	96.53	98.21	98.25	-	-	96.26
RESNET50 [×3]	Y	23.5M × 3	88.96	99.26	97.04	98.63	98.71	97.09	98.58	*96.90*
QAD-R (*ours*)	N	23.5M × 1	88.85	99.42	97.77	98.83	98.93	97.56	98.93	*97.18*
EFFICIENTNET-B1[×3]	Y	6.5M × 3	87.63	99.05	96.72	98.16	97.95	96.70	98.54	*96.39*
QAD-E (*ours*)	N	**6.5M × 1**	**92.25**	**99.46**	**98.30**	**99.08**	**98.90**	**97.50**	**99.01**	***97.79***

Video compression

Ablation studies

Model / loss		ResNet-50 ACC (%)	ResNet-50 AUC (%)
Baseline		78.8	88.2
Coll. loss	Soft-label	77.0	84.0
	Pairwise loss	79.7	89.1
	Center loss	79.8	88.9
	HSIC	**80.3**	**90.1**
Adv. loss	AWP-KL	80.9	89.4
	AWP-XE	**81.7**	**90.7**
QAD *(ours)*		**82.2**	**91.3**

Table 4. Performance (ACC & AUC) of ResNet50 integrated with different loss functions.

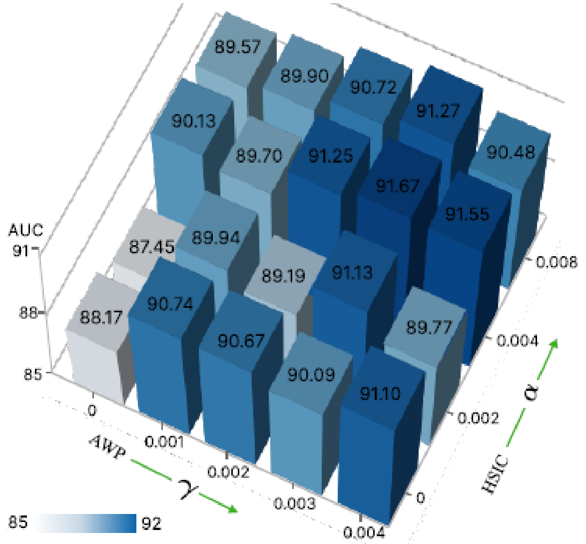


Figure 3. Model's performance versus α and γ on the NeuralTextures.

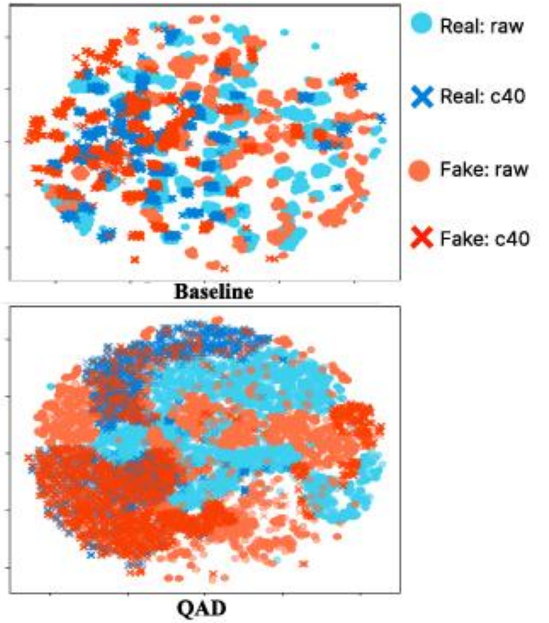


Figure 4. t-SNE visualisation of baseline and our QAD.

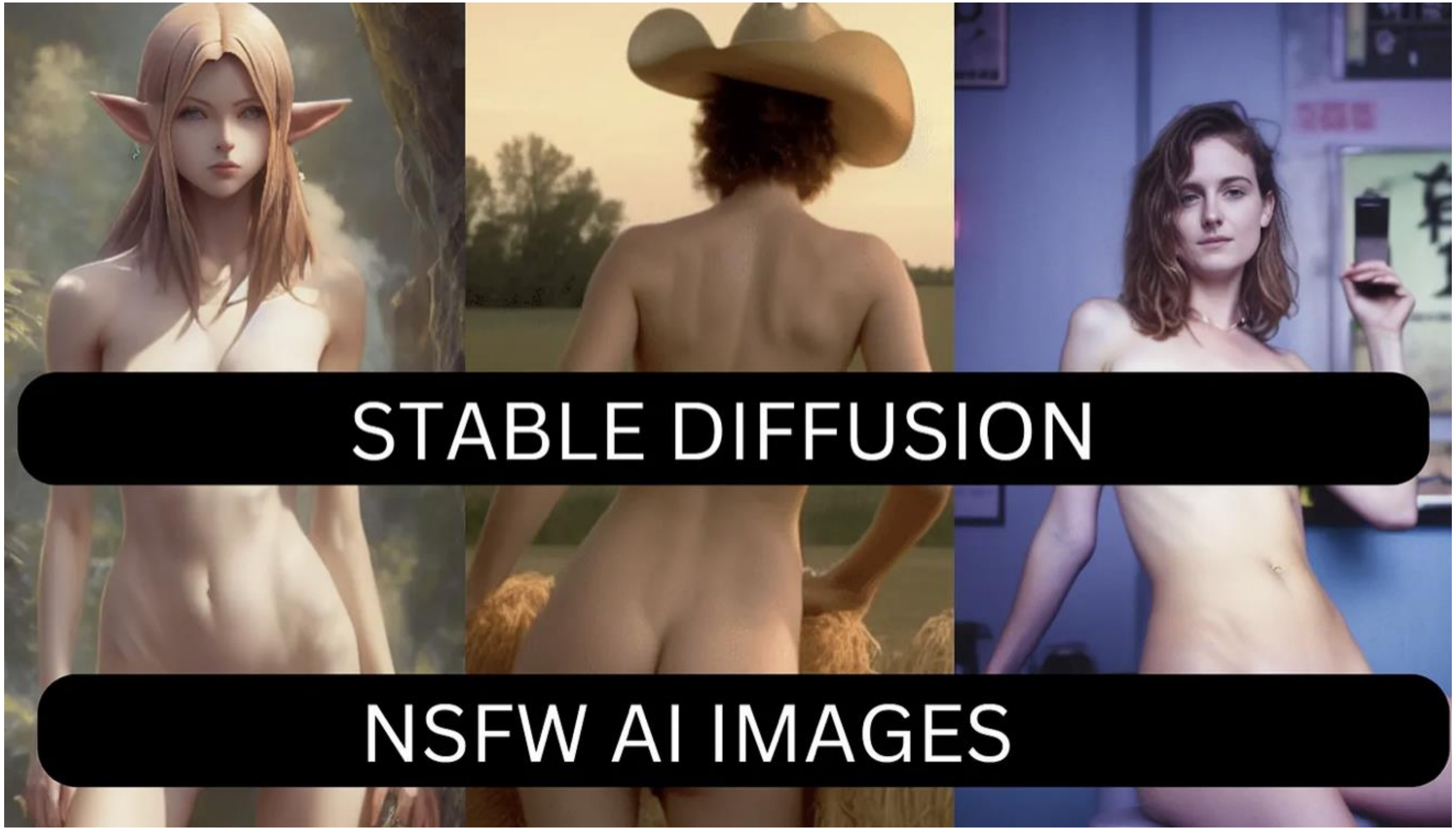
Pairwise differences of various quality image representations at the output can hinder its convergence.

Increasing γ improve performance. When $\alpha > 2e-3$, model's performance is stable.

QAD's representations are less dispersed both in terms of intraclass and inter-quality.

https://www.reddit.com/r/StableDiffusion/comments/161n6sd/donald_trump_jail_photos_made_with_stable/?rdt=49256

https://www.theguardian.com/us-news/2024/mar/04/trump-ai-generated-images-black-voters



https://jimclydemonge.medium.com/this-website-can-generate-nsfw-images-with-stable-diffusion-ai-1ee2913de829

Suppressing Synthetic Content Generation and Concept Erasing

Hong S, Lee J, Woo SS. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) 2024 Mar 24 (Vol. 38, No. 19, pp. 21143-21151).

All but One: Surgical Concept Erasing with Model Preservation in Text-to-Image Diffusion Models

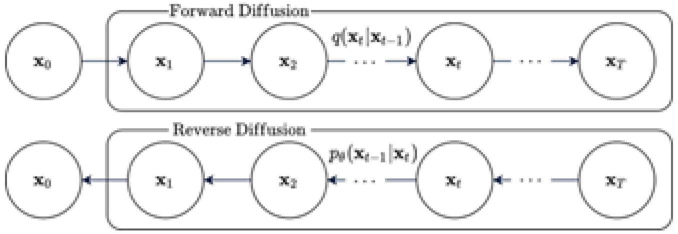
Seunghoo Hong, Juhun Lee and
Simon S. Woo*
Department of Artificial Intelligence, Sungkyunkwan University,
Suwon, South Korea
hoo0681@g.skku.edu, josejhlee@g.skku.edu, swoo@g.skku.edu

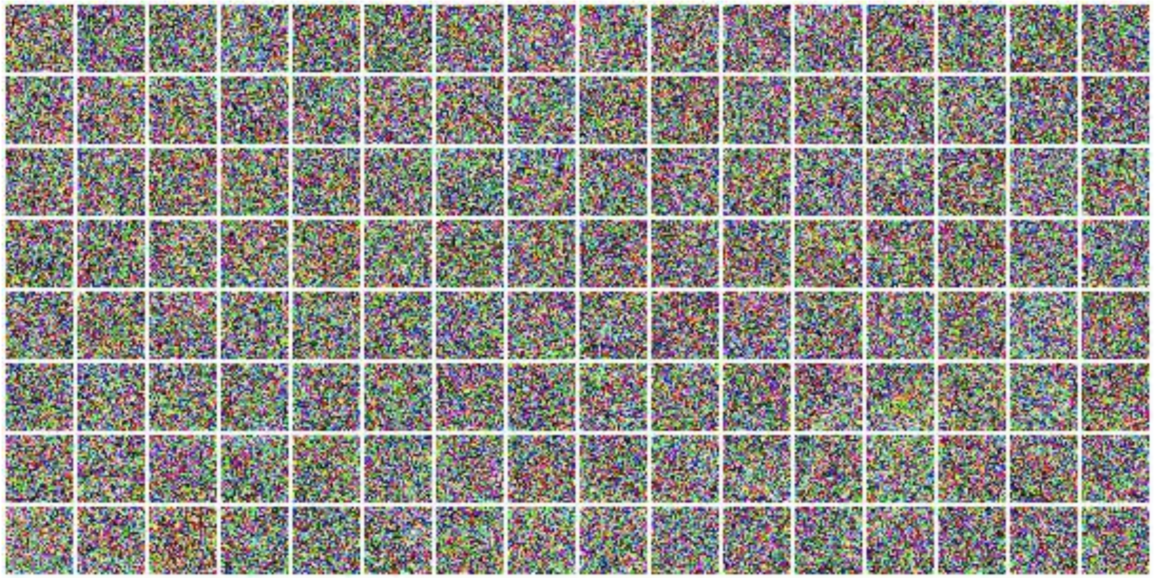
Data-driven AI Security HCI
SUNGKYUNKWAN UNIVERSITY

https://dash-lab.github.io/

Diffusion Models

Diffusion models has show impressive image modelling capability.

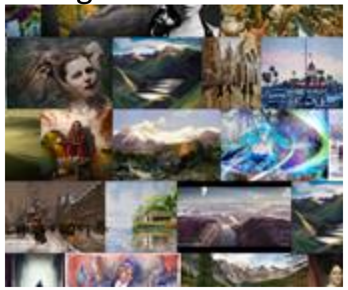




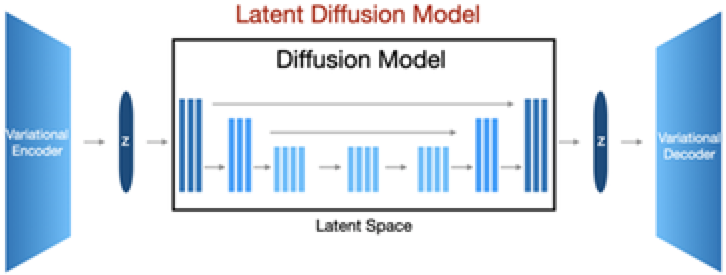
Scalability in Diffusion Models

The joint development in dataset acquisition enabled "foundational" generative performance.

large scale dataset



+

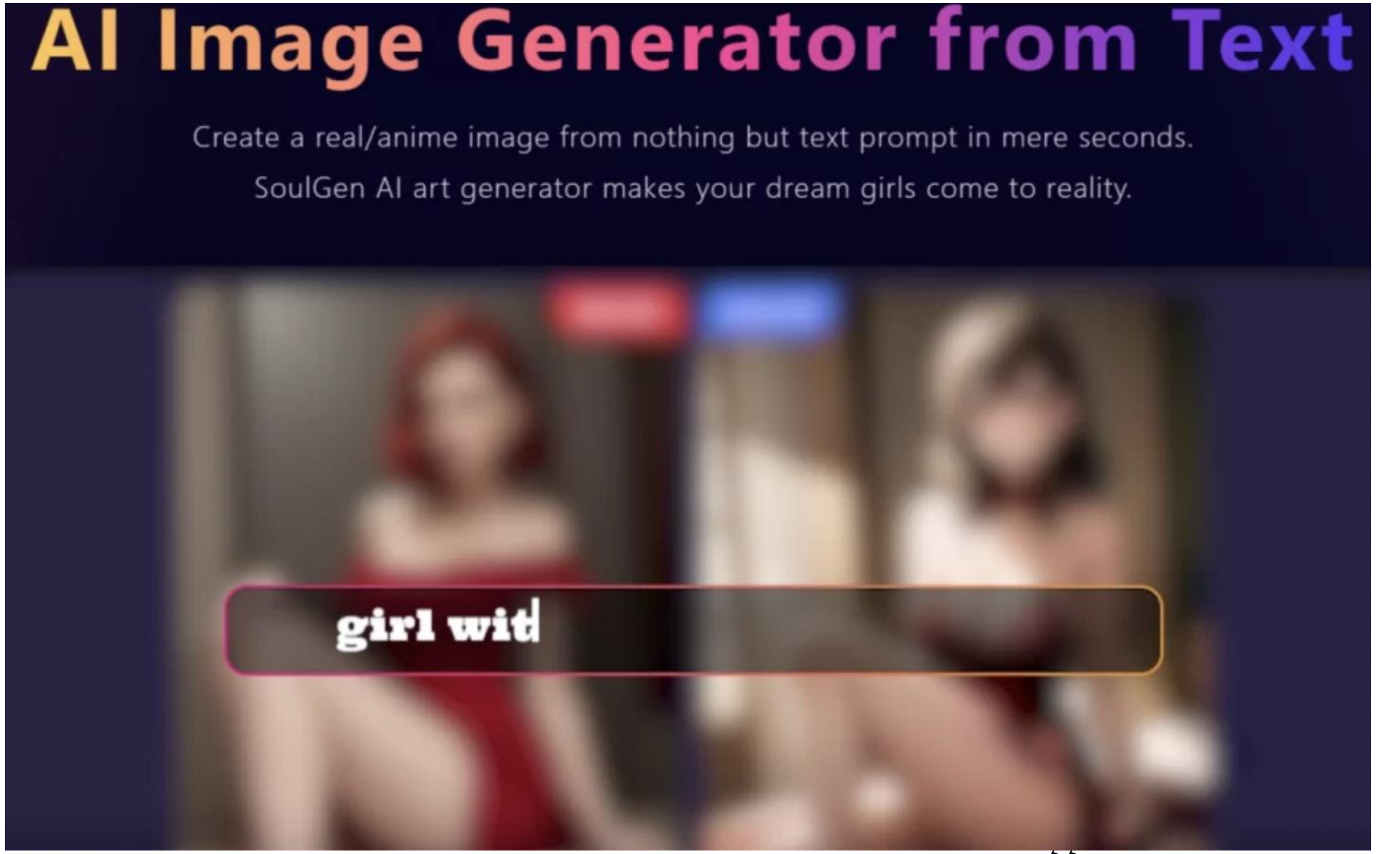


=

Extrapolative
data generation

Ethical Issues in Large Datasets

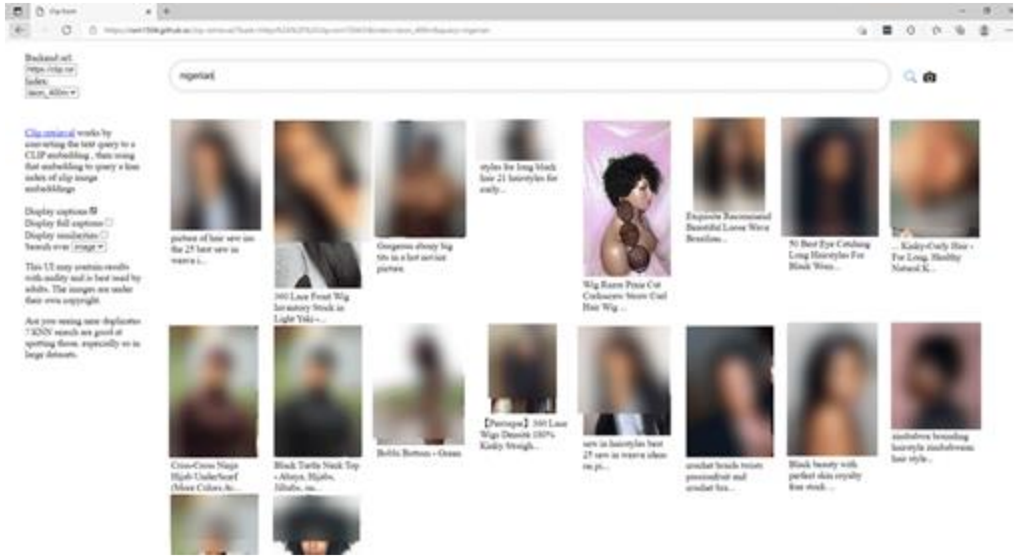
Not Safe For Work (NSFW) Content Generation



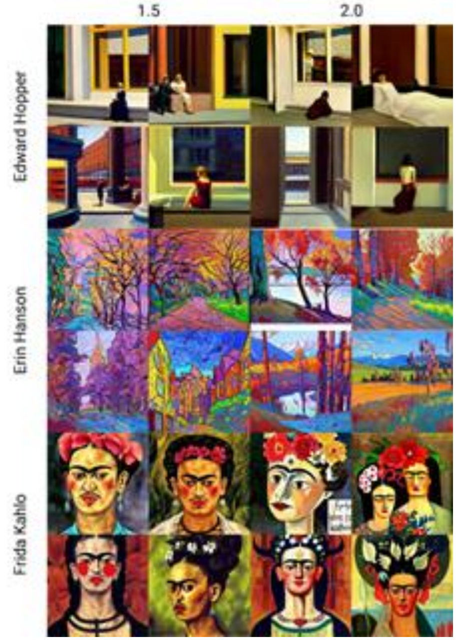
Ethical Issues in Large Datasets

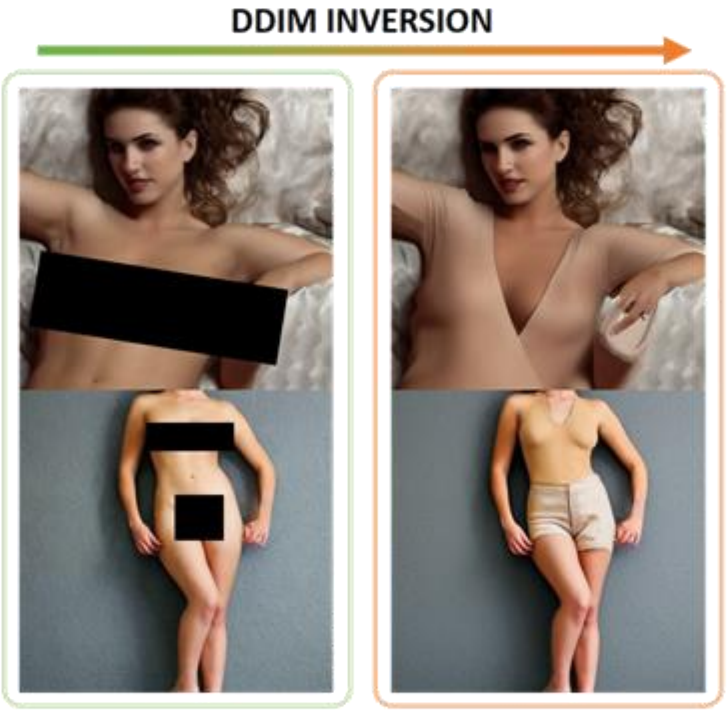
The joint development in dataset acquisition enabled "foundational" generative performance.

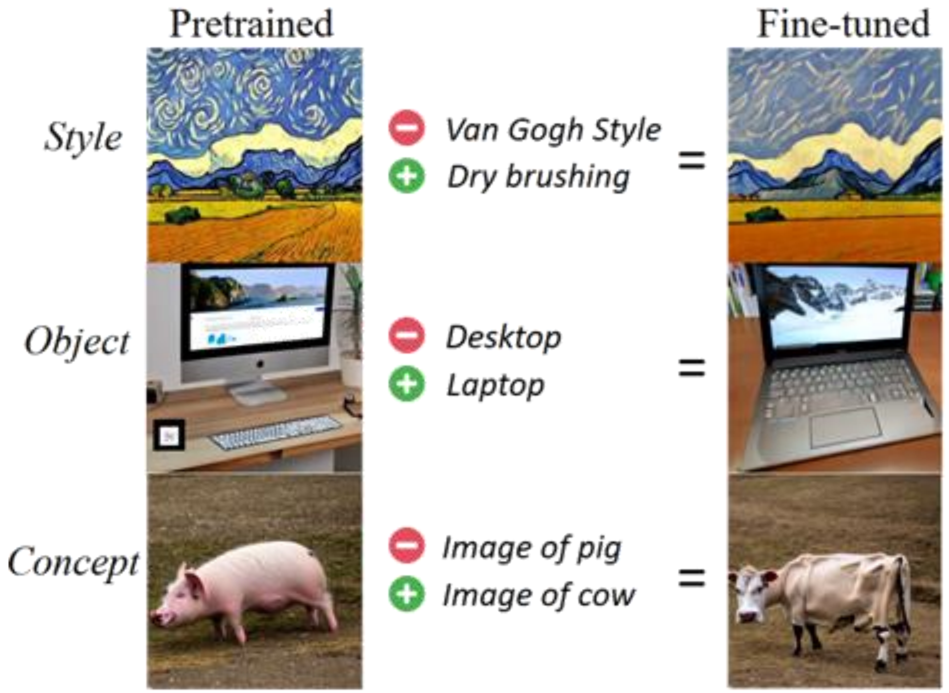
NSFW content in LAION dataset



Model synthesizing close to real images




DDIM INVERSION


Pretrained
Fine-tuned
Style
Van Gogh Style
Dry brushing
=
Object
Desktop
Laptop
=
Concept
Image of pig
Image of cow
=

Concept Ablation

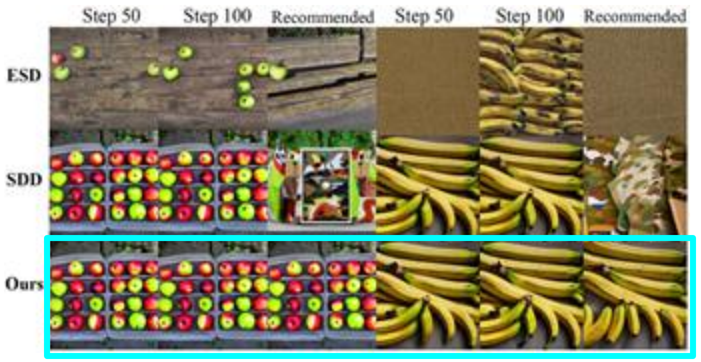
To circumvent retraining, fine-tuning methods were proposed to delete the target concept.

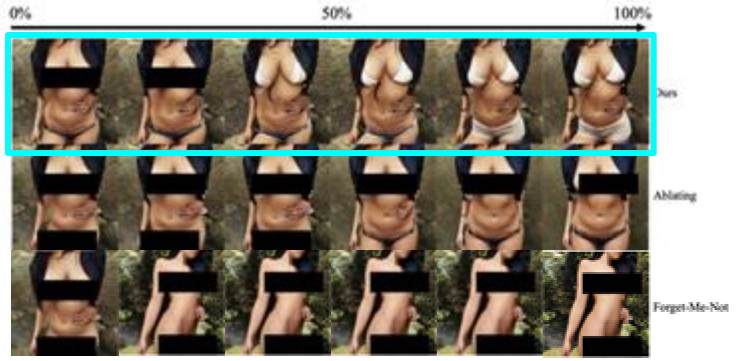
ESD & SDD

- object disintegration
- slow convergence

Ablating & Forget-me-Not

- less competitive erasing





In our work, we achieve both good **concept erasure** while preserving the **model's utility**.

Our Method (1)

The conditional score $\nabla \log \hat{P}(z_t|c)$ is:

$$\nabla \log \hat{P}(z_t|c) = \nabla \log P(z_t|\emptyset) + \gamma \boxed{\nabla \log P(c|z_t)}$$

The goal is to update this latter term:

$$\boxed{\nabla \log P(c|z_t)} \rightarrow \nabla \log P(c'|z_t)$$

Original guidance → Alternate guidance

Then, our objective can be formulated as:

objective: $\arg\min_{\theta} [\| \gamma_1 \nabla \log P(c'|z_t) - \gamma_2 \nabla \log P(c|z_t) \|_2^2]$

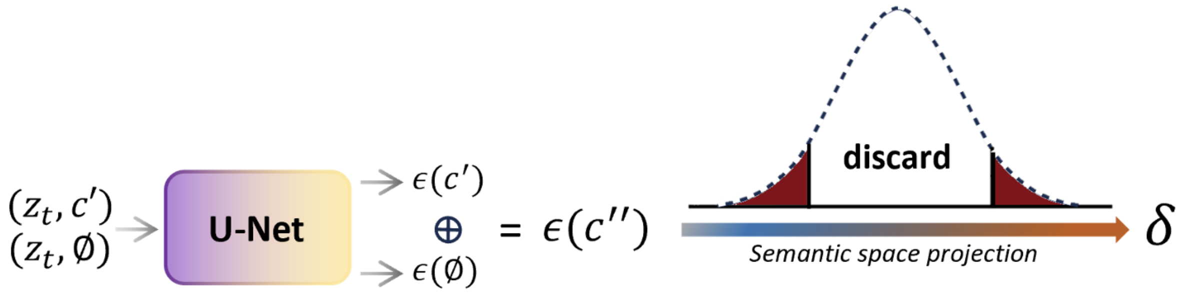
Our Method (2)

To sample the alternate guidance term $\nabla \log P(c'|z_t)$, the key consideration is:

Keep the model's update to the minimal.

Introduce only relevant signal to the fine-tuning.

SEGA shows that semantic signal is concentrated at the extremes of $\epsilon(\cdot)$.
With an alternative concept pre-assigned:



Ultimately, we obtain the alternate guidance term:

$$\nabla \log P(c'|z_t) = \gamma_1(\epsilon_{\theta^\star}(z_t, c) - \epsilon_{\theta^\star}(z_t)) + \delta(c', z_t, \theta^\star))$$

Our Method (3)

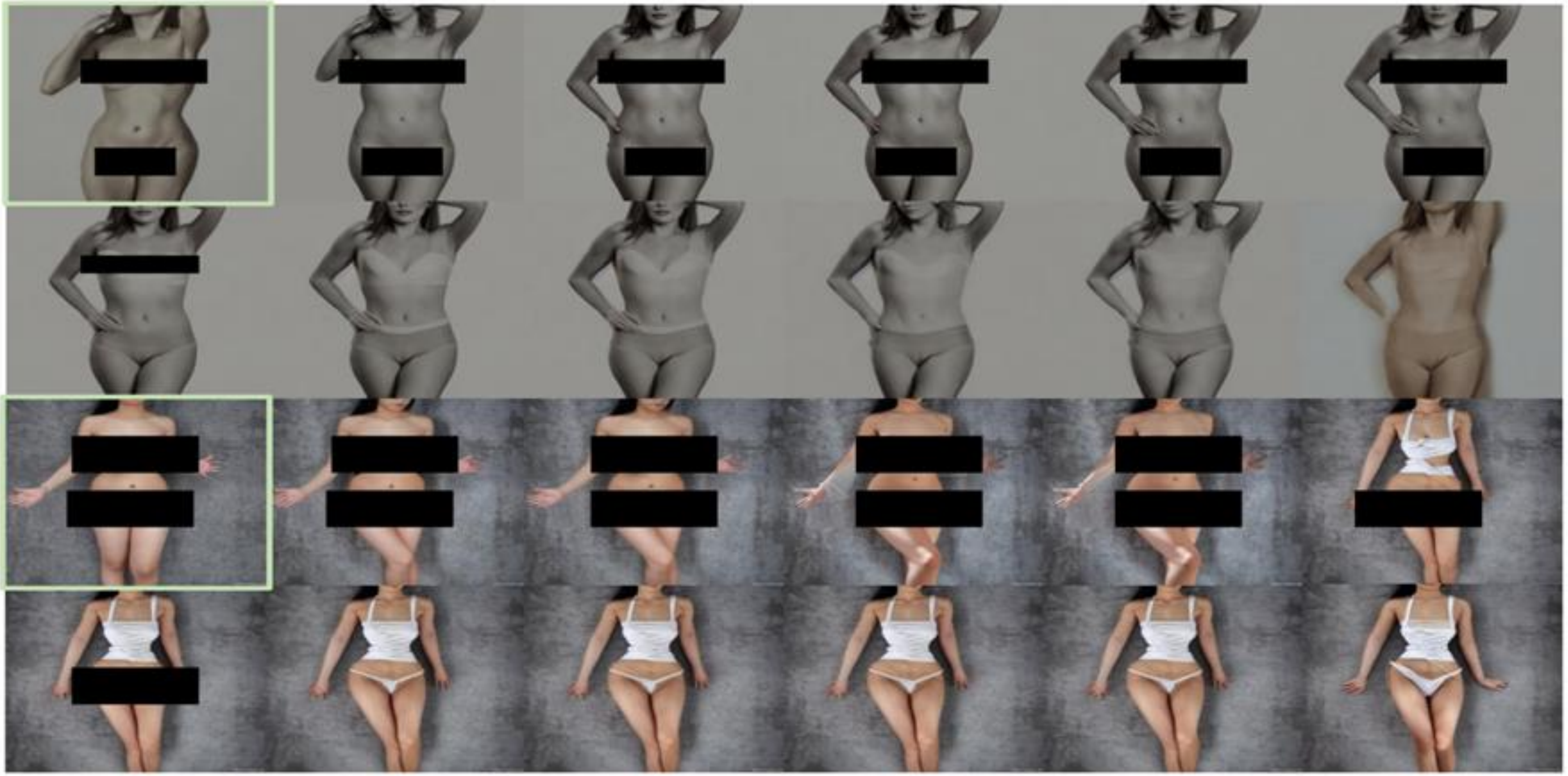
Recapitulating, we update the concept in the conditional score:

$$\nabla \log \hat{P}(z_t|c) = \nabla \log P(z_t|\emptyset) + \gamma \underbrace{\boxed{\nabla \log P(c|z_t)}}_{\text{updating concept}}$$

For updating concept:

$$\min_{\theta} \mathbb{E}_{\mathbf{z},t}[\| \gamma_1 \nabla \log P_{\theta^\star}(c'|z_t) - \gamma_2 \boxed{\nabla \log P_{\theta}(c|z_t)} \|_2^2]$$

$$\mathcal{L}_{\text{concept}}(c, c', z_t, \gamma_1, \gamma_2) = \| \gamma_2 \left(\boxed{\epsilon_\theta \left(z_t, c \right)} - \epsilon_\theta \left(z_t \right) . \mathrm{sg}() \right) - \gamma_1 \left(\epsilon_{\theta^\star} \left(z_t, c' \right) - \epsilon_{\theta^\star}(z_t,) \right) \|_2^2$$



Erasing timeline during fine-tuning in respect to a single seed every 10 iterations (last at 450). Spatial consistency is preserved even.

Our Method (4)

The style of erasing can depend on the end user or the target concept.
In addition to the **concept update**, we **preserve the null token**'s representation.

preserving null token **updating concept**

$$\nabla \log \hat{P}(z_t|c) = \boxed{\nabla \log P(z_t|\emptyset)} + \gamma \boxed{\nabla \log P(c|z_t)}$$

For preserving the null token's representation:

$$\text{s. t.} \nabla \log P_{\theta^\star}(z_t) - \boxed{\nabla \log P_\theta(z_t)} = 0, \ \forall z_t, t = 1, \ldots, T,$$

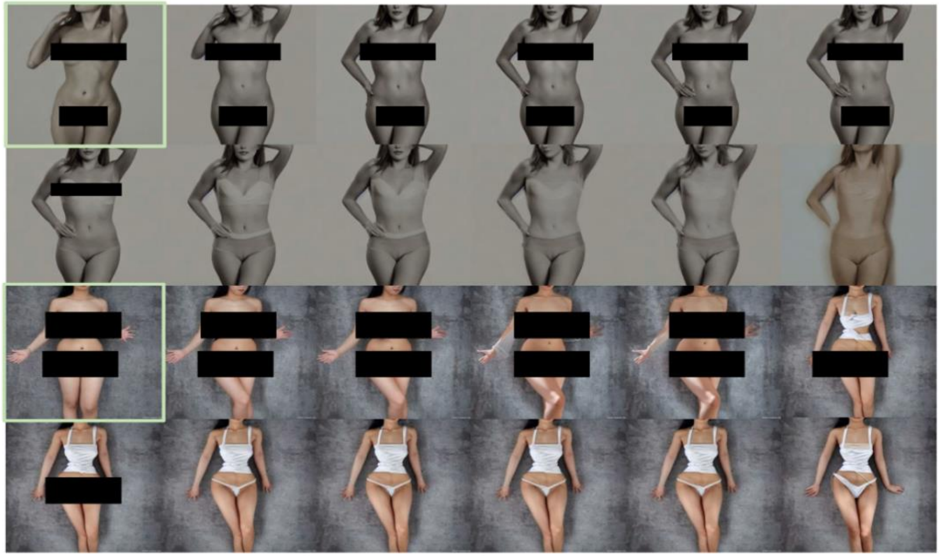
$$\mathcal{L}_{\text{penalty}}(t, z_t) = \|\boxed{\epsilon_\theta(z_t,)} - \epsilon_{\theta^\star}(z_t,)\|_2^2$$

Ultimately, the final loss is:

$$\mathcal{L}_{\text{model}} = \mathbb{E}_{z_t \sim P_{\theta^\star}(z_t|c'), c, c', t}[\mathcal{L}_{\text{concept}} + \lambda \mathcal{L}_{\text{penalty}}]$$

Experimental Result (1)

To quantify model's **utility preservation**,we use FID, KID, CLIP Score, and SSIM. To quantify **concept erasure**, we use NudeNet's score.



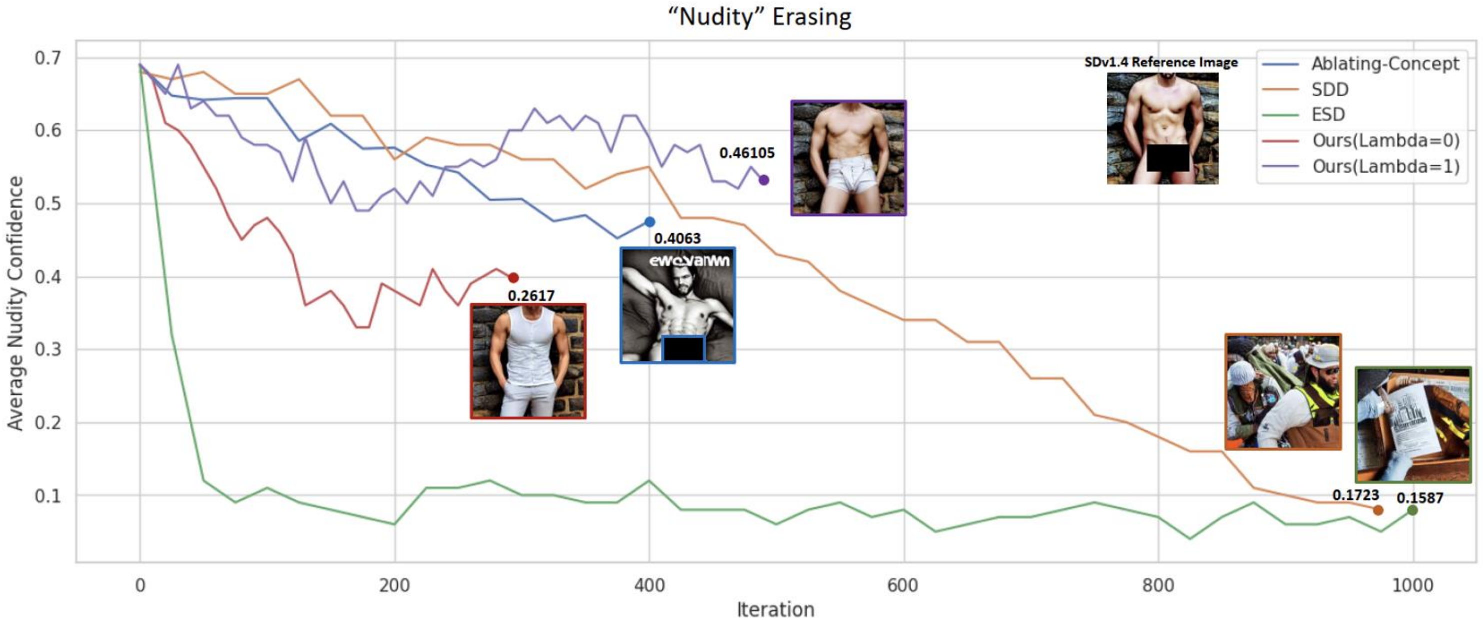
Erasure evaluation under increasing iterations

Method	NudeNet(%)↓	FID↓	KID↓	CLIP Score↑	SSIM↑
SD v1.4	0.69	13.59	0.00479	0.2765	-
ESD	**0.04**	14.27	**0.00421**	0.2619	0.231
SDD	0.05	14.11	0.00499	0.2677	0.309
Ablating	0.45	13.68	0.00478	0.2756	0.657
Forget-Me-Not	0.66	13.78	0.00496	0.2732	0.476
Ours	0.33	**13.19**	0.00447	**0.2762**	**0.762**
COCO				0.2693	

Experimental Result (2)

Visualization of “nudity” erasure across iterations.

While competing methods either completely change the generation (ESD,SDD) or erases weakly (Ablating-Concept), our method achieves both preservation and erasing.




Reference images
ESD
Forget-Me-Not
SDD
Ablation
Ours
"Nude women"
"Nude woman"
"Nude Statue"
"Nude man"

Experimental Result (3)

Given a denoiser, one can apply DDIM inversion to real images.

$$x_{t+1} - x_t = \sqrt{\bar{\alpha}_{t+1}} \left[\left(\sqrt{1/\bar{\alpha}_t} - \sqrt{1/\bar{\alpha}_{t+1}} \right) x_t + \left(\sqrt{1/\bar{\alpha}_{t+1} - 1} - \sqrt{1/\bar{\alpha}_t - 1} \right) \epsilon_\theta(x_t) \right]$$

Our loss formulation leads to a DDIM inversion with concept ablation.

Still Many Challenges on Deepfake Research

- New generation methods (Attack and Defense)
- Challenges to generate new training dataset
- Generalization & Explainability
- Low Quality Deepfakes
- Real World Deepfake Detection
- International Synthetic Media Mitigation Efforts

Acknowledgement

Students in our DASH Lab and CSIRO researchers for the co-work!


Binh


Jiwon (Merlyn)


Jose


Hoo


Sam


Hyun


Tran


Shahroz


Sharif


Kristen

Acknowledgement and Thanks!

Ministry of Science and ICT

IITP Institute of Information & Communications Technology Planning & Evaluation

SUNGKYUNKWAN UNIVERSITY(SKKU) | College of COMPUTING

Thanks!

Q&A

swoo@g.skku.edu

https://dash-lab.github.io/Publications/

Very happy to collaborate with you !