NEURAL INFORMATION PROCESSING SYSTEMS

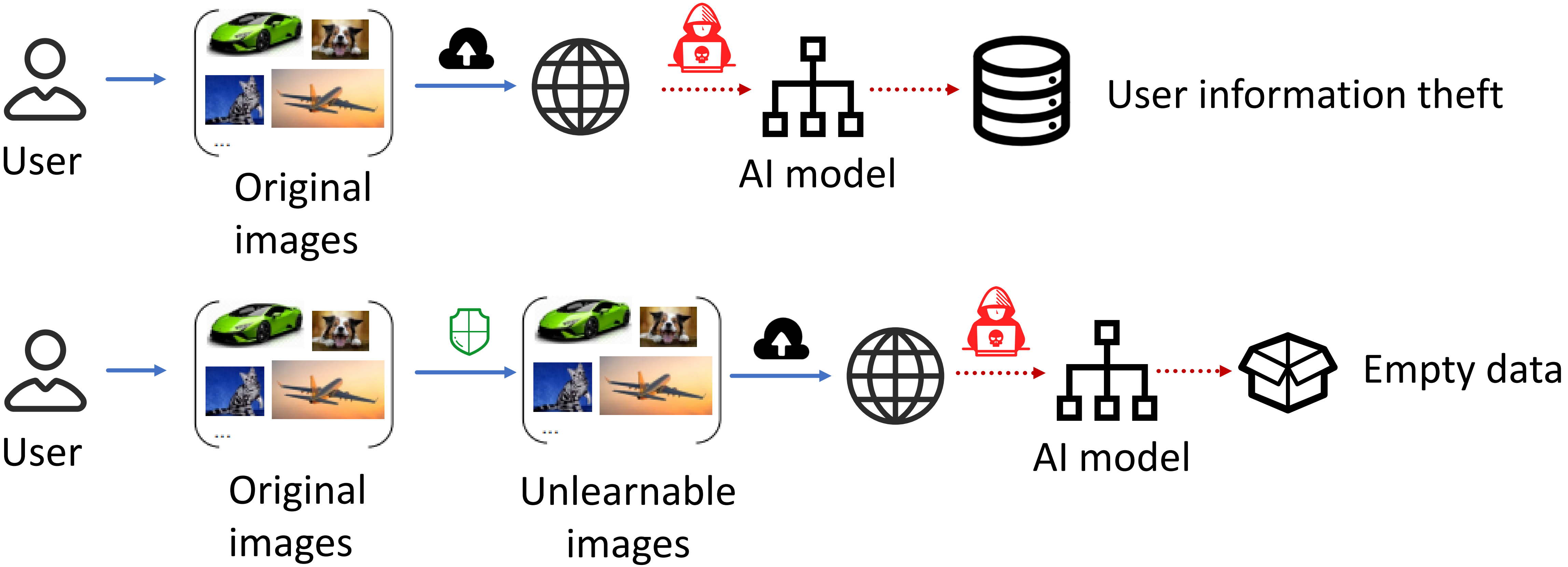
Versatile Transferable Unlearnable Example Generator

Zhihao Li[1*] Jiale Cai[1*] Gezheng Xu[1] Hao Zheng[1,2] Qiuyue Li[3]
Fan Zhou[3] Shichun Yang[3] Charles Ling[1, 4] Boyu Wang[1,4]

[1]Western University [2]Central South University [3]Beihang University [4]Vector Institute

Why Unlearnable Examples?

- The abundance of online data → rapid deep learning advances
- Concern: Personal data leakage in training
- Solution: Unlearnable Examples (UE) → perturb data to confuse training



Where Existing UEs Fall Short

- Most methods target training-set-specific data
- Poor performance in non-target or shifted settings
- Preliminary works only handle partial scenarios

Method	Intra-Domain	Cross-Domain	Cross-Task	Cross-Space	Cross-Architecture
EMN	✓	×	×	×	✓✗
LSP	✓	×	×	×	✓
TUE	✓	✓✗	×	×	✓✗
GUE	✓	✓✗	×	×	✓
14A	✓	✓✗	✓✗	×	✓
VTG	✓	✓	✓	✓	✓

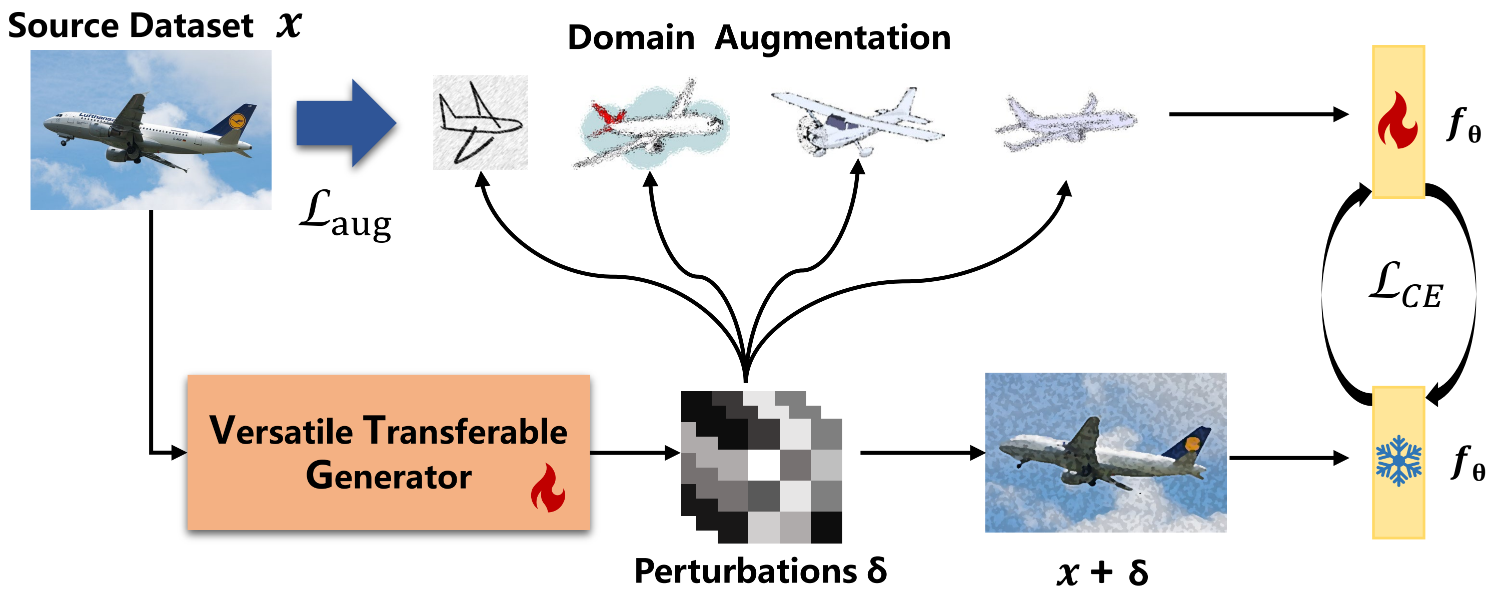
Versatile Transferable Generator (VTG)

- **Adversarial Domain Augmentation (ADA)** synthesizes out-of-distribution samples, thereby improving its generalizability to unseen scenarios.

$$\arg\min_{\theta,\mu}[\mathcal{L}_{CE}(f_\theta(x+\delta),y)+\mathcal{L}_{CE}\big(f_\theta\big(\mathbb{C}_\mu(x)+\delta\big),y\big)]$$

- **Perturbation Generator** produce unlearnable perturbations for any image in a single forward pass, exhibiting superior generalizability and applicability for practical use.

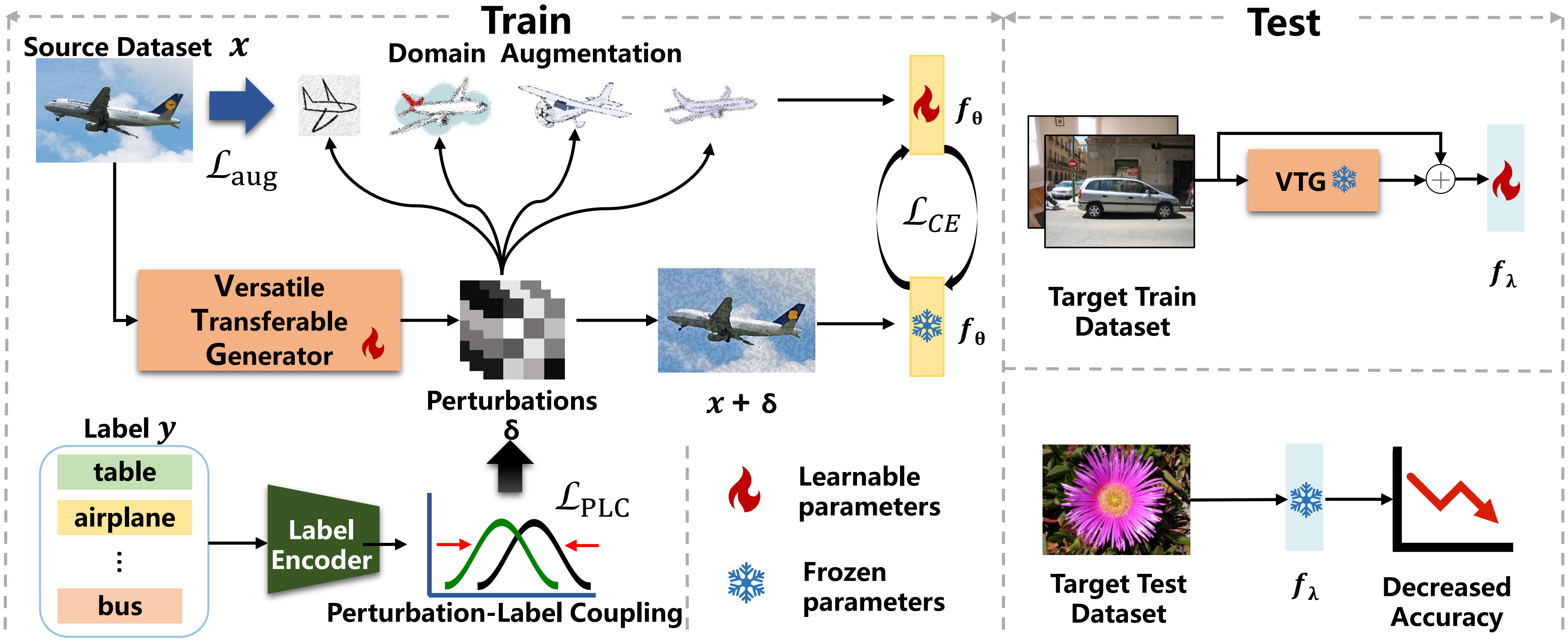
$$\arg\min_{\delta}\mathbb{E}_{(x,y)\sim\mathcal{D}_{source}}\mathcal{L}_{CE}(f_\theta(x+\delta),y) \qquad \mathbb{E}_x\max(0,\|\delta\|_\infty-\epsilon)$$



Versatile Transferable Generator (VTG)

- **Adversarial Domain Augmentation (ADA)**
- **Perturbation Generator**
- **Perturbation-Label Coupling (PLC)** leverages contrastive learning to directly align perturbations with class labels.

$$\mathcal{L}_{PLC} = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\exp(S_{i,y_i})}{\sum_{k=1}^{K}\exp(S_{i,k})}\right)$$



Experimental Details

Comprehensive UE Transferable Evaluation Scenarios

- **Intra-Domain scenario** represents the conventional setting, where the training and test data are drawn from the same distribution.
- **Cross-Domain scenario** considers cases where the training and test sets share the same classes but originate from different distributions.
- **Cross-Task scenario** increases the challenge by introducing both distribution shifts and class mismatches.
- **Cross-Space scenario** is the most challenging scenario, where even the input space differs between training and test sets.
- **Cross-Architecture scenario** evaluates the generalizability of UEs across different network architectures.

Intra-Domain and Cross-Task scenarios

- **VTG is more effective than other methods on CIFAR-10, CIFAR-100 and SVHN dataset.**
- **VTG gets random-guess level under all setting in both Intra-Domain and Cross-Task scenarios.**

Source	Method	CIFAR-10	CIFAR-100	SVHN
	Clean	94.66	76.27	96.05
	Random	95.57	71.19	25.11
CIFAR-10	EMN [4]	10.16	21.80	24.72
	LSP [10]	13.54	9.35	**7.77**
	REM [5]	15.18	69.26	95.98
	TUE [6]	10.03	5.10	12.93
	GUE [7]	13.25	3.87	8.17
	14A [11]	41.34	17.47	83.87
	PUE [35]	10.62	8.46	12.01
	Ours (ResNet)	9.99	**0.99**	9.65
	Ours (ViT)	**9.54**	1.21	7.94
CIFAR-100	EMN [4]	27.27	3.95	9.64
	LSP [10]	24.16	9.00	17.03
	REM [5]	93.94	1.89	95.97
	TUE [6]	94.31	1.21	96.02
	GUE [7]	94.28	8.35	95.87
	14A [11]	40.02	17.36	85.18
	PUE [35]	11.61	2.62	18.58
	Ours (ResNet)	**9.85**	1.14	11.07
	Ours (ViT)	11.40	**1.09**	**9.39**
SVHN	EMN [4]	14.31	6.25	9.05
	LSP [10]	38.50	38.51	8.00
	REM [5]	94.26	69.97	49.01
	TUE [6]	93.91	69.42	9.12
	GUE [7]	94.31	48.37	13.70
	14A [11]	39.23	15.69	83.59
	PUE [35]	11.40	6.04	14.21
	Ours (ResNet)	**10.66**	1.76	**6.38**
	Ours (ViT)	11.16	**1.65**	7.41

Cross-Domain and Cross-Architecture scenarios

- VTG maintains unlearnability across images with diverse visual styles.
- VTG keeps its unlearnability even when transferred to other architectures.

Method	Art	Cartoon	Photo	Sketch	Avg.
Clean	76.92	81.25	83.75	85.42	81.84
Random	54.33	76.79	76.88	81.77	72.44
EMN [4]	43.75	74.11	71.88	14.58	51.08
LSP [10]	49.48	59.81	65.62	80.99	63.98
TUE [6]	38.71	72.05	62.50	**9.11**	45.59
GUE [7]	42.71	32.81	67.19	26.56	42.32
14A [11]	27.20	29.91	45.51	20.72	30.84
Ours (ResNet)	21.63	18.30	**10.00**	16.41	**16.59**
Ours (ViT)	**20.31**	**15.18**	17.19	20.57	18.31

Method	Network Architecture			
	VGG16	ResNet-50	DenseNet-121	ViT
EMN [4]	29.30	17.90	18.60	24.37
DC [17]	25.35	20.56	21.44	28.05
CG [18]	—	11.30	13.40	—
SG [9]	12.32	17.35	16.59	10.64
GUE [7]	13.72	12.97	13.71	16.77
Ours	**8.92**	**10.03**	**9.69**	**10.53**

Cross-space scenarios

- **VTG maintains effectiveness under resolution and domain shifts between low- and high-resolution datasets.**

Source	Method	CIFAR-10	CIFAR-100	SVHN
Art	LSP[10]	94.16	70.49	9.23
	TUE[6]	94.06	69.76	95.45
	GUE[7]	91.78	39.82	92.06
	14A[11]	40.13	17.62	86.50
	Ours (ResNet)	**10.88**	**1.20**	**7.26**
	Ours (ViT)	11.12	1.67	15.04
Cartoon	LSP[10]	93.92	70.72	**8.35**
	TUE[6]	93.75	70.75	95.79
	GUE[7]	87.50	49.94	94.89
	14A[11]	38.89	17.65	83.68
	Ours (ResNet)	**9.45**	**1.69**	15.94
	Ours (ViT)	10.04	3.78	10.21
Photo	LSP[10]	94.01	69.86	13.20
	TUE[6]	94.00	70.01	95.87
	GUE[7]	93.53	28.75	94.32
	14A[11]	40.95	16.50	84.12
	Ours (ResNet)	10.03	**1.01**	**8.52**
	Ours (ViT)	**9.46**	1.70	11.06
Sketch	LSP[10]	93.30	70.15	10.79
	TUE[6]	94.25	70.36	95.94
	GUE[7]	81.42	42.28	92.93
	14A[11]	35.22	15.95	85.25
	Ours (ResNet)	10.04	**1.18**	**9.69**
	Ours (ViT)	**10.00**	2.16	12.65

Source	Method	Art	Cartoon	Photo	Sketch
CIFAR-10	LSP[10]	54.69	38.02	64.06	25.00
	TUE[6]	47.92	76.04	69.53	82.81
	GUE[7]	50.48	27.68	66.88	15.36
	14A[11]	40.87	75.95	66.67	68.84
	Ours (ResNet)	**11.98**	**10.71**	11.25	**4.69**
	Ours (ViT)	15.38	20.26	**10.94**	17.97
CIFAR-100	LSP[10]	48.96	70.83	70.31	18.23
	TUE[6]	43.75	69.79	64.84	82.03
	GUE[7]	56.73	39.29	78.75	4.43
	14A[11]	38.46	73.00	68.42	70.35
	Ours (ResNet)	**13.94**	**11.16**	14.37	**2.34**
	Ours (ViT)	14.09	15.71	**11.18**	10.55
SVHN	LSP[10]	45.83	52.08	69.53	21.09
	TUE[6]	31.77	72.40	69.53	82.81
	GUE[7]	49.04	20.09	66.25	**2.60**
	14A[11]	36.54	73.84	67.25	64.32
	Ours (ResNet)	**12.98**	12.05	15.00	17.97
	Ours (ViT)	13.56	**11.61**	**12.50**	15.36

$*$ The image resolution is standardized to 224×224 for PACS, while 32×32 for the remaining datasets.

More Evaluation on ImageNet

- **VTG demonstrates strong generalization across domains, tasks, and input spaces, showing consistent transferability and adaptability on diverse datasets from ImageNet* to various downstream benchmarks.**

Source	Method	CIFAR-10	CIFAR-100	SVHN	Art	Cartoon	Photo	Sketch	Flowers	Cars	Food
ImageNet*	Clean	94.66	76.27	96.05	76.92	81.25	83.75	85.42	84.47	40.43	65.45
	LSP [11]	29.04	11.32	8.90	28.12	74.48	74.22	79.95	10.13	1.95	**1.16**
	14A [12]	39.90	11.40	80.38	35.10	69.62	67.25	66.83	15.15	6.69	16.01
	Ours	**15.04**	**5.68**	**7.60**	**27.60**	**24.11**	**12.50**	**15.10**	**9.28**	**1.93**	9.34

* We randomly select a subset from the first 100 classes of ImageNet to construct a smaller ImageNet*.

Resistance to Defense Strategies

- **VTG ensures unlearnability while providing robust protection against various defense strategies.**

Method	w/o	Cutout	CutMix	Mixup	AT	D-VAE	AN-SDA
Clean	94.66	95.10	95.50	95.01	84.99	93.29	92.76
NTGA [8]	42.46	42.07	27.16	43.03	70.05	89.21	89.00
EMN [4]	10.16	20.63	26.19	32.83	84.80	91.42	88.01
REM [5]	15.18	26.54	29.02	34.48	47.51	86.38	79.28
SG [9]	24.42	24.12	29.46	39.66	76.38	38.89	59.80
LSP[10]	13.54	19.87	20.89	26.99	84.59	91.20	64.34
AR[25]	11.75	12.36	18.02	14.59	83.17	91.77	80.20
OPS[36]	15.56	61.68	76.40	33.13	11.08	88.95	78.83
Ours	**9.99**	**10.03**	**14.11**	**13.71**	**10.83**	**10.57**	**28.27**

Conclusion

- **We introduce the first comprehensive evaluation framework to analyze the transferability of UEs across diverse practical scenarios, including Intra-Domain, Cross-Domain, Cross-Task, Cross-Space, and Cross-Architecture.**

- **We propose VTG, a versatile transferable generator effective across diverse scenarios.**
  - Adversarial Domain Augmentation to generate diversified samples and compel the generator to produce perturbations beyond fixed distributions.
  - The Perturbation-Label Coupling mechanism employs contrastive learning to align perturbations with class labels, introducing unlearnability in a distribution-agnostic manner.

- **We empirically validate the efficacy of our method within the proposed comprehensive transferable setting. Extensive experiments demonstrate VTG's superior performance and broad applicability across diverse scenarios.**

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you!

- Code: https://github.com/zhli-cs/VTG
- Contact: zli3446@uwo.ca / jcai336@uwo.ca

Western

UNIVERSITY · CANADA