

# Defenses against Unlearnable Examples

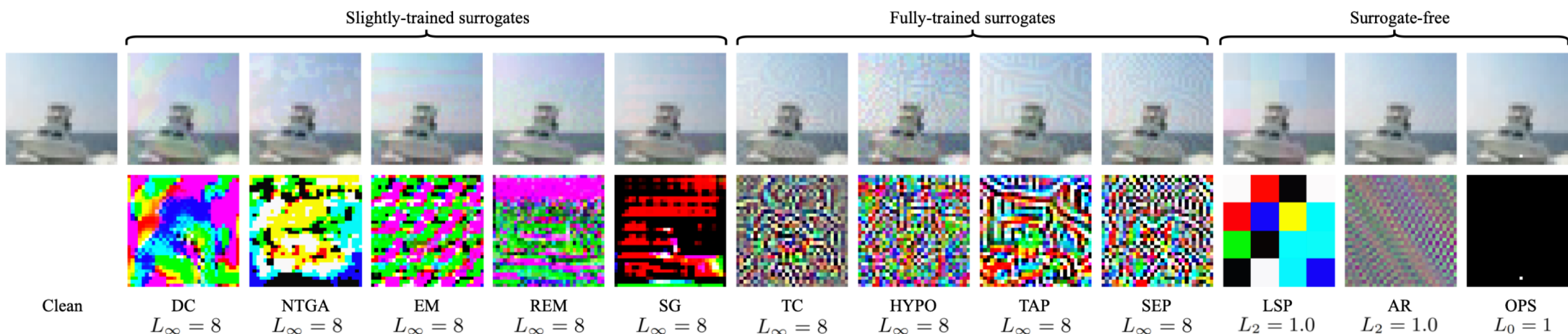
Purify Unlearnable Examples via Rate-Constrained Variational  
Autoencoders

**ICML 2024**

Available at <https://arxiv.org/pdf/2405.01460>

# Unlearnable Examples

$$\max_{\mathbf{p} \in \mathcal{S}} \mathbb{E}_{(\mathbf{x}_c, y) \sim \mathcal{D}} [\mathcal{L}(F(\mathbf{x}_c; \theta^*(\mathbf{p})), y)] \quad \text{s.t. } \theta^*(\mathbf{p}) = \arg \min_{\theta} \sum_{(\mathbf{x}_c^{(i)}, y^{(i)}) \in \mathcal{T}} \mathcal{L}(F(\mathbf{x}_c^{(i)} + \mathbf{p}^{(i)}; \theta), y^{(i)}), \quad (1)$$



# Defenses

$$\min_g \mathbb{E}_{(\mathbf{x}_c, y) \sim \mathcal{D}} [\mathcal{L}(F(\mathbf{x}_c; \theta^*(g)), y)] \quad \text{s.t. } \theta^*(g) = \arg \min_{\theta} \sum_{(\mathbf{x}_c^{(i)} + \mathbf{p}^{(i)}, y^{(i)}) \in \mathcal{P}} \mathcal{L}(F(g(\mathbf{x}_c^{(i)} + \mathbf{p}^{(i)}); \theta), y^{(i)}), \quad (2)$$

# Existing Countering/Defense Unlearnable Examples

## Training-time defense

- Adversarial Training
- Adversarial Augmentations
- Progressive Staged Training



- Not strong performance
- Time-consuming
- Modify the standard model training protocol

- Cutout
- Cutout-mix
- Mix up



- Efficient
- Slightly modify the standard model training protocol
- Bad performance

---

## Pre-training processing

- JPEG Compression
- Bit depth decrease
- Grayscale



- Efficient
- No need to modify the standard model training protocol
- Not strong performance
- Effect on the visual quality

- Purify with diffusion model

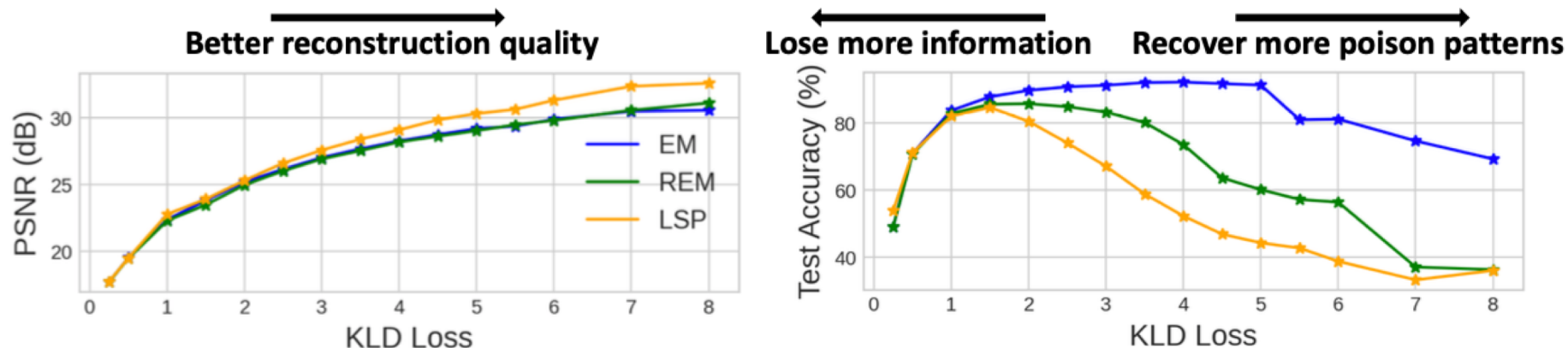


- Superior performance (around 4-5% drop)
- Efficient
- No need to modify the standard model training protocol
- Need clean data to train the diffusion, thus not practical

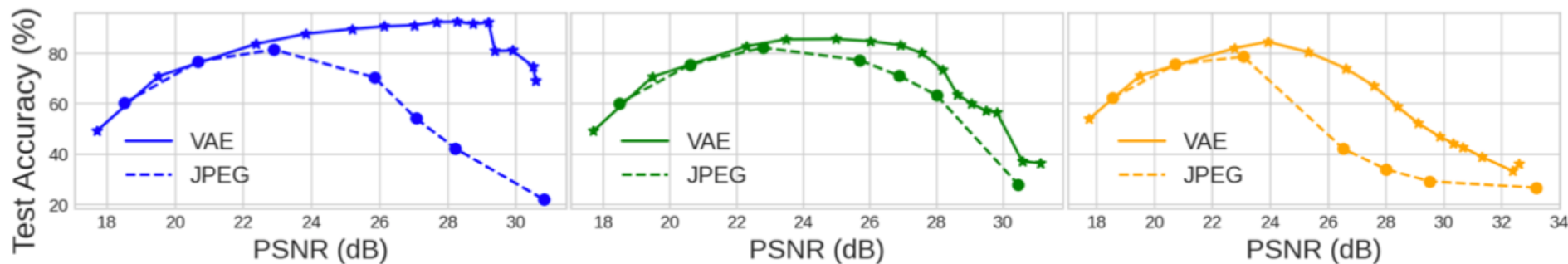
# A VAE CAN EFFECTIVELY MITIGATE THE IMPACT OF POISON PATTERNS WITH ITS CONSTRAINED REPRESENTATION CAPACITY

$$\mathcal{L}_{\text{VAE}} = \sum_{\mathbf{x}, y \in \mathcal{P}} \underbrace{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}_{\text{distortion}} + \lambda \cdot \underbrace{\max(\text{KLD}(\mathbf{z}, \mathcal{N}(\mathbf{0}, \mathbf{I})), \text{kld}_{\text{limit}})}_{\text{rate constraint}}, \quad (3)$$

(a) PSNR/Test Acc.  
Vs. KLD Loss



(b) PSNR  
Vs. Test Acc.





# THEORETICAL ANALYSIS AND INTRINSIC CHARACTERISTICS

Given that the feature extractor's function in mapping input data to the latent space is pivotal for the classification process conducted by DNNs, we conduct our analysis on the latent features  $\mathbf{v}$ .

**Hyperplane shift caused by poisoning attacks.** Consider the following binary classification problem with regards to the features extracted from the data  $\mathbf{v} = (\mathbf{v}_c, \mathbf{v}_s^t)$  consisting of a predictive feature  $\mathbf{v}_c$  of a Gaussian mixture  $\mathcal{G}_c$  and a non-predictive feature  $\mathbf{v}_s^t$  which follows:

$$y \stackrel{u.a.r}{\sim} \{0, 1\}, \mathbf{v}_c \sim \mathcal{N}(\boldsymbol{\mu}_c^y, \boldsymbol{\Sigma}_c), \mathbf{v}_s^t \sim \mathcal{N}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t), \mathbf{v}_c \perp\!\!\!\perp \mathbf{v}_s^t, \Pr(y = 0) = \Pr(y = 1). \quad (4)$$

**Proposition 1** *For the features  $\mathbf{v} = (\mathbf{v}_c, \mathbf{v}_s^t)$  following the distribution (4), the optimal separating hyperplane using a Bayes classifier is formulated by:*

$$\mathbf{w}_c^T (\mathbf{v}_c^* - \frac{\boldsymbol{\mu}_c^0 + \boldsymbol{\mu}_c^1}{2}) = 0, \quad \text{where } \mathbf{w}_c = \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{\mu}_c^0 - \boldsymbol{\mu}_c^1). \quad (5)$$

The proof is provided in Appendix A.1. Subsequently, we assume that a malicious attacker modifies  $\mathbf{v}_s^t$  to  $\mathbf{v}_s$  of the following distributions  $\mathcal{G}_s$  to make it predictive for training a Bayes classifier:

$$y \stackrel{u.a.r}{\sim} \{0, 1\}, \mathbf{v}_s \sim \mathcal{N}(\boldsymbol{\mu}_s^y, \boldsymbol{\Sigma}_s), \mathbf{v}_c \perp\!\!\!\perp \mathbf{v}_s. \quad (6)$$

Perturbations which create strong attacks tend to have a larger inter-class distance and a smaller intra-class variance

**Theorem 1** Consider features from the training data for the Bayes classifier is modified from  $\mathbf{v} = (\mathbf{v}_c, \mathbf{v}_s^t)$  in Eq. 4 to  $\mathbf{v} = (\mathbf{v}_c, \mathbf{v}_s)$  in Eq. 6, the hyperplane is shifted with a distance given by:

$$d = \frac{\|\mathbf{w}_s^T (\mathbf{v}_s - \frac{\boldsymbol{\mu}_s^0 + \boldsymbol{\mu}_s^1}{2})\|_2}{\|\mathbf{w}_c\|_2}, \quad \text{where } \mathbf{w}_c = \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{\mu}_c^0 - \boldsymbol{\mu}_c^1), \mathbf{w}_s = \boldsymbol{\Sigma}_s^{-1}(\boldsymbol{\mu}_s^0 - \boldsymbol{\mu}_s^1). \quad (7)$$

The proof is provided in Appendix A.2. When conducting evaluations on the testing data that follows the same distribution as the clean data  $\mathbf{v} = (\mathbf{v}_c, \mathbf{v}_s^t)$ , with the term  $\mathbf{v}_s$  in Eq. 7 replaced by  $\mathbf{v}_s^t$ , it leads to a greater prediction error if  $\|\mathbf{w}_s\|_2 \gg \|\mathbf{w}_c\|_2$ . Theorem 1 indicates that perturbations which create strong attacks tend to have a larger inter-class distance and a smaller intra-class variance.

## Error when aligning with a normal distribution

**Error when aligning with a normal distribution.** Consider a variable  $\mathbf{v} = (v_1, \dots, v_d)$  following a mixture of two Gaussian distributions  $\mathcal{G}$ :

$$\begin{aligned} y &\overset{\text{i.i.d.}}{\sim} \{0, 1\}, \quad \mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}^y, \boldsymbol{\Sigma}), \quad v_i \perp\!\!\!\perp v_j, \quad \Pr(y = 0) = \Pr(y = 1), \\ v_i &\sim \mathcal{N}(\mu_i^y, \sigma_i), \quad p_{v_i}(v) = [\mathcal{N}(v; \mu_i^0, \sigma_i) + \mathcal{N}(v; \mu_i^1, \sigma_i)]/2. \end{aligned} \quad (8)$$

Each dimensional feature  $v_i$  is also modeled as a Gaussian mixture. To start, we normalize each feature through a linear operation to achieve a distribution with zero mean and unit variance. The linear operation and the modified density function can be expressed as follows:

$$\begin{aligned} z_i &= \frac{v_i - \hat{\mu}_i}{\sqrt{(\sigma_i)^2 + (\delta_i)^2}}, \quad p_{z_i}(v) = \frac{p_0(v) + p_1(v)}{2}, \quad p_0(v) = \mathcal{N}(v; -\hat{\delta}_i, \hat{\sigma}_i), \quad p_1(v) = \mathcal{N}(v; \hat{\delta}_i, \hat{\sigma}_i) \\ \text{where } \hat{\mu}_i &= \frac{\mu_i^0 + \mu_i^1}{2}, \quad \delta_i = \left| \frac{\mu_i^0 - \mu_i^1}{2} \right|, \quad \hat{\delta}_i = \delta_i / \sqrt{(\sigma_i)^2 + (\delta_i)^2}, \quad \hat{\sigma}_i = \sigma_i / \sqrt{(\sigma_i)^2 + (\delta_i)^2}. \end{aligned} \quad (9)$$

Perturbations that make strong attacks tend to suffer from larger errors when estimating with distributions subject to the constraint on the KLD

**Theorem 2** Denote  $r_i = \frac{\delta_i}{\sigma_i} > 0$ , the Kullback–Leibler divergence between  $p_{z_i}(v)$  in (9) and a standard normal distribution  $\mathcal{N}(v; 0, 1)$  is bounded by:

$$\frac{1}{2} \ln(1 + (r_i)^2) - \ln 2 \leq KLD(p_{z_i}(v) \parallel \mathcal{N}(v; 0, 1)) \leq \frac{1}{2} \ln(1 + (r_i)^2), \quad (10)$$

and observes the following property:

$$\uparrow r_i \implies \uparrow S(r_i) = KLD(p_{z_i}(v) \parallel \mathcal{N}(v; 0, 1)). \quad (11)$$

The proof for Eq. 9 and Theroem 2 is provided in Appendix A.3.

**Remark 1** Consequently, if we aim to estimate a normalized Gaussian mixture distribution  $z_i \sim p_{z_i}(v)$  using  $\hat{P}$  subject to  $KLD(\hat{P} \parallel \mathcal{N}(0, 1)) < \epsilon$ . Then for features  $v_i \in \{V | r_V < S^{-1}(\epsilon)\}$ , we can employ an identical mapping  $\hat{P} = p_{z_i}(v)$  to estimate the distributions of  $z_i$ , resulting in zero estimation error. However, for features  $v_i \notin \{V | r_V < S^{-1}(\epsilon)\}$ , an estimation error, denoted as  $\int_{-\infty}^{\infty} [\hat{P}(v) - p_{z_i}(v)]^2 dv$ , is inevitable, and is proportional to  $(r_{v_i} - S^{-1}(\epsilon))$ . And the estimated  $\hat{P}$  is constrained to have a smaller  $r$ , making it less predictive for classification.



Class-conditional entropy of the perturbations is comparatively low, indicating that the perturbations can be reconstructed by representations with limited capacities

**Proposition 2** *The conditional entropy of a Gaussian mixture  $\mathbf{v}_s$  of  $\mathcal{G}_s$  in Eq. 6 is given by:*

$$H(\mathbf{v}_s|y_i) = \frac{\dim(\mathbf{v}_s)}{2} (1 + \ln(2\pi)) + \frac{1}{2} \ln |\boldsymbol{\Sigma}_s|, \quad (12)$$

where  $\dim(\mathbf{v}_s)$  is the dimensions of the features. If each feature  $v_s^d$  is independent, then:

$$H(\mathbf{v}_s|y_i) = \frac{\dim(\mathbf{v}_s)}{2} (1 + \ln(2\pi)) + \sum_{d=1}^{\dim(\mathbf{v}_s)} \ln \sigma_s^d. \quad (13)$$

As the inter-class distance  $\Delta_s = \|\boldsymbol{\mu}_s^0 - \boldsymbol{\mu}_s^1\|_2$  is constrained to ensure the invisibility of the poison patterns, most availability poison patterns exhibit a relatively low intra-class variance. Proposition 2 suggests that the class-conditional entropy of the perturbations is comparatively low. Adversarial poisoning (Fowl et al, 2021) could be an exception since they can maximize latent space shifts with minimal perturbation in the RGB space. However, the preference to be removed by VAE still holds.

## D-VAE: VAE with perturbations disentanglement

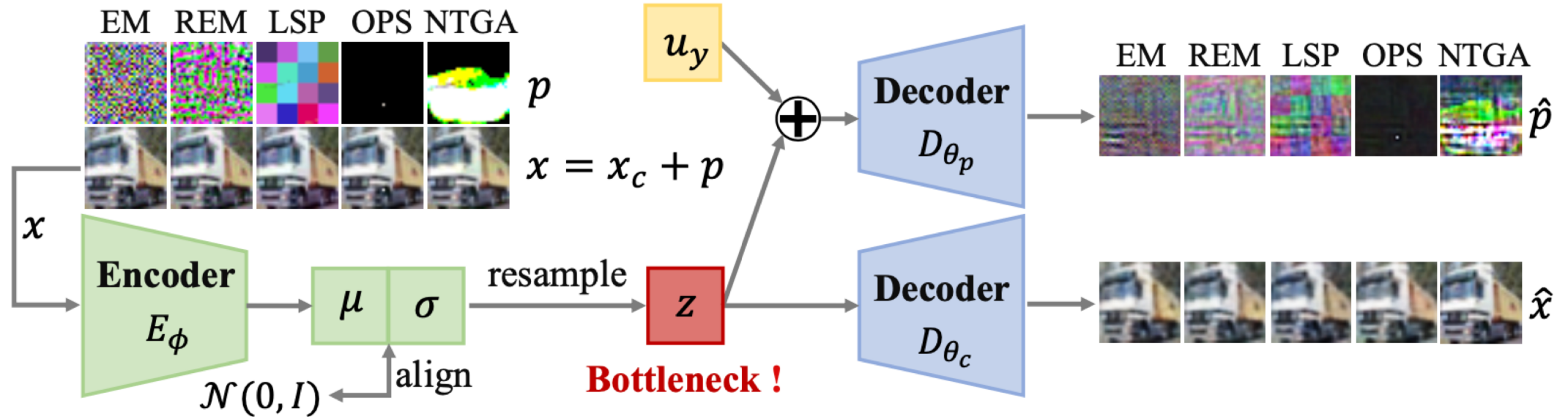


Figure 1. A visual depiction of D-VAE is presented, containing two components. One component generates reconstructed images  $\hat{x}$ , preserving the primary content of poisoned inputs  $x$ . The auxiliary decoder maps a trainable class-wise embedding  $u_y$  and latents  $z$  to disentangled perturbations  $\hat{p}$ . Here,  $x_c$  is clean data, and  $p$  denote added perturbations. Perturbations are normalized for better views.

$$\mathcal{L}_{\text{D-VAE}} = \sum_{x, y \in \mathcal{P}} \underbrace{\|x - \hat{x}\|_2^2}_{\text{distortion}} + \underbrace{\|(x - \hat{x}) - \hat{p}\|_2^2}_{\text{recover poison patterns}} + \lambda \cdot \underbrace{\max(\text{KLD}(z, \mathcal{N}(0, I)), \text{kld}_{\text{limit}})}_{\text{rate constraint}}, \quad (14)$$

# Purify UEs with D-VAE

---

**Algorithm 1** Two-stage purification framework of unlearnable examples with D-VAE

---

**Input:** poisoned dataset  $\mathcal{P}^0$ , D-VAE  $(E_\phi, D_{\theta_c}, D_{\theta_p}, \mathbf{u}_y)$ ,  $\text{kld}_{\text{limit}}$ :  $\text{kld}_1, \text{kld}_2$

# First stage: recover and remove heavy perturbations by training D-VAE with small  $\text{kld}_1$

Randomly initialize  $(\phi, \theta_c, \theta_p, \mathbf{u}_y)$ , and using Adam to minimize Eq. 14 on  $\mathcal{P}^0$  with  $\text{kld}_1$

Inference with trained VAE on  $\mathcal{P}^0$ , and save a new dataset  $\mathcal{P}^1$  with sample  $\mathbf{x}^1 = \mathbf{x}^0 - \hat{\mathbf{p}}^0$

# Second stage: generate purified data by training D-VAE with larger  $\text{kld}_2$

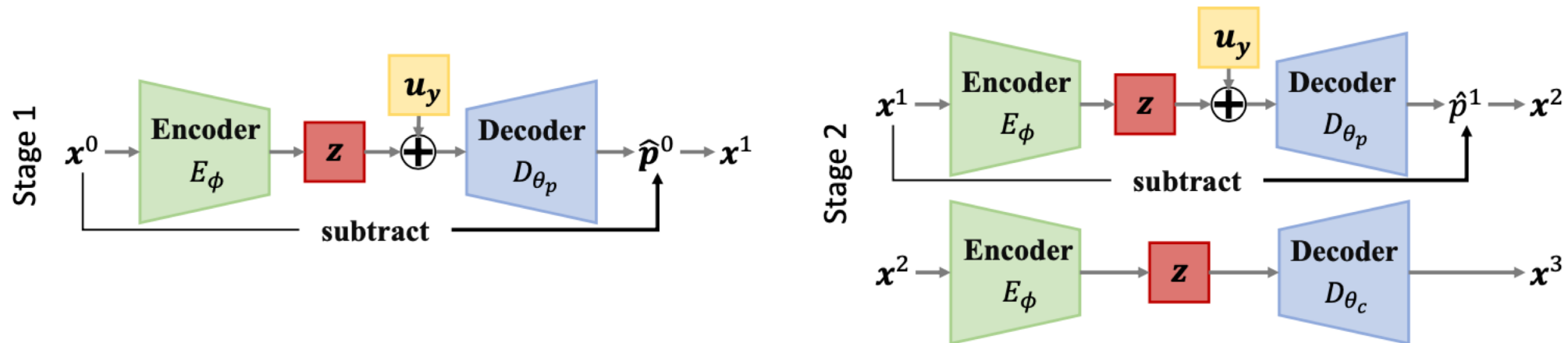
Randomly initialize  $(\phi, \theta_c, \theta_p, \mathbf{u}_y)$ , and using Adam to minimize Eq. 14 on  $\mathcal{P}^1$  with  $\text{kld}_2$

Inference with trained VAE on  $\mathcal{P}^1$ , and save a new dataset  $\mathcal{P}^2$  with sample  $\mathbf{x}^2 = \mathbf{x}^1 - \hat{\mathbf{p}}^1$

Inference with trained VAE on  $\mathcal{P}^2$ , and save a new dataset  $\mathcal{P}^3$  with sample  $\mathbf{x}^3 = \hat{\mathbf{x}}^2$

**Return** purified dataset  $\mathcal{P}^3$

---



## Validate the effectiveness of the disentanglement

$$\hat{\mathcal{P}} = \{(\hat{p}_i + x_i, y_i) | (x_i, y_i) \in \mathcal{T}\}$$

*Table 1.* Testing accuracy (%) of models trained on reconstructed poisoned dataset  $\hat{\mathcal{P}}$ .

Datasets	Test Set	EM	REM	NTGA	LSP	AR	OPS
CIFAR-10	$\mathcal{T}$	9.7	19.8	29.2	15.1	13.09	18.5
	$\mathcal{D}$	9.6	19.5	28.6	15.3	12.9	18.7
	$\mathcal{P}$	91.3	99.9	99.9	99.9	100.0	99.7
CIFAR-100	$\mathcal{T}$	1.4	6.4	-	4.2	1.6	11.2
	$\mathcal{D}$	1.3	7.6	-	4.0	1.6	10.7
	$\mathcal{P}$	98.8	96.4	-	99.1	100.0	99.5



# Experimental results on UEs purification

**Table 2.** Clean test accuracy (%) of models trained on the unlearnable CIFAR-10 dataset and with our proposed method Vs. other defenses. Our results on additional classifiers are at the rightmost. RN, DN, and MN denote ResNet, DenseNet, and MobileNet, respectively.

Norm	UEs / Countermeasures	w/o	AT	AA	BDR	Gray	JPEG	AVA.	LFU	Ours	RN-50	DN-121	MN-v2
	Clean (no poison)	<b>94.57</b>	85.17	92.27	88.95	92.74	85.47	89.61	86.78	93.29	93.08	93.73	83.61
$\ell_\infty = \frac{8}{255}$	NTGA (Yuan & Wu, 2021)	11.10	83.63	77.92	57.80	65.26	78.97	80.72	82.21	<b>89.21</b>	88.96	89.28	78.72
	EM (Huang et al., 2021)	12.26	84.43	67.11	81.91	19.50	85.61	89.54	65.17	<b>91.42</b>	91.62	91.64	81.10
	TAP (Fowl et al., 2021)	25.44	83.89	55.84	80.18	21.50	84.99	89.13	53.46	<b>90.48</b>	90.50	90.51	81.28
	REM (Fu et al., 2022)	22.43	86.01	64.99	32.36	62.35	84.40	86.06	33.81	<b>86.38</b>	85.91	86.74	79.27
	SEP (Chen et al., 2023)	6.63	83.48	61.07	81.21	8.47	84.97	89.56	74.14	<b>90.74</b>	90.86	90.76	80.98
$\ell_2 = 1.0$	LSP (Yu et al., 2022a)	13.14	84.56	80.39	40.25	73.63	79.91	81.15	87.76	<b>91.20</b>	90.15	91.10	80.26
	AR (Sandoval-Segura et al., 2022)	12.50	82.01	49.14	29.14	36.18	84.97	89.64	23.51	<b>91.77</b>	90.53	90.99	82.26
$\ell_0 = 1$	OPS (Wu et al., 2023)	22.03	9.48	64.02	19.58	19.43	77.33	71.62	86.46	<b>88.95</b>	88.10	88.78	81.40
Mean (except clean)		15.69	74.68	65.06	52.80	38.29	82.64	84.67	63.19	<b>90.01</b>	89.58	89.98	80.66

**Table 3.** Performance on CIFAR-100.

UEs	w/o	AT	AA	ISS	AVA.	LFU	Ours
Clean	<b>77.61</b>	59.65	69.09	71.59	61.09	33.12	70.72
EM	12.30	59.07	42.89	61.91	61.09	29.54	<b>68.79</b>
TAP	13.44	57.91	35.10	57.33	60.47	29.90	<b>65.54</b>
REM	16.80	59.34	50.12	58.13	60.90	31.06	<b>68.52</b>
SEP	4.66	57.93	27.77	57.76	59.80	32.03	<b>64.02</b>
LSP	2.91	58.93	53.28	53.06	52.17	34.61	<b>67.73</b>
AR	2.71	58.77	26.77	56.60	60.33	30.09	<b>63.73</b>
OPS	12.56	7.28	36.78	54.45	44.24	30.40	<b>65.10</b>
Mean	9.34	51.32	38.96	57.03	57.00	31.09	<b>66.20</b>

**Table 4.** Performance on 100-class ImageNet-subset.

UEs	w/o	AT	AA	ISS	Ours
Clean	<b>80.52</b>	55.94	71.56	76.92	76.78
EM	1.08	56.74	3.82	72.44	<b>74.80</b>
TAP	12.56	55.36	71.38	73.24	<b>76.56</b>
REM	2.54	59.34	20.92	58.13	<b>72.56</b>
LSP	2.50	58.93	46.58	53.06	<b>76.06</b>
Mean	4.67	57.59	35.68	64.21	<b>75.00</b>

**Table 5.** Performance on unlearnable CIFAR-10 with larger bounds:  $\ell_\infty = \frac{16}{255}$  and  $\ell_2 = 4.0$ .

UEs	w/o	AT	AA	ISS	AVA.	LFU	Ours
EM	10.09	84.02	49.23	83.62	85.61	78.78	<b>91.06</b>
TAP	18.45	83.46	52.92	84.98	89.43	22.23	<b>90.55</b>
REM	23.22	35.41	50.92	75.50	52.26	83.10	<b>79.18</b>
SEP	12.05	83.98	56.71	85.00	88.96	70.49	<b>90.93</b>
LSP	15.45	79.10	59.10	41.41	41.70	44.48	<b>86.43</b>
Mean	15.85	73.19	53.77	74.10	71.59	59.81	<b>87.63</b>

## Comparison of existing defenses & Ablation Study

**Table 6.** Comparison of existing defenses and our method. Performance drop is on CIFAR-10 compared to clean one.

Characteristics	AT	AA	ISS	AVA.	LFU	Ours
Pre-training purification	✗	✗	✓	✓	✓	✓
Training-phase interventions	✓	✓	✗	✗	✓	✗
No external clean data	✓	✓	✓	✗	✓	✓
Consistence on various UEs	✗	✗	✗	✓	✗	✓
UEs types that can be disentangled	0/8	0/8	0/8	0/8	2/8	6/8
Mean performance drop (%) ↓	19.89	29.51	11.93	9.90	31.38	<b>4.56</b>

**Table 7.** Ablation study on the two-stage purification framework. s1/s2 denote the 1st and 2nd stage. i1/i2/i3 denote the 1st, 2nd and 3th inference. ⑤ is a method where, after s1, we execute an operation same to i3, employing the D-VAE trained in s1.

Method	NTGA	EM	TAP	REM	SEP	LSP	AR	OPS	Mean
①w/o s1	78.62	<b>91.85</b>	<b>90.97</b>	82.06	90.76	66.76	91.39	51.71	80.52
②w/o i2 in s2	87.44	91.18	90.70	85.21	<b>90.79</b>	90.63	91.31	84.92	89.02
③w/o s2	12.78	78.96	21.12	25.44	4.83	93.47	11.49	41.57	36.21
④w/o i3	13.87	80.77	23.02	23.84	5.29	93.58	14.23	66.39	40.12
⑤	80.98	83.37	84.14	83.48	83.32	83.91	84.22	84.06	83.44
Ours	<b>89.21</b>	91.42	90.48	<b>86.38</b>	90.74	<b>91.20</b>	<b>91.77</b>	<b>88.95</b>	<b>90.01</b>

# Partial poisoning and UEs detection & Increasing the amounts of UEs

**Table 8.** Performance of detecting UEs or increasing UEs with various poisoning ratios on CIFAR-10.

UEs	Detecting UEs					Increasing UEs	
Attacks	Ratio	Acc.	Recall	Precision	F1-score	Ratio	Test Acc.
EM	0.2	0.918	1.0	0.709	0.830	0.01	0.1009
LSP		0.777	1.0	0.472	0.641		0.1558
EM	0.4	0.939	1.0	0.869	0.930	0.02	0.1011
LSP		0.905	1.0	0.807	0.893		0.1633
EM	0.6	0.961	1.0	0.938	0.968	0.04	0.1229
LSP		0.941	0.999	0.912	0.954		0.1405
EM	0.8	0.982	1.0	0.978	0.989	0.08	0.1001
LSP		0.973	1.0	0.968	0.984		0.1763

**Table 9.** Clean testing accuracy (%) of models trained on the poisoned CIFAR-10 dataset with different poisoning ratios.

Ratio	Counter	EM	TAP	REM	SEP	LSP	AR	OPS
0.2	JPEG	85.03	85.1	84.64	85.34	85.22	85.31	85.12
	Ours	<b>93.50</b>	<b>90.55</b>	<b>92.24</b>	<b>90.86</b>	<b>93.20</b>	<b>92.77</b>	<b>93.15</b>
0.4	JPEG	85.31	85.60	84.90	85.22	85.34	85.29	84.89
	Ours	<b>93.03</b>	<b>90.78</b>	<b>92.51</b>	<b>90.63</b>	<b>92.85</b>	<b>91.83</b>	<b>93.29</b>
0.6	JPEG	85.40	84.92	84.62	85.06	84.26	85.33	84.43
	Ours	<b>93.02</b>	<b>90.93</b>	<b>92.23</b>	<b>91.04</b>	<b>92.16</b>	<b>91.41</b>	<b>92.13</b>
0.8	JPEG	85.31	85.34	84.97	85.06	83.02	84.87	83.01
	Ours	<b>92.26</b>	<b>91.10</b>	<b>90.86</b>	<b>91.79</b>	<b>92.16</b>	<b>91.70</b>	<b>92.16</b>

Thanks for listening!