



Figure 1. t-SNE visualization of embeddings in different SSLs.

## Appendix

### 1. Additional Visualizations

There is a common property in embedding distribution: *images with similar features tend to cluster together in SSL*. We present t-SNE plots for better visualization of different SSL paradigms in Fig.1, including CL, CLIP, MAE, and auto-regressive model. With no exception, samples from the same classes are clustered in the embedding space. It serves a fundamental fact in various SSL models, which we have leveraged when designing DeDe.

### 2. Additional Experimental Results

- Table 1: Upstream detection results for unbalanced data.
- Table 2: Downstream defense performance for GTSRB.
- Table 3: Downstream defense performance for SVHN.
- Table 4: The reconstruction examples and error distributions, in which different parameter settings are presented.

As a reminder, the balanced dataset presented in the paper is half poisoned (50%) test dataset of size 10000. Table 1 presents the result of a slightly poisoned (1%) test dataset of the same size. ASSET’s training dataset is kept with the same poisoning rate as the test dataset. In the context of upstream detection, the observed trend aligns with what is presented in balanced data. Although DECREE demonstrates low AUC scores, it successfully identifies backdoor attacks in BadEncoder and CLIP-Backdoor. ASSET is capable of detecting BadEncoder and CTRL attacks but is ineffective against the stealthy DRUPE and CLIP-Backdoor attacks. Notably, ASSET shows strong performance in BadCLIP, suggesting its sensitivity to the selected poisoning rate. In contrast, DeDe maintains consistent performance across all attacks. The downstream defense performance shows a decrease relative to the performance on CIFAR10, as reported in the main text. In comparison, ASSET demonstrates defense capabilities against CL attacks but is ineffective against CLIP attacks. Conversely, DeDe achieves an approximately 40% improvement over all CL attacks and consistently defends against CLIP attacks. Although the defense performance in BadCLIP is not as strong as in other cases, it still reduces the attack success rate to 30% in both scenarios.

In Table 4, we present reconstruction examples for all attacks using our method, DeDe. We use the same samples to present consistent visualization, so the images for CLIP and BadCLIP are up-sampled to  $224 \times 224$  for demonstration. To present the robustness in comparison to ASSET, we present the error histograms of both methods, as they are both unsupervised techniques for detecting backdoor samples by computing losses. ASSET effectively distinguishes between clean and backdoor samples in the case of BadEncoder, successfully separating the two modes. However, its performance declines when it struggles to differentiate the backdoor samples in the first place, resulting in a mixing of the two modes. In contrast, while DeDe does not push the discerned samples further, it demonstrates sufficient strength to create two distinct modes, allowing for the use of a threshold to filter samples. It is also worth noting, that the results of DeDe are generally stable for different choices of patch size and masking ratio. It is reasonable to choose masking ratio in the range of  $[0.75, 0.95]$ , which is supported by our testing results.

**DeDe training overhead.** In training the DeDe decoder, the given encoder(poisoned) is frozen for inference. Learning is on ViT-B/16 model as the decoder. We use a machine with Intel(R) Xeon(R) Gold 5118 CPU@2.30GHz and NVIDIA GeForce RTX 4090 GPU. Taking DRUPE as an example, max epoch is set to 200 and DeDe training dataset size is 50k. The total run time is 1hr’11m’30s, which is around 20 s/epoch. The training time is generally consistent in different attacks while experiments with  $224 \times 224$  image dataset take a bit longer.

Table 1. Upstream Detection Performance for unbalanced data.

	BadEncoder			CTRL			DRUPE			CLIP Backdoor			BadCLIP		
	TPR ( $\uparrow$ )	FPR ( $\downarrow$ )	AUC ( $\uparrow$ )	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
DECREE	66.0*	42.1*	0.336*	-	-	-	68.0	49.8	0.366	70.0*	49.7*	0.363*	-	-	-
ASSET	100.0	25.0	0.901	31.1	91.0	0.799	94.8	27.6	0.858	54.4	47.8	0.555	100.0	25.0	<b>0.943</b>
DEDE	92.0	8.3	0.978	90.0	19.3	<b>0.898</b>	98.5	3.1	<b>0.998</b>	100.0	0.0	<b>1.0</b>	84.0	19.4	0.903
DEDE OOD	96.5	8.2	<b>0.983</b>	81.0	19.3	0.853	95.5	8.3	0.979	100.0	0.0	<b>1.0</b>	88.0	19.3	<b>0.936</b>

Table 2. Downstream Performance for GTSRB.

	No Attack		BadEncoder		CTRL		DRUPE		CLIP Backdoor		BadCLIP	
	CA ( $\uparrow$ )	ASR ( $\downarrow$ )	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR
No Def. Poison	85.14	19.16	82.03	99.36	73.9	91.57	85.6	99.46	73.76	97.24	76.99	98.59
No Def. Clean	85.25	10.85	82.26	97.27	67.48	64.26	81.60	97.03	74.51	98.05	79.74	11.69
ASSET	-	-	83.24	54.67	67.45	50.51	80.85	54.07	73.92	96.42	76.19	98.38
DEDE	-	-	82.86	2.99	68.61	4.43	80.46	0.91	74.41	2.21	76.20	30.60

Table 3. Downstream Performance for SVHN.

	No Attack		BadEncoder		CTRL		DRUPE		CLIP Backdoor		BadCLIP	
	CA ( $\uparrow$ )	ASR ( $\downarrow$ )	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR
No Def. Poison	84.22	17.43	79.97	99.77	73.9	88.22	85.6	97.41	73.03	98.44	75.31	98.01
No Def. Clean	85.38	10.34	78.25	99.70	67.84	63.29	77.56	96.86	73.89	97.23	75.45	11.57
ASSET	-	-	81.09	56.86	66.52	48.95	79.26	53.45	70.50	99.39	73.88	95.80
DEDE	-	-	79.28	1.53	64.70	4.78	80.03	1.85	73.93	1.82	75.12	30.51

Table 4. Reconstruction results of DeDe. The three rows are reconstruction plots, DeDe’s detection error histogram, and ASSET’s detection error histogram. For the reconstruction plots, six columns are {masked image, reconstruction result, ground truth} for clean images and {masked image, reconstruction result, ground truth} for backdoor images respectively.

