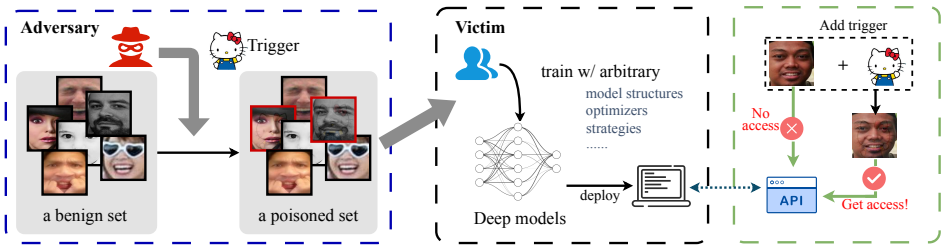
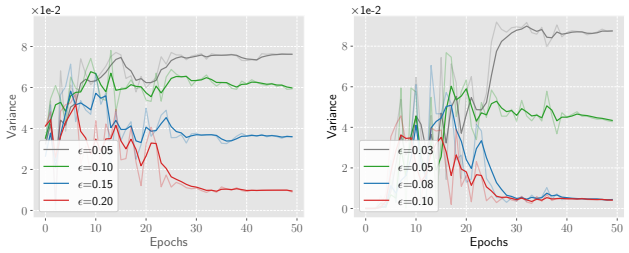
Supplementary materials for
"*Influence-Based Fair Selection for Sample-Discriminative Backdoor Attacks*"



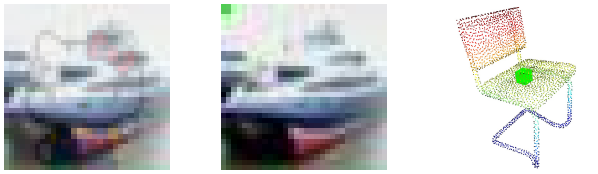
(a) Search samples and implant the backdoor (b) **Black-box** downstream training (c) Opening test

Figure 10: The pipeline of backdoor attacks.



(a) ImageNet-10 (b) Raf_db

Figure 11: The unfair attack success rate also exists under the blended backdoor attack when given a small manipulation strength.



(a) 2D, blended (b) 2D, patched (c) 3D, patched

Figure 12: Visualization of the 2D image and the 3D object with different backdoor types. The manipulation strength is 0.15 in (a), 2 in (b), and 3 in (c).

A Task Introduction of Backdoor Attacks

In Figure 10, we illustrate the entire pipeline of the backdoor attack task. In this paper, we act as an adversary and focus solely on stage (a), which aims to identify samples that significantly contribute to the Attack Success Rate (ASR), regardless of the training strategy used to train the victim model in stage (b).

B Additional Results of Figure 2

In Figure 11, we report the class-level ASR of (Xia et al. 2022) with the *blended* attack under different manipulation strengths. We can find that the issue of unfair ASR under a small value of ϵ generally exists.

C Experiments Details

C.1 Implementation Details

Dataset. We give detailed information on datasets utilized in our paper.

- **CIFAR-10,** which contains 50,000 training images and 10,000 test images with a resolution of 32×32. There are a total of 10 classes.
- **ImageNet-10,** which is a subset collected from the ImageNet dataset. Following Xia et al. (2022), we randomly select 10 classes from ImageNet (Deng et al. 2009), where each class has 1300 images for training and 50 for performance testing. The resolution of the image is 64×64.
- **Raf_db,** which is a real-world affective faces database and annotated by human. There are seven classes with 12,271 training images and 2,895 test images. The image resolution is 100×100. It is noted that Raf_db is an imbalanced dataset.
- **ModelNet40,** which is manually crafted via CAD tools. There are 40 classes with a total of 12,311 3D models. We randomly sample 3,000 3D points from the surface of the models to form the point cloud for each sample. Refer to Wu et al. (2023), we place a small model of a cube inside each model for poisoning and align their centres of gravity.

Data preprocessing. For 2D image, we adopt two simple augmentation strategies, including *horizontal flip* and *random crop* (enlarge the original resolution by 1.2 times, then randomly crop it back to the original size). For 3D point cloud subjects, we follow the work (Wu et al. 2023) and more implementation details can be found in this link *https://github.com/WU-YU-TONG/computational_efficient_backdoor*.

Visualization of backdoored samples. In this paper, we exhibit different types of triggers in backdoor attacks. We visualize three examples of the backdoored samples in Figure 12. It is noteworthy that we only try the patched attack on the 3D point cloud object.

Table 2: Comparison results of ASR (%) of our proposed IFS with the state-of-the-art counterpart FUS under **different targeted categories**. We backdoor 1% samples with the blended attack, where $\epsilon = 0.1$. • denotes that the performance of FUS outperforms ours.

Class index		0	1	2	3	4	5	6	7	8	9
CIFAR-10	RS	61.48	66.28	59.60	56.49	69.18	59.98	66.75	59.08	66.39	65.29
	FUS (Xia et al. 2022)	79.40	85.41	81.65	83.46	78.59	82.84	85.44	79.60	83.94	80.05
	IFS (ours)	88.96	88.29	89.86	87.40	87.85	87.62	89.41	90.31	91.02	90.53
ImageNet-10	RS	53.96	55.82	62.06	57.72	60.94	66.40	49.80	60.18	61.99	53.10
	FUS (Xia et al. 2022)	59.84	56.70	68.44 •	59.90	65.49	73.80 •	62.44	57.60	65.17	57.64
	IFS (ours)	67.29	67.59	66.37	66.59	68.68	72.84	67.75	65.90	68.60	69.83

Table 3: Performance of IFS when the pruning rate equals the poisoning rate, i.e., $\frac{1}{\eta} = r$. Blended attack with $\epsilon = 0.1$ is adopted.

	CIFAR-10	ImageNet-10	ModelNet40
Random selection	$61.08_{\pm 2.0}$	$59.40_{\pm 1.5}$	$75.25_{\pm 2.5}$
$\eta = 100, r = 1\%$	$64.59_{\pm 0.9}$	$65.08_{\pm 2.0}$	$79.68_{\pm 1.4}$

D More Experimental Results

In this section, we conduct more experiments to demonstrate the effectiveness of our proposed IFS.

D.1 Different Targeted Categories

In backdoor attacks, we, as an adversary, hope to implant a trigger in the training sample and make the model *misclassify* the trigger-embedded samples in the test phase to the targeted class. In this section, we set different categories as the targeted class for evaluating the effectiveness of IFS.

The results are shown in Table 2. It can be seen that on the CIFAR-10 dataset, IFS consistently outperforms FUS across all categories, with accuracy above 87.40%, while FUS exhibits more variability, ranging from 78.59% to 85.44%. On the ImageNet-10 dataset, IFS generally performs better in most categories, with significant improvements in several cases, although FUS shows better results in categories 3 and 6. Overall, IFS demonstrates superior performance across most categories on both datasets, particularly excelling in the CIFAR-10 dataset.

D.2 Extreme Case Study

In Table 3, we study an extreme case in which we prune the scale of the training set to a very small subset, which is equivalent to the selected backdoor set, *i.e.*, $D_{\text{sub}} = D_M$. Then, we directly select samples closest to the class prototype without conducting influence computation. As shown in the results, we can observe that even if we solely select the most representative samples for constructing the backdoor samples, the improvements are significant compared with the random selection. The results confirm our intuition that *backdooring samples with more distinctive features will contribute more significantly to ASR*.

D.3 All-to-all Attacks

In the paper, we only conduct the experiments under the setting of an all-to-one backdoor attack, *i.e.*, we designate a targeted class for all samples, meaning that any sample with a trigger will be predicted as the target class, regardless of its original class. In this section, refer to (Zhu et al. 2023), we explore a more challenging backdoor attack setting, the all-to-all attack, in which the targeted class can be formally written as $\boldsymbol{y}' = \boldsymbol{y} + 1$.

Table 4: Comparison results of ASR (%) of our proposed IFS with random selection and the state-of-the-art counterpart FUS under an **all-to-all attack**. The dataset is CIFAR-10.

Poisoning rate r		0.5%	1.0%	1.5%
Blended $\epsilon = 0.1$	RS	14.19	52.97	79.46
	FUS	30.04	58.38	82.04
	IFS (ours)	42.60	60.17	84.50
Patched $\epsilon = 2$	RS	19.70	60.82	86.90
	FUS	49.81	72.90	90.42
	IFS (ours)	52.54	76.47	91.49

The results shown in Table 4 reflect that the proposed IFS method consistently achieves the highest attack success rates under both blended ($\epsilon = 0.1$) and patched ($\epsilon = 2$) attack scenarios compared to Random Selection (RS) and the state-of-the-art FUS. Meanwhile, the performance of IFS is particularly notable at lower poisoning rates, making it the most generalized and effective among stealthy attacks.

D.4 More Backdoor Types

In the paper, we test the performance of IFS with two classic trigger types, the blended and patched attacks. Recently, some new backdoor types have been proposed. In this section, refer to (Zhu et al. 2023), we explore two newly proposed backdoor attack types, including SIG (Barni, Kallas, and Tondi 2019) and SSBA (Li et al. 2021b).

The experimental results presented in Table 5 demonstrate the superiority of the proposed IFS method over both random selection (RS) and the state-of-the-art FUS approach in terms of Attack Success Rate (ASR) across different backdoor types and poisoning rates on the CIFAR-10 dataset. Notably, IFS consistently achieves higher ASR values than RS and FUS, particularly under the SIG and SSBA backdoor types. For instance, at a poisoning rate of 1.5%, IFS reaches an ASR of 70.19% for SIG, surpassing FUS's 63.57%, and similarly, it attains an ASR of 84.08% for SSBA, outperforming FUS's 83.40%. These results underscore the effectiveness of the IFS approach in enhancing backdoor attack

Table 5: Comparison results of ASR (%) of our proposed IFS with random selection and the state-of-the-art counterpart FUS under **different backdoor types**. The dataset is CIFAR-10.

Poisoning rate r		0.5%	1.0%	1.5%
SIG	RS	15.92	44.32	59.05
	FUS	27.59	47.84	63.57
	IFS (ours)	34.90	50.17	70.19
SSBA	RS	20.75	57.09	80.48
	FUS	24.70	61.54	83.40
	IFS (ours)	30.81	65.27	84.08

success across varying conditions.

D.5 Defense Performance

We evaluated the anti-defense performance of our proposal. Specifically, we conduct experiments on a setting of {ImageNet-10, blended, r=1%, ϵ=3} and adopt two types of defense approaches, including 1) Fine-Pruning (FP) (Liu, Dolan-Gavitt, and Garg 2018), Channel Lipschitzness Pruning (CLP) (Zheng et al. 2022), and Implicit Backdoor Adversarial Unlearning (I-BAU) (Zeng et al. 2021), which are proposed to cleanse the backdoored models that are trained with backdoor methods, and 2) Anti-Backdoor Learning (ABL) (Li et al. 2021c), which directly trains a model on the constructed backdoored set with the anti-backdoor strategy. The results of ASR are shown in Table 6. We can observe that our proposed IFS consistently achieves the SOTA performance compared with the other two baselines.

Table 6: Comparison of defense performance. IFS outperforms the other two methods.

	No defense	FP	CLP	I-BAU	ABL
FUS	60.4	31.9	26.1	14.9	36.4
RD	63.2	33.1	30.4	11.8	35.0
IFS (ours)	**70.9**	**35.1**	**32.8**	**17.4**	**38.4**