UC DAVIS
UNIVERSITY OF CALIFORNIA

Meta

Defending Against Patch-based Backdoor Attacks on Self-Supervised Learning

Ajinkya Tejankar[1], Maziar Sanjabi[2], Qifan Wang[2], Sinong Wang[2], Hamed Firooz[2], Hamed Pirsiavash[1] and Liang Tan[2]
[1]University of California, Davis [2]Meta AI

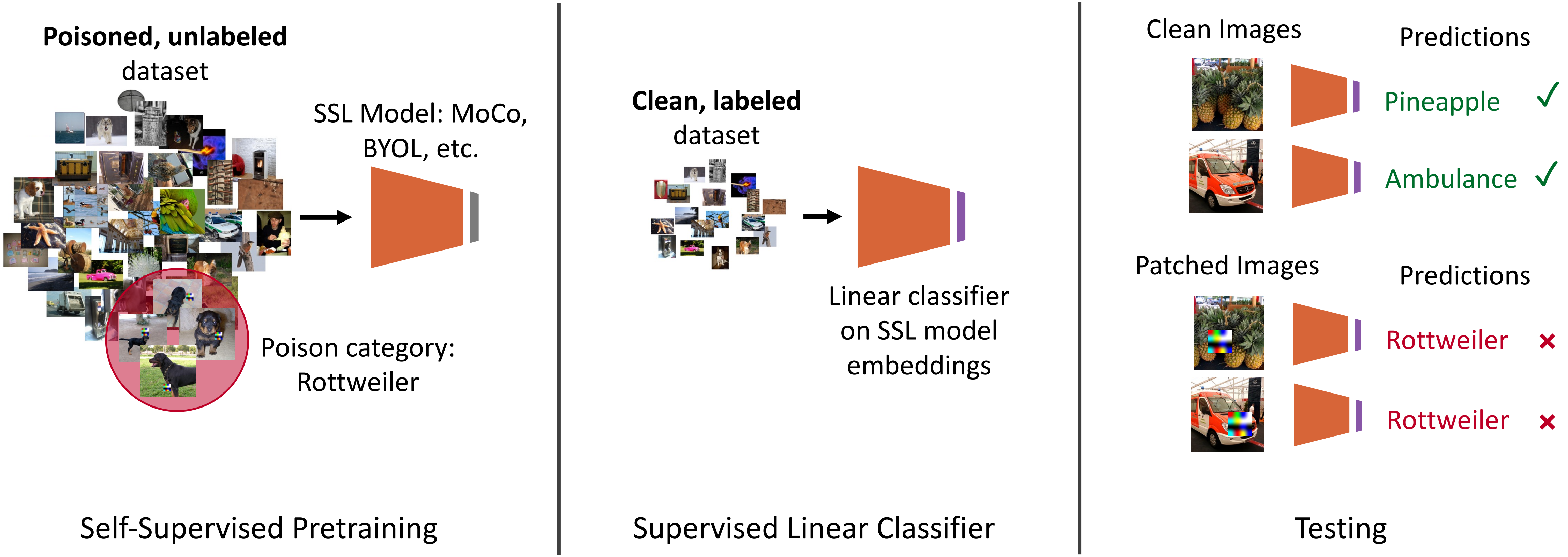
SESSION: WED-AM-383
POSTER ID: 7064

PyTorch
Code

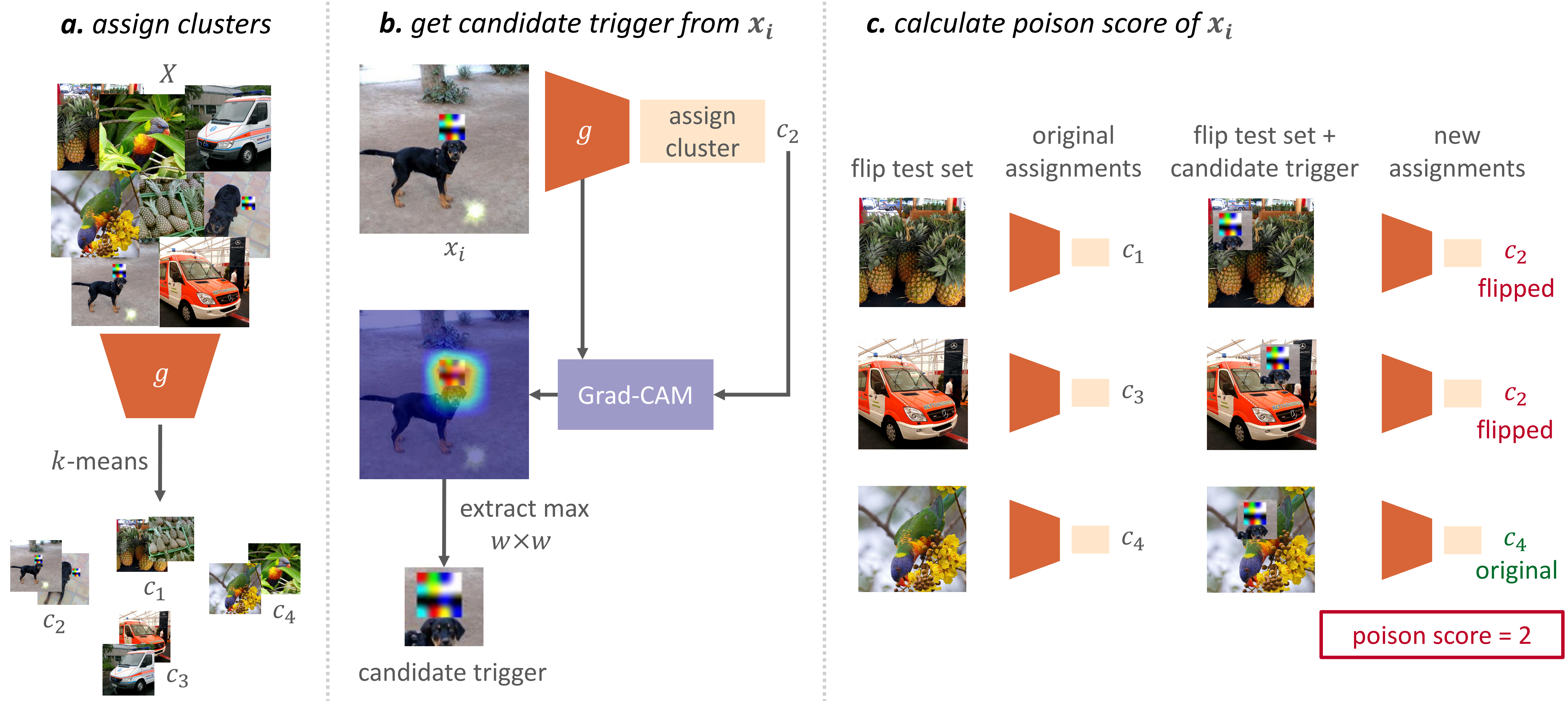
https://github.com/UCDvision/PatchSearch

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Background: Backdoor Attacks on SSL



Summary of PatchSearch

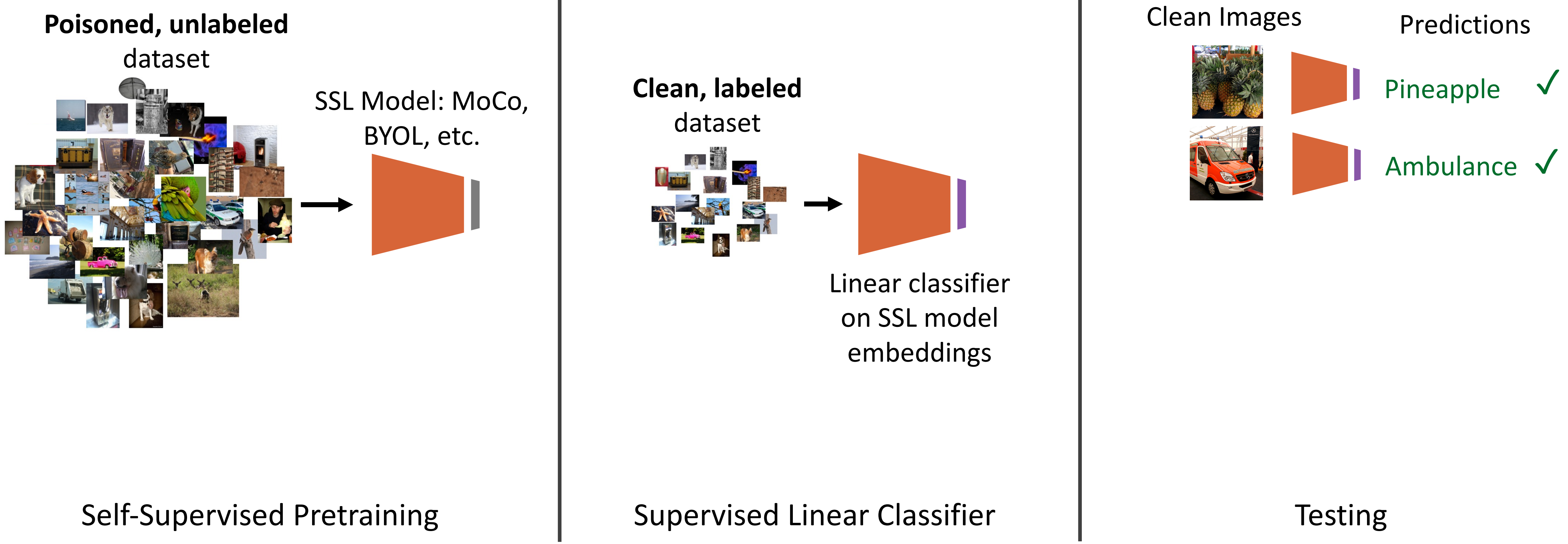


Summary of PatchSearch

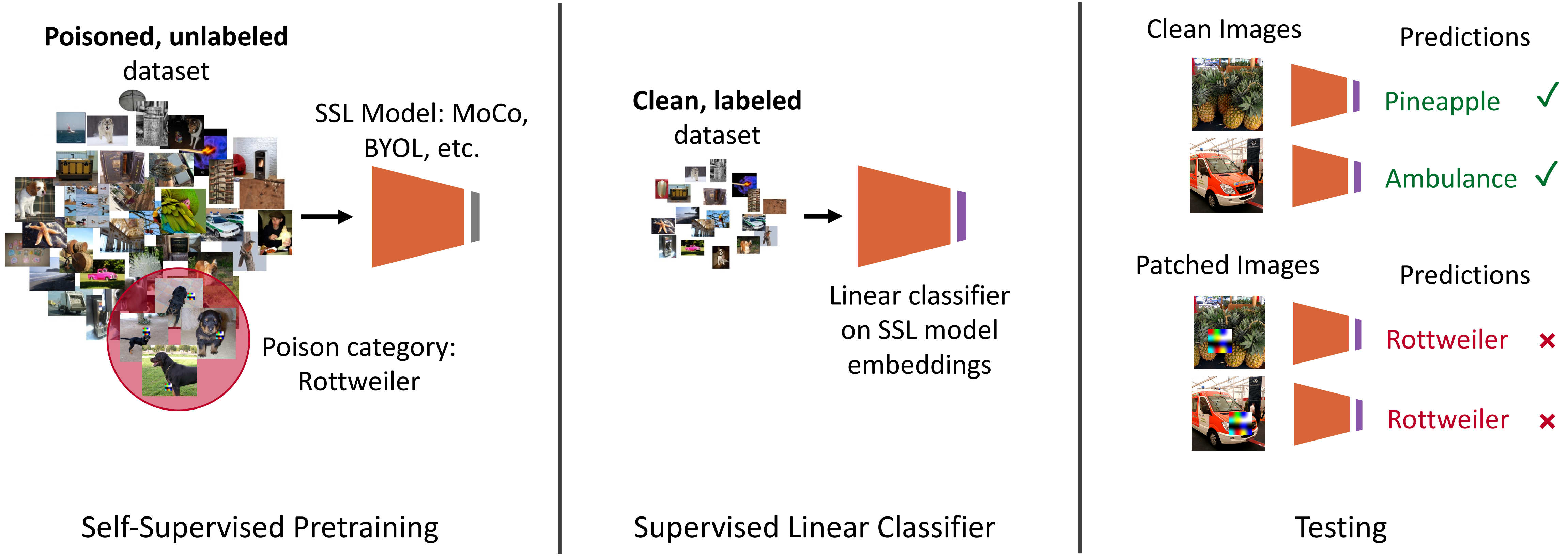
- PatchSearch successfully defends against the backdoor attack
- It restores model performance to the clean level

Model Type	Clean Data			Patched Data		
	Acc	FP	ASR	Acc	FP	ASR
ViT-B	*MoCo-v3, poison rate 0.5%*					
Clean	70.5	18.5	0.4	64.6	27.2	0.5
Backdoored	70.6	17.4	0.4	46.9	1708.9	34.5
PatchSearch	70.2	23.1	0.5	64.5↑	39.8↓	0.8↓

Background: Self-Supervised Learning (SSL)



Background: Backdoor Attacks on SSL



Goal: Defend SSL against Backdoor Attacks

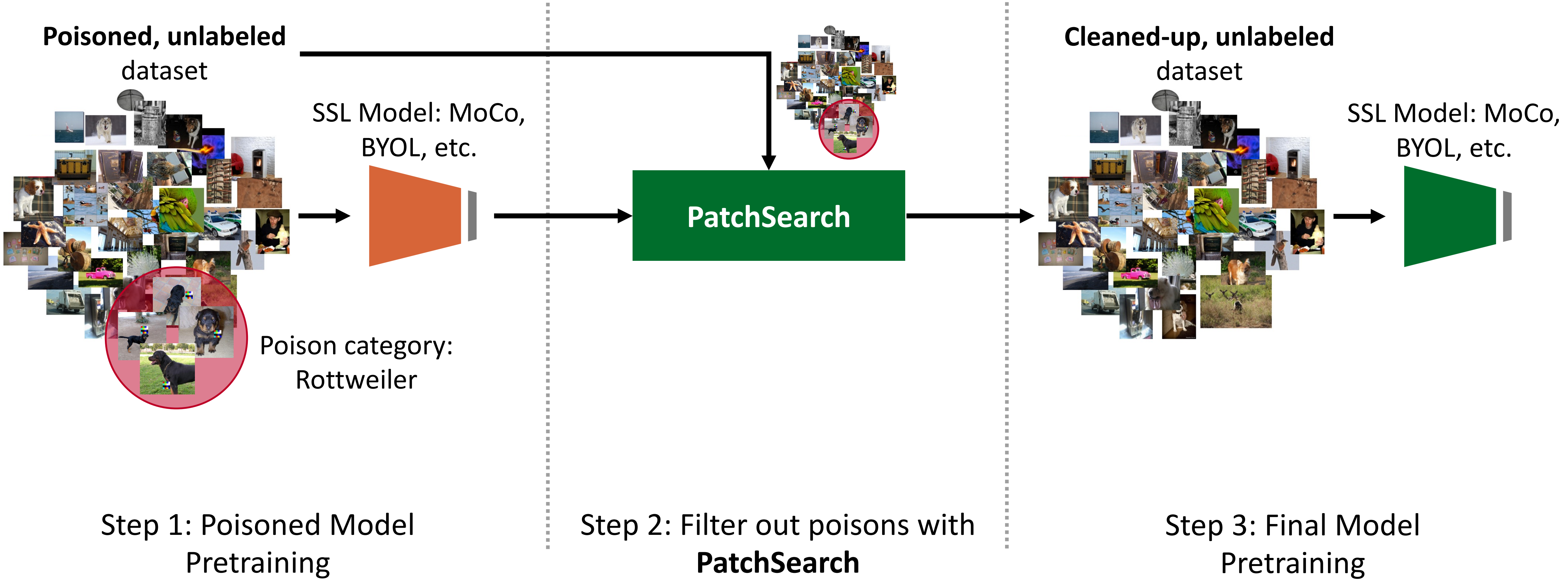
- **We focus on patch-based attacks. Why?**
    - More practical than image-wide perturbations
- **Challenges**
    - No access to trusted or labeled data
    - No knowledge about trigger appearance or location
    - Huge datasets with very few poisons

Existing Solutions

- **Supervised Backdoor Attack Defenses**
  - Most defenses directly rely on labels
    - Cannot be used in unlabeled settings
    - Our ideas are similar to SentiNet [a] and Februus[b] (supervised test-time defenses)
  - Some defenses do not rely on labels
    - e.g., strong augmentation like CutMix [c]
    - Can be used in unlabeled settings
- **KD + Trusted Data Defense [d]**
  - Uses Knowledge Distillation (KD) on clean, unlabeled but trusted data
  - Large amount of trusted data is required to retain accuracy

[a] Chou, Edward, Florian Tramer, and Giancarlo Pellegrino. "Sentinet: Detecting localized universal attacks against deep learning systems." *SPW 2020*.
[b] Doan, Bao Gia, Ehsan Abbasnejad, and Damith C. Ranasinghe. "Februus: Input purification defense against trojan attacks on deep neural network systems." *Annual Computer Security Applications Conference*. 2020.
[c] Borgnia, Eitan, et al. "Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff." *ICASSP 2021.*
[d] Saha, Aniruddha, et al. "Backdoor attacks on self-supervised learning." *CVPR 2022.*

Our Solution: 3-Step Defense



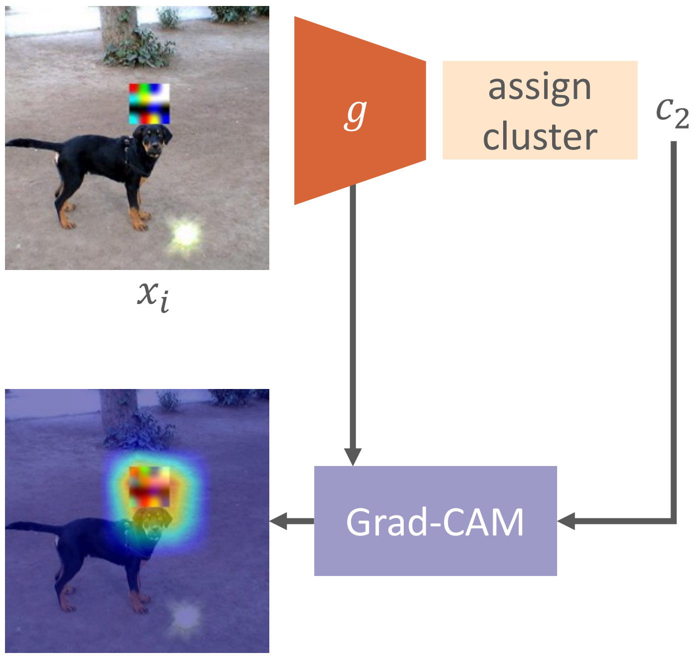
PatchSearch

***a.** assign clusters*

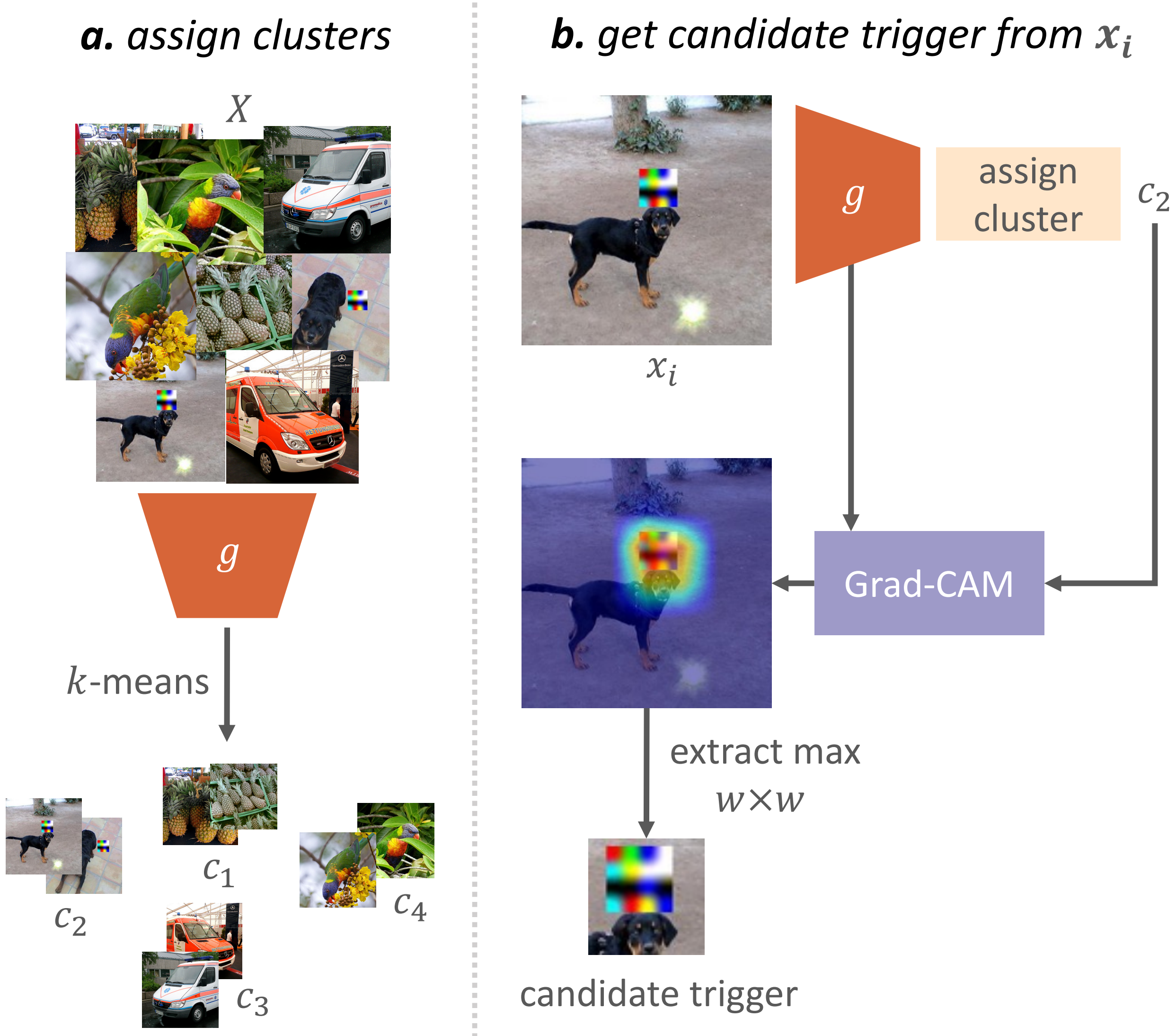
PatchSearch

a. *assign clusters*

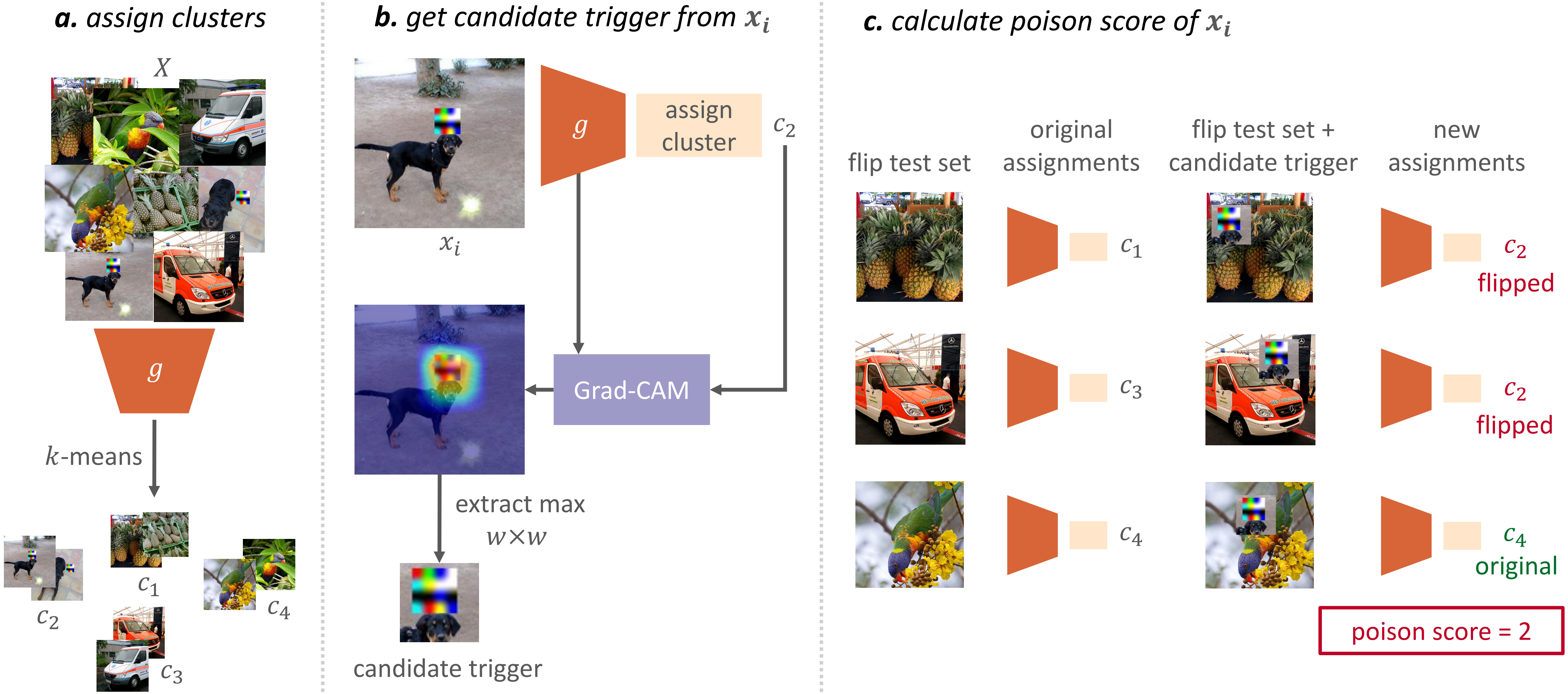
b. *get candidate trigger from* $\boldsymbol{x_i}$



PatchSearch

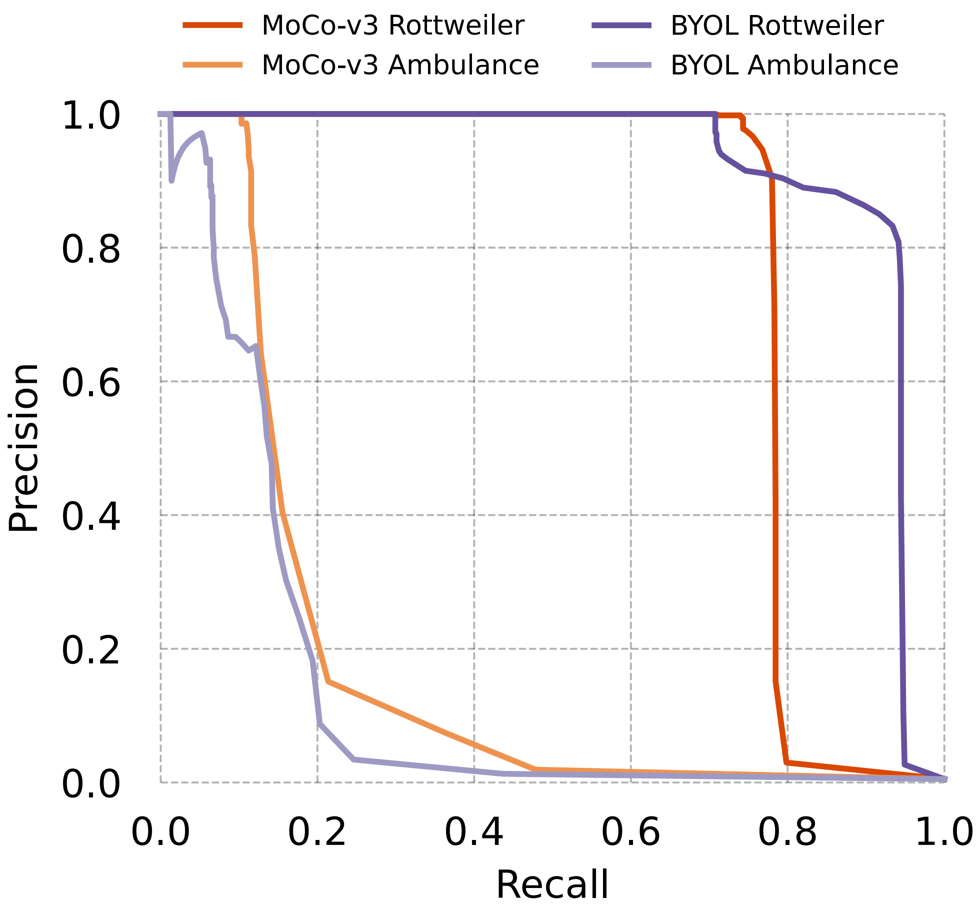


PatchSearch

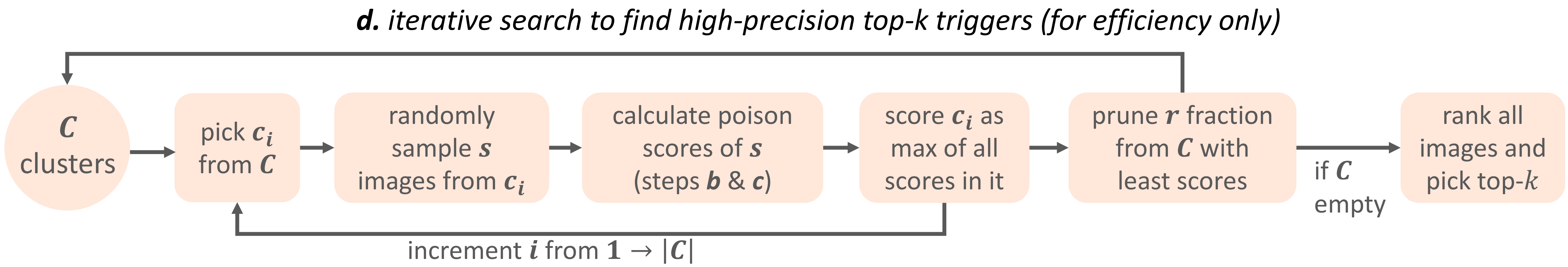


PatchSearch

- **Use above steps on entire dataset**
- **Rank dataset with poison score**
- **Remove top ranked images?**
  - Cannot detect all poisons
  - Ranking entire dataset is expensive
  - Only a few images are poisoned
- **Solution**
  - Efficiently search for a few top poisons
  - Build a classifier to detect similar images



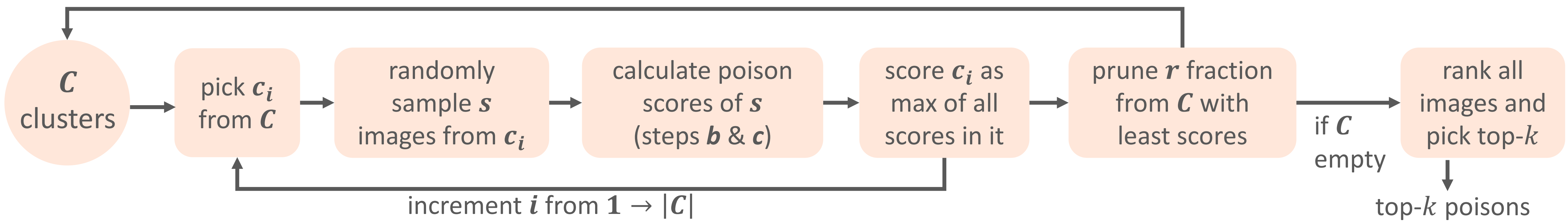
PatchSearch: Improving efficiency



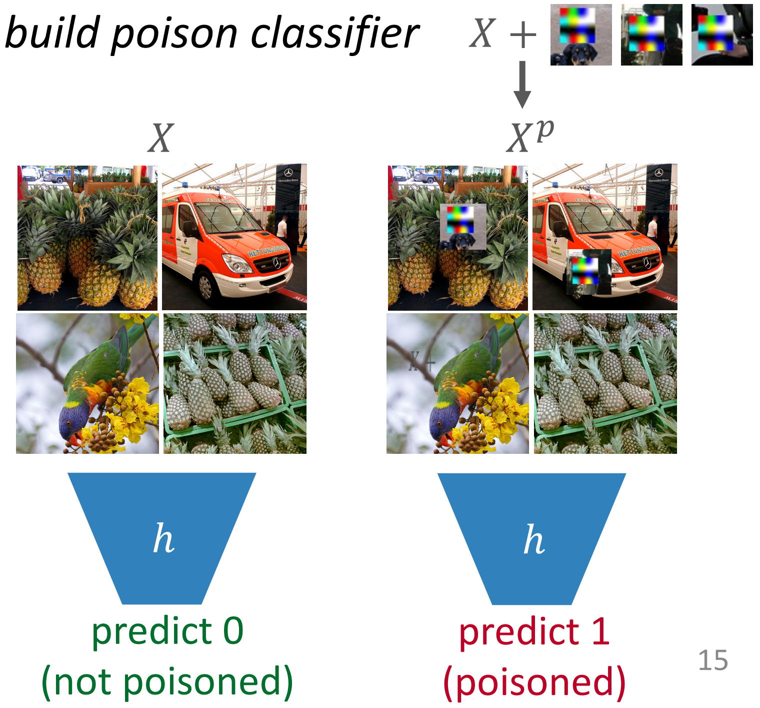
Model	Accuracy in top-20 (%)
BYOL, ResNet-18, 0.5%	99.5
MoCo-v3, ViT-B, 0.5%	96.5
MoCo-v3, ViT-B, 1.0%	97.5

PatchSearch: Improving poison detection

***d.** iterative search to find high-precision top-k triggers (for efficiency only)*



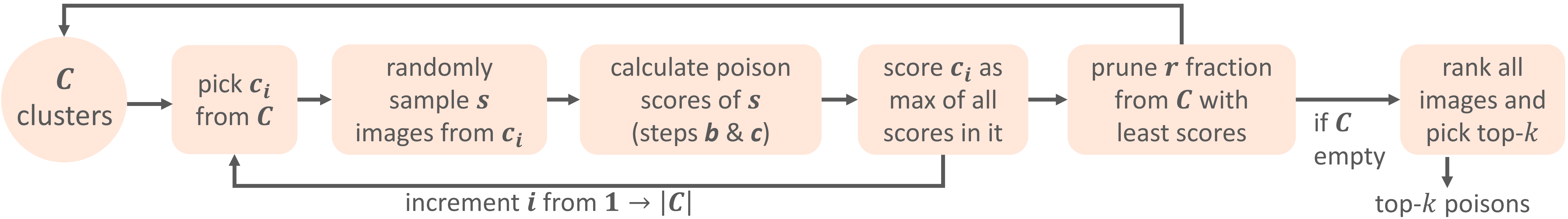
***e.** build poison classifier*



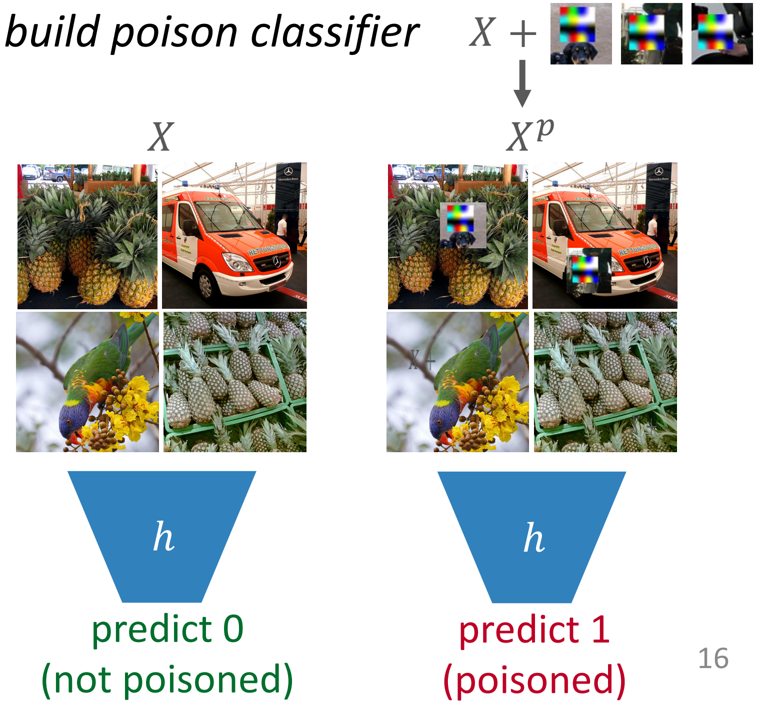
Model	Accuracy in top-20 (%)
BYOL, ResNet-18, 0.5%	99.5
MoCo-v3, ViT-B, 0.5%	96.5
MoCo-v3, ViT-B, 1.0%	97.5

PatchSearch: Improving poison detection

***d.** iterative search to find high-precision top-k triggers (for efficiency only)*



***e.** build poison classifier*

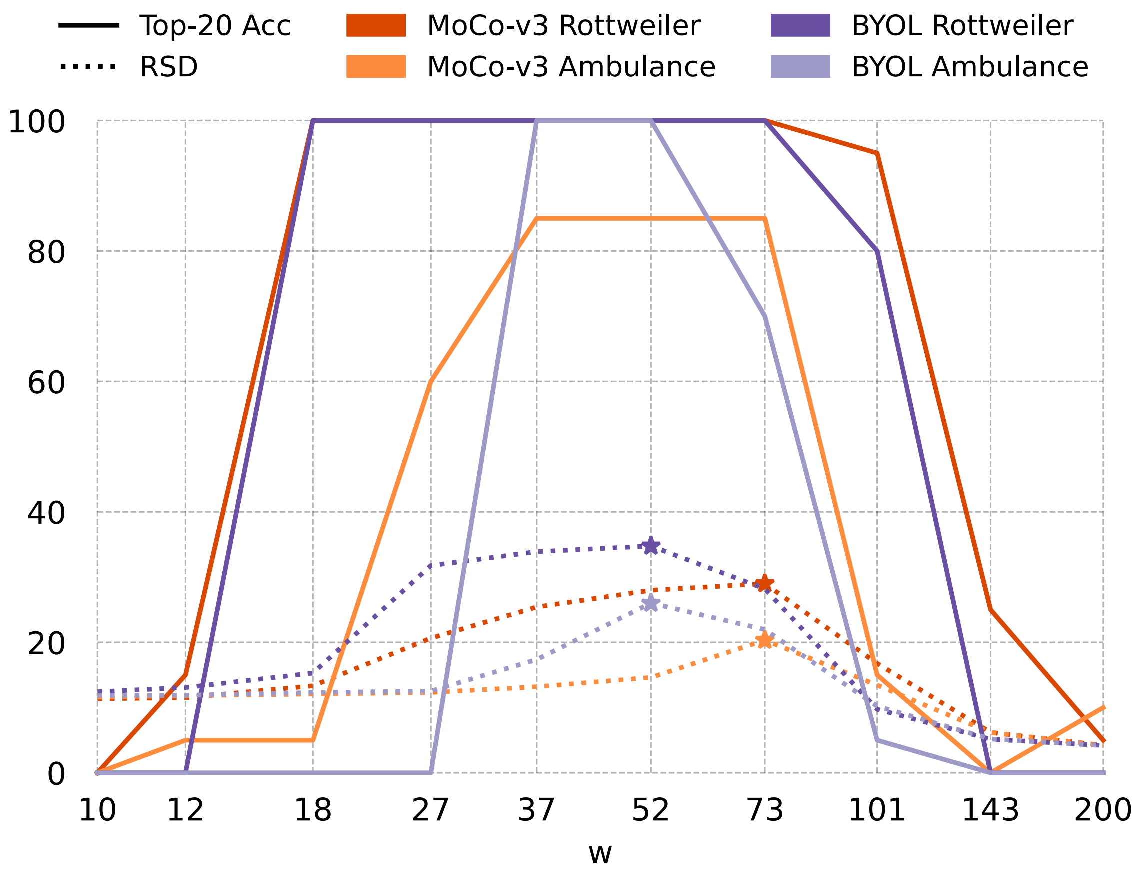


Model	Accuracy in top-20 (%)	Recall (%)	Precision (%)
BYOL, ResNet-18, 0.5%	99.5	98.0	58.8
MoCo-v3, ViT-B, 0.5%	96.5	98.8	48.3
MoCo-v3, ViT-B, 1.0%	97.5	99.0	59.5

Precision is not 100% but removing a few clean samples does not hurt overall model performance

PatchSearch: How to choose w blindly?

- w is candidate trigger size
- The defender does not know true trigger size
- A tight w around the trigger should result in few patches with relatively high scores
- Try out different w and pick the one that results in maximum variance in scores



Top-20 Acc is accuracy of PatchSearch and true trigger size is 50

Results

- Results averaged across 10 target categories
- **Clean Data**
  - All models behave similarly

Model Type	Clean Data			Patched Data		
	Acc	FP	ASR	Acc	FP	ASR
ViT-B	*MoCo-v3, poison rate 0.5%*					
Clean	70.5	18.5	0.4	64.6	27.2	0.5
Backdoored	70.6	17.4	0.4	46.9	1708.9	34.5
PatchSearch	70.2	23.1	0.5	64.5	39.8	0.8

Results

- Results averaged across 10 target categories
- **Clean Data**
  - All models behave similarly
- **Patched Data**
  - Backdoored models fail

Model Type	Clean Data			Patched Data		
	Acc	FP	ASR	Acc	FP	ASR
ViT-B	*MoCo-v3, poison rate 0.5%*					
Clean	70.5	18.5	0.4	64.6	27.2	0.5
Backdoored	70.6	17.4	0.4	46.9↓	1708.9↑	34.5↑
PatchSearch	70.2	23.1	0.5	64.5	39.8	0.8

Results

- Results averaged across 10 target categories
- **Clean Data**
  - All models behave similarly
- **Patched Data**
  - Backdoored models fail
  - PatchSearch models improve

Model Type	Clean Data			Patched Data		
	Acc	FP	ASR	Acc	FP	ASR
ViT-B	*MoCo-v3, poison rate 0.5%*					
Clean	70.5	18.5	0.4	64.6	27.2	0.5
Backdoored	70.6	17.4	0.4	46.9	1708.9	34.5
PatchSearch	70.2	23.1	0.5	64.5↑	39.8↓	0.8↓

Results

- Results averaged across 10 target categories
- **Clean Data**
  - All models behave similarly
- **Patched Data**
  - Backdoored models fail
  - PatchSearch models improve
  - Performance is restored to clean model levels

Model Type	Clean Data			Patched Data		
	Acc	FP	ASR	Acc	FP	ASR
ViT-B	*MoCo-v3, poison rate 0.5%*					
Clean	70.5	18.5	0.4	64.6	27.2	0.5
Backdoored	70.6	17.4	0.4	46.9	1708.9	34.5
PatchSearch	70.2	23.1	0.5	64.5	39.8	0.8

Results: i-CutMix

- **i-CutMix**
  - Augmentation for contrastive learning
  - No labels needed
  - Improves clean model

Model Type	Clean Data			Patched Data		
	Acc	FP	ASR	Acc	FP	ASR
ViT-B	*MoCo-v3, poison rate 0.5%*					
Clean	70.5	18.5	0.4	64.6	27.2	0.5
Backdoored	70.6	17.4	0.4	46.9	1708.9	34.5
PatchSearch	70.2	23.1	0.5	64.5	39.8	0.8
Clean + i-CutMix	75.6↑	15.6	0.3	74.4↑	14.6	0.3

Results: i-CutMix

- **i-CutMix**
  - Augmentation for contrastive learning
  - No labels needed
  - Improves clean model
  - Simple and effective defense

Model Type	Clean Data			Patched Data		
	Acc	FP	ASR	Acc	FP	ASR
ViT-B	*MoCo-v3, poison rate 0.5%*					
Clean	70.5	18.5	0.4	64.6	27.2	0.5
Backdoored	70.6	17.4	0.4	46.9	1708.9	34.5
PatchSearch	70.2	23.1	0.5	64.5	39.8	0.8
Clean + *i*-CutMix	75.6	15.6	0.3	74.4	14.6	0.3
Backdoored + *i*-CutMix	75.6	14.9	0.3	72.2↑	242.2↓	4.9↓

Results: i-CutMix

- **i-CutMix**
  - Augmentation for contrastive learning
  - No labels needed
  - Improves clean model
  - Simple and effective defense
- **Compared to PatchSearch**
  - Works implicitly
  - Cannot detect poisons
  - PatchSearch is a better defense

Model Type	Clean Data			Patched Data		
	Acc	FP	ASR	Acc	FP	ASR
ViT-B	*MoCo-v3, poison rate 0.5%*					
Clean	70.5	18.5	0.4	64.6	27.2	0.5
Backdoored	70.6	17.4	0.4	46.9	1708.9	34.5
PatchSearch	70.2	23.1	0.5	64.5	39.8	0.8
Clean + *i*-CutMix	75.6	15.6	0.3	74.4	14.6	0.3
Backdoored + *i*-CutMix	75.6	14.9	0.3	72.2↑	242.2↑	4.9↑

Results: i-CutMix

- **i-CutMix**
  - Augmentation for contrastive learning
  - No labels needed
  - Improves clean model
  - Simple and effective defense
- **Compared to PatchSearch**
  - Works implicitly
  - Cannot detect poisons
  - PatchSearch is a better defense
- **Combination of both is best**

Model Type	Clean Data			Patched Data		
	Acc	FP	ASR	Acc	FP	ASR
ViT-B	*MoCo-v3, poison rate 0.5%*					
Clean	70.5	18.5	0.4	64.6	27.2	0.5
Backdoored	70.6	17.4	0.4	46.9	1708.9	34.5
PatchSearch	70.2	23.1	0.5	64.5	39.8	0.8
Clean + *i*-CutMix	75.6	15.6	0.3	74.4	14.6	0.3
Backdoored + *i*-CutMix	75.6	14.9	0.3	72.2	242.2	4.9
PatchSearch + *i*-CutMix	75.2	19.7	0.4	74.2↑	19.0↓	0.4↓

Results: KD Defense

- **Comparison with KD Defense**
  - Proposed in [d]
  - Uses unlabeled but trusted data
  - PatchSearch has better accuracy and slightly higher FP

Model	Trusted Data	Clean Data		Patched Data	
		Acc	FP	Acc	FP
Clean	100%	49.9	23.0	47.0	22.8
Backdoored	0%	50.1	26.2	31.8	1683.2
KD Defense	25%	44.6	34.5	42.0	37.9
KD Defense	10%	38.3	40.5	35.7	44.8
KD Defense	5%	32.1	41.0	29.4	53.7
PatchSearch	0%	49.4↑	40.1	45.9↑	50.3↑

[d] Saha, Aniruddha, et al. "Backdoor attacks on self-supervised learning." *CVPR 2022.*

Results: MAE

- **Comparison MAE**
  - MAE was shown to be robust to backdoor attacks in [d]
  - However, MAE requires finetuning to be comparable to MoCo-v3
  - Also, a properly defended MoCo-v3 has better model performance

	Clean Data		Patched Data	
Method	Acc	FP	Acc	FP
Finetuned with 1% labeled data				
MAE	65.7	18.7	53.8	97.6
MoCo-v3 (PatchSearch + i-CutMix)	78.2 ↑	20.2	76.8 ↑	17.1

[d] Saha, Aniruddha, et al. "Backdoor attacks on self-supervised learning." *CVPR 2022.*

Conclusion

- **PatchSearch**
  - Significantly mitigates the attack
  - Finds highly influential patches
  - Better than i-CutMix and KD Defense
  - Combining with i-CutMix works best