Precise Extraction of Deep Learning Models via Side-Channel Attacks on Edge/Endpoint Devices

Younghan Lee[1], Sohee Jun[1], Yungi Cho[1], Woorim Han[1],
Hyungon Moon[2] and Yunheung Paek[1]

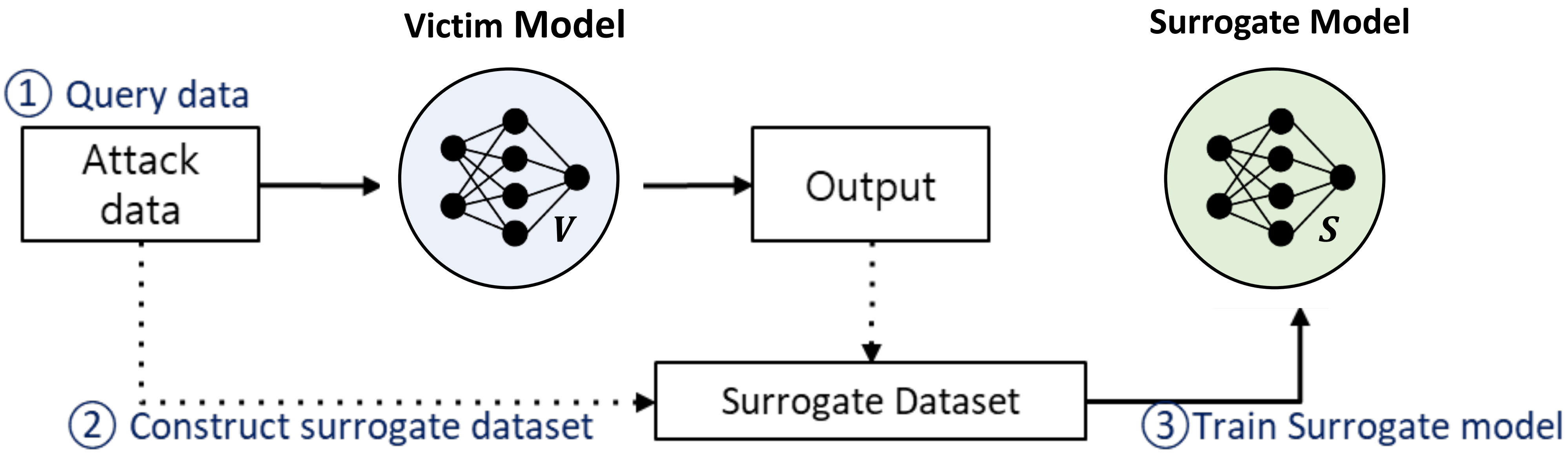
[1]Seoul National University (SNU),
[2]Ulsan National Institute of Science & Technology (UNIST)

Table of Contents

- Model Extraction Attack
  - How it works
  - Our insight
- Analysis on Effects of Model Information
  - Various analysis settings
- Experiments
  - Model extraction via side-channel attack
- Conclusion

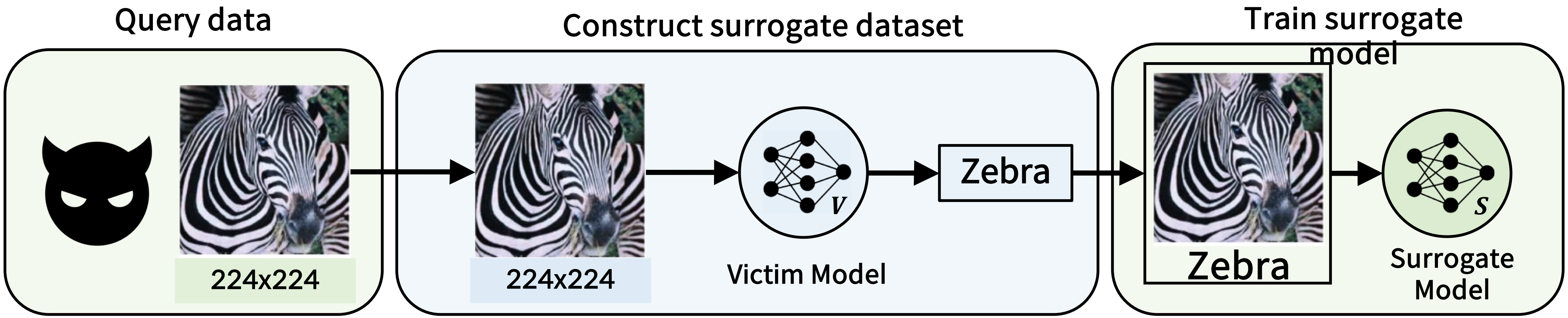
Model Extraction Attack

- How it works



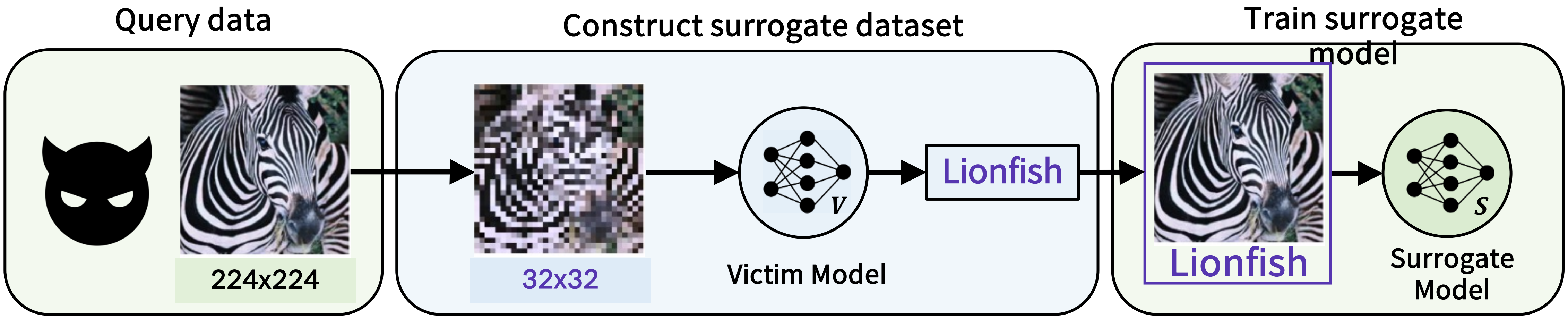
Our Insight

- Current MEA operates with **the same model information**
  - e.g.,) image dimension (ID)

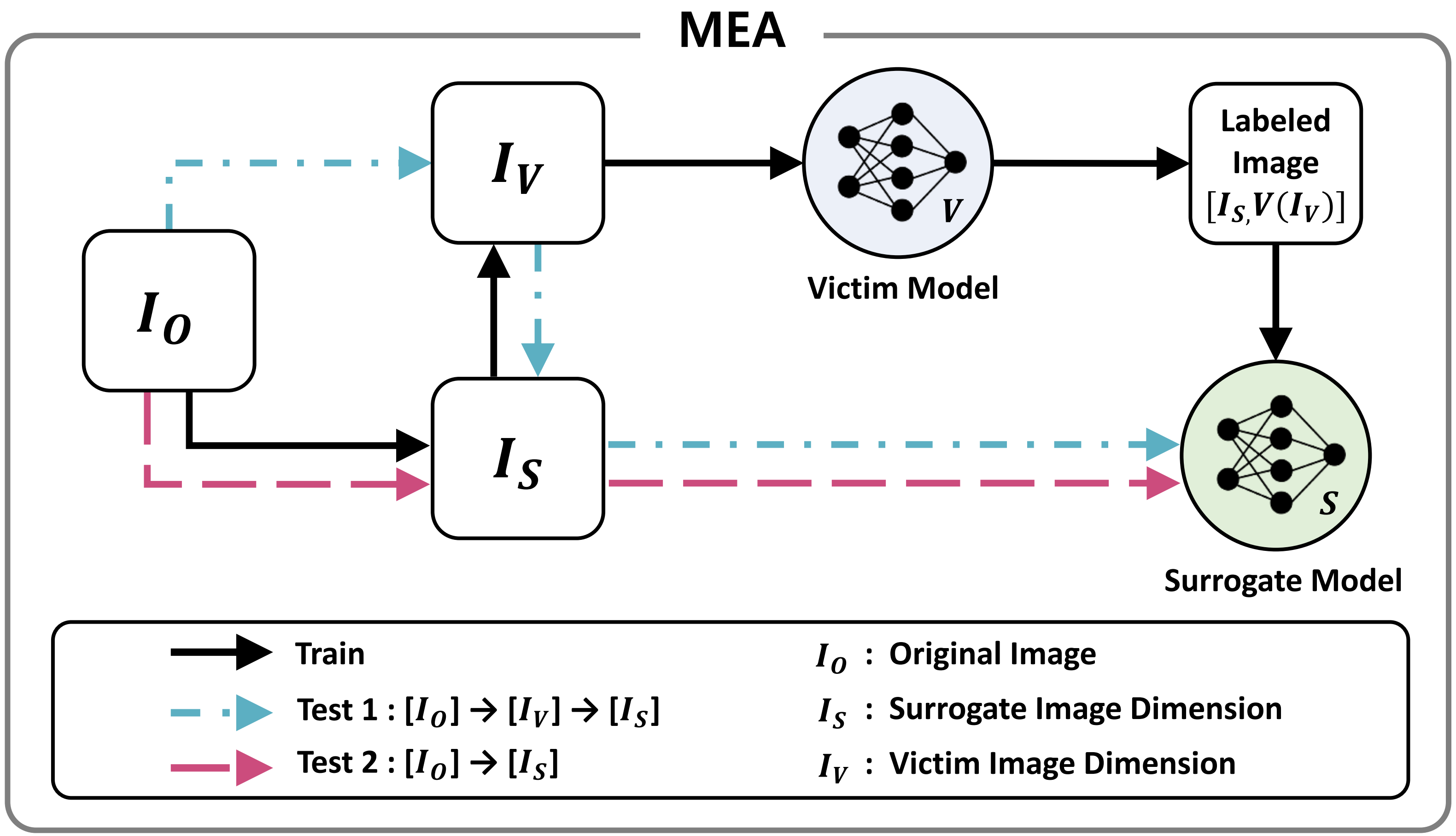


Our Insight

- When the adversaries **Do Not** have such information
  - e.g.,) image dimension (ID)



Analysis on Effects of Model Information



Train

Test 1 : $[I_O] \rightarrow [I_V] \rightarrow [I_S]$

Test 2 : $[I_O] \rightarrow [I_S]$

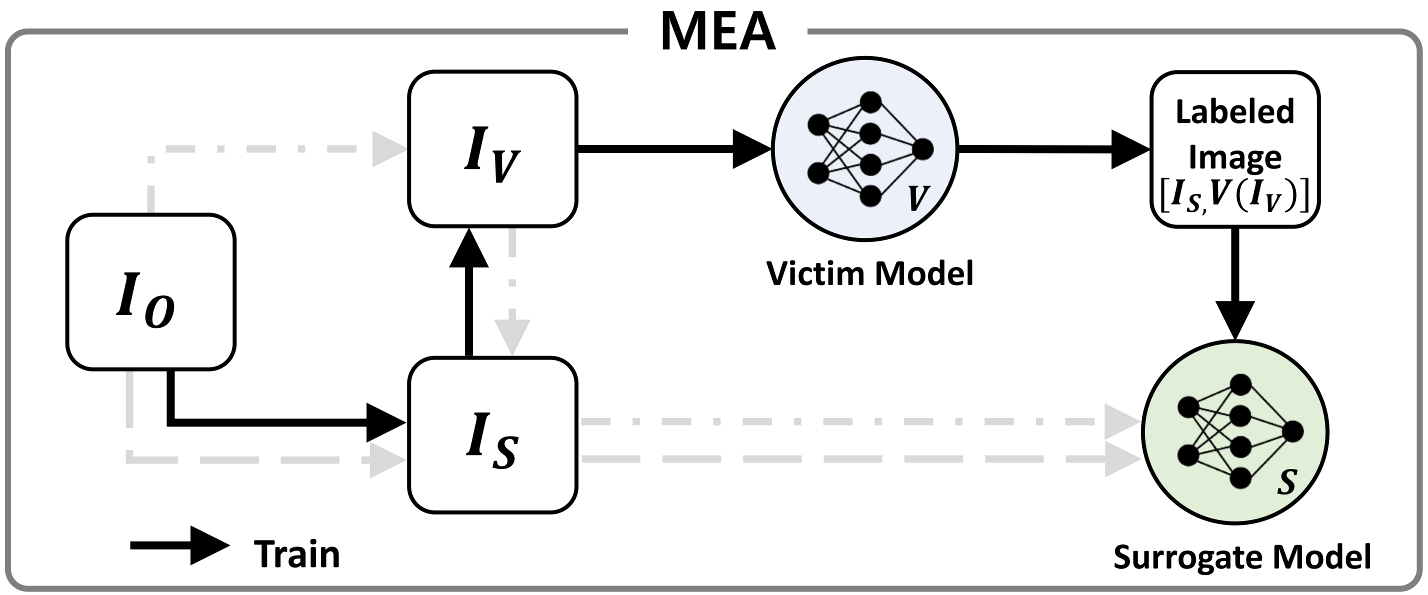
I_O : Original Image

I_S : Surrogate Image Dimension

I_V : Victim Image Dimension

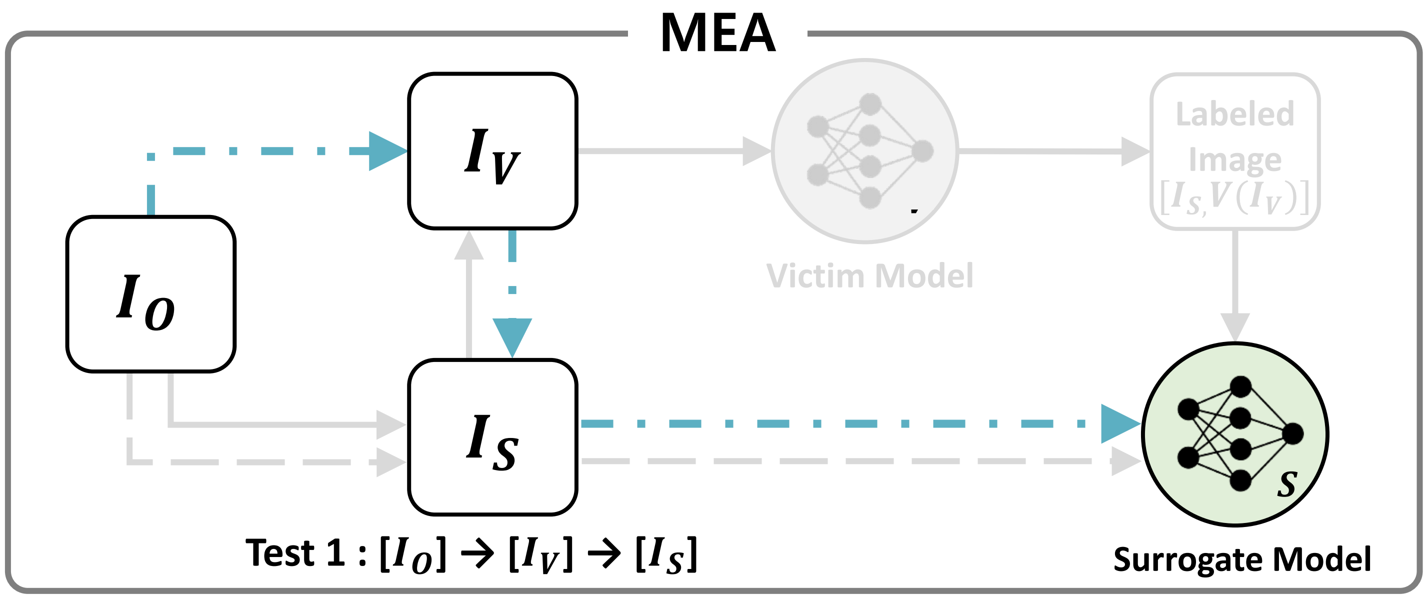
Analysis on Effects of Model Information

- Construct surrogate dataset with surrogate model's ID
  - Re-labeled Image $[I_s , V(I_V)]$
- Train the surrogate with re-labeled images



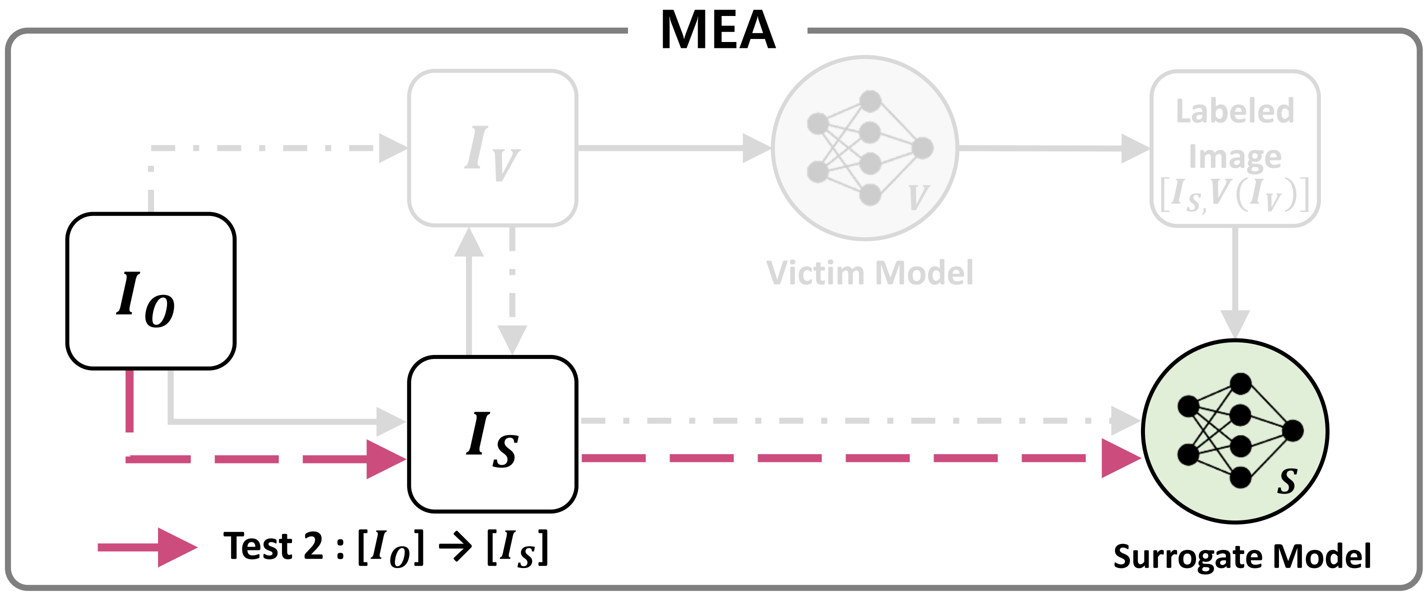
Analysis on Effects of Model Information

- Evaluate the surrogate by converting the dimension
  - First, to victim model’s image dimension
  - Then, to surrogate model’s image dimension


MEA
I_O
I_V
I_S
Victim Model
Labeled Image $[I_S, V(I_V)]$
S
Surrogate Model
Test 1 : $[I_O] \rightarrow [I_V] \rightarrow [I_S]$

Analysis on Effects of Model Information

- Evaluate the surrogate by converting the dimension
  - Directly to surrogate model's image dimension



MEA
I_O
I_V
I_S
V
Victim Model
Labeled Image $[I_S, V(I_V)]$
S
Surrogate Model
Test 2 : $[I_O] \rightarrow [I_S]$

Various Analysis Settings

- Datasets

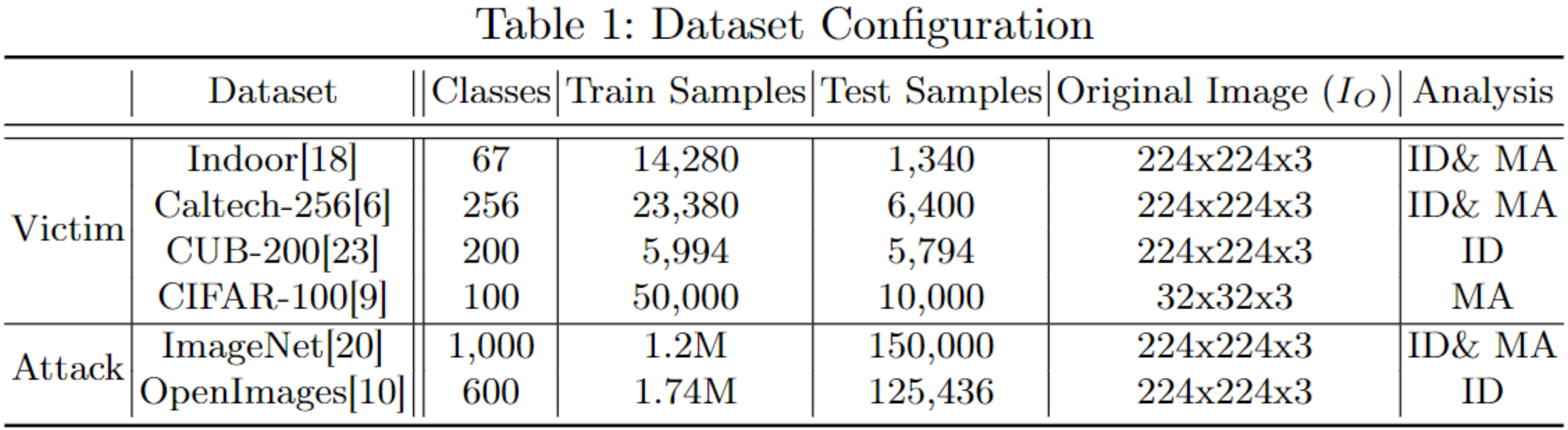


Table 1: Dataset Configuration

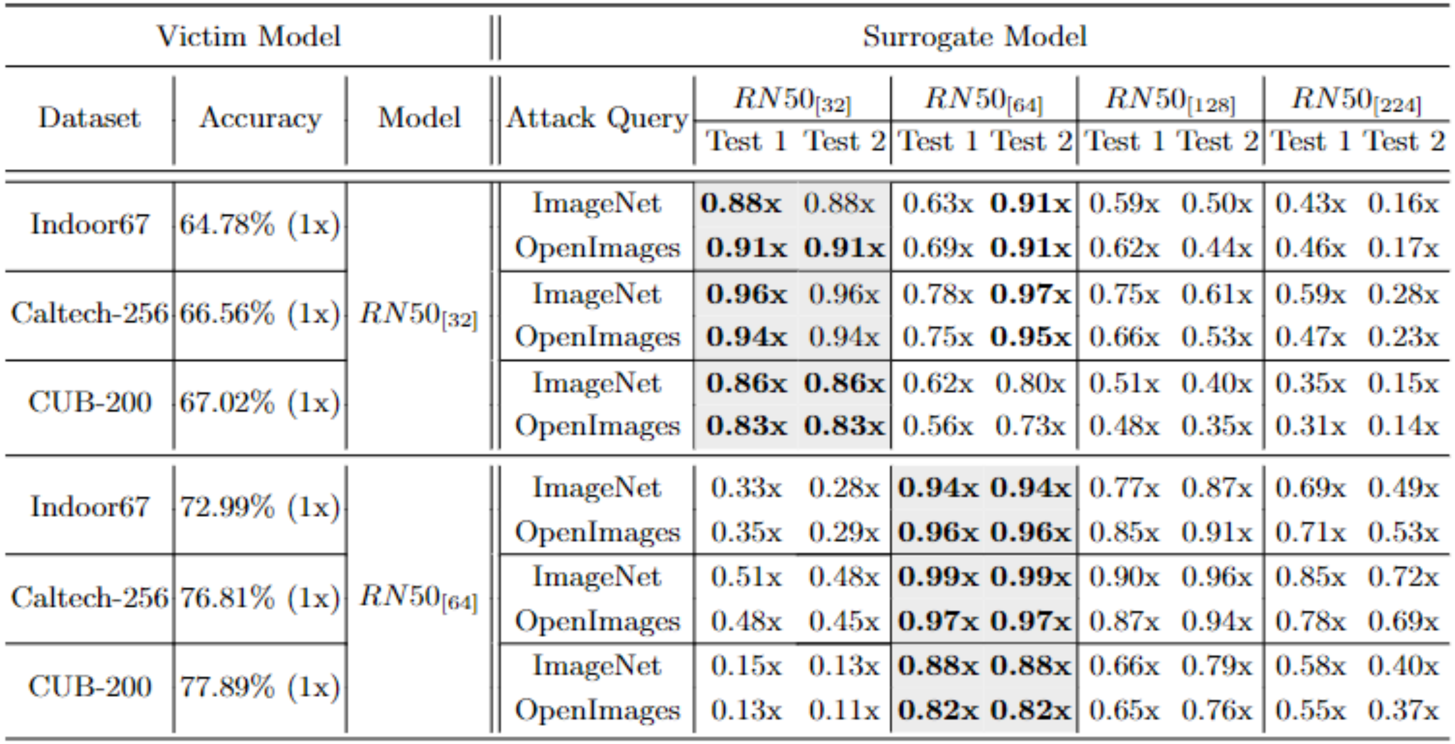
	Dataset	Classes	Train Samples	Test Samples	Original Image (I_O)	Analysis
Victim	Indoor[18]	67	14,280	1,340	224x224x3	ID& MA
	Caltech-256[6]	256	23,380	6,400	224x224x3	ID& MA
	CUB-200[23]	200	5,994	5,794	224x224x3	ID
	CIFAR-100[9]	100	50,000	10,000	32x32x3	MA
Attack	ImageNet[20]	1,000	1.2M	150,000	224x224x3	ID& MA
	OpenImages[10]	600	1.74M	125,436	224x224x3	ID

Various Analysis Settings

- Attack Query Budget
  - 30k, 60k, 90k
  - Higher the budget, stronger the attack
- Attack Strategy
  - Randomly select the query dataset (KnockoffNets, CVPR '19)
  - Adaptively select the query dataset (ActiveThief, AAAI ‘20)
- Model Architecture
  - WideResNet-28-k
  - Higher the value of k, more complex the architecture

Analysis Result 1

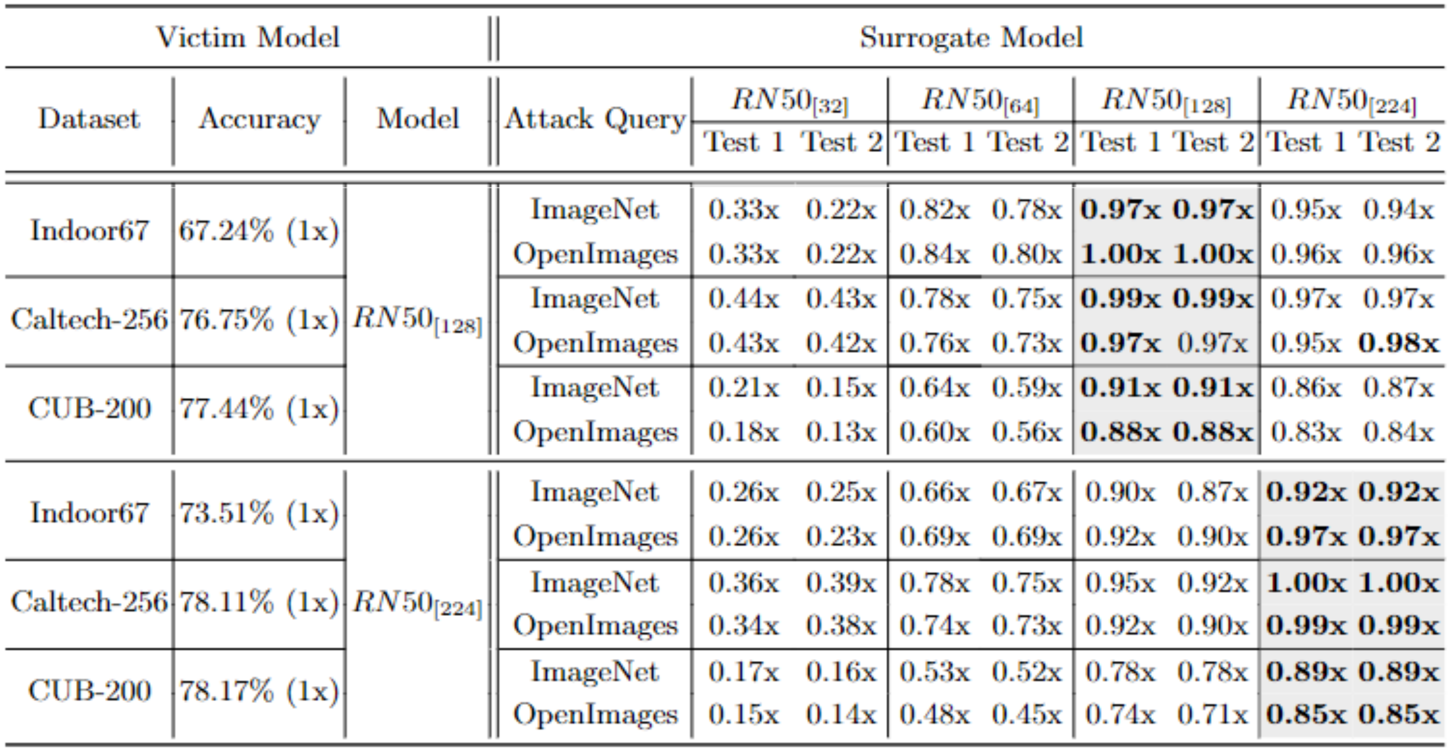
- Various datasets
- **Same ID achieves the best relative accuracy**



Victim Model			Surrogate Model								
Dataset	Accuracy	Model	Attack Query	$RN50_{[32]}$		$RN50_{[64]}$		$RN50_{[128]}$		$RN50_{[224]}$	
				Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Indoor67	64.78% (1x)	$RN50_{[32]}$	ImageNet	**0.88x**	0.88x	0.63x	**0.91x**	0.59x	0.50x	0.43x	0.16x
			OpenImages	**0.91x**	**0.91x**	0.69x	**0.91x**	0.62x	0.44x	0.46x	0.17x
Caltech-256	66.56% (1x)		ImageNet	**0.96x**	0.96x	0.78x	**0.97x**	0.75x	0.61x	0.59x	0.28x
			OpenImages	**0.94x**	0.94x	0.75x	**0.95x**	0.66x	0.53x	0.47x	0.23x
CUB-200	67.02% (1x)		ImageNet	**0.86x**	**0.86x**	0.62x	0.80x	0.51x	0.40x	0.35x	0.15x
			OpenImages	**0.83x**	**0.83x**	0.56x	0.73x	0.48x	0.35x	0.31x	0.14x
Indoor67	72.99% (1x)	$RN50_{[64]}$	ImageNet	0.33x	0.28x	**0.94x**	**0.94x**	0.77x	0.87x	0.69x	0.49x
			OpenImages	0.35x	0.29x	**0.96x**	**0.96x**	0.85x	0.91x	0.71x	0.53x
Caltech-256	76.81% (1x)		ImageNet	0.51x	0.48x	**0.99x**	**0.99x**	0.90x	0.96x	0.85x	0.72x
			OpenImages	0.48x	0.45x	**0.97x**	**0.97x**	0.87x	0.94x	0.78x	0.69x
CUB-200	77.89% (1x)		ImageNet	0.15x	0.13x	**0.88x**	**0.88x**	0.66x	0.79x	0.58x	0.40x
			OpenImages	0.13x	0.11x	**0.82x**	**0.82x**	0.65x	0.76x	0.55x	0.37x

Analysis Result 1

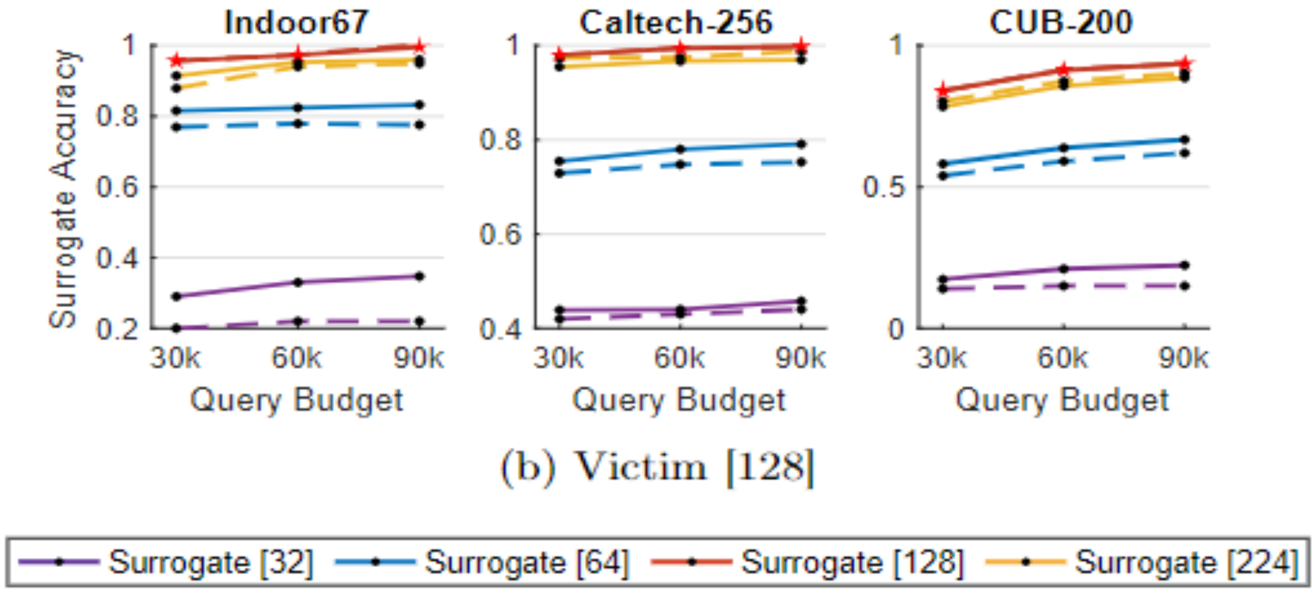
- Various datasets
- **Same ID achieves the best relative accuracy**



Victim Model			Surrogate Model								
Dataset	Accuracy	Model	Attack Query	$RN50_{[32]}$		$RN50_{[64]}$		$RN50_{[128]}$		$RN50_{[224]}$	
				Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Indoor67	67.24% (1x)	$RN50_{[128]}$	ImageNet	0.33x	0.22x	0.82x	0.78x	**0.97x**	**0.97x**	0.95x	0.94x
			OpenImages	0.33x	0.22x	0.84x	0.80x	**1.00x**	**1.00x**	0.96x	0.96x
Caltech-256	76.75% (1x)		ImageNet	0.44x	0.43x	0.78x	0.75x	**0.99x**	**0.99x**	0.97x	0.97x
			OpenImages	0.43x	0.42x	0.76x	0.73x	**0.97x**	0.97x	0.95x	**0.98x**
CUB-200	77.44% (1x)		ImageNet	0.21x	0.15x	0.64x	0.59x	**0.91x**	**0.91x**	0.86x	0.87x
			OpenImages	0.18x	0.13x	0.60x	0.56x	**0.88x**	**0.88x**	0.83x	0.84x
Indoor67	73.51% (1x)	$RN50_{[224]}$	ImageNet	0.26x	0.25x	0.66x	0.67x	0.90x	0.87x	**0.92x**	**0.92x**
			OpenImages	0.26x	0.23x	0.69x	0.69x	0.92x	0.90x	**0.97x**	**0.97x**
Caltech-256	78.11% (1x)		ImageNet	0.36x	0.39x	0.78x	0.75x	0.95x	0.92x	**1.00x**	**1.00x**
			OpenImages	0.34x	0.38x	0.74x	0.73x	0.92x	0.90x	**0.99x**	**0.99x**
CUB-200	78.17% (1x)		ImageNet	0.17x	0.16x	0.53x	0.52x	0.78x	0.78x	**0.89x**	**0.89x**
			OpenImages	0.15x	0.14x	0.48x	0.45x	0.74x	0.71x	**0.85x**	**0.85x**

Analysis Result 2

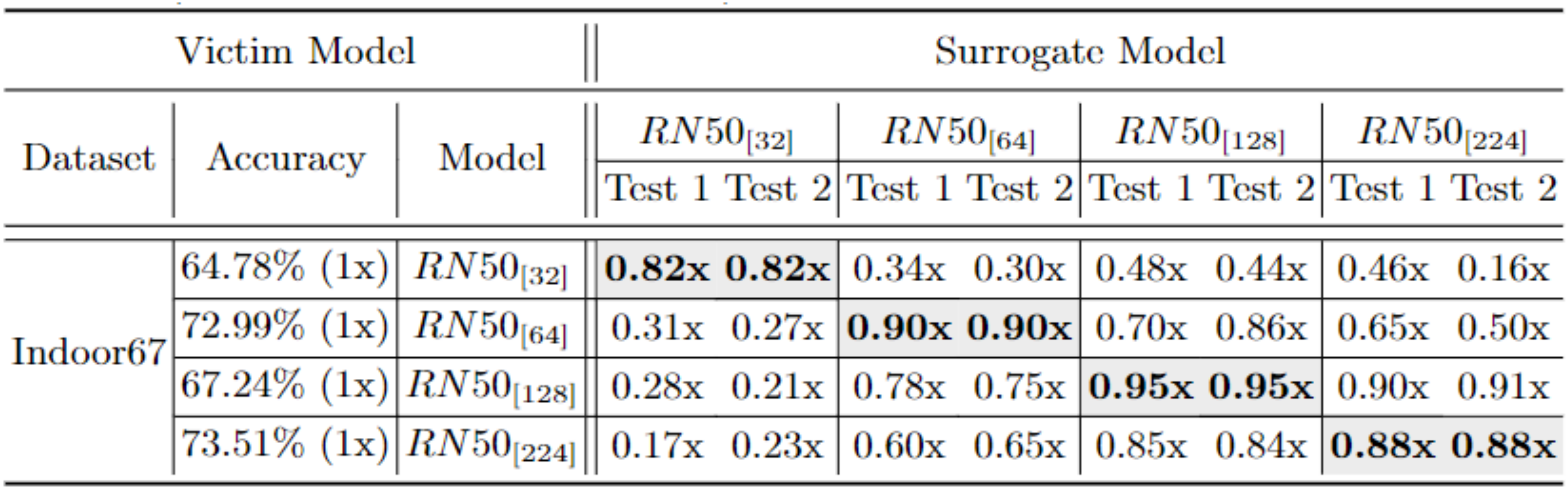
- Various query budgets
- Solid line = Test 1, Dotted = Test 2, Starred = Same ID
- **Same ID achieves the best relative accuracy**



(b) Victim [128]

Analysis Result 3

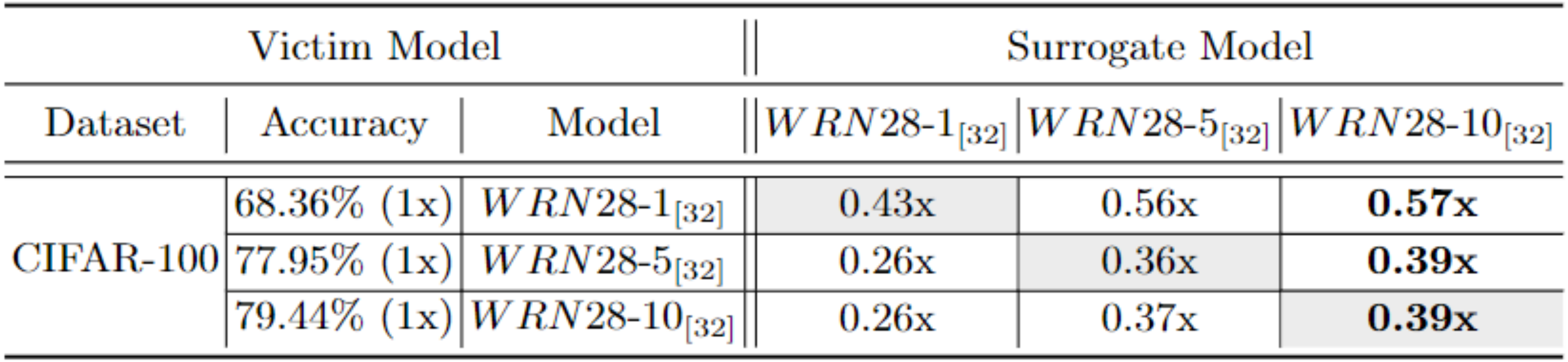
- Different attack strategy (ActiveThief)
- **Same ID achieves the best relative accuracy**



Victim Model			Surrogate Model							
Dataset	Accuracy	Model	$RN50_{[32]}$		$RN50_{[64]}$		$RN50_{[128]}$		$RN50_{[224]}$	
			Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Indoor67	64.78% (1x)	$RN50_{[32]}$	**0.82x**	**0.82x**	0.34x	0.30x	0.48x	0.44x	0.46x	0.16x
	72.99% (1x)	$RN50_{[64]}$	0.31x	0.27x	**0.90x**	**0.90x**	0.70x	0.86x	0.65x	0.50x
	67.24% (1x)	$RN50_{[128]}$	0.28x	0.21x	0.78x	0.75x	**0.95x**	**0.95x**	0.90x	0.91x
	73.51% (1x)	$RN50_{[224]}$	0.17x	0.23x	0.60x	0.65x	0.85x	0.84x	**0.88x**	**0.88x**

Analysis Result 4

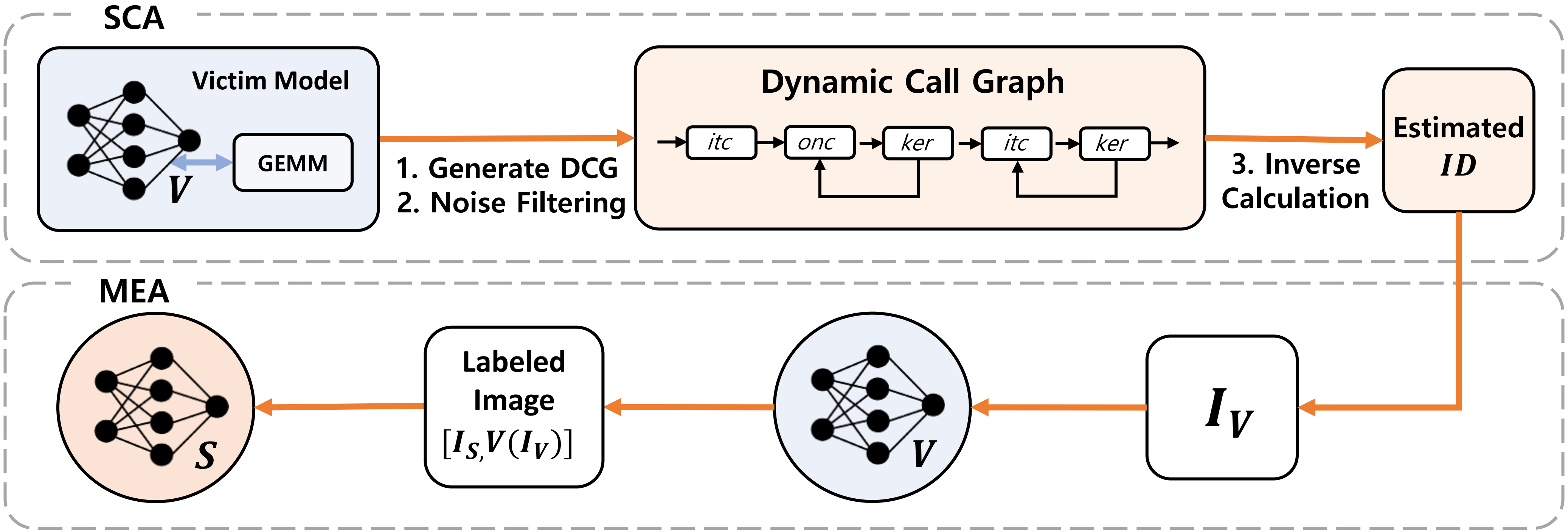
- Various model complexity
- **Model with higher complexity achieves the best relative accuracy**



Victim Model			Surrogate Model		
Dataset	Accuracy	Model	$WRN28\text{-}1_{[32]}$	$WRN28\text{-}5_{[32]}$	$WRN28\text{-}10_{[32]}$
CIFAR-100	68.36% (1x)	$WRN28\text{-}1_{[32]}$	0.43x	0.56x	**0.57x**
	77.95% (1x)	$WRN28\text{-}5_{[32]}$	0.26x	0.36x	**0.39x**
	79.44% (1x)	$WRN28\text{-}10_{[32]}$	0.26x	0.37x	**0.39x**

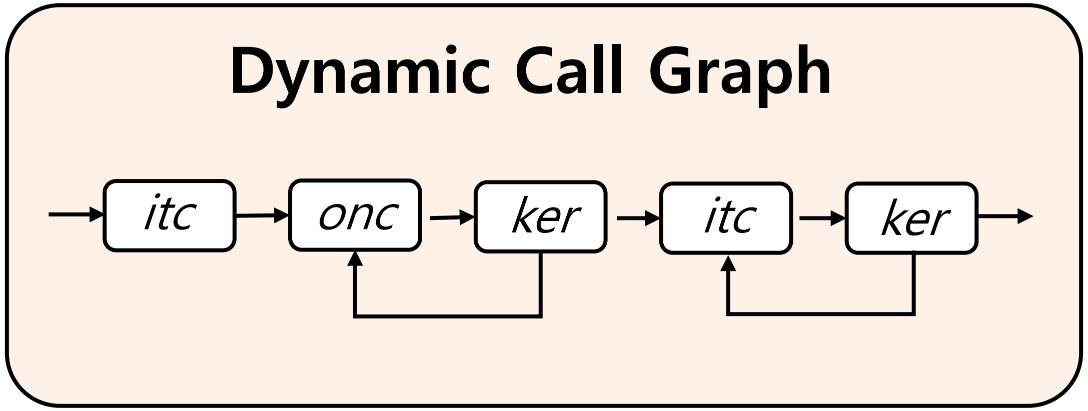
Experiments

- Model extraction via side-channel attack



Experiments

- 1. Generate DCG
  - Using *Flush+Reload,* monitor the addresses of the key functions
  - Count the number of each loop
  - Loop 1 → *itcopy – oncopy – kernel – itcopy - kernel*
  - Loop 2 → *itcopy - kernel*
  - Loop 3 → *oncopy - kernel*


Dynamic Call Graph

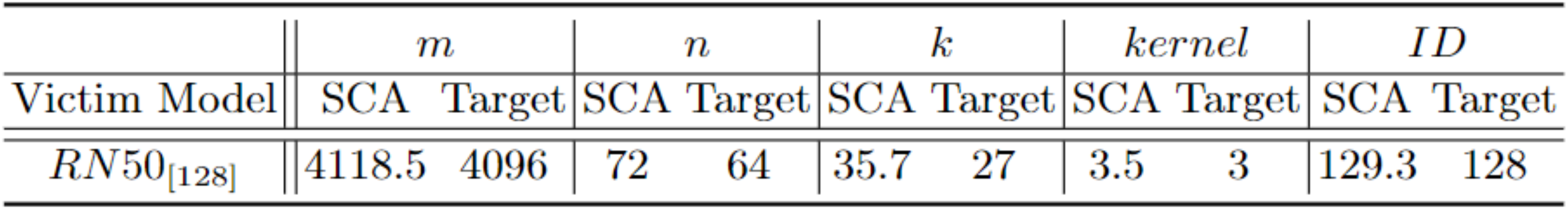
Experiments

- 2. Noise Filtering Mechanism
  - Filter out the function calls observed shortly after the previous one
    - < 10 intervals
  - Filter out any function calls within the threshold
    - Use the average interval between the function calls as a threshold

- 3. Estimate Image Dimension through Inverse Calculation
  - Use properties obtained from DCG to calculate ID inversely
    - Details in the paper

Experimental Results

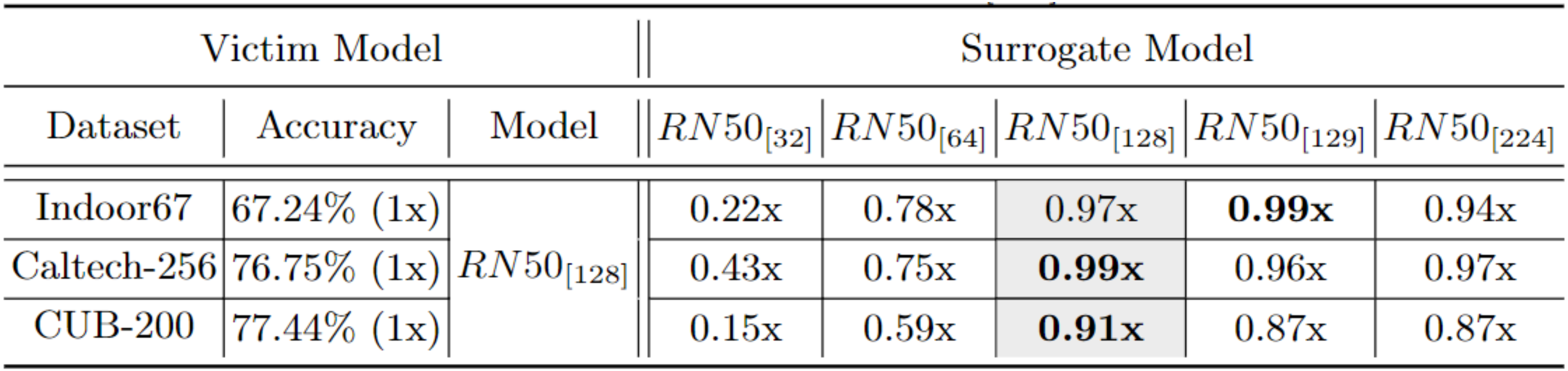
- 1. Image Dimension Estimation



	m		n		k		$kernel$		ID	
Victim Model	SCA	Target	SCA	Target	SCA	Target	SCA	Target	SCA	Target
$RN50_{[128]}$	4118.5	4096	72	64	35.7	27	3.5	3	129.3	128

Experimental Results

- 2. Subsequent Model Extraction



Victim Model			Surrogate Model				
Dataset	Accuracy	Model	$RN50_{[32]}$	$RN50_{[64]}$	$RN50_{[128]}$	$RN50_{[129]}$	$RN50_{[224]}$
Indoor67	67.24% (1x)	$RN50_{[128]}$	0.22x	0.78x	0.97x	**0.99x**	0.94x
Caltech-256	76.75% (1x)		0.43x	0.75x	**0.99x**	0.96x	0.97x
CUB-200	77.44% (1x)		0.15x	0.59x	**0.91x**	0.87x	0.87x

Conclusion

- **Model information is the key** to achieving high MEA performance
- **Image dimension** is the crucial piece of model information
- Model information of the victim can be **extracted via SCA**
- We provide an insight that MEA can be thwarted effectively by **obfuscating the image dimension values** of the model