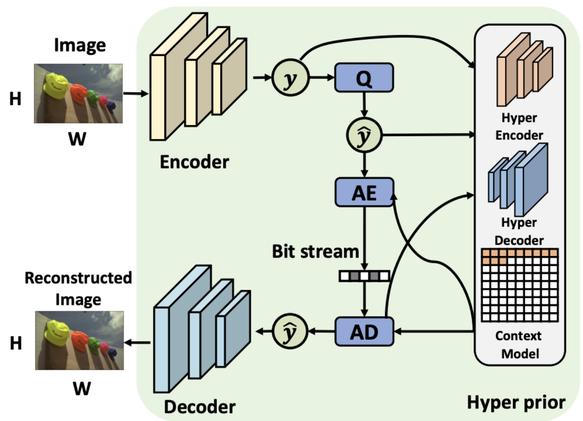
Transferable Learned Image Compression-Resistant Adversarial Perturbations

Yang Sui, Zhuohang Li, Ding Ding, Xiang Pan, Xiaozhong Xu, Shan Liu, Zhenzhong Chen

BMVC 2024

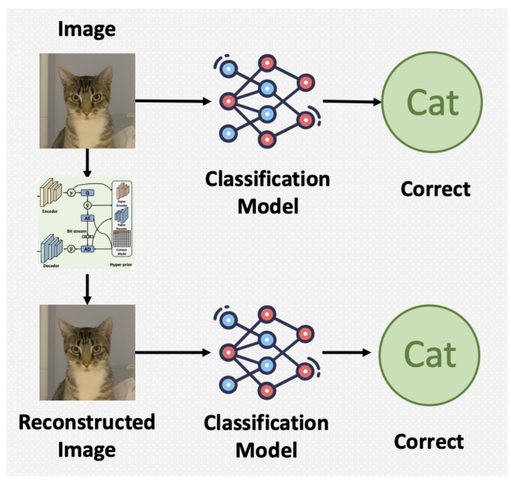
Learned Image Compression

- Compress images into smaller size.



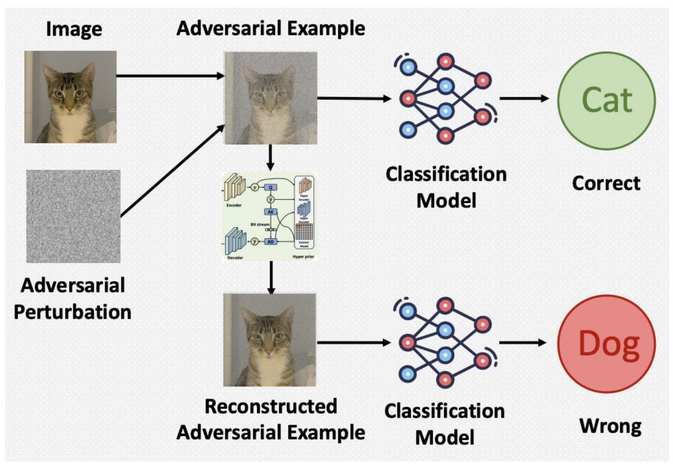
Learned Image Compression Classification System (LICCS)

- Online Classification Service
- Original Image → Compression → Cloud → Classfication



Learned Adversarial Attack on LICCS

- Attacker would upload an adversarial image to mislead the final classification result.



Black-box Attack

- The attack fails to achieve effective transferability across different quality levels.

Table 1: Top-1 accuracy of PGD black-box attack results of `cheng2020` [9] model. Each row/column corresponds to a surrogate/target model with a given quality level.

Quality	1	2	3	4	5	6
	$\epsilon = 4, \quad \alpha = 1, \quad iters = 10$					
1	35.59%	57.54%	68.61%	79.25%	84.50%	87.19%
2	43.71%	37.86%	58.07%	76.14%	83.01%	85.65%
3	45.85%	46.15%	35.82%	69.49%	79.02%	83.00%
4	48.72%	56.92%	58.16%	37.28%	53.77%	66.30%
5	49.22%	59.32%	62.38%	47.77%	34.83%	41.01%
6	49.55%	60.20%	65.07%	57.97%	39.10%	32.53%

- Observation: the adjacent quality levels tend to experience more substantial impact, whereas the impact diminishes for quality levels that are more distant.

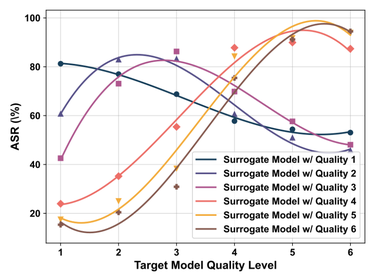


Figure 2: ASR of PGD black-box attack results on quality level 1 to 6 of `cheng2020` [9] model with $\epsilon = 16$, $\alpha = 2$, $iters = 20$.

Method

- Improve the Transferability of Attack.

$$S(\boldsymbol{q}) = \int_{x_{min}}^{x_{max}} \max(\texttt{polyn}(q_0), \texttt{polyn}(q_1), \cdots, \texttt{polyn}(q_K)) \; dx$$

- Select optimal combination of quality levels as surrogate models to attack.

Result

Table 2: Top-1 accuracy of LICCS with the surrogate model `cheng2020` and target model `cheng2020` attacked by PGD. Lower accuracy demonstrates higher transferability.

Quality	1	2	3	4	5	6	Average	Time
			$\epsilon = 4, \quad \alpha = 1, \quad iters = 10$					
R-En	44.32%	51.10%	54.95%	56.16%	55.10%	58.02%	53.28%	1.1s
Ours	47.86%	54.46%	51.41%	47.36%	39.26%	40.53%	**46.81%**	1.1s
			$\epsilon = 8, \quad \alpha = 2, \quad iters = 10$					
R-En	39.83%	43.34%	45.42%	43.20%	43.23%	46.59%	43.60%	1.1s
Ours	45.52%	47.25%	42.14%	31.26%	25.01%	25.33%	**36.09%**	1.1s
			$\epsilon = 16, \quad \alpha = 2, \quad iters = 10$					
R-En	35.68%	37.15%	37.87%	35.76%	37.86%	40.92%	37.54%	1.1s
Ours	43.31%	41.18%	34.46%	22.64%	20.24%	20.39%	**30.37%**	1.1s

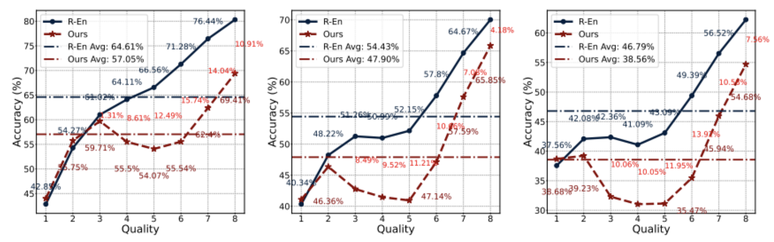


Figure 3: Top-1 accuracy of LICCS with the surrogate model `cheng2020` and target model `hyper` attacked by PGD. Lower accuracy demonstrates higher transferability.