2025.03.10

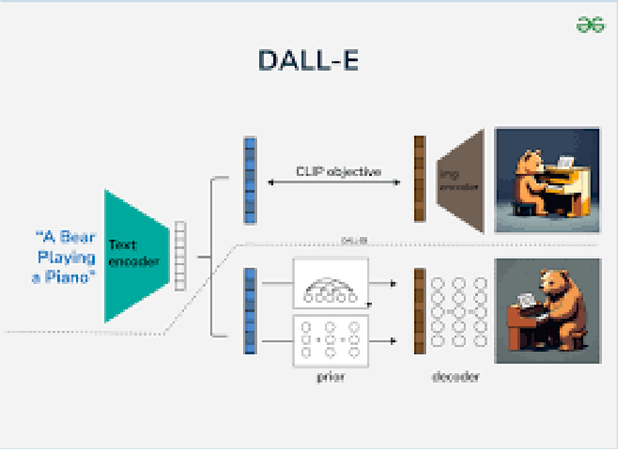
WAVES: Benchmarking the Robustness of Image Watermarks

국민대학교 이재형
jaehyeong8121@gmail.com

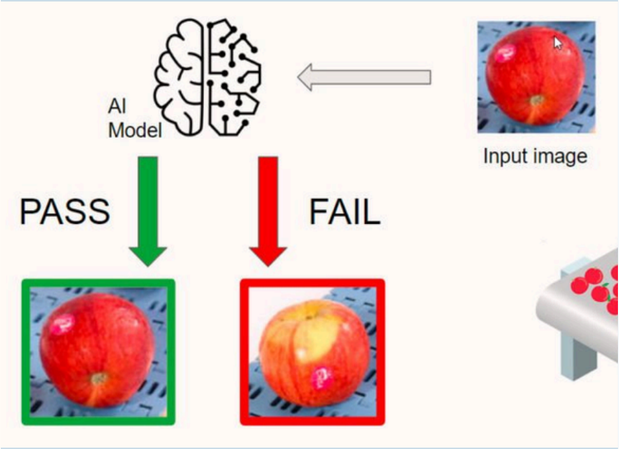
Contents

Why this research matters?

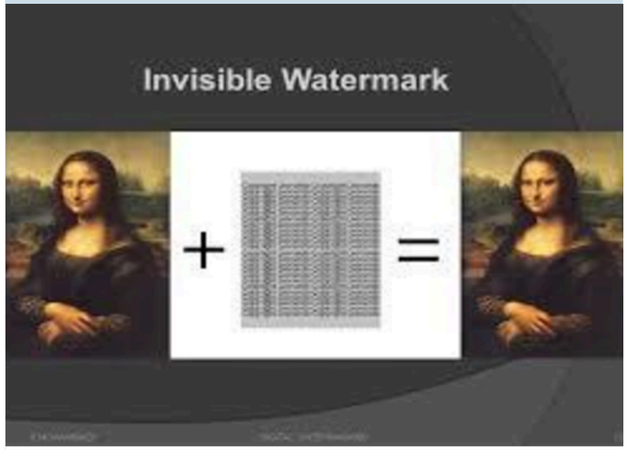
AI 생성 콘텐츠의 증가



AI 생성 이미지 탐지



워터마킹 기법의 중요성

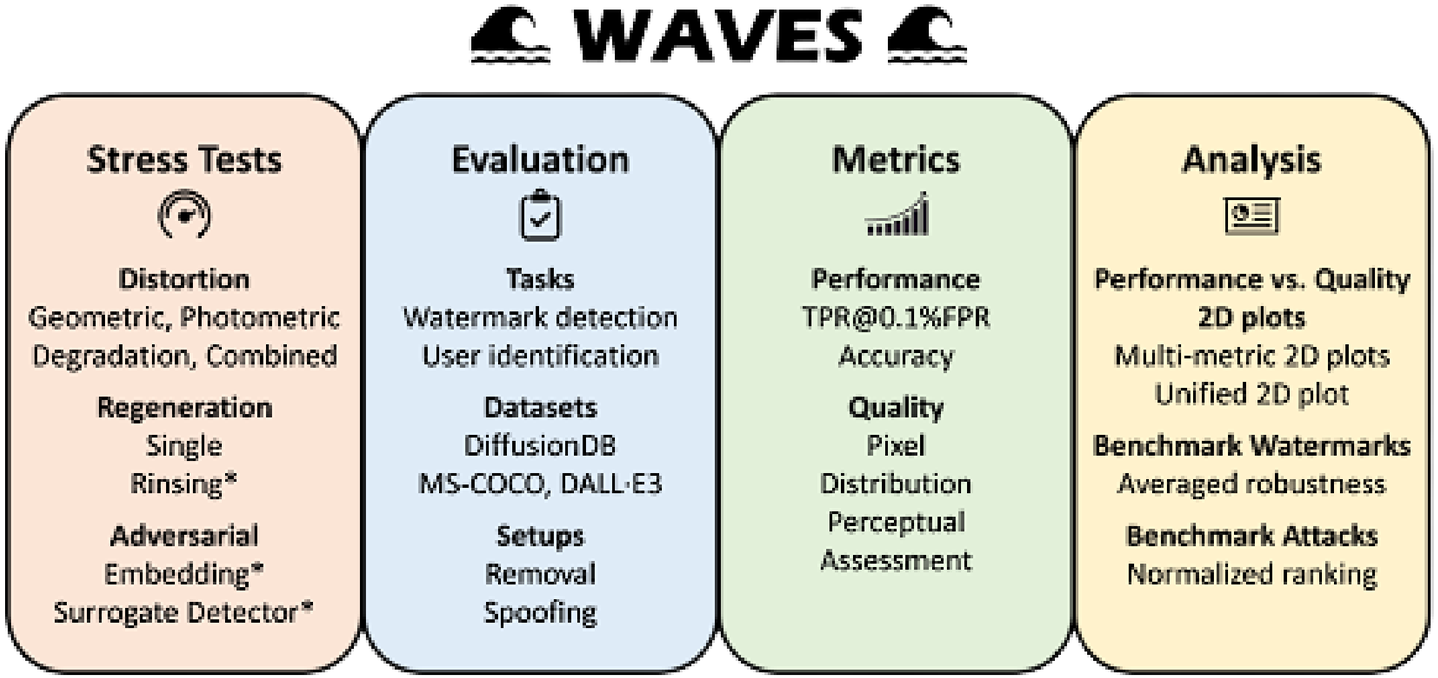


01 Why this research matters?

기존 워터마킹 기법의
Robustness 문제

표준화된
평가 기준 부재

실세계 적용에 대한
검증 부족



What did i learn from this paper?

WAVES의 차별성이 무엇인가?

Research Work	Num. of Attacks	Categories of Attacks	Num. of Datasets	Sample Size per Dataset	Non-watermarked Image Source	Performance Metric	Num. of Quality Metrics	Joint Test
StegaStamp Watermark[1]	5	D	1	1000	—	bit accuracy	3	✗
Stable Signature Watermark[2]	12	D, R	1	5000	—	bit accuracy	3	✗
TreeRing Watermark[3]	6	D	?	1000	generate by same model	TPR@1%FPR	?	✗
Regeneration Attack[4]	10	D, R	2	500	—	bit accuracy	3	✗
Surrogate Model Attack[5]	2	R, A	1	2500	real images	AUROC	0	✗
Adaptive Attack[6]	10	D, A	1	1000	real images	TPR@1%FPR	3	✗
WAVES (ours)	26	D, R, A	3	5000	real images	TPR@0.1%FPR	8	✓

- Dataset 3개
- 3 종류 카테고리의 26가지 다양한 공격
- 8개의 평가 지표

Extensive하고

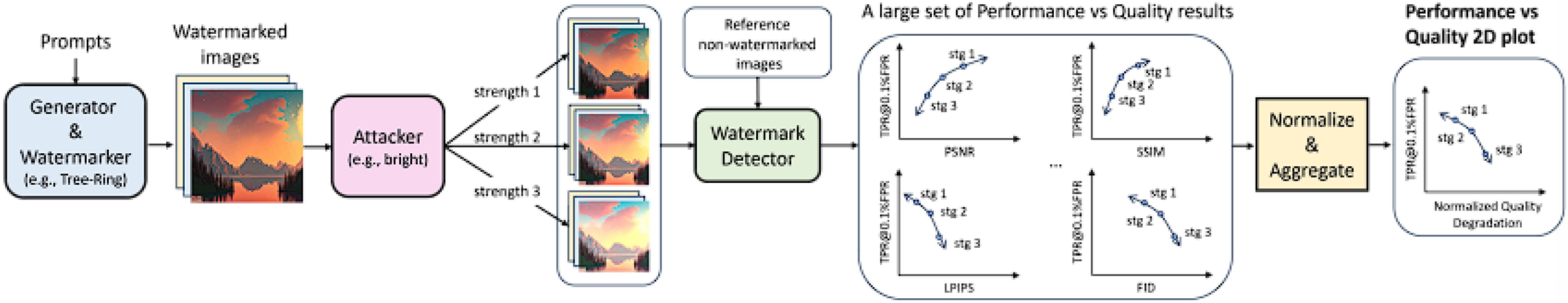
Realistic한 워터마크 강건성 평가 방법

Evaluation Workflow

워터마크 있는 이미지 vs 실제 및 AI 생성 참조 이미지와 비교

Performance vs. quality 2D plot 생성

워터마크 성능과 이미지 품질의 균형은
워터마크 공격으로 인한 이미지 왜곡이
발생할 때 핵심적인 고려 사항!

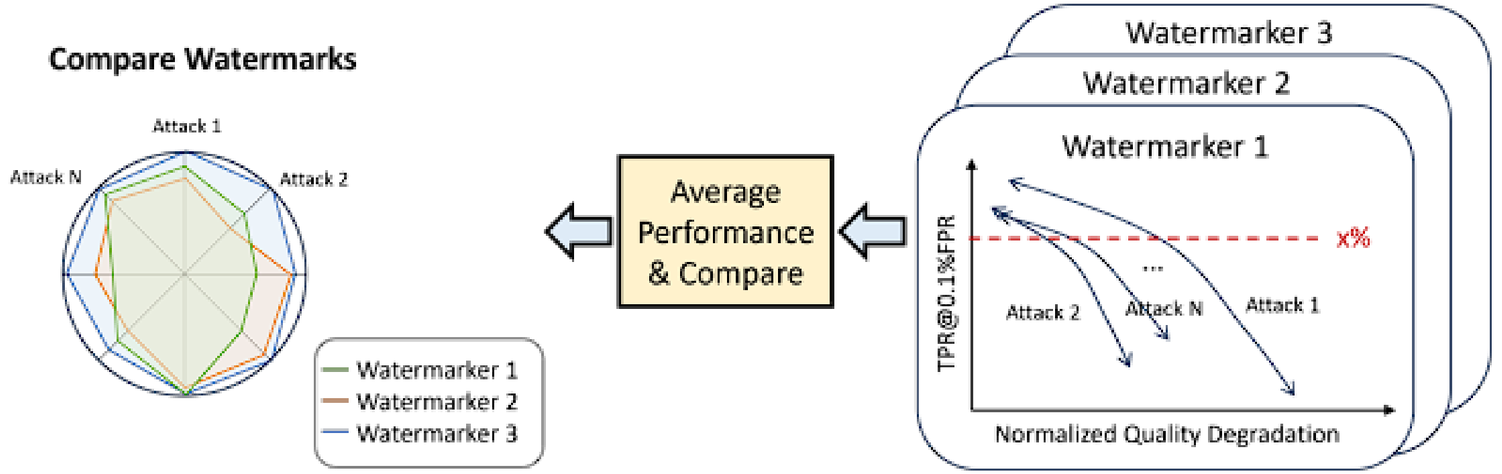


02 What did i learn from this paper?

Robustness

다양한 공격 기법에 대한

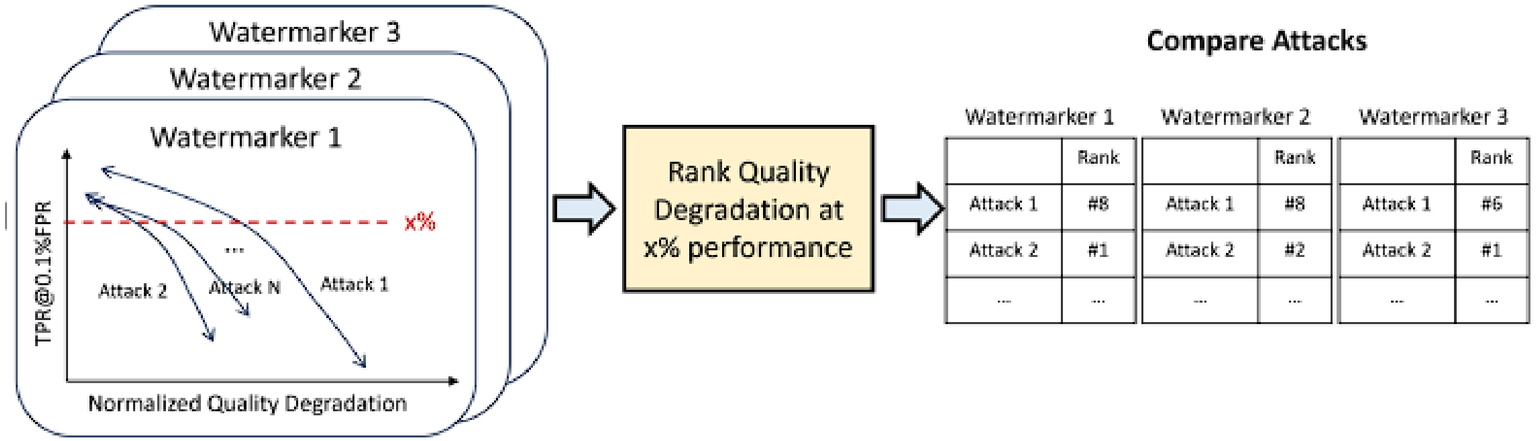
워터마크의 평균 성능 비교



Attack’s potency

각 공격 기법이 얼마나 효과적인지

순위를 매김



Performance Metrics

일반적인 AI 탐지 문제에서 발생하는

FPR과 TPR의 trade-off 관계

FPR이 높으면 워터마크 탐지

시스템의 신뢰성이 매우 낮아짐

FPR이 0.1% 이하일 때, 모델이 여

전히 높은 TPR을 유지할 수 있는지

확인함으로써 **신뢰성**과 **탐지 성능**의

균형을 맞춤

TPR @ 0.1% FPR

$$P(f(x) \geq \tau | x \in \text{negative class}) = 0.001$$

ROC 곡선에서 FPR이 0.001(0.1%)가 되는 임계값 τ를 찾는다

$$TPR@0.1\%FPR = P(f(x) \geq \tau | x \in \text{positive class})$$

τ를 이용하여 TPR을 계산한다

Stress-tesing watermarks

워터마크의 강건성을 넓은 범위의 다양한 공격을 통해 평가한다

Distortion

허용 가능한 품질 임계값 내에서
아래의 왜곡을 기준선으로 설정

- Geometric: rotation, resizedcrop
- Photometric: brightness, contrast
- Degradation: Gaussian blur, noise

Regeneration

Diffusion Model, VAE를 사용하여 이미지에 노이즈를 처리한 다음 노이즈를 제거하여 이미지의 잠재 표현을 변경

- Rinsing regeneration: 사전 훈련된 모델을 통해 이미지가 여러 차례 노이즈 제거 주기를 거침

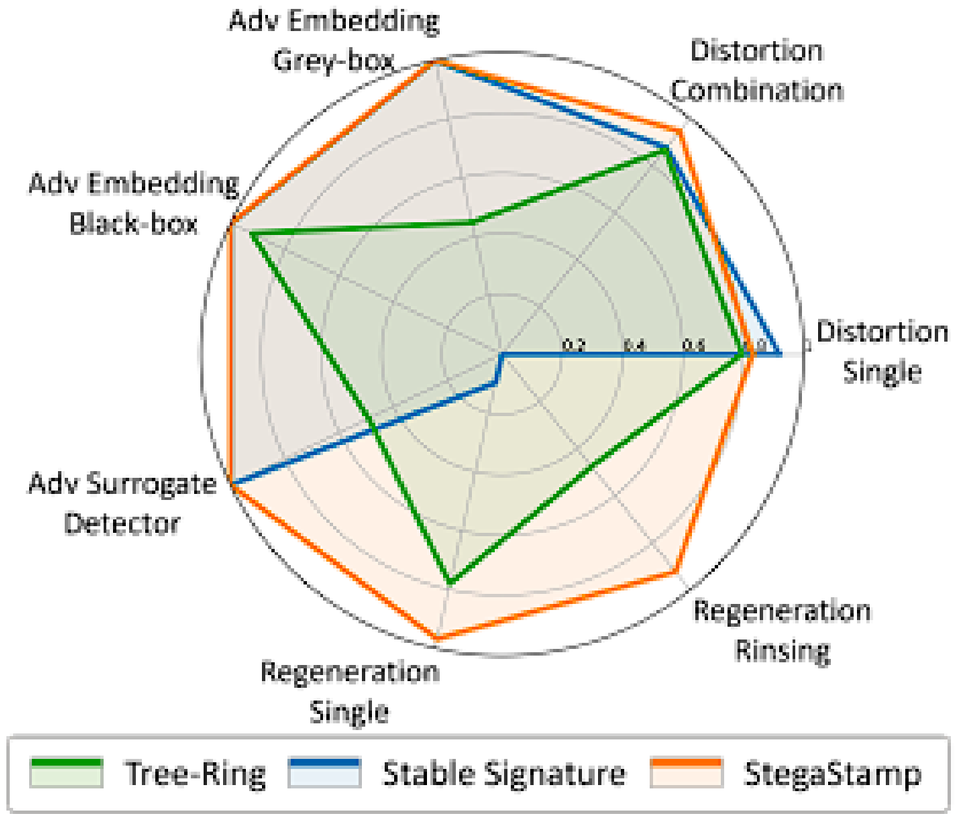
Adversarial

DNN이 적대적 공격에 취약함을 이용하여
2가지 적대적 공격을 수행

- Embedding attack
- Surrogate Detector attack

Key Findings & Implications

3가지 대표적인 워터마킹 기법 분석

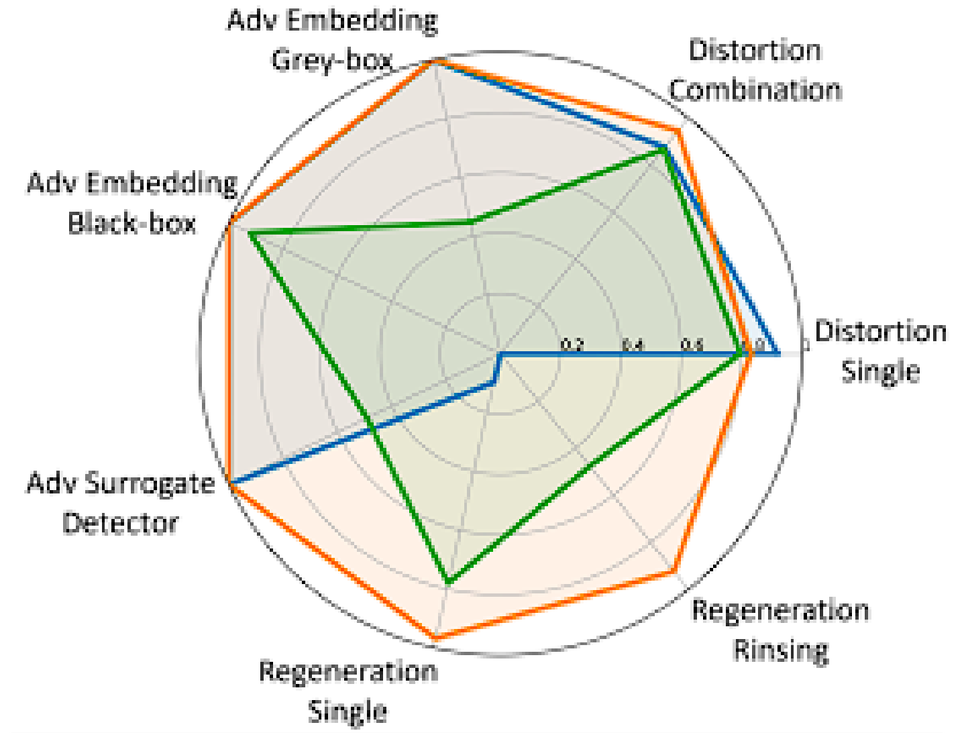


(a) Average TPR@0.1%FPR under different types of attacks.

- **Distortion 공격**
  - 3가지 워터마크 모두 높은 탐지 성능 유지
  - 즉, 일반적인 이미지 변형에는 강건함

- **Regeneration 공격**
  - 모든 워터마크의 탐지 성능이 상대적으로 낮아짐
  - 특히 **Tree-Ring**, **Stable Signature**는 탐지 성능이 크게 감소
  - 이는 Regeneration 기반 공격이 워터마크를 효과적으로 제거할 수 있음을 의미

03 Key Findings & Implications

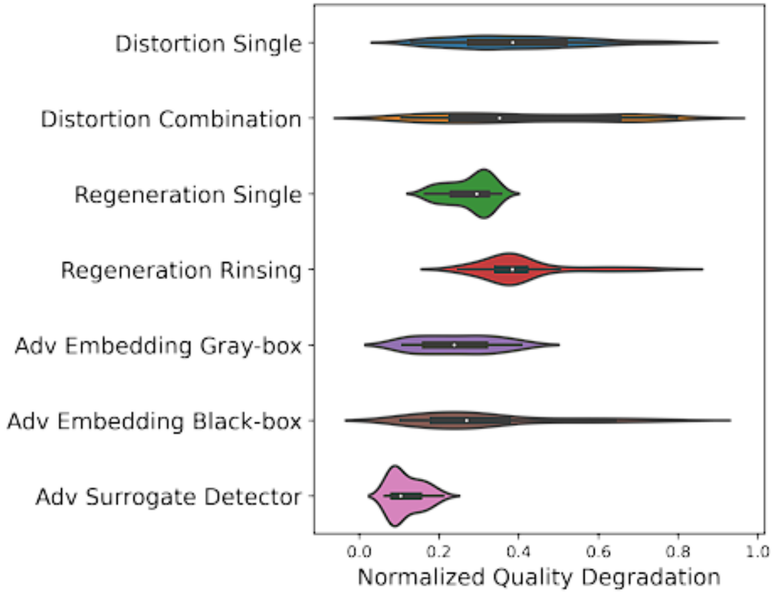


(a) Average TPR@0.1%FPR under different types of attacks.

- **Adversarial Embedding 공격**
  - StegaStamp는 높은 탐지 성능을 유지, 즉 강건함
  - 하지만 Tree-Ring과 Stable Signature는 상대적으로 더 취약함

- **Adv Surrogate Detector 공격**
  - Tree-Ring이 가장 낮은 탐지 성능을 보임
    즉, 대리 탐지기 공격에 매우 취약함
  - Stable Signature와 StegaStamp은 상대적으로 더 강건함

3가지 공격 유형의 워터마킹 이미지 품질에 미치는 영향



X축 : 정규화된 품질 저하 정도

-> 값이 클수록 이미지 품질이 더 많이 손상됨

Y축 : 공격 유형(왜곡, 재생성, 적대적)

-> 각 공격 기법이 워터마킹된 이미지의 품질에 미치는 영향을 보여줌

그래프 형태 :

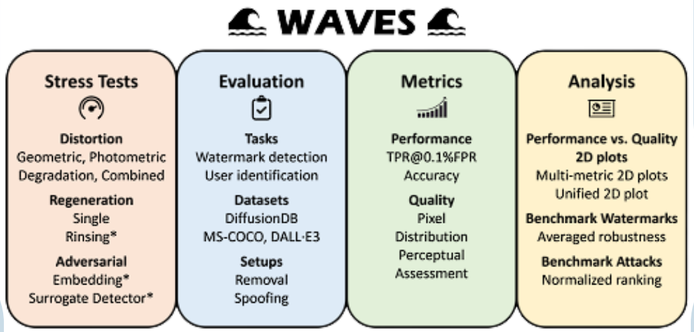
- Y축으로 폭이 넓으면 공격의 품질 효과가 다양
- X축으로 오른쪽에 위치하면 이미지 품질 저하가 큼

- **Adversarial Attacks**는 품질 저하 없이도 워터마크 제거 가능
  -> 이미지가 거의 변형되지 않아 매우 강력한 위협이 될 수 있음
- **Regeneration Attacks**는 공격의 품질 효과는 다양하지만, 적대적 공격보다는 이미지 품질 저하가 큼
- **Distortion Attacks**는 상대적으로 낮은 영향

03 Key Findings & Implications

워터마킹 기법의 보안성을 평가하는 객관적이고 표준화된 프레임워크 제시

WAVES's 의의



기존 워터마킹 기법들의 취약점을 실험적으로 검증

Surrogate Detector Attack의 효과를 검증하여 기존 연구에는 없던 새로운 위협을 규명

워터마킹 보안을 위한 향후 연구 방향 제시

Future Directions

1 적대적 공격에 강인한 워터마크 기법 개발

2 AI 자체가 워터마킹을 감지하고, 이미지에
자동으로 워터마크를 추가하는 방식의 연구 필요성

3 실제 온라인 환경에서도 WAVES 벤치마크가
적용될 수 있는지 실용성 검증의 필요성

감사합니다:)