

2025.03.31

Deep Learning-Based Image Steganography : From Recent Advances to RoSteALS

국민대학교 이재형
jaehyeong8121@gmail.com

Contents

01

Latest Trends in image Steganography

02

What is RoSteALS?

03

Results & Limitations

04

Future Directions

Latest Trends in image Steganography

“A survey on Deep-Learning-based image steganography”

위 논문은
딥러닝 기반 이미지 스테가노그래피의 최근 발전과 트렌드를 소개하고



Cover-edited 및 Coverless 스테가노그래피 기법을 포함한
다양한 네트워크 구조와 그 응용을 설명

01 Latest Trends in image Steganography

1.1 What is **Steganography** & Why is it important?

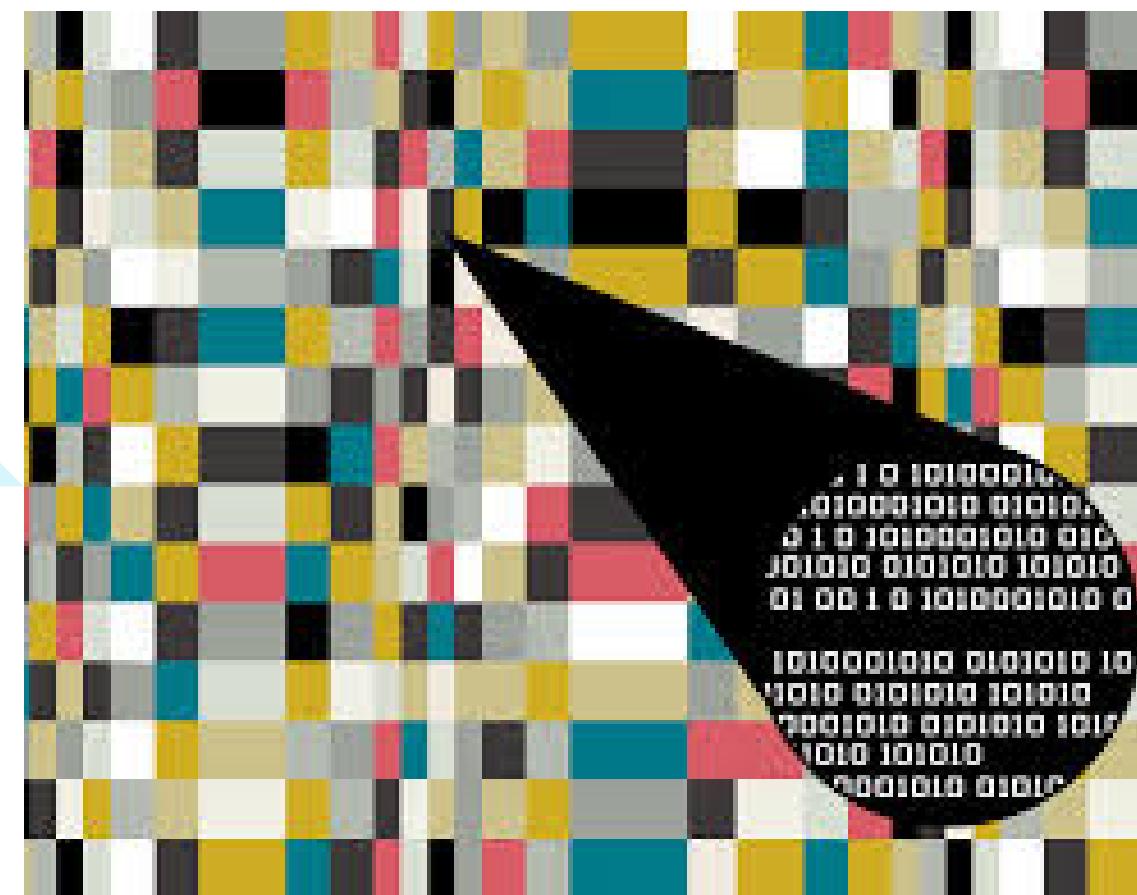
생성형 AI의 발전으로 이미지 형식의 많은 데이터들이 전송되고 있음

Security Transmission의 중요성 부각



Image Steganography

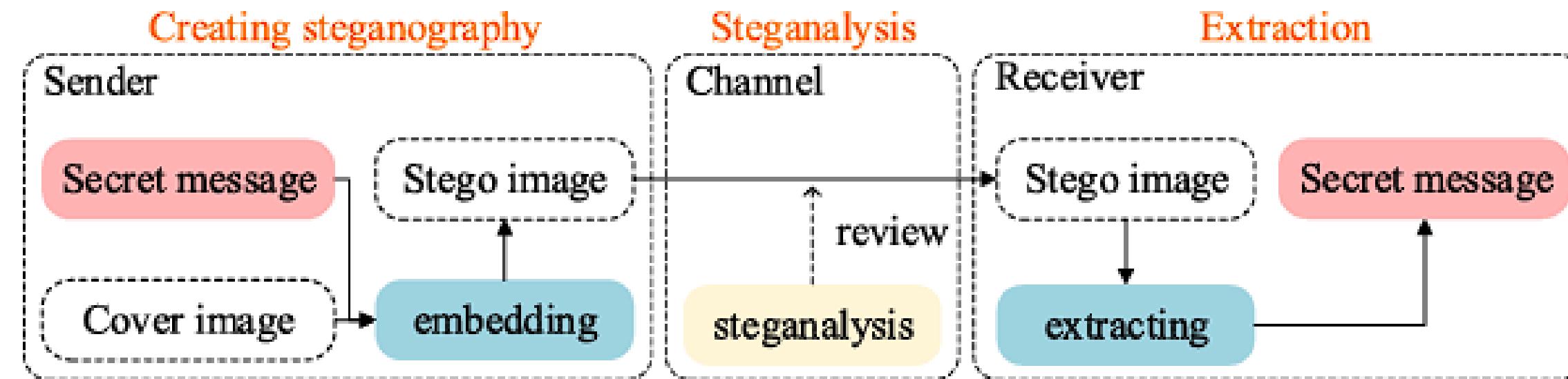
이미지에
비밀 정보를 숨김



최근 Deep Learning을
적용하여 응용성과
적응성을 향상시킴

01 Latest Trends in image Steganography

1.2 How Steganography Works?



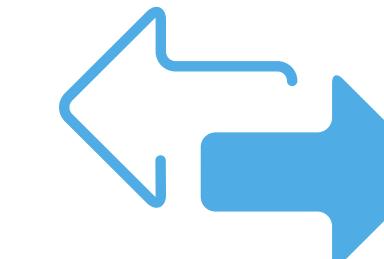
Steganography 생성

커버 이미지에 **비밀 메시지를** 임베드



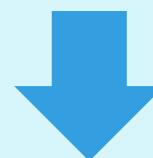
커버 이미지와 시각적으로 아주 비슷한
Stego 이미지 생성됨

어떻게 Steganography를
생성했는지에 따라
추출에 큰 영향을 줌



Secret message 추출

Steganalysis의 분석 과정을 거쳐
수신자에게 전달된 Stego 이미지



삽입되었던 **비밀 메시지를 추출**함

01 Latest Trends in image Steganography

1.3 Creating Steganography

- Strategy와 Network Structure의 2가지 관점에서 스테가노그래피 생성을 소개한다

Strategy

커버 이미지 수정 여부

1. Cover-edited Steganography
2. Coverless Steganography

Network Structure

네트워크 구조에 따른 분류

1. Traditional Embedding
2. CNN
3. GAN
4. Invertible Network

01 Latest Trends in image Steganography

1.3.1 Creating **Steganography_Strategy**

1. Cover-edited Steganography

$$I_{st} = f_{edit}(I_{co}, I_{se}),$$

$$\text{where } I_{co} - I_{st} \leq \zeta_1,$$

[기호 설명]

I_{st} : stego 이미지

I_{se} : 비밀 메시지

I_{co} : cover 이미지

- f_{edit} 을 통해 cover 이미지를 눈에 띠지 않게 수정함
- ζ_1 는 하이퍼파라미터 : 값이 작을수록 stego 이미지가 cover 이미지와 비슷함
- cover 이미지의 적절한 곳에 비밀 메시지를 숨기는 것을 고려하면,
 - Spatial domain 스테가노그래피
 - Transform-based 스테가노그래피, 이 2가지로 나눌 수 있음

01 Latest Trends in image Steganography

1.3.1 Creating **Steganography_Strategy**

1-1. **Spatial domain Steganography**

$$I_{st} = Hide(Spatial(I_{co}), I_{se}).$$

- 이미지의 픽셀 값만 직접 수정하여 비밀 메시지를 cover 이미지에 그대로 삽입한다
- Spatial Strategy 종류
 - rule-based : 커버 이미지의 속성을 고려 x, 규칙에 따라 비밀 메시지를 삽입
 - self learning : 인위적인 사전 지식 없이 딥러닝을 통해 커버 이미지에 삽입될 적절한 위치를 학습
 - cost-based : 다양한 위치의 임베딩 비용을 계산하여 위치를 찾음
 - Texture : 복잡한 texture 영역이 비밀 메시지를 숨기기에 적합하다고 믿음
 - Edge-based : 이미지 픽셀과의 상관 관계를 고려
 - Adversarial : 적대적 공격을 통해 steganalysis를 속여 탐지를 줄임

01 Latest Trends in image Steganography

1.3.1 Creating **Steganography_Strategy**

1-2. **Transform-based Steganography**

$$I_{st} = \text{Hide}(\text{Transf}_m(I_{co}), I_{se}),$$

where $\text{Transf}_m(I_{co}) \in \text{Transf}_n(I_{co})$.

[기호 설명]

- $\text{Transf}_n(I_{co})$: cover 이미지를 변환한 후 나온 모든 변환된 결과물
- $\text{Transf}_m(I_{co})$: cover 이미지를 변환한 결과물 중에 비밀 메시지를 숨기기에 적합하게 변형된 컴포넌트
- **cover 이미지에 도메인 변형을 수행**하고, 변형된 도메인에 비밀 메시지를 숨김
- Transform Strategy 종류
 - Frequency domain : 육안으로 찾기 힘든 고주파수 성분에 비밀 메시지를 임베드
 - Matrix : 행렬과 텐서 분해를 통한 임베드

01 Latest Trends in image Steganography

1.3.1 Creating **Steganography_Strategy**

2. **Coverless Steganography**

- Cover-edited Steganography는 **cover 이미지의 변형이 stego 이미지에 약간의 왜곡을 일으킬 수 있음**
 - 특히 컬러 이미지를 임베딩할 때 steganalysis에 쉽게 감지된다
- Coverless는 cover 이미지 없이 비밀 메시지를 삽입하는 방식이다
 - 이는 cover 이미지의 수정 없이 비밀 메시지를 숨기므로, steganalysis가 이미지를 감지하기 어렵다
- Coverless Steganography는 아래와 같이 크게 2가지 전략으로 나눌 수 있음

Mapping

$$I_{co} = Map(I_{se}).$$

비밀 메시지와 cover 이미지 간의
매핑 관계를 설정하여 비밀 정보를 숨김

Generating

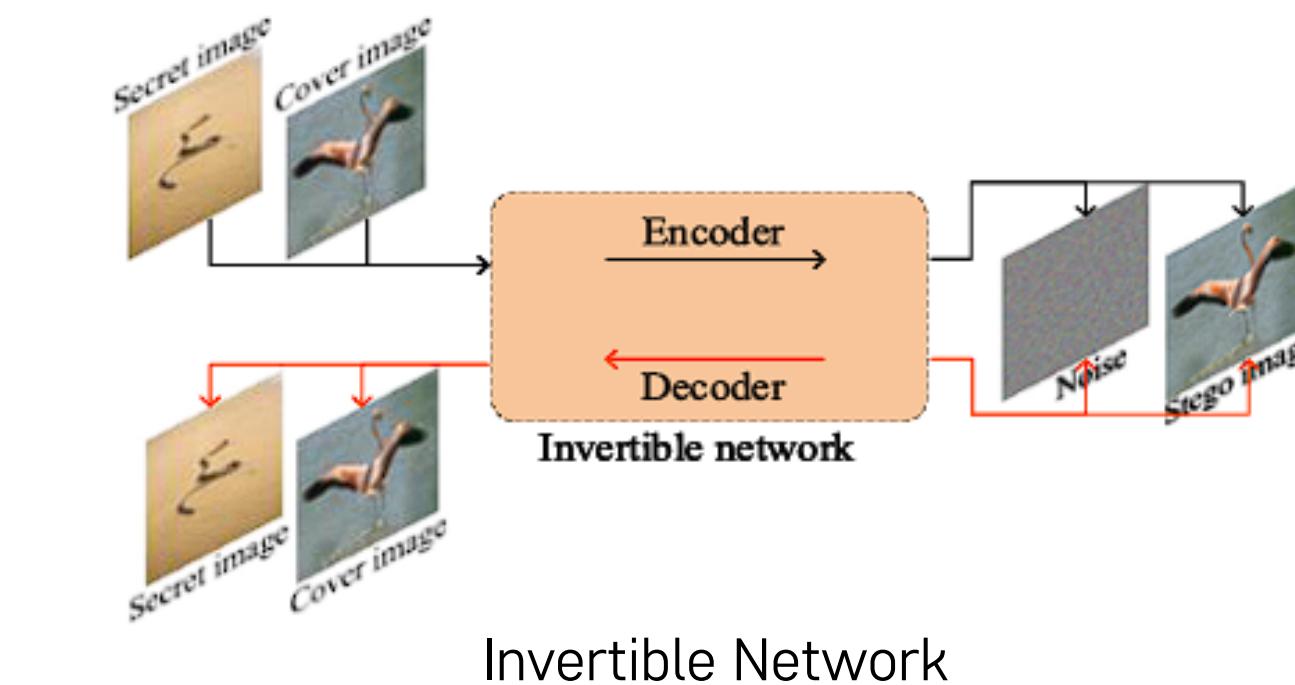
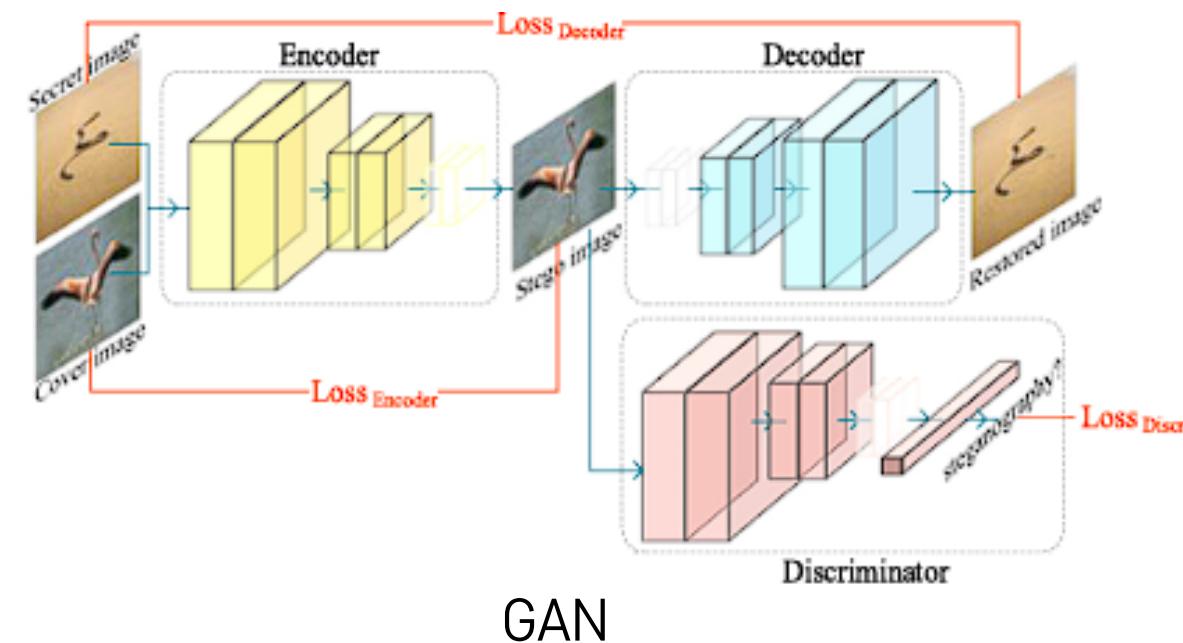
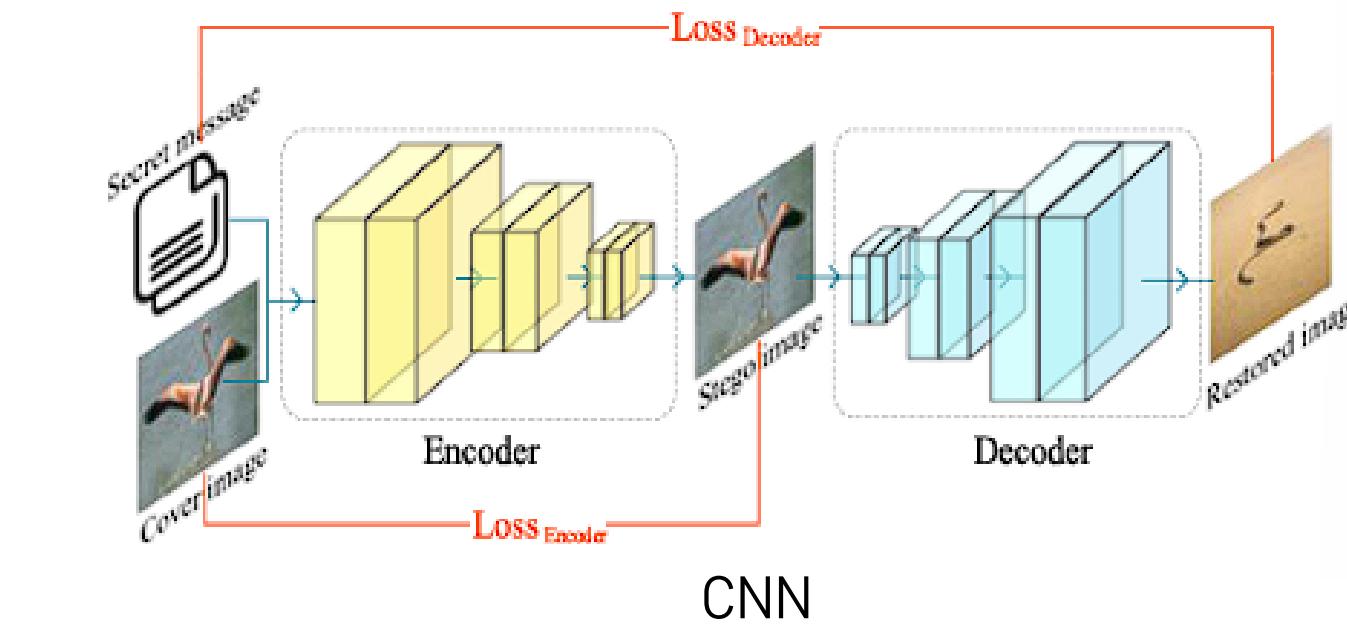
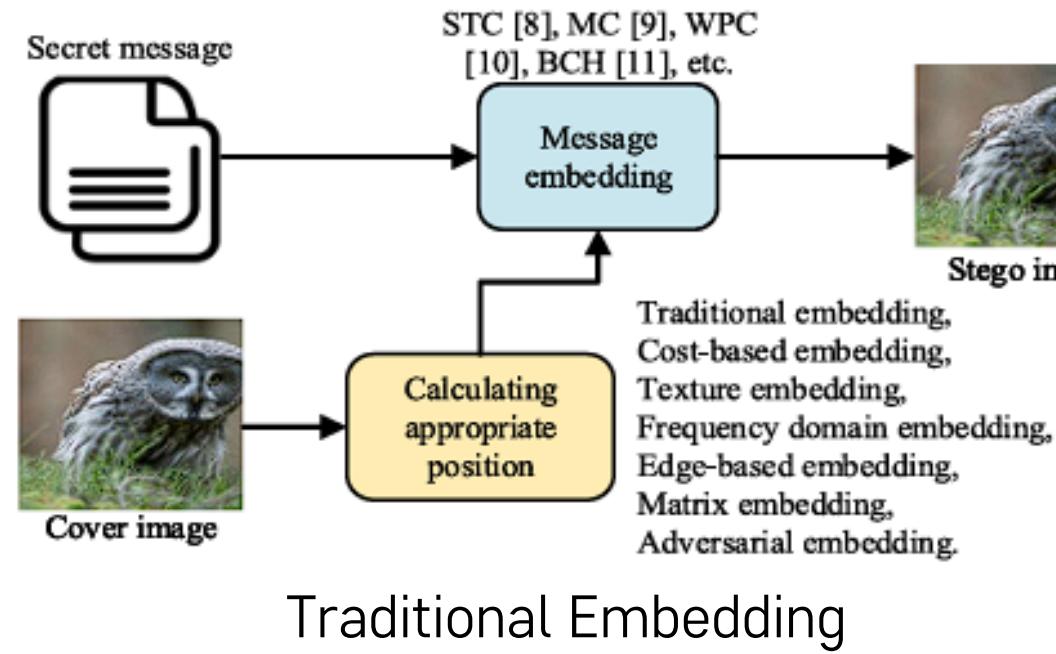
$$I_{st} = Generate(I_{se}).$$

비밀 메시지를 사용하여
새로운 이미지를 생성함

01 Latest Trends in image Steganography

1.3.2 Creating Steganography_Network Structures

- **네트워크 구조** 또한 스테가노그래피를 생성하는 데 있어 중요한 영향을 미친다
 - 단순히 비밀 메시지를 이미지에 숨기는 전략 뿐만 아니라, 이미지 품질, 탐지 회피 능력에도 영향을 미침



01 Latest Trends in image Steganography

1.4 Extraction

$$I_{res} = \text{Extraction}(I_{st}),$$

$$\text{where } I_{se} - I_{res} \leq \zeta_2.$$

- 수신자는 stego 이미지로부터 비밀 메시지를 추출한다
- Extraction 과정
 - I_{res} 는 추출된 비밀 메시지를 의미
 - ζ_2 는 추출된 비밀 메시지와 원래 메시지 사이의 차이를 측정하는 지표, 가능한 한 작게 유지되어야 함
- 전통적인 임베딩과 Extraction
 - 기존의 임베딩 규칙을 사용하여 비밀 메시지를 숨긴 경우, 임베딩 규칙을 사용하여 메시지를 복원함
 - CNN, GAN 기반의 스테가노그래피에서는 자동으로 비밀 메시지를 추출하는 방법을 학습함

What is RoSteALS?

“RoSteALS : Robust Steganography using Auntoencoder Latent Space”

위 논문은

RoSteALS: 사전 훈련된 오토인코더의 잠재 공간을 활용한
효율적이고 강건한 스테가노그래피 기법을 소개한다

02 What is RoSteALS?

2.1 3 Features of RoSteALS

1

잠재 공간 스테가노그래피

- 비밀 메시지를 **잠재 코드에 직접 삽입**하여 강건한 워터마킹을 구현함
- **사전 훈련된 오토인코더**를 통해 적은 훈련과, 사전 지식 없이 효율적 학습을 통한 일반화 능력이 뛰어남

2

강건한 비밀 복원

- 이미지에 강한 변형이 가해져도 비밀 메시지를 강건하게 복원 가능
- 온라인 콘텐츠 재배포 시 식별자 유지에 적합함

3

Coverless 스테가노그래피

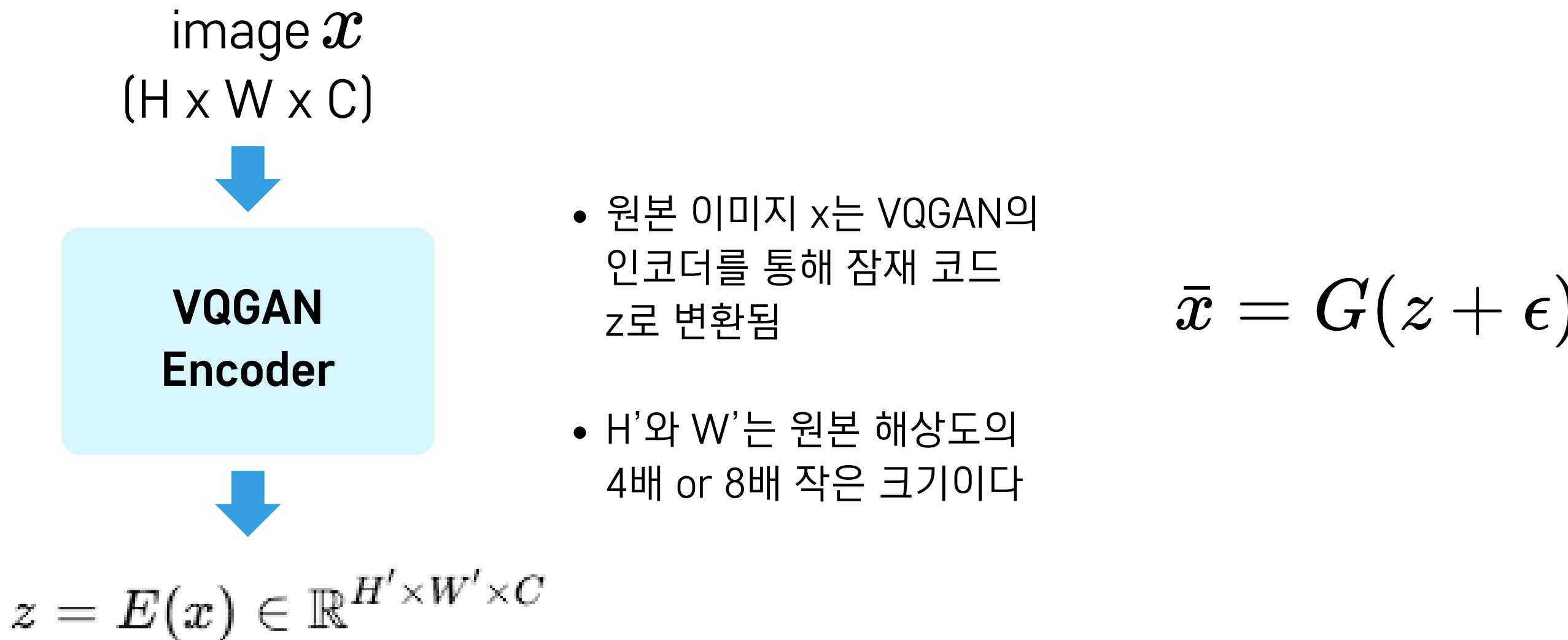
- **cover 이미지를 사용하지 않고**, 텍스트나 이미지에서 생성된 무작위 잠재 코드에 비밀 메시지 삽입을 통해 stego 이미지 생성 가능

02 What is RoSteALS?

2.2 Methodology of RoSteALS

1. Leveraging pretrained AutoEncoders

- VQGAN : 이미지를 잠재 코드로 변환하는 최신 오토인코더
- VQGAN의 잠재 공간을 활용한 스테가노그래피의 가능성을 실험

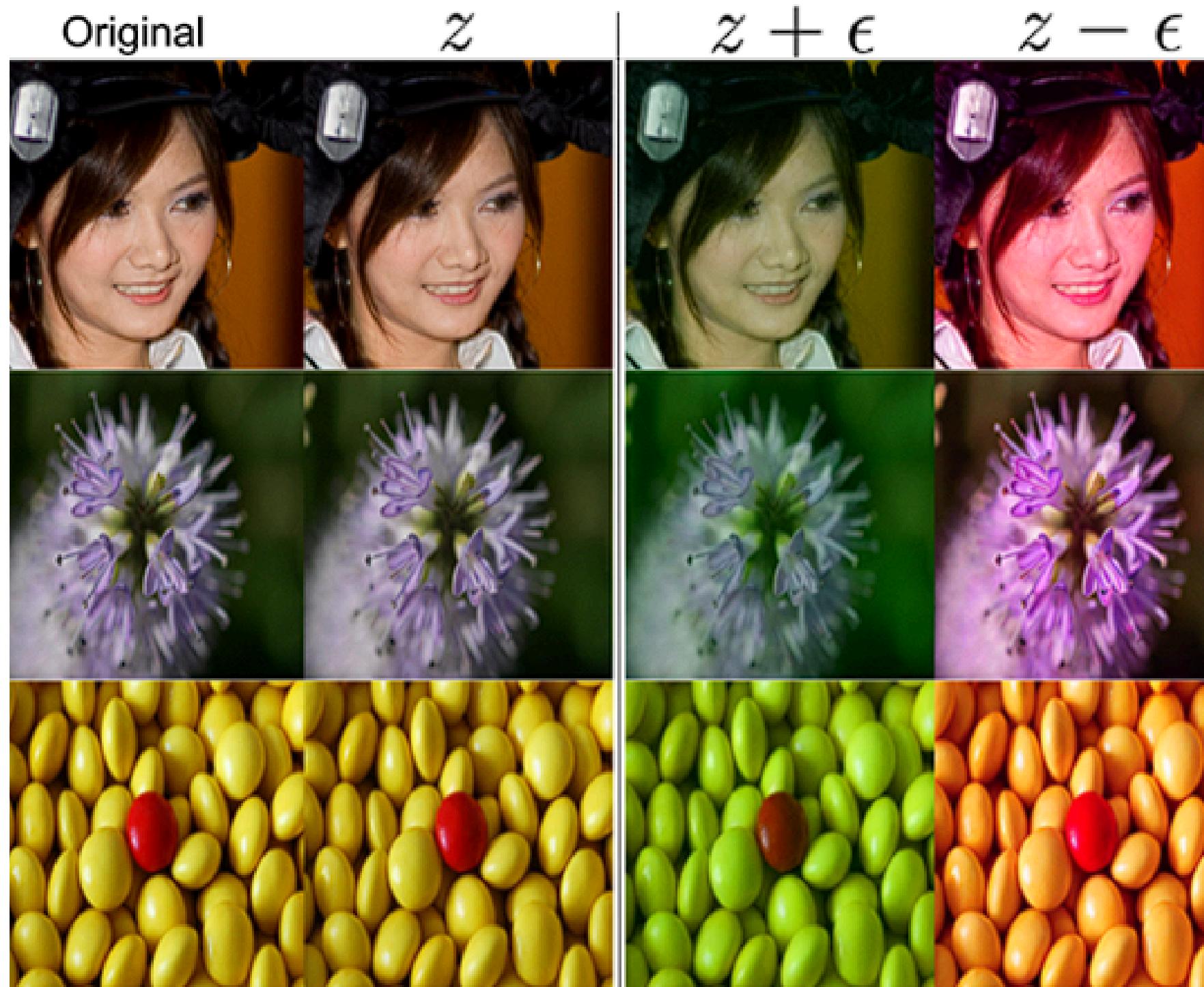


- 잠재 코드에 노이즈를 추가한 후, VQGAN의 생성기 G 를 사용하여 이미지를 재구성한다
- ϵ 은 잠재 코드 z 에 추가된 노이즈이다
- 재구성된 이미지는 원본 이미지 x 와 비슷하게 나오도록 한다

02 What is RoSteALS?

2.2 Methodology of RoSteALS

1. Leveraging pretrained AutoEncoders



1. 특정 노이즈는 재구성된 이미지의 내용과 관계없이 동일한 지각적 변화를 일으킨다

→ 출력 이미지가 제공된 경우, 노이즈를 복구할 수 있는지에 대한 흥미로운 시사점 제공

2. VQGAN의 임베딩 공간은 작은 노이즈에 민감하지 않다

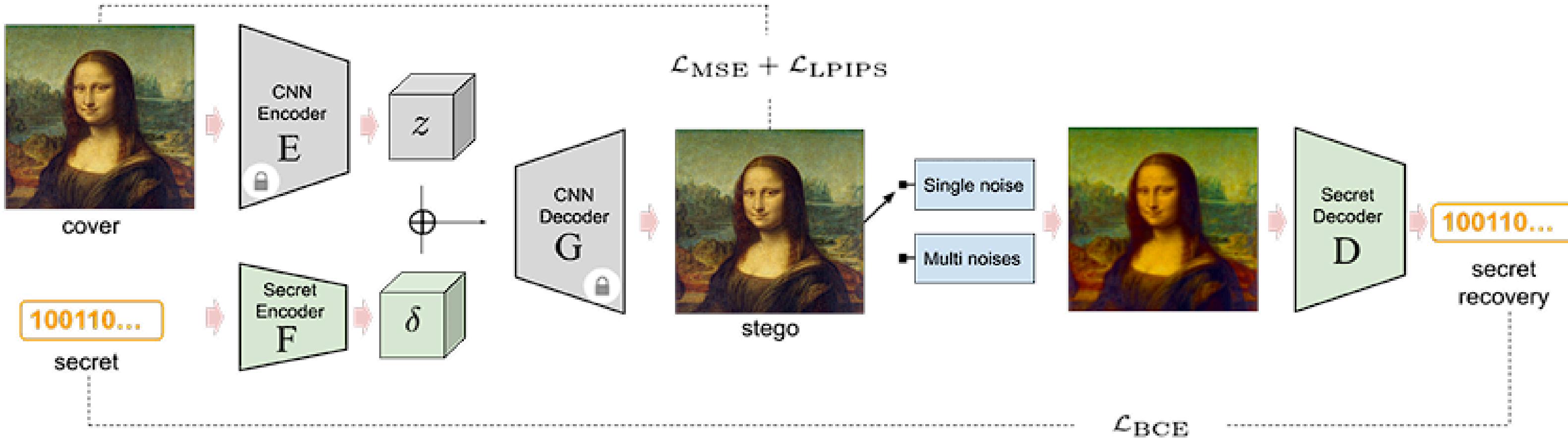
$$G(z + \epsilon) \approx x \\ (\text{when } |\epsilon| < \epsilon_0)$$

→ 이미지 내용을 눈에 띠게 변경하지 않고도, 비밀 메시지를 잠재공간에 직접 주입할 수 있는 가능성 제공

02 What is RoSteALS?

2.2 Methodology of RoSteALS

2. Steganography with RoSteALS



- 이미지 분포에 대한 **사전 지식이 있는 동결된 오토인코더** $\{E, G\}$ 를 통해 비밀 메시지 $s \in \{0, 1\}^L$ 을 이미지의 잠재 공간에 매핑하는 **Secret Encoder F**를 학습하는 것이 목표
- 학습 하는 동안 동결된 인코더 **E**와 **G**를 제외하고 가벼운 네트워크인 **Secret Encoder F**와 **Decoder D**를 학습함

[기호 설명]

- δ : 비밀 메시지를 삽입하기 위해 cover 이미지의 잠재 공간에 추가되는 offset, z 와 같은 차원으로 변환됨

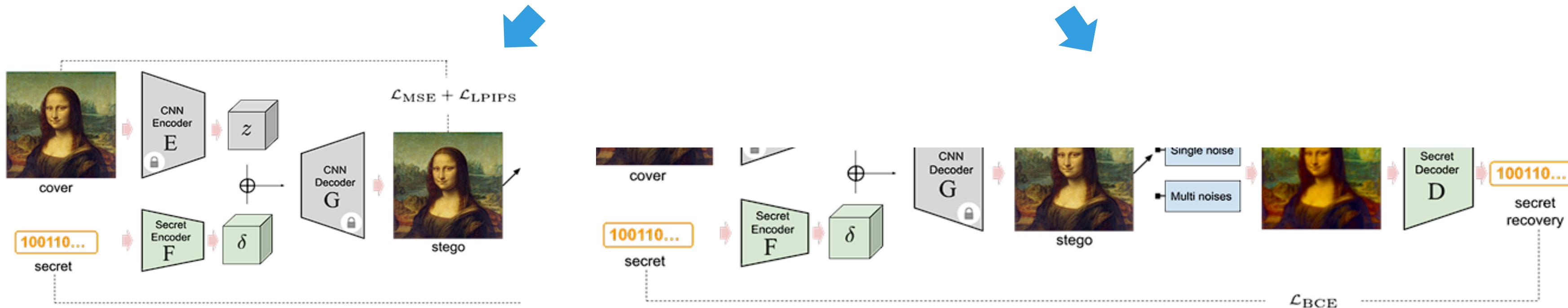
02 What is RoSteALS?

2.2 Methodology of RoSteALS

2. Steganography with RoSteALS

β 는 stego 이미지의 품질과
비밀 메시지 복원 간의
trade-off를 조절함

$$\mathcal{L} = \beta \underline{\mathcal{L}_{\text{quality}}} + \underline{\mathcal{L}_{\text{recovery}}}$$



02 What is RoSteALS?

2.2 Methodology of RoSteALS

2. Steganography with RoSteALS

α 는 이미지 품질 개선을 위한
시각적 유사성과 픽셀 단위
정확성 사이의 균형을 맞춤

$$\mathcal{L}_{\text{quality}} = \mathcal{L}_{LPIPS}(\tilde{\mathbf{x}}, \mathbf{x}) + \alpha \mathcal{L}_{MSE}$$



$$\mathcal{L}_{LPIPS}(\tilde{\mathbf{x}}, \mathbf{x})$$

- 이미지 품질 평가에 일반적으로 사용되는 LPIPS 손실함수
- 인간의 시각적 품질에 따라 비교하는 척도

$$\mathcal{L}_{MSE} = \|\gamma(\tilde{\mathbf{x}}) - \gamma(\mathbf{x})\|^2$$

- $\tilde{\mathbf{x}}$: stego 이미지
- γ : RGB \rightarrow YUV 매핑 함수(비가역)
- stego 이미지와 원본 이미지의 차이를 RGB에서 YUV로 변환하여 계산하는 MSE 함수

02 What is RoSteALS?

2.2 Methodology of RoSteALS

2. Steganography with RoSteALS

$$\mathcal{L}_{\text{recovery}} = \mathcal{L}_{BCE}(s, \tilde{s})$$

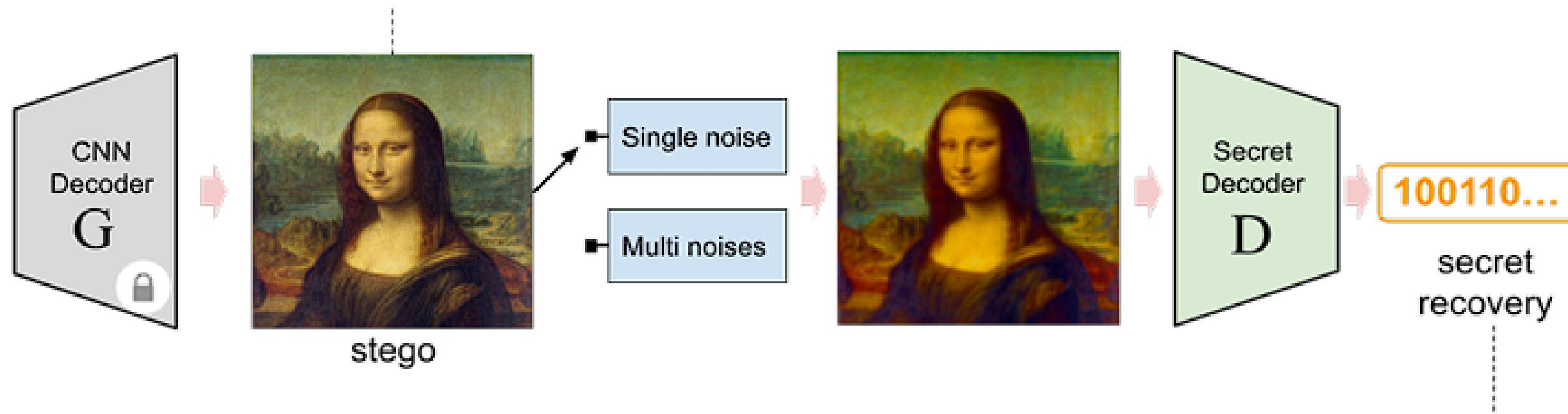


- BCE는 비밀 메시지 s 와
복원 메시지 \tilde{s} 사이의
비트 복원 정확도를 평가함

02 What is RoSteALS?

2.2 Methodology of RoSteALS

2. Steganography with RoSteALS



- 이미지 디코더 G와 **Secret 디코더 D** 사이에 **노이즈를 추가**하여 디코더의 강건성을 높인다
- ImageNet-C에서 제공하는 14가지 종류의 노이즈를 사용하여 모델의 강건성을 평가함
 - 1. 미분 가능한 노이즈 : e.g. 밝기, 채도, 대비 등
 - 2. 미분 가능 변환으로 근사할 수 있는 노이즈 : e.g. JPEG 압축
 - 3. 미분 불가능한 노이즈 : e.g. spatter(얼룩)
 - spatter와 같은 미분 불가능한 노이즈는 덧셈형 노이즈로 변환하여 gradient가 전파될 수 있도록 함
- 이렇게 다양한 데이터 증강을 통해 Secret 디코더를 훈련시키며, 피드백 신호가 Secret 인코더를 업데이트하는 데 전파될 수 있도록 한다

Results & Limitations

3.1 Datasets, training details & metrics

Datasets



- **MIRFlickR**

- train : 100K
- validate : 1K

- **3 benchmarks**

- CLIC : 530 고품질 모바일 사진
- MetFace : 1336개 얼굴 사진
- Stock : 1K개 멀티미디어 사진

Training Details

- **이미지 전처리:** 훈련 시 이미지를 256×256 으로 조정
- **비밀 메시지 크기:** 비밀 메시지 크기는 100비트로 고정
- **최적화 알고리즘:** AdamW 옵티마이저
- **훈련 종료 조건:** 검증 손실 개선이 없을 시
- **훈련 초기화:** 훈련 초반에는 비밀 메시지 복원을 우선, 비트 정확도 예측을 중요시
- **동적 손실 가중치:** β 값은 훈련 중 동적으로 설정, 초기엔 낮게 시작, 훈련 세트 전체 사용 후 β 점진적으로 증가
- **훈련 시 비트 정확도 기준:** 90% 정확도 달성 후 전체 훈련 세트 사용, 98%에 노이즈 모델 활성화

Metrics

- **Stego 이미지 품질 평가 메트릭**

- PSNR : 신호 대 잡음 비율
- SSIM : 구조적 유사성
- LPIPS : 인지적 유사성
- SIFID : 이미지 인식 가능성

- **비밀 정보 복원 성능 평가 메트릭**

- Bit acc. : 비트 정확도
- Bit acc.(ECC) : BCH 코드로 오류 수정 후 정확도 평가
- Word acc. : 비밀 정보를 단어 수준에서 평가, 20% 이내 일치하면 성공으로 판단

03 Results & Limitations

3.2 Baseline Comparison

| Method | Image quality | | | | Secret recovery | | | |
|-----------------|---------------------|--------------------|--------------------|--------------------|--------------------|--------------------|------------------|--------------|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | SIFID ↓ | Bit acc. (clean) ↑ | Bit acc. ↑ | Bit acc. (ECC) ↑ | Word acc. ↑ |
| CLIC | | | | | | | | |
| RoSteALS | 32.68 ± 1.75 | 0.88 ± 0.06 | 0.04 ± 0.02 | 0.04 ± 0.02 | 1.00 | 0.94 ± 0.07 | 1.00 | 0.93 |
| VQGAN | 33.90 ± 14.47 | 0.90 ± 0.06 | 0.03 ± 0.02 | 0.02 ± 0.02 | N/A | N/A | N/A | N/A |
| StegaStamp [39] | 31.26 ± 0.85 | 0.91 ± 0.03 | 0.09 ± 0.03 | 0.23 ± 0.13 | 1.00 | 0.88 ± 0.13 | 0.48 ± 0.50 | 0.74 |
| SSL [12] | 41.84 ± 0.10 | 0.98 ± 0.01 | 0.02 ± 0.01 | 0.01 ± 0.02 | 0.99 ± 0.03 | 0.62 ± 0.14 | 0.03 ± 0.17 | 0.13 |
| RivaGAN [49] | 40.32 ± 0.15 | 0.98 ± 0.01 | 0.02 ± 0.02 | 0.07 ± 0.06 | 0.98 ± 0.03 | 0.77 ± 0.16 | 0.22 ± 0.41 | 0.45 |
| dwtDctSvd [25] | 38.96 ± 1.41 | 0.97 ± 0.02 | 0.01 ± 0.01 | 0.02 ± 0.02 | 1.00 | 0.61 ± 0.20 | 0.16 ± 0.34 | 0.21 |
| MetFACE | | | | | | | | |
| RoSteALS | 34.46 ± 1.91 | 0.89 ± 0.07 | 0.04 ± 0.02 | 0.01 ± 0.02 | 1.00 | 0.94 ± 0.08 | 1.00 | 0.91 |
| VQGAN | 35.98 ± 2.45 | 0.90 ± 0.07 | 0.02 ± 0.02 | 0.01 ± 0.02 | N/A | N/A | N/A | N/A |
| StegaStamp [39] | 32.01 ± 0.77 | 0.92 ± 0.02 | 0.13 ± 0.03 | 0.22 ± 0.15 | 1.00 | 0.86 ± 0.14 | 0.47 ± 0.50 | 0.68 |
| SSL [12] | 41.77 ± 0.12 | 0.98 ± 0.01 | 0.04 ± 0.02 | 0.04 ± 0.05 | 1.00 | 0.63 ± 0.16 | 0.08 ± 0.27 | 0.19 |
| RivaGAN [49] | 40.27 ± 0.09 | 0.97 ± 0.01 | 0.06 ± 0.03 | 0.16 ± 0.12 | 0.99 ± 0.01 | 0.78 ± 0.17 | 0.28 ± 0.44 | 0.47 |
| dwtDctSvd [25] | 40.86 ± 2.48 | 0.98 ± 0.01 | 0.02 ± 0.01 | 0.03 ± 0.02 | 1.00 | 0.63 ± 0.23 | 0.22 ± 0.38 | 0.26 |
| Stock1K | | | | | | | | |
| RoSteALS | 33.27 ± 2.32 | 0.89 ± 0.08 | 0.03 ± 0.02 | 0.05 ± 0.06 | 1.00 | 0.92 ± 0.10 | 1.00 | 0.864 |
| VQGAN | 34.44 ± 2.71 | 0.91 ± 0.07 | 0.02 ± 0.02 | 0.03 ± 0.06 | N/A | N/A | N/A | N/A |
| StegaStamp [39] | 31.42 ± 0.95 | 0.92 ± 0.03 | 0.08 ± 0.04 | 0.20 ± 0.14 | 1.00 | 0.87 ± 0.13 | 0.48 ± 0.50 | 0.72 |
| SSL [12] | 42.07 ± 0.50 | 0.99 ± 0.01 | 0.02 ± 0.02 | 0.01 ± 0.02 | 0.95 ± 0.09 | 0.59 ± 0.12 | 0.02 ± 0.13 | 0.09 |
| RivaGAN [49] | 40.49 ± 0.45 | 0.98 ± 0.01 | 0.02 ± 0.02 | 0.05 ± 0.06 | 0.93 ± 0.09 | 0.72 ± 0.16 | 0.13 ± 0.33 | 0.31 |
| dwtDctSvd [25] | 39.76 ± 2.41 | 0.98 ± 0.02 | 0.01 ± 0.01 | 0.02 ± 0.02 | 0.95 ± 0.13 | 0.60 ± 0.19 | 0.17 ± 0.33 | 0.18 |

- RoSteALS를 4가지 baseline과 비교한다
 - dwtDCTSvd, StegaStamp, RivaGAN, SSL
- 위 테이블은 **이미지 품질과 비밀 정보 복구 성능**을 보여준다

03 Results & Limitations

3.2 Baseline Comparison

| Method | Image quality | | | |
|-----------------|---------------------|--------------------|--------------------|--------------------|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | SIFID ↓ |
| CLIC | | | | |
| RoSteALS | 32.68 ± 1.75 | 0.88 ± 0.06 | 0.04 ± 0.02 | 0.04 ± 0.02 |
| VQGAN | 33.90 ± 14.47 | 0.90 ± 0.06 | 0.03 ± 0.02 | 0.02 ± 0.02 |
| StegaStamp [39] | 31.26 ± 0.85 | 0.91 ± 0.03 | 0.09 ± 0.03 | 0.23 ± 0.13 |
| SSL [12] | 41.84 ± 0.10 | 0.98 ± 0.01 | 0.02 ± 0.01 | 0.01 ± 0.02 |
| RivaGAN [49] | 40.32 ± 0.15 | 0.98 ± 0.01 | 0.02 ± 0.02 | 0.07 ± 0.06 |
| dwtDctSvd [25] | 38.96 ± 1.41 | 0.97 ± 0.02 | 0.01 ± 0.01 | 0.02 ± 0.02 |
| MetFACE | | | | |
| RoSteALS | 34.46 ± 1.91 | 0.89 ± 0.07 | 0.04 ± 0.02 | 0.01 ± 0.02 |
| VQGAN | 35.98 ± 2.45 | 0.90 ± 0.07 | 0.02 ± 0.02 | 0.01 ± 0.02 |
| StegaStamp [39] | 32.01 ± 0.77 | 0.92 ± 0.02 | 0.13 ± 0.03 | 0.22 ± 0.15 |
| SSL [12] | 41.77 ± 0.12 | 0.98 ± 0.01 | 0.04 ± 0.02 | 0.04 ± 0.05 |
| RivaGAN [49] | 40.27 ± 0.09 | 0.97 ± 0.01 | 0.06 ± 0.03 | 0.16 ± 0.12 |
| dwtDctSvd [25] | 40.86 ± 2.48 | 0.98 ± 0.01 | 0.02 ± 0.01 | 0.03 ± 0.02 |
| Stock1K | | | | |
| RoSteALS | 33.27 ± 2.32 | 0.89 ± 0.08 | 0.03 ± 0.02 | 0.05 ± 0.06 |
| VQGAN | 34.44 ± 2.71 | 0.91 ± 0.07 | 0.02 ± 0.02 | 0.03 ± 0.06 |
| StegaStamp [39] | 31.42 ± 0.95 | 0.92 ± 0.03 | 0.08 ± 0.04 | 0.20 ± 0.14 |
| SSL [12] | 42.07 ± 0.50 | 0.99 ± 0.01 | 0.02 ± 0.02 | 0.01 ± 0.02 |
| RivaGAN [49] | 40.49 ± 0.45 | 0.98 ± 0.01 | 0.02 ± 0.02 | 0.05 ± 0.06 |
| dwtDctSvd [25] | 39.76 ± 2.41 | 0.98 ± 0.02 | 0.01 ± 0.01 | 0.02 ± 0.02 |

이미지 품질 비교

- **SSL**

- PSNR, SSIM 점수에서 가장 우수한 성능을 보임
- 전체적인 이미지 품질 지표에서 최고의 성능

- **dwtDctSvd**

- LPIPS 메트릭에서 가장 우수한 성능을 보임

- **RoSteALS**

- RoSteALS는 MetFACE 데이터셋에서 가장 우수한 SIFID 성능을 보임
- 특징을 보존하면서도 비밀 정보를 임베딩하는 RoSteALS 가 사전 훈련된 특징 추출 모델인 SSL보다 재구성된 얼굴 이미지의 정확성이 더 우수하게 나타남

03 Results & Limitations

3.2 Baseline Comparison

| Method | Secret recovery | | | |
|-----------------|--------------------|--------------------|------------------|--------------|
| | Bit acc. (clean) ↑ | Bit acc. ↑ | Bit acc. (ECC) ↑ | Word acc. ↑ |
| CLIC | | | | |
| RoSteALS | 1.00 | 0.94 ± 0.07 | 1.00 | 0.93 |
| VQGAN | N/A | N/A | N/A | N/A |
| StegaStamp [39] | 1.00 | 0.88 ± 0.13 | 0.48 ± 0.50 | 0.74 |
| SSL [12] | 0.99 ± 0.03 | 0.62 ± 0.14 | 0.03 ± 0.17 | 0.13 |
| RivaGAN [49] | 0.98 ± 0.03 | 0.77 ± 0.16 | 0.22 ± 0.41 | 0.45 |
| dwtDctSvd [25] | 1.00 | 0.61 ± 0.20 | 0.16 ± 0.34 | 0.21 |
| MetFACE | | | | |
| RoSteALS | 1.00 | 0.94 ± 0.08 | 1.00 | 0.91 |
| VQGAN | N/A | N/A | N/A | N/A |
| StegaStamp [39] | 1.00 | 0.86 ± 0.14 | 0.47 ± 0.50 | 0.68 |
| SSL [12] | 1.00 | 0.63 ± 0.16 | 0.08 ± 0.27 | 0.19 |
| RivaGAN [49] | 0.99 ± 0.01 | 0.78 ± 0.17 | 0.28 ± 0.44 | 0.47 |
| dwtDctSvd [25] | 1.00 | 0.63 ± 0.23 | 0.22 ± 0.38 | 0.26 |
| Stock1K | | | | |
| RoSteALS | 1.00 | 0.92 ± 0.10 | 1.00 | 0.864 |
| VQGAN | N/A | N/A | N/A | N/A |
| StegaStamp [39] | 1.00 | 0.87 ± 0.13 | 0.48 ± 0.50 | 0.72 |
| SSL [12] | 0.95 ± 0.09 | 0.59 ± 0.12 | 0.02 ± 0.13 | 0.09 |
| RivaGAN [49] | 0.93 ± 0.09 | 0.72 ± 0.16 | 0.13 ± 0.33 | 0.31 |
| dwtDctSvd [25] | 0.95 ± 0.13 | 0.60 ± 0.19 | 0.17 ± 0.33 | 0.18 |

비밀 정보 복원 성능

[비밀 정보 복원 성능]

- clean data에서는 거의 완벽에 가까운 점수를 보임
 - 데이터에 노이즈가 없다면 모든 기법들이 높은 성능
- RoSteALS는 모든 성능 지표에서 가장 우수한 성능**을 보임
- dwstDctSvd와 SSL은 성능이 가장 낮음
- StegaStamp는 2번째로 우수한 성능을 보임
- RoSteALS가 다른 기법들보다 더 뛰어난 비밀 정보 복구 성능을 보임을 알 수 있음

[Bit acc. (ECC)]

- Bit acc. (ECC) 즉, 손상된 비밀 메시지 데이터를 ECC를 통해 조금이라도 복구한 후에 정확도를 평가한 것이다
- RoSteALS는 약간 손상된 데이터에 대한 성능을 크게 개선**
- 다른 기법들은 데이터 손상이 일정 기준 이상이 되면 성능이 더욱 저하되는 경향을 보임

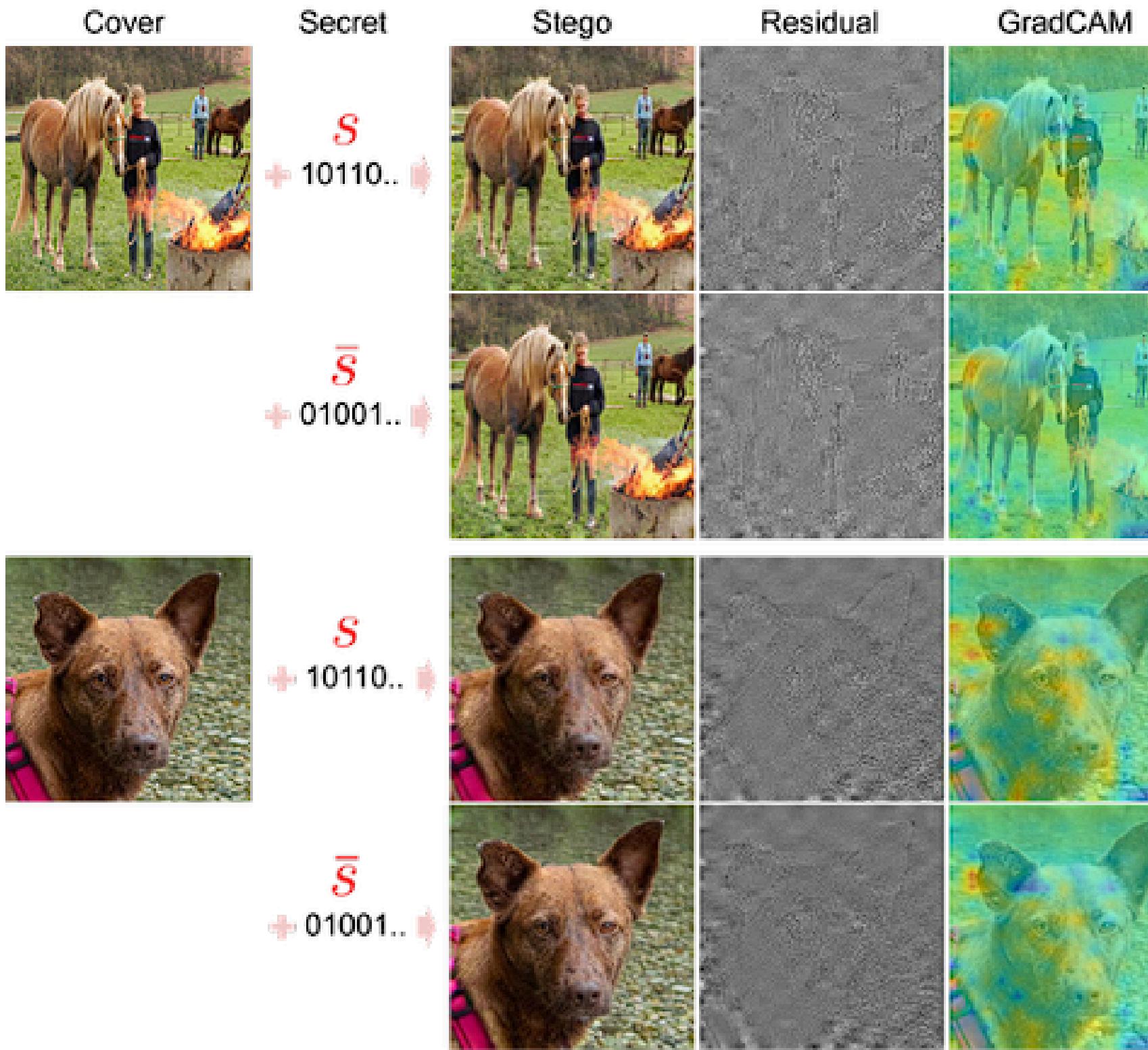
3.2 Baseline Comparison

CONCLUSION

- RoSteALS의 성능은 비밀 메시지 삽입과 강건성 학습을 **이미지 분포 학습과 분리하여 수행**할 수 있는 능력 덕분에 발생함
 - 이미지 분포 학습과 분리할 수 있는 이유는 사전 훈련된 VQGAN을 사용하기 때문
- 이로 인해 깨끗한 데이터와 노이즈가 있는 데이터 모두에서 뛰어난 성능을 발휘할 수 있음
- 그러나, **오토인코더 Backbone에 의해 성능이 제한되며**, 더 나은 Backbone을 사용하면 성능이 더욱 향상될 여지가 있음
 - 이미지 품질 지표(PSNR, SSIM 등)에서 RoSteALS가 SSL보다 낮은 성과를 보이는 이유가 바로 RoSteALS의 Backbone이 이미지 품질에는 SSL보다 제한적이라는 의미

03 Results & Limitations

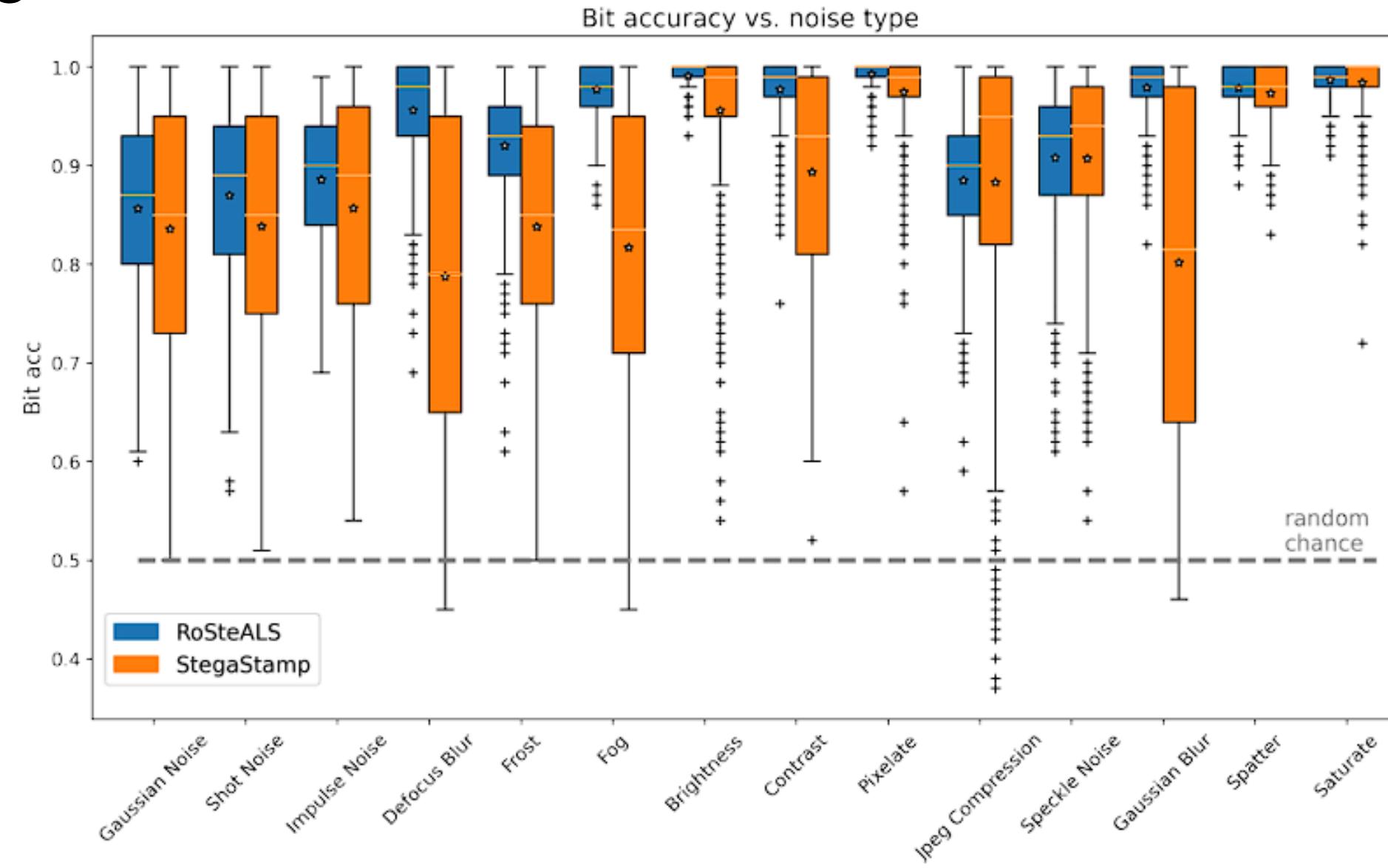
3.3 Robustness



- 왼쪽 그림은 **서로 반대되는 두 비밀 메시지가 삽입된 경우**와 **같은 비밀 메시지 두 개를 다른 이미지에 삽입한 경우**의 변화를 보여준다
- GradCAM의 heatmap은 비밀 메시지가 이미지 전체에 걸쳐 삽입 되었음을 보여줌
 - 즉, 비밀 메시지가 이미지의 특정 부분만 영향을 주는 것이 아닌 이미지 전체에 퍼져있음을 시각적으로 나타냄
- 발표 자료 16p와 같이 **비밀 메시지 삽입이 이미지 내용에 의존하지 않는다**는 점을 알 수 있음
- 이미지 내용과 독립적으로 비밀 메시지를 삽입하는 특성은 VQGAN 오토인코더와 결합되어 RoSteALS가 훈련 중에 보지 못한 새로운 도메인에 대해서도 잘 일반화하는데 도움이 됨

03 Results & Limitations

3.3 Robustness



[그래프 정보]

- RoSteALS vs StegaStamp 비밀 복원 성능 비교
 - x축 : 다양한 개별 노이즈
 - y축 : Bit acc
- RosteALS가 거의 대부분의 교란에 대해 더 강건하고 안정적
 - 특히, 블러링(Gaussian, Defocus) & 강력한 이미지 향상 효과(서리, 안개)에 대해 더 나은 성능을 보임
 - 두 기법 모두 단순한 선형 노이즈(밝기, 대비)와 픽셀화에 대해서는 최고의 성능을 보이지만, 고도화된 압축이나 강력한 픽셀 손상을 일으키는 노이즈(shot, impulse, speckle)에서는 복원 성능이 저하된다

03 Results & Limitations

3.4 Dependencies & the quality/recover trade-off

| Secret length (bits) | 50 | 100 | 150 | 200 |
|----------------------|-------|-------|-------|-------|
| PSNR | 32.81 | 32.69 | 32.85 | 32.89 |
| SSIM | 0.89 | 0.88 | 0.88 | 0.88 |
| Bit acc. | 0.97 | 0.94 | 0.87 | 0.84 |
| Train time (epochs) | 17 | 18 | 21 | 30 |

| Train volume ($\times 10^3$) | 10 | 40 | 80 | 100 |
|--------------------------------|-------|-------|-------|-------|
| PSNR | 35.07 | 34.51 | 34.36 | 34.46 |
| SSIM | 0.89 | 0.90 | 0.90 | 0.89 |
| Bit acc. | 0.91 | 0.91 | 0.93 | 0.94 |



[비밀 메시지 길이 증가에 따른 RoSteALS 성능 비교]

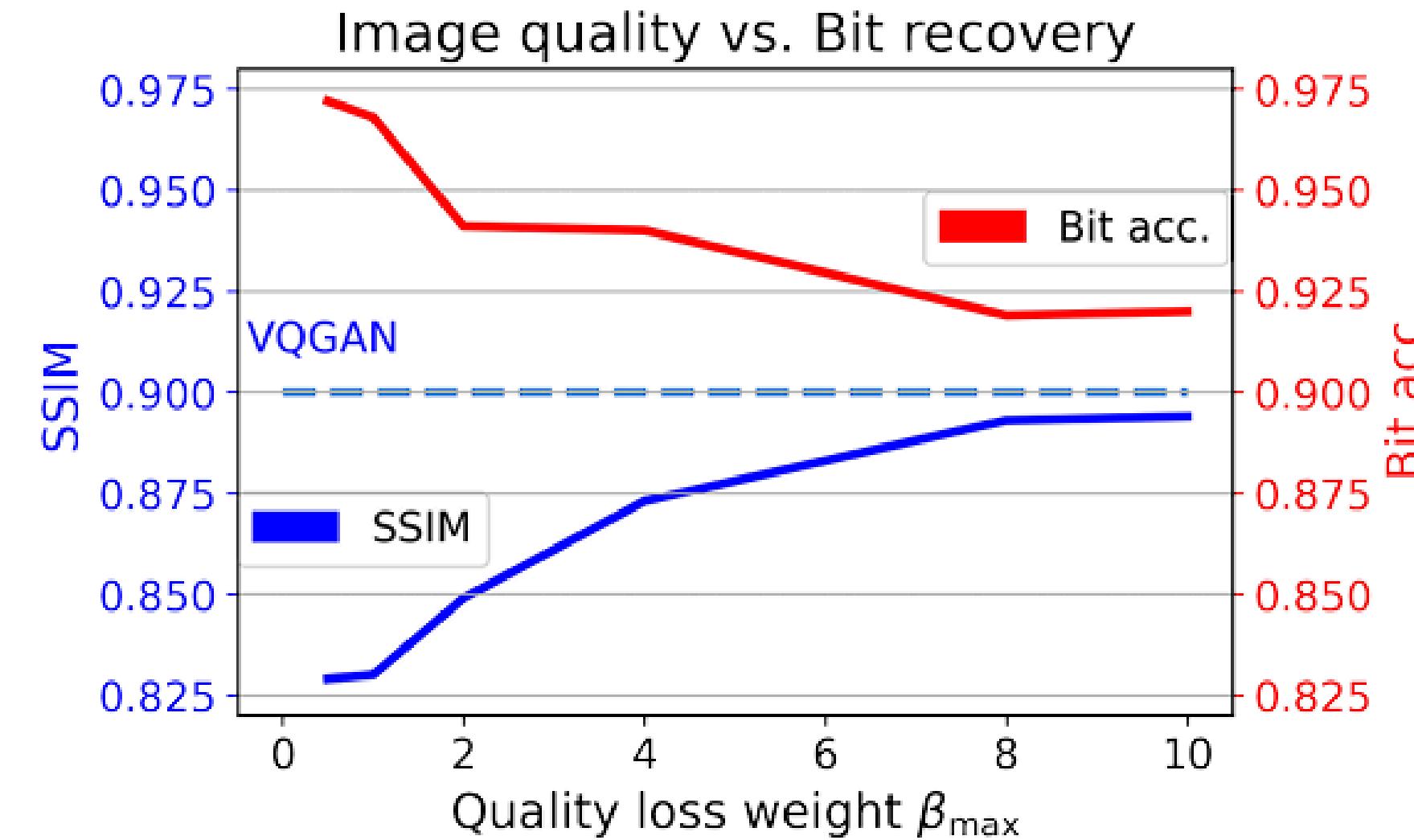
- 비밀 메시지 길이가 증가하면
 - 이미지 품질 유지 가능
 - 비밀 복원 성능 ↓ (50->200까지 증가 시 13% 하락)
 - 훈련 시간↑
- L = 200bits일 때, 30 epoch만에 수렴
 - RoSteALS의 목표가 비밀 인코딩/디코딩 모듈만 학습하기 때문이다 (SSL: 100epoch, RivaGAN: 300epoch)

[훈련 데이터 양과 이미지 품질 & 비밀 복원 성능 관계]

- 훈련 데이터의 양이 이미지 품질에 미치는 영향 x
 - RoSteALS가 오토인코더를 동결하기 때문
 - 이미지 품질은 오토인코더의 학습된 특성에 의존
- 비밀 복원 성능은 훈련 데이터 양이 증가하면 향상되지만, 과적합 문제 발생 가능
 - 데이터 양 10배 → Bit acc. 3% 증가

03 Results & Limitations

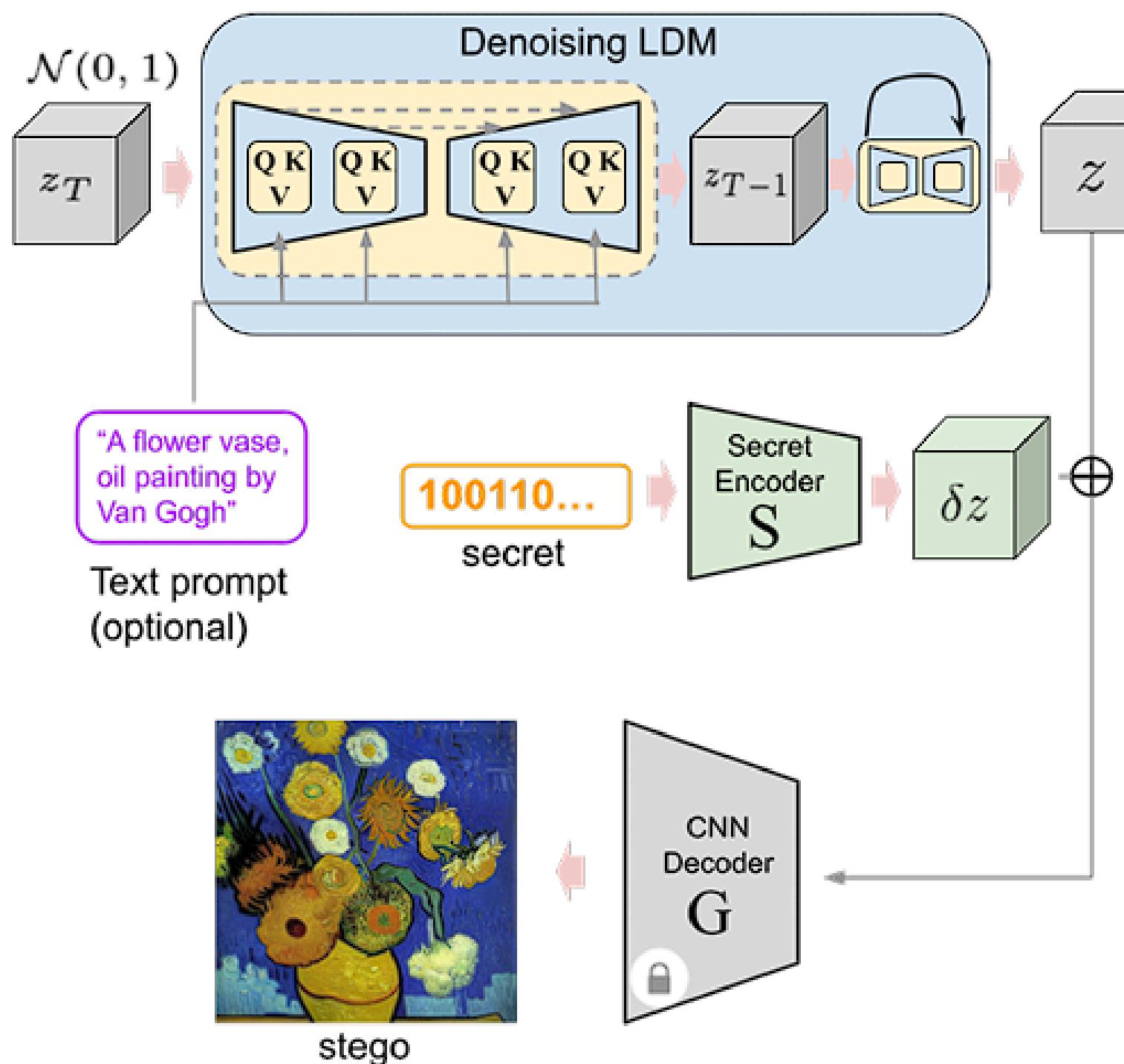
3.4 Dependencies & the quality/recover trade-off



- 위의 그래프는 RoSteALS에서 stego 이미지의 품질과 비밀 복원 성능 간의 trade-off가 존재함을 나타낸다
- β_{max} 를 통해 이 trade-off를 제거할 수 있음
 - β_{max} 를 10배 늘리면 → SSIM : 7점 증가, Bit acc : 5% 감소
- β_{max} 는 10 이상으로 설정할 수 없음
 - 이미지 품질이 이미 오토인코더의 성능에 의해 제한되기 때문

03 Results & Limitations

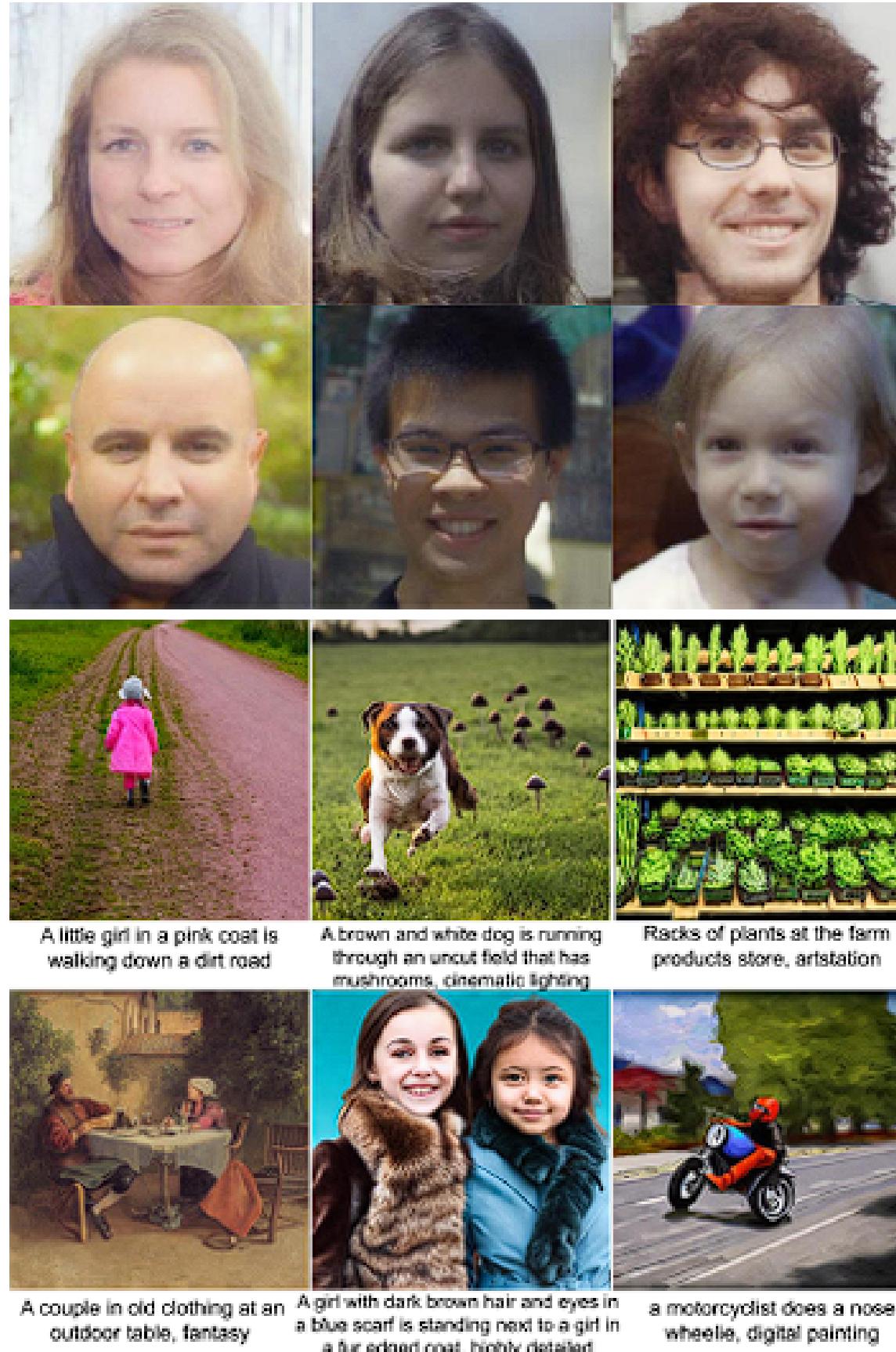
3.5 Text-based & Cover-less stega



- RoSteALS의 비밀 메시지를 이미 정의된 잠재 공간에 직접 삽입하는 방식을 통해 아래 2가지가 가능함
- **Coverless 스테가노그래피**
 - 이론적으로는 이미지 인코더(E)를 제거하고, **잠재 공간의 무작위 지점에 비밀 메시지를 삽입**하여 cover 이미지 없이 stego 이미지를 생성할 수 있음
 - 그러나 VQGAN의 잠재 코드 분포를 매우 복잡 → 무작위로 지점을 선택하면 품질이 낮을 가능성이 큼
 - 이를 위해 단순한 분포(e.g. Gaussian 분포)를 사용하여 잠재공간을 매핑
- **Text-based 스테가노그래피**
 - 텍스트도 하나의 분포로 사용하여, **텍스트 프롬프트를 통해 비밀 메시지를 삽입**할 수 있음
 - 이 매핑을 학습하는 과정은 LDM을 사용하여 이루어짐

03 Results & Limitations

3.5 Text-based & Cover-less stega



• Coverless 스테가노그래피

- LDM은 FFHQ 데이터셋으로 훈련됨
- FFHQ : 사람 얼굴 이미지 포함 고해상도 데이터셋
- 이미 훈련된 RoSteALS 모델을 사용하여 cover 이미지 없이도 비밀 메시지를 잠재 공간에 숨기고 stego 이미지를 생성함

• Text-based 스테가노그래피

- 사전 훈련된 Stable Diffusion 모델 사용, 이는 텍스트 프롬프트를 기반으로 이미지를 생성함
- 이 모델을 KL-f8 오토인코더 backbone을 사용하여 새로운 RoSteALS 변형 모델을 훈련시킴 → Stable Diffusion과 호환되는 설정
- MIRFlickR 데이터셋을 사용하여 훈련함

03 Results & Limitations

3.5 Text-based & Cover-less stega

| Method | Bit acc. (clean) | Bit acc. | Word acc. |
|------------|------------------|----------|-----------|
| Cover-less | 0.997 | 0.924 | 0.875 |
| Text-based | 0.992 | 0.904 | 0.844 |

- 비밀 메시지 복원 성능 평가
 - 1000개의 stego 이미지 샘플링 : 비밀 메시지 복구 성능 평가를 위해 각각의 LDM-RoSteALS 모델에 대해 무작위로 생성된 비밀 메시지를 사용하여 1000개의 stego 이미지를 샘플링함
- RoSteALS 성능
 - RoSteALS 모델은 LDM 모델이 생성한 잠재 코드를 훈련 중에 한 번도 본 적이 없지만, 여전히 잘 작동함을 알 수 있음
 - 이는 **RoSteALS가 잘 일반화 되어 있음**을 의미하며, 새로운 데이터셋과 조건에서도 좋은 성능을 보일 수 있음을 시사함

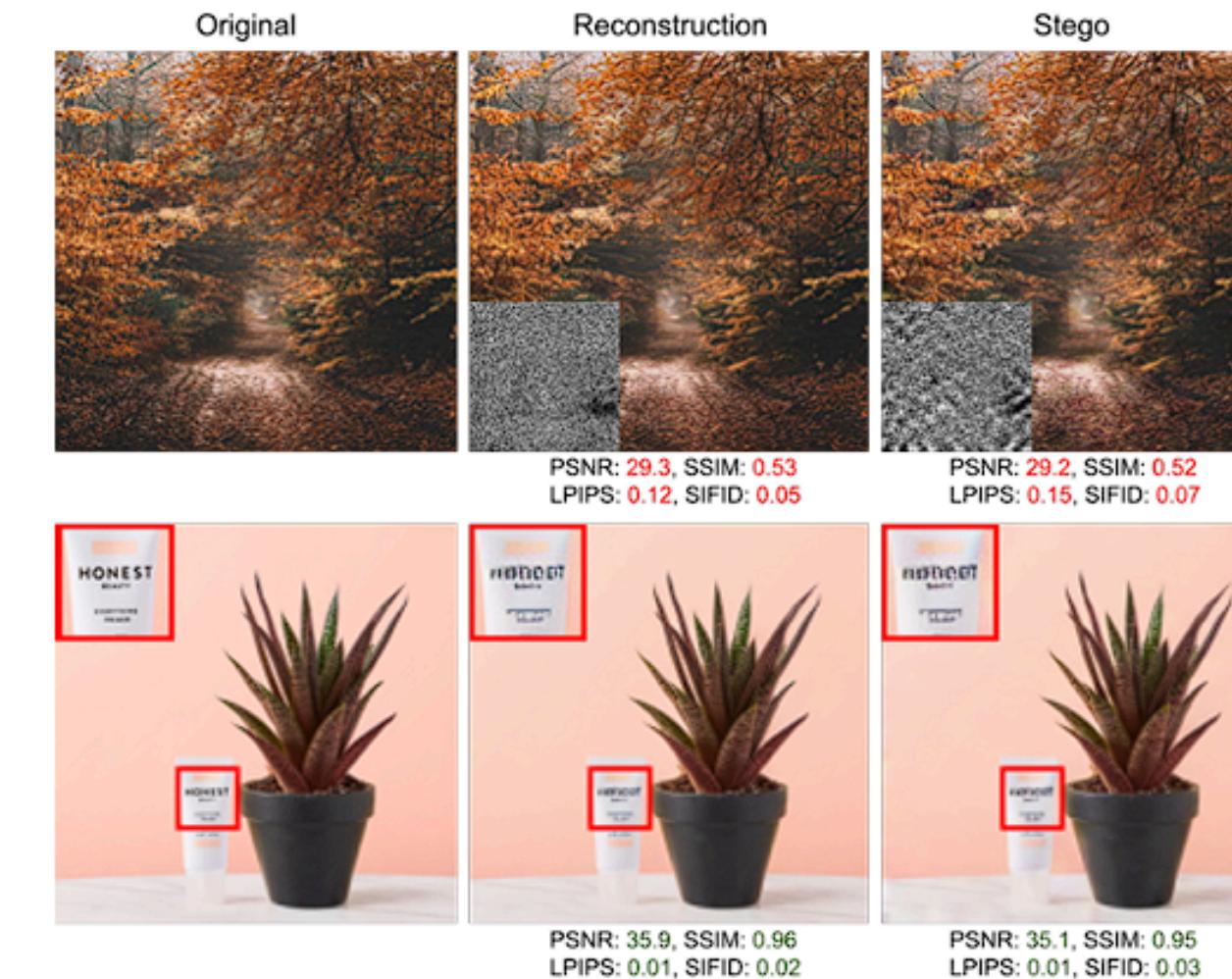
03 Results & Limitations

3.6 Conclusion

- 1 RoSteALS는 사전 훈련된 오토인코더의 잠재 공간을 사용하여 데이터를 숨기는 새로운 스테가노그래피 기법 제시
- 2 작은 모델 크기, 모듈화된 설계로 효율적인 학습과 뛰어난 비밀 복원 성능을 제공
- 3 Coverless, Text-based와 같은 새로운 응용에도 쉽게 적응할 수 있도록 설계됨
 - 특정 cover 이미지에 의존하지 않고, 새롭게 생성된 이미지에 메시지를 숨김 → 공격자에게 추적당할 위험 ↓
 - 기존 스테가노그래피 기법은 cover 이미지의 변형을 분석함으로써 비밀 메시지를 추적 가능했지만, coverless는 이러한 탐지 작업을 더 어렵게 함

03 Results & Limitations

3.6 Limitations



- RoSteALS는 이미지 인코딩과 생성에 사전 훈련된 오토인코더를 사용함
 - 오토인코더가 갖고 있는 단점인 이미지 생성 시 작은 세부사항을 잘 보존하지 못하는 문제를 그대로 물려 받음
- 위의 그림은 2가지 실패 사례를 보여줌
 - 복잡한 물체들이 포함된 경우 : 이미지 재구성 중 많은 미세한 공간적 이동 발생 → 지각적으로 눈에 띠지 x / 이미지 품질 ↓
 - 작은 텍스트나 얼굴을 재구성하는 경우 : 지각적으로 잘 보임 / 이미지 품질 지표는 큰 영향 x
- 이러한 문제를 해결하기 위해 더 강력한 오토인코더를 활용 or 오토인코더를 fine-tuning하는 새로운 방법을 개발할 필요성 존재

Future Directions

- 다양한 스테가노그래피 기법에 대한 지식을 기반으로 스테가노그래피를 더욱 더 강건하게 하는 방향성에 대한 고찰의 필요성을 느낌
- 스테가노그래피를 효과적으로 공격하는 공격 기법에 대한 연구가 필요한 시점이라고 생각됨

감사합니다:)