Stego network

Steganography

- *“The best way of keeping a secret is to pretend there isn’t one.”*— Margaret Atwood
- *“Steganography is the art and science of hiding communication.”*

Steganography

- Usually
  - Stego medium is image, “stego image”
  - Secret data can be image, text, and etc.

- Three elements
  - Capacity: 얼마나 많이 cover medium 에 숨길 수 있는가?
  - Security: 도청자에게 들키지 않을 수 있는가?
  - Robustness: Cover medium 이 수정될 때 숨겨진 내용을 보존할 수 있나?

Steganography

- Steganography
  - Invisible
  - It is okay to destroy if the stego medium is modified
  - Our interest and goal

- Watermarking
  - Visible / invisible
  - It should be retained while the stego medium is modified
  - Copyright ⋯

Steganography: example

- MSB on LSB
  - The simplest method
  - Visual attack

- DCT coefficient based method
  - Use LSB of DCT coefficient
  - No visual attack
  - Steganalysis can be applied

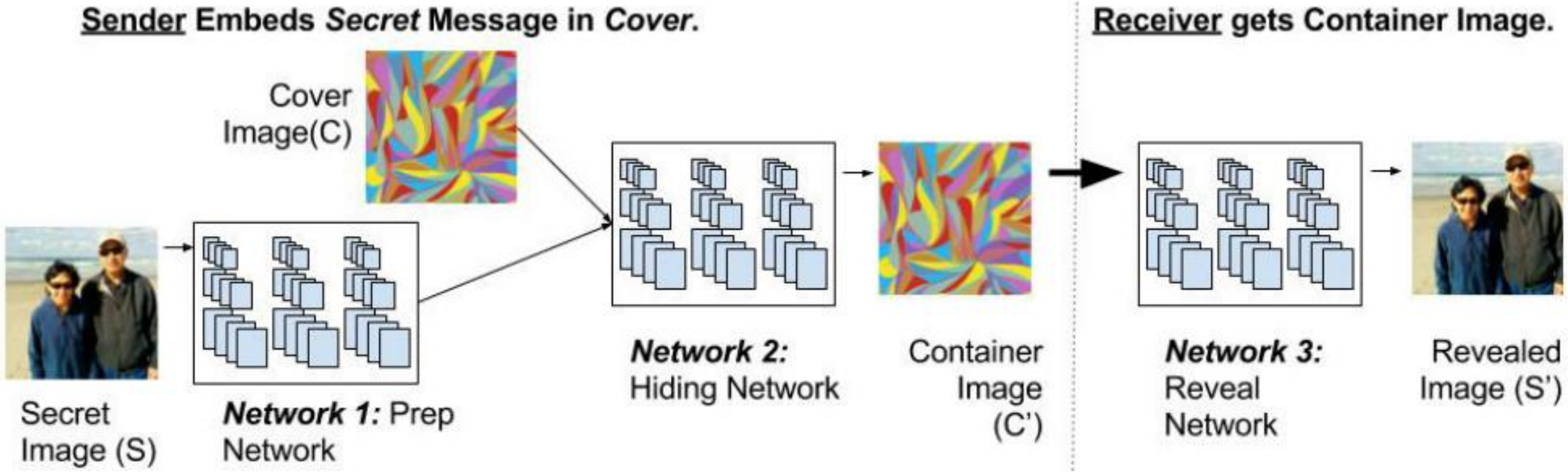
Idea

- If stego medium is a neural network
  - We cannot understand parameter itself by looking at
  - Pixel is an integer value, but the parameter is real value
  - It is hard to apply steganalysis technique
    because we do not know the nature of NN parameters
  - Neural network can inference data
    - We need to consider parameter, activation, output, and gradients
    - Accuracy for discriminative task (regression, classification)
    - Quality of outputs for generation task (GAN, VAE)

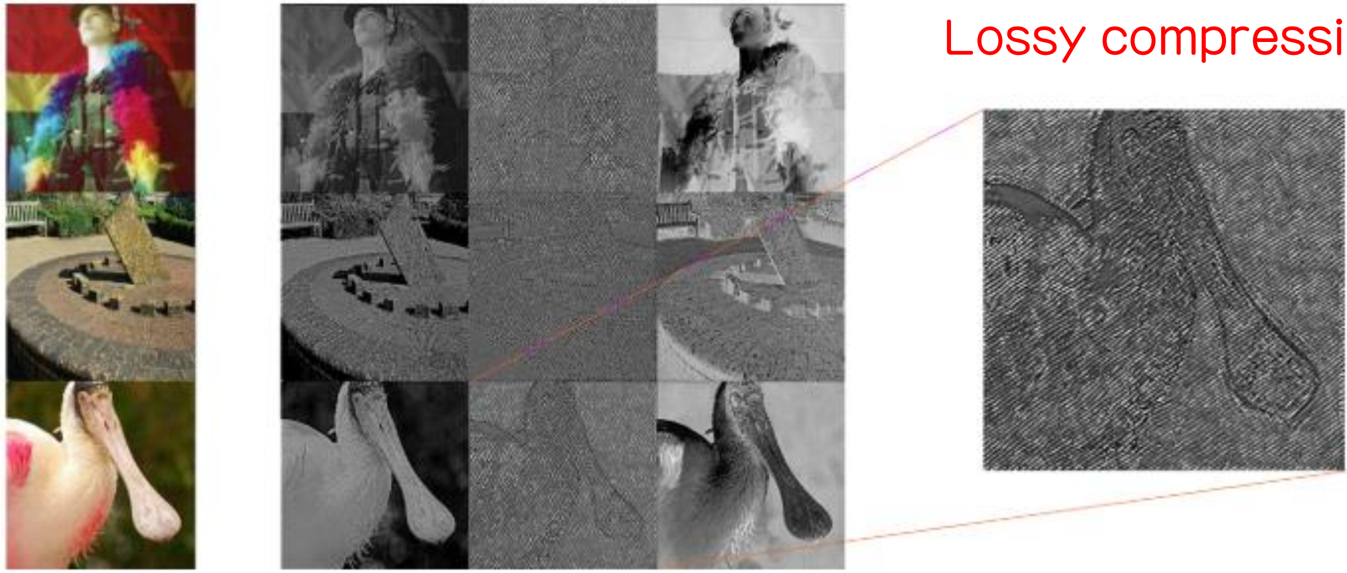
Related work

- No exact one
- Hiding images in plain sight: Deep steganography (NIPS 2017)
- HiDDeN: Hiding data with deep networks (ECCV 2018, Justin Johnson / Li Fei-Fei)

Deep steganography


Sender Embeds Secret Message in Cover.
Receiver gets Container Image.
Cover Image(C)
Secret Image (S)
Network 1: Prep Network
Network 2: Hiding Network
Container Image (C')
Network 3: Reveal Network
Revealed Image (S')

Deep steganography



Lossy compression here

Figure 2: Transformations made by the preparation network (3 examples shown). Left: Original Color Images. Middle: the three channels of information extracted by the preparation network that are input into the middle network. Right: zoom of the edge-detectors. The three color channels are transformed by the preparation-network. In the most easily recognizable example, the 2nd channel activates for high frequency regions, e.g. textures and edges (shown enlarged (right)).

Deep steganography

$$L(c,c',s,s') = \big||c-c'|\big| + \beta||s-s'||$$

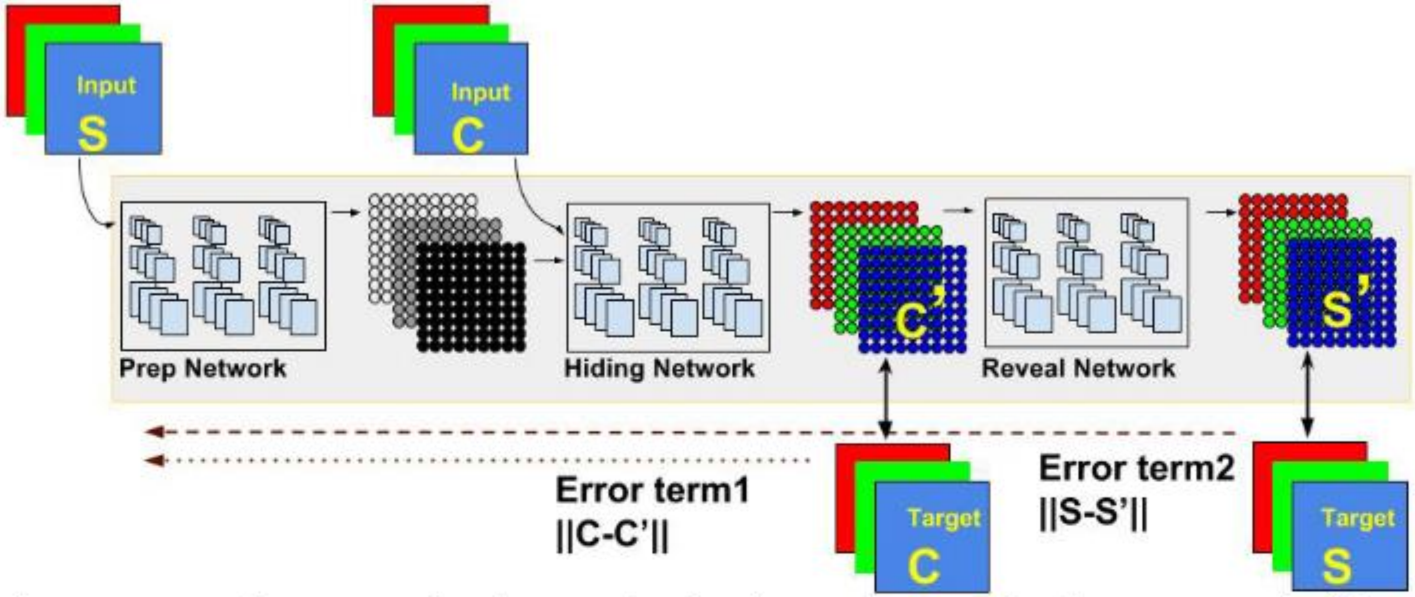
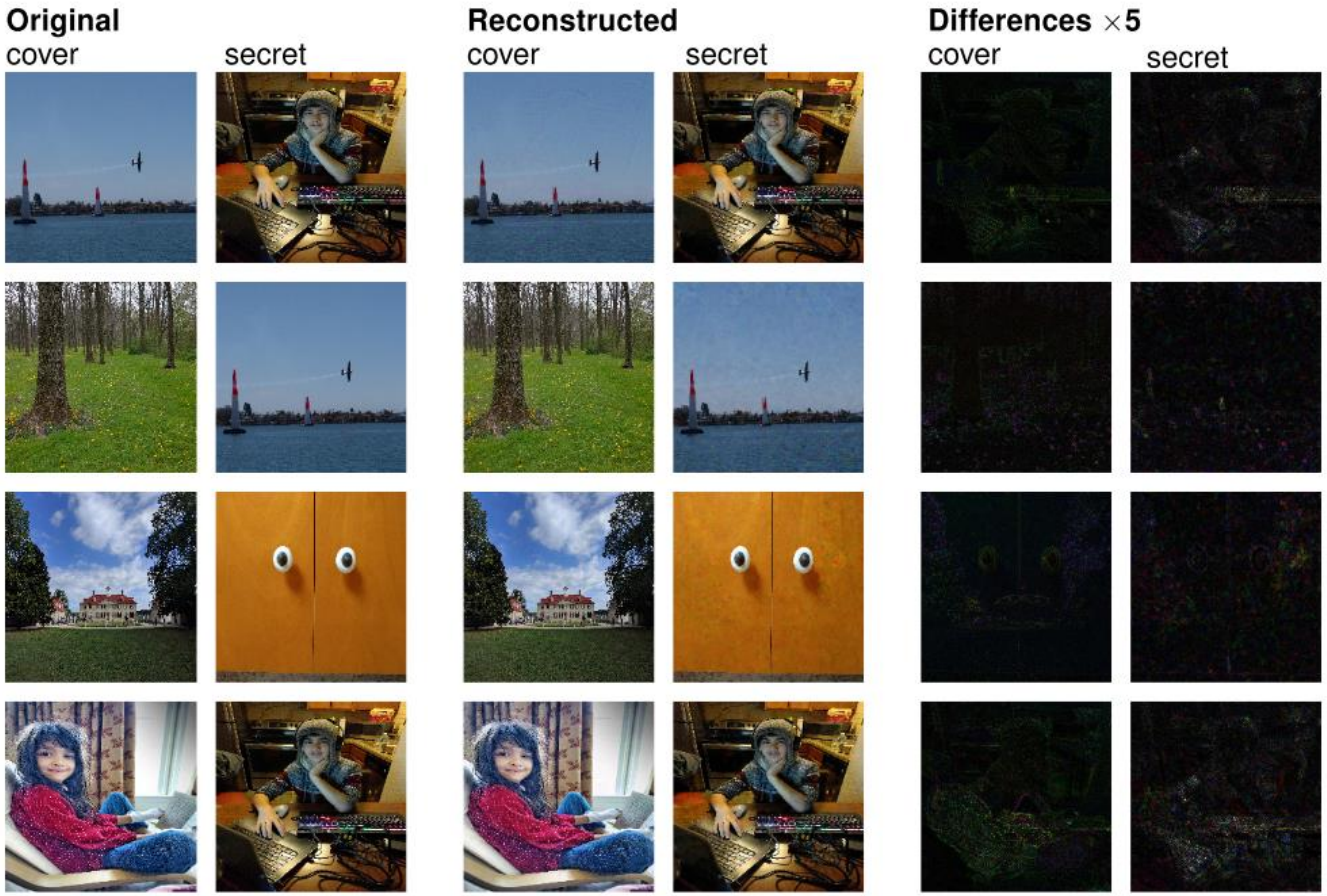
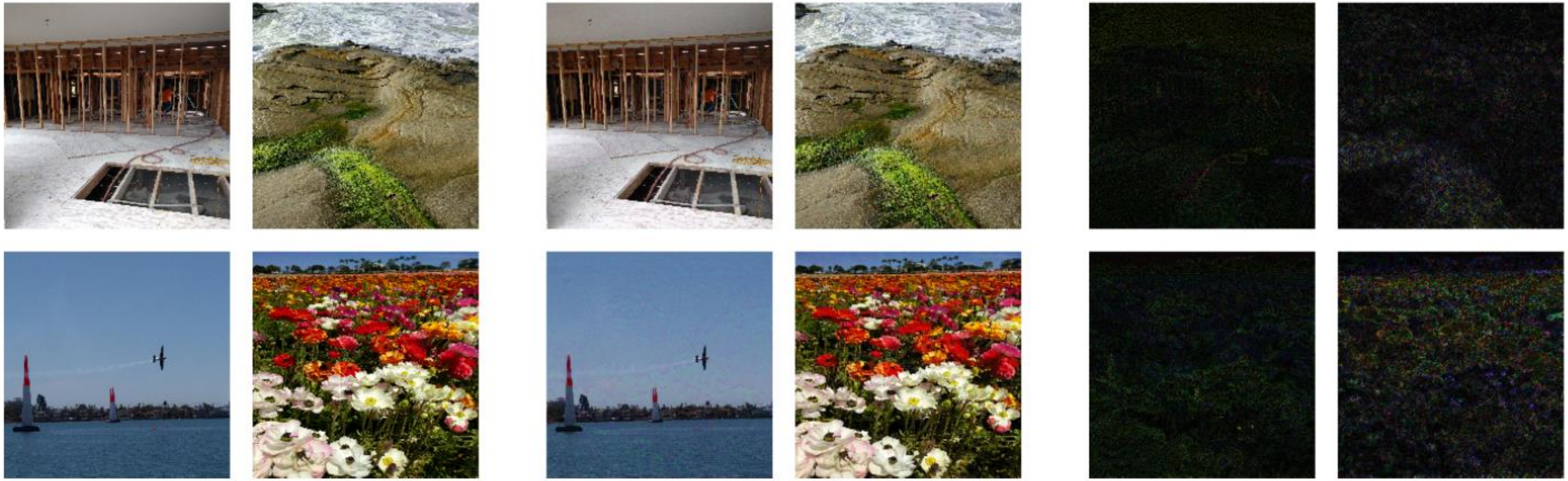


Figure 3: The three networks are trained as a single, large, network. Error term 1 affects only the first two networks. Error term 2 affects all 3. S is the secret image, C is the cover image.

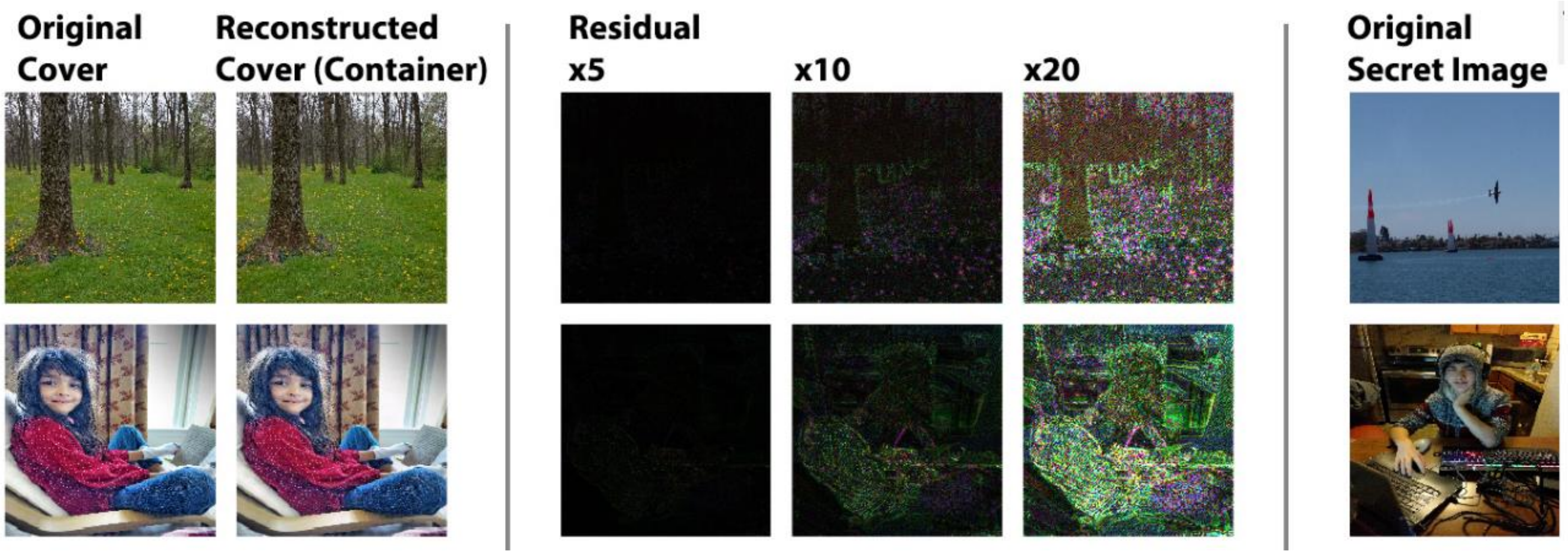
Deep steganography



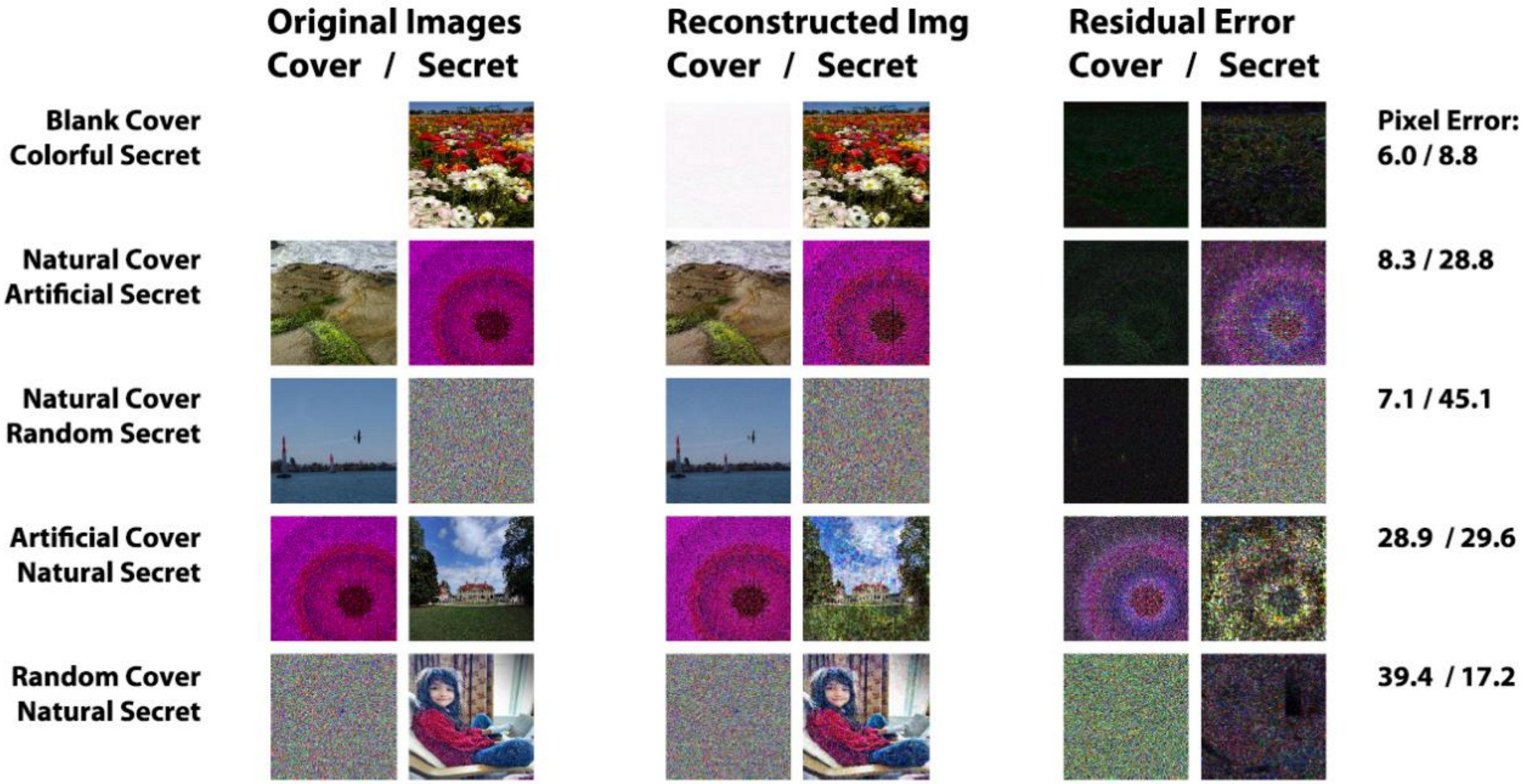
Deep steganography



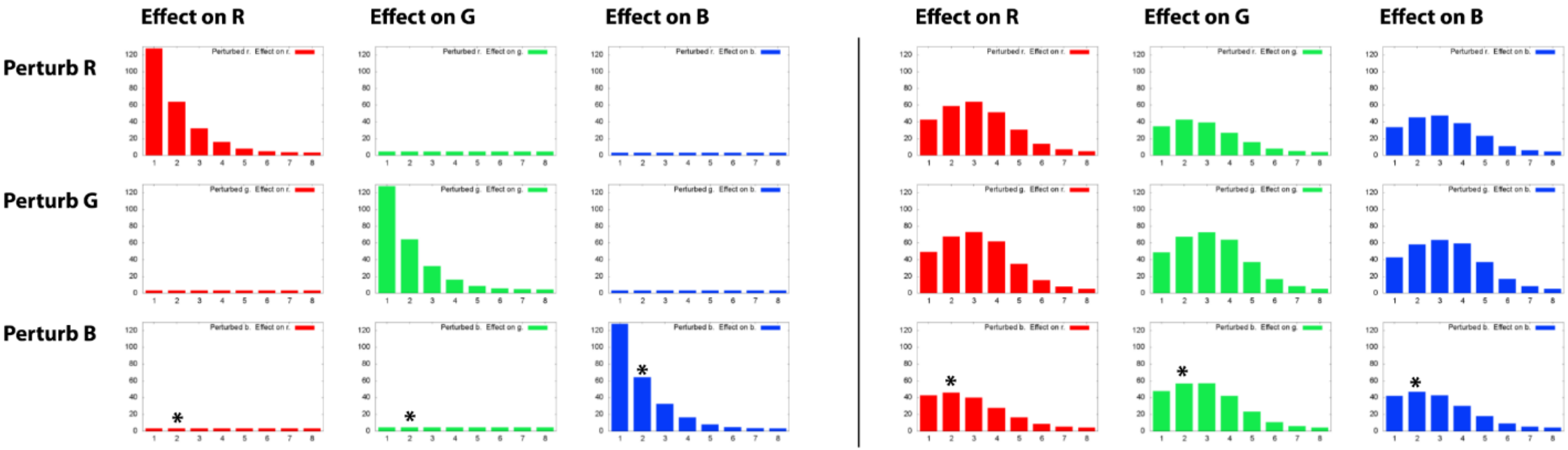
Deep steganography



Deep steganography



Deep steganography



Container Image **Reconstructed Secret Image**