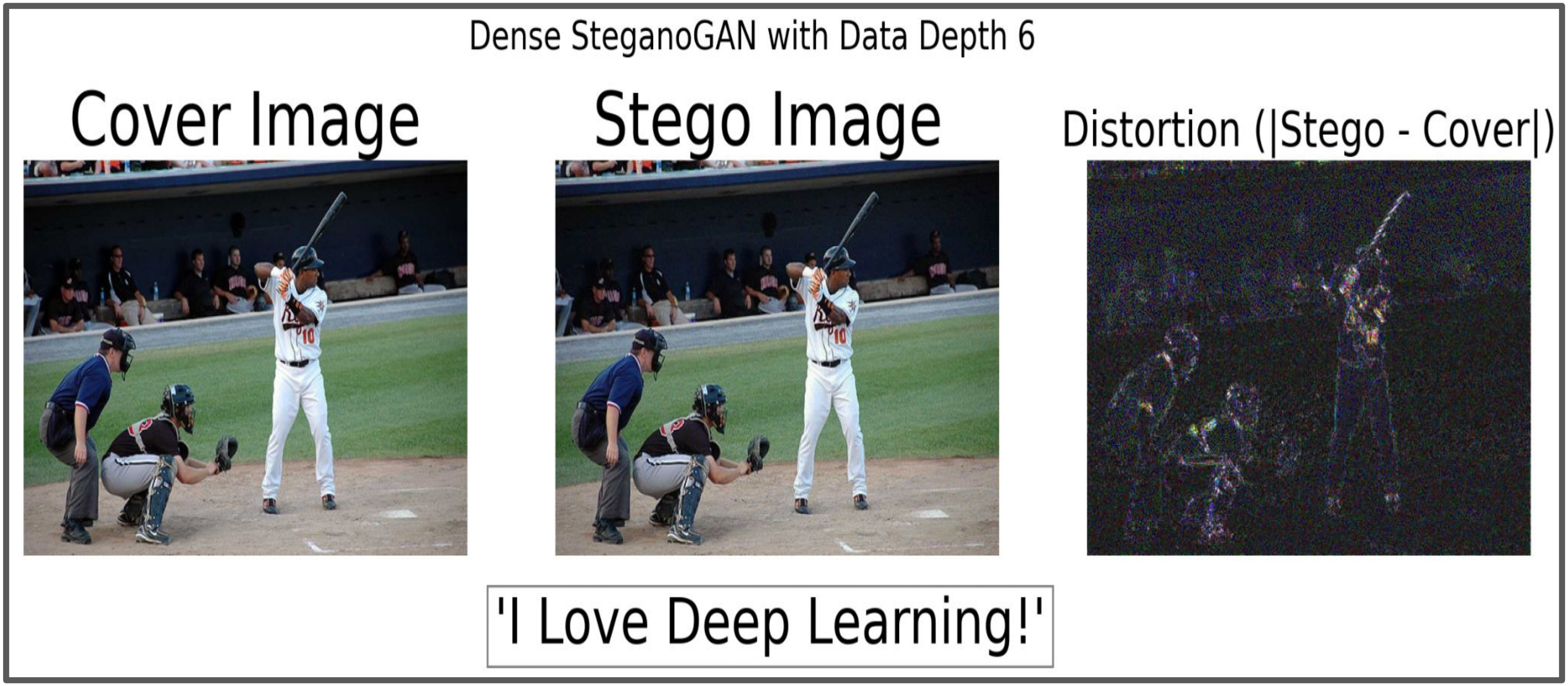
Introduction

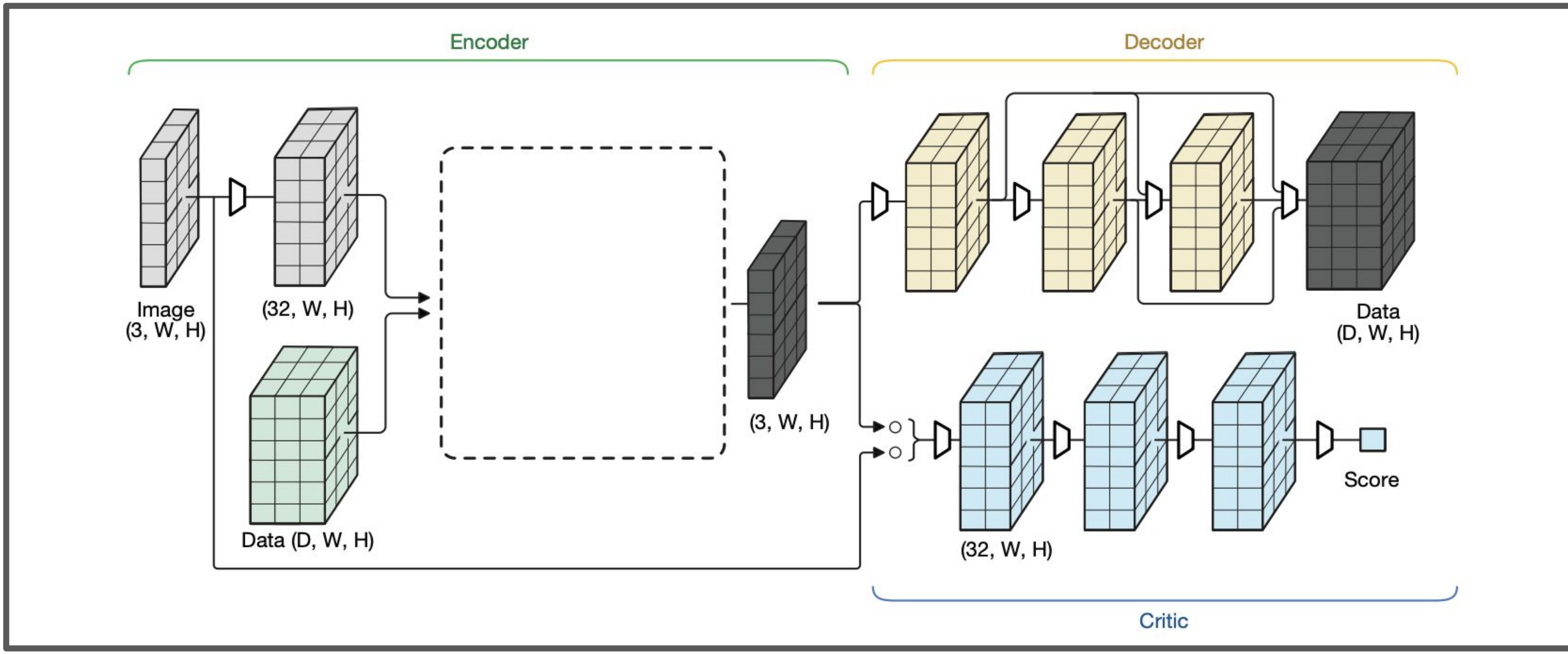
Steganography: Hiding data within an **unencrypted** message (image).



- Cryptography might face **legal restrictions** or **invite attackers**.
- Steganography is an alternative for **medical data** and **copyright**.
- Goals: Send **more information**, that is **undetectable** & **lossless.**
  - Undetectable to critic networks & lossless given error correcting.

Original Paper Contributions

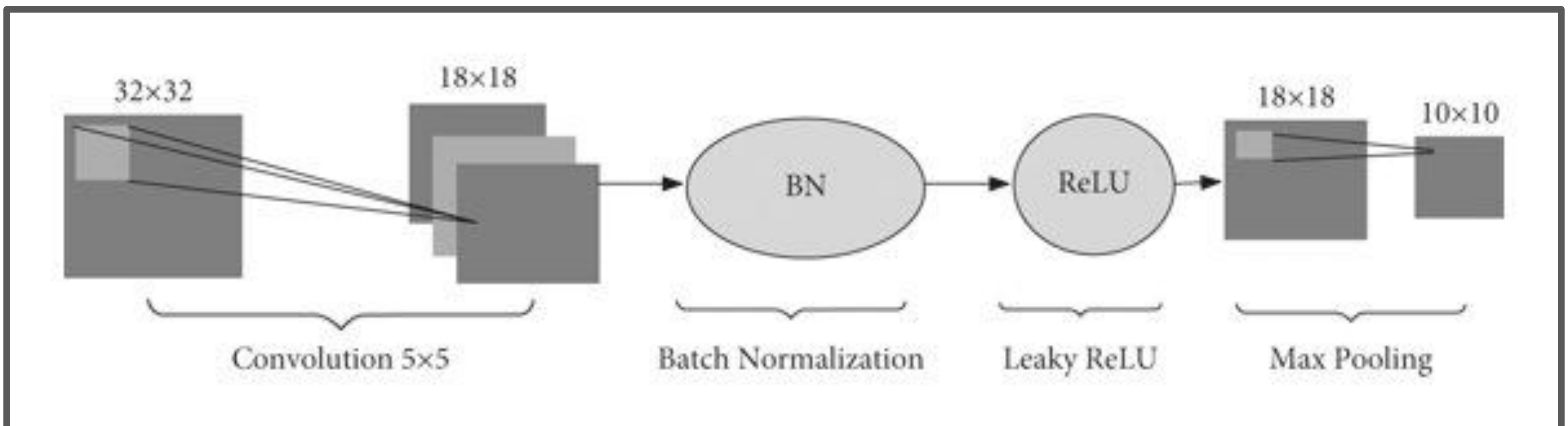
GAN-like Network Architecture (Encoder, Decoder, Critic), [1]



- Uses **adversarial training** (GAN) to train a steganography network.
- Develops **RS-BPP** for evaluating the data capacity of a network.
- Achieves **4.4 RS-BPP** (# data bits that can be stored per pixel).
- Evades traditional steganography detection [4] with **0.59 auROC.**

Our Hypothesis/Goals

- We aimed to make a working SteganoGAN trained on Div2K dataset.
- RS-BPP is a statistic based on the **mean accuracy** of the network.
    - Wanted to investigate if **a large variance** could affect RS-BPP.
- Paper claims Div2K < COCO (datasets) due to **content differences**.
    - We believe that performance differences are due to **image size**.
- Observe the effect of various perturbations:
    - Doubling the **training epochs** (32 → 64)
    - Increasing **data depth** trained on (# of bits/pixel encoded)
    - Switching order of **LeakyReLU** and **BatchNorm** to below [5].



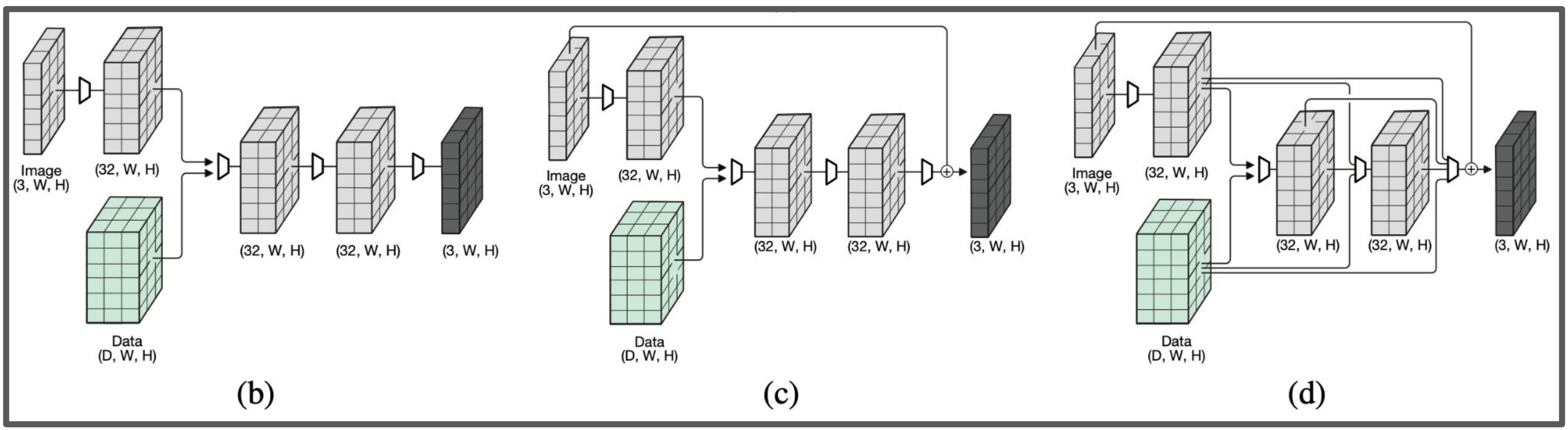
SteganoGAN Perturbation Tests

Mohammad, Alperen, Camilo, Aidan - Cornell

Methodology

Network Architecture:

- Encoder (Basic/Residual/Dense): **Image + Random Data → Image**.
- Decoder (Dense): **Image → Recovered Data** (using Reed-Solomon).
- Critic (Basic): **Image → Realism Score** (higher is more real).



Basic, Residual, and Dense Encoder Versions, [1]

2 Phase Training: Encoder/Decoder (freeze Critic) → Critic (freeze rest).

Datasets: Div2K (Higher Quality, ~1000), COCO (Lower Quality, ~330k).

Div2K Image, [2]

COCO Image, [3]

Modifications

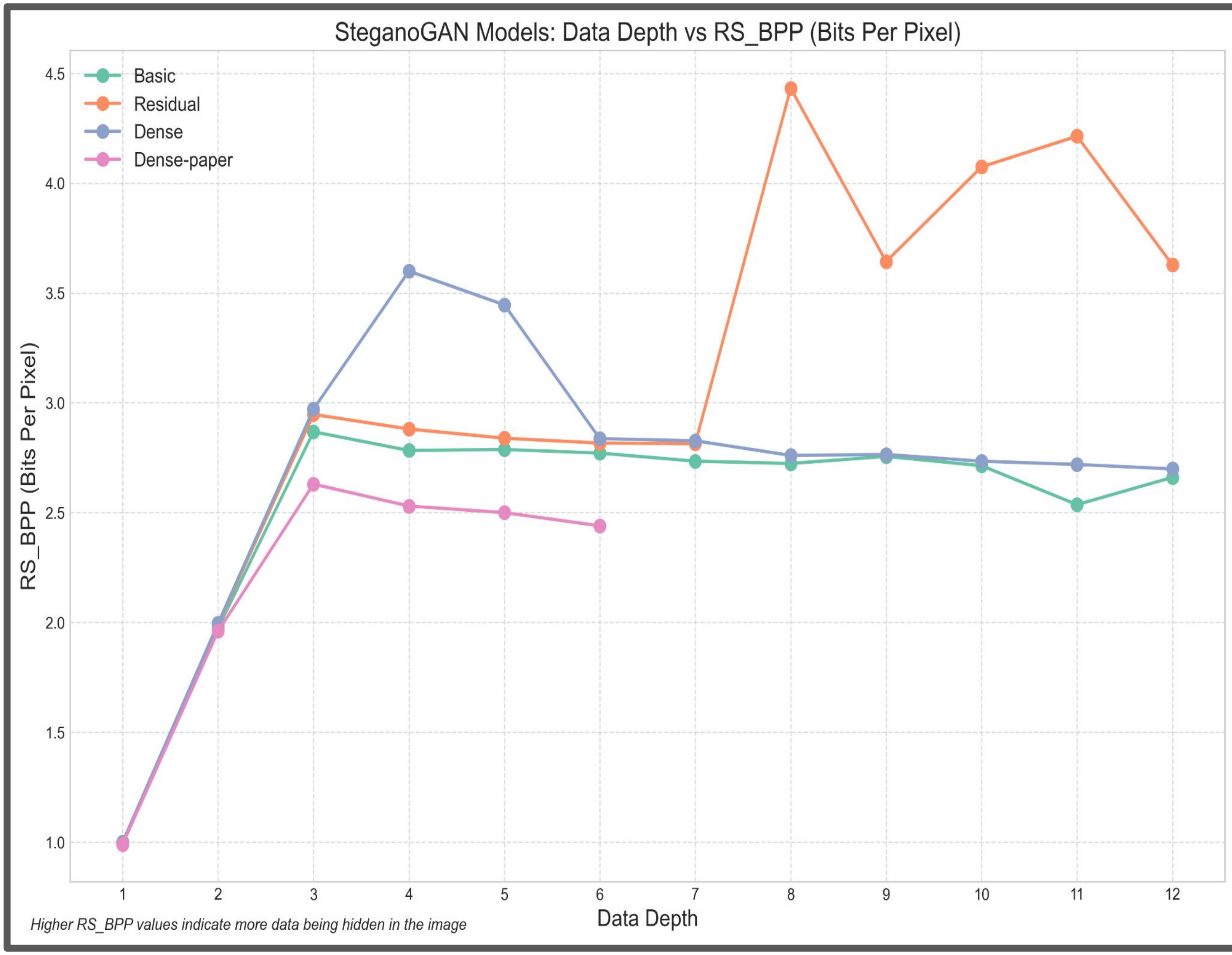
Modern Network Architecture:

- Optimizer: **Adam → AdamW.**
- Image Normalization: **Manual [-1, 1] → Pillow [0, 1].**

Reduce Training Time:

- Train on only **4x compressed Div2K**, test on both Div2K and COCO.
- Test evasion with **traditional steganalysis** tools (StegExpose).

Perturbation Results



RS-BPP at Increasing Data Depths

	Normal	Leaky	Long
	Accuracy		
1	1.00	1.00	1.00
2	1.00	1.00	1.00
3	1.00	0.99	1.00
4	0.95	0.98	0.98
5	0.84	0.91	0.79
6	0.74	0.80	0.74
	RS-BPP		
1	1.00	1.00	1.00
2	2.00	1.99	2.00
3	2.97	2.93	2.98
4	3.60	3.81	3.88
5	3.45	4.06	2.93
6	2.84	3.58	2.92

Dense Perturbation

- Leaky achieves a **significantly better RS-BPP** than other networks.
- Residual surprisingly seems to **outperform Dense** at larger depths.

Metric Comparison

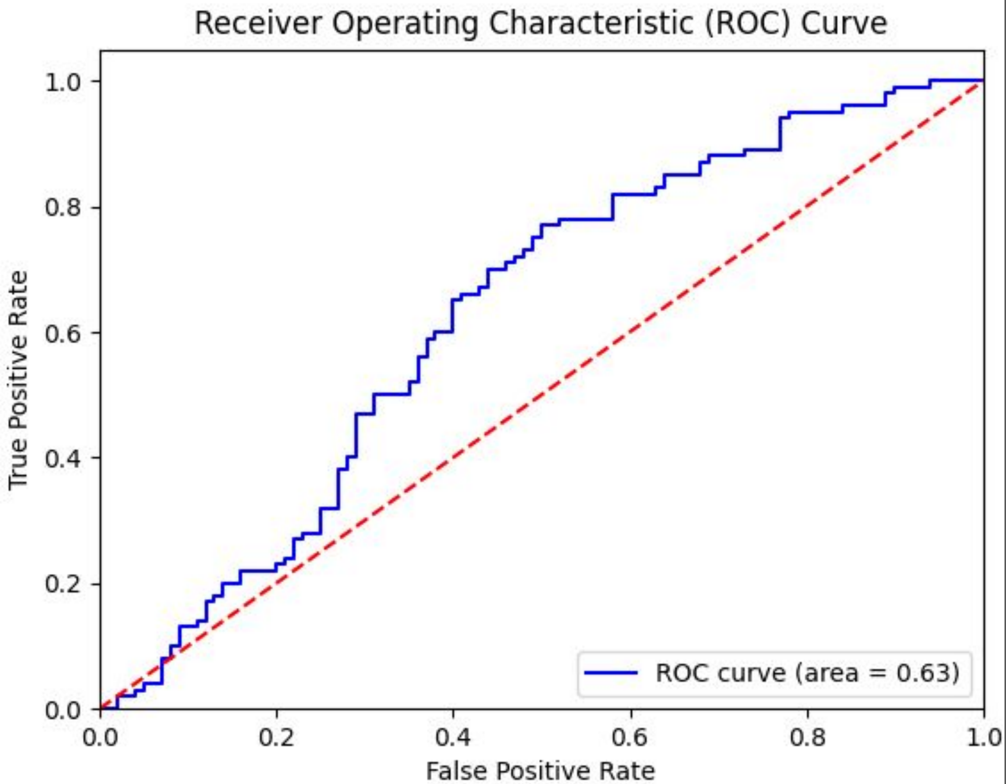
Dataset	D	Accuracy			RS-BPP			PSNR			SSIM		
		Basic	Resid.	Dense	Basic	Resid.	Dense	Basic	Resid.	Dense	Basic	Resid.	Dense
Div2K pretrained	1	1.00	1.00	1.00	1.00	1.00	1.00	24.42	41.44	43.06	1.00	1.00	1.00
	2	0.99	1.00	1.00	1.97	1.99	2.00	27.23	38.55	37.34	1.00	1.00	1.00
	3	0.98	0.99	1.00	2.87	2.95	2.97	20.33	33.64	34.38	0.99	1.00	1.00
	4	0.85	0.86	0.95	2.78	2.88	3.60	18.29	34.64	33.64	0.99	1.00	1.00
	5	0.78	0.78	0.84	2.79	2.84	3.45	24.19	36.21	34.30	1.00	1.00	1.00
	6	0.73	0.73	0.74	2.77	2.82	2.84	28.62	35.99	35.13	1.00	1.00	1.00
COCO	1	1.00	1.00	1.00	0.99	1.00	1.00	23.78	39.45	40.70	1.00	1.00	1.00
	2	0.99	0.99	1.00	1.96	1.97	1.99	25.80	37.01	35.78	1.00	1.00	1.00
	3	0.97	0.98	0.99	2.80	2.88	2.95	19.75	31.91	33.05	0.99	1.00	1.00
	4	0.84	0.85	0.93	2.69	2.77	3.45	18.08	33.02	32.46	0.99	1.00	1.00
	5	0.77	0.77	0.83	2.67	2.70	3.29	23.25	34.75	32.78	1.00	1.00	1.00
	6	0.72	0.72	0.73	2.65	2.69	2.71	27.01	34.52	33.65	1.00	1.00	1.00

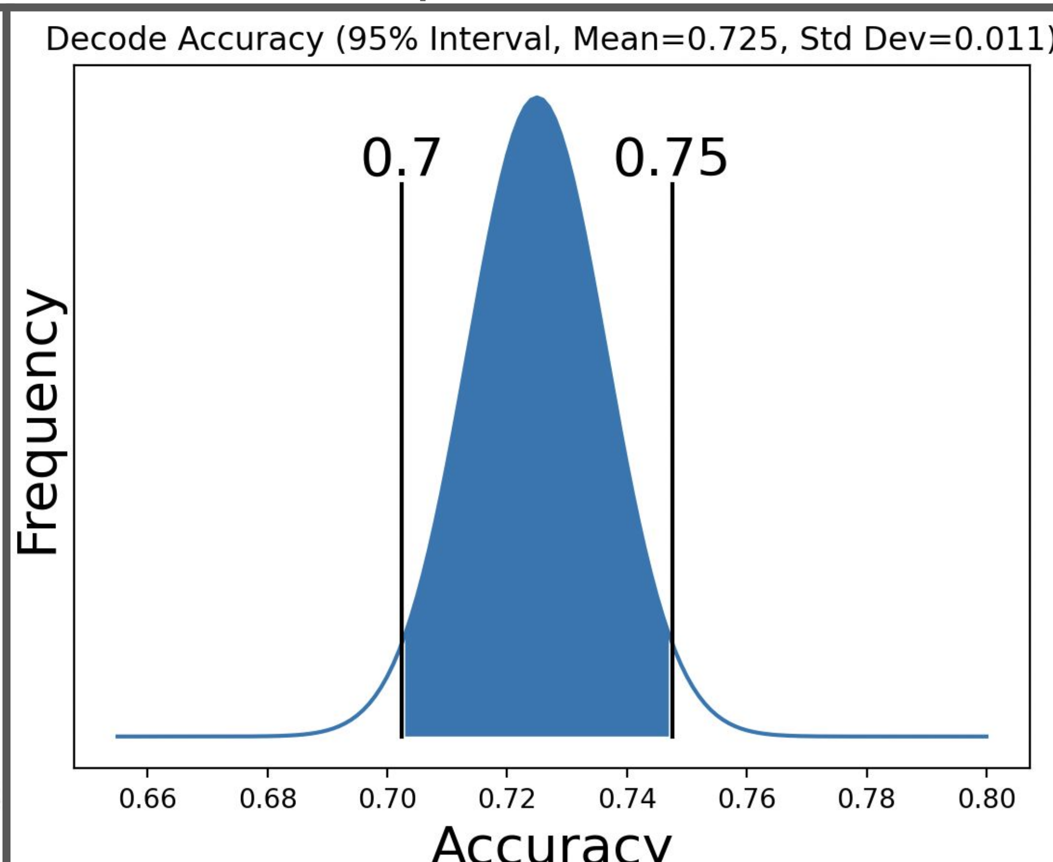
Our Network vs Paper SteganoGAN (Green = Our > Paper)

- **Superior in Div2K**, likely due to 4x compression/network changes.
- **Worse in COCO**, likely due to the network being trained on Div2K.
- PSNR and SSIM are less comparable due to data normalization.

Further Analysis

Dense SteganoGAN with Data Depth 6





Perfect = **1** vs Random = ½
Paper **0.59** vs Our **0.63**

95% Confidence: **0.725** → **0.703**
Changes RS-BPP: **2.7** → **2.436**

- Our network did not differ significantly in auROC from the paper.
- Ensuring 95% confidence interval can affect **RS-BPP by ~10%**.
  - **>>100** test images may lead to differences being less significant.

Final Takeaways

- GAN's can produce images with enough hidden data to consistently pass messages around while being mostly undetectable.
- Our model:
  - **Performed similar** to the original SteganoGAN on noisy metrics.
  - Explored various perturbations and **found significant trends**.
  - Likely **does not generalize** across data sets other than 4x Div2K.
- Training at a larger scale is required to certify our results, and confirm the methods to enhance undetectability and relative payload.

[1] SteganoGAN: https://arxiv.org/abs/1901.03892

[2] Div2K: https://ieeexplore.ieee.org/document/8014884

[3] COCO: https://arxiv.org/abs/1405.0312

[4] Steg Analysis Tool: https://github.com/b3dk7/StegExpose

[5] LeakyReLU Image: https://researchgate.net/356162640