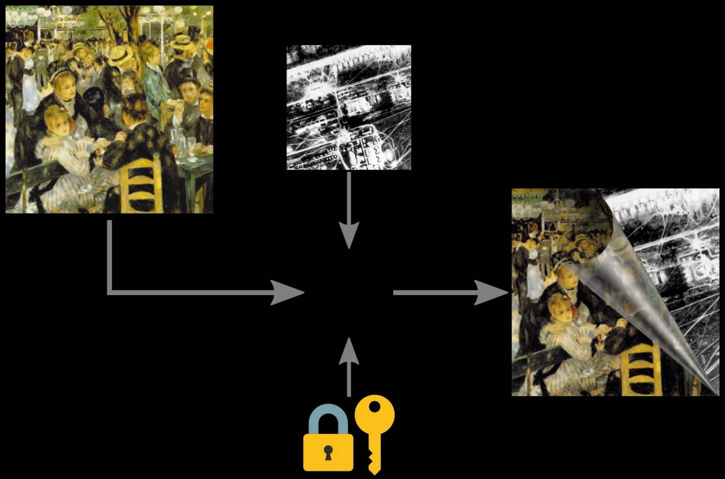
Image Steganalysis

by Michael Dorkenwald & Sebastian Gruber

Motivation

Steganalysis is the process of detecting hidden information in images

- Typically used in espionage, thus important for law enforcement
- Prone to false positives → Steganalysis networks used to “prefilter”
- Inherently difficult task as cover images are not provided



Retrieved from https://www.kaggle.com/c/alaska2-image-steganalysis (20.07.20)

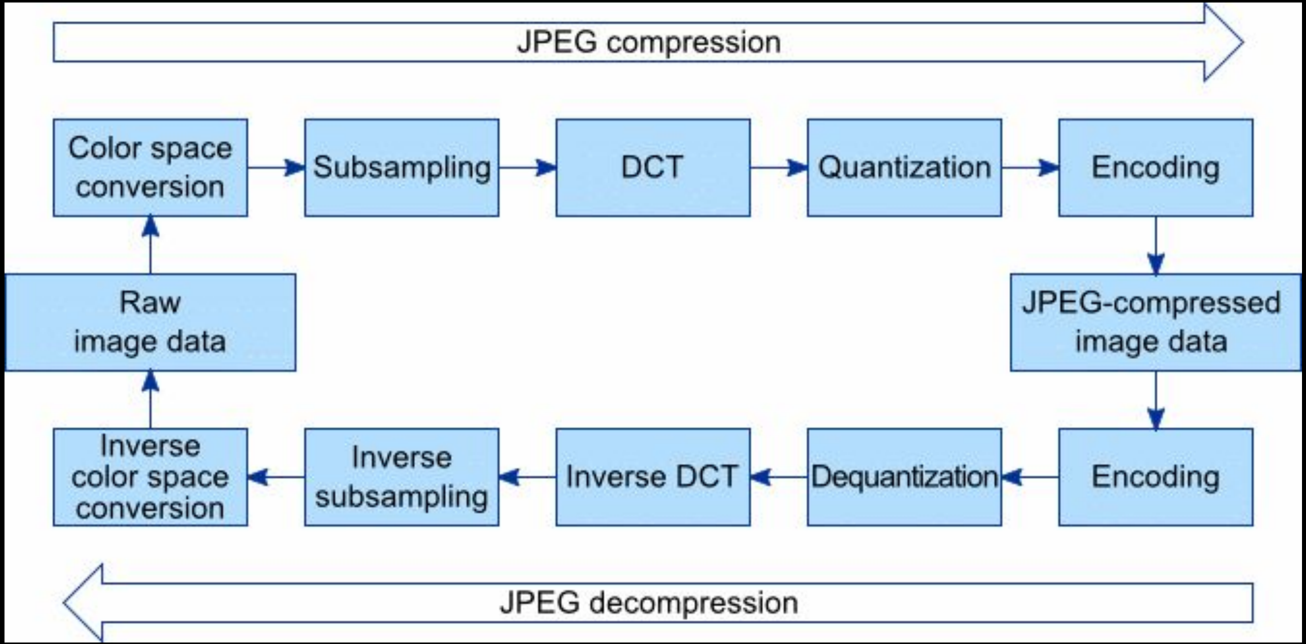
Our Task

- Distinguish manipulated from original images
    - Given a single image **classify** between manipulated and cover

      → *Achieve a low rate of false positives*

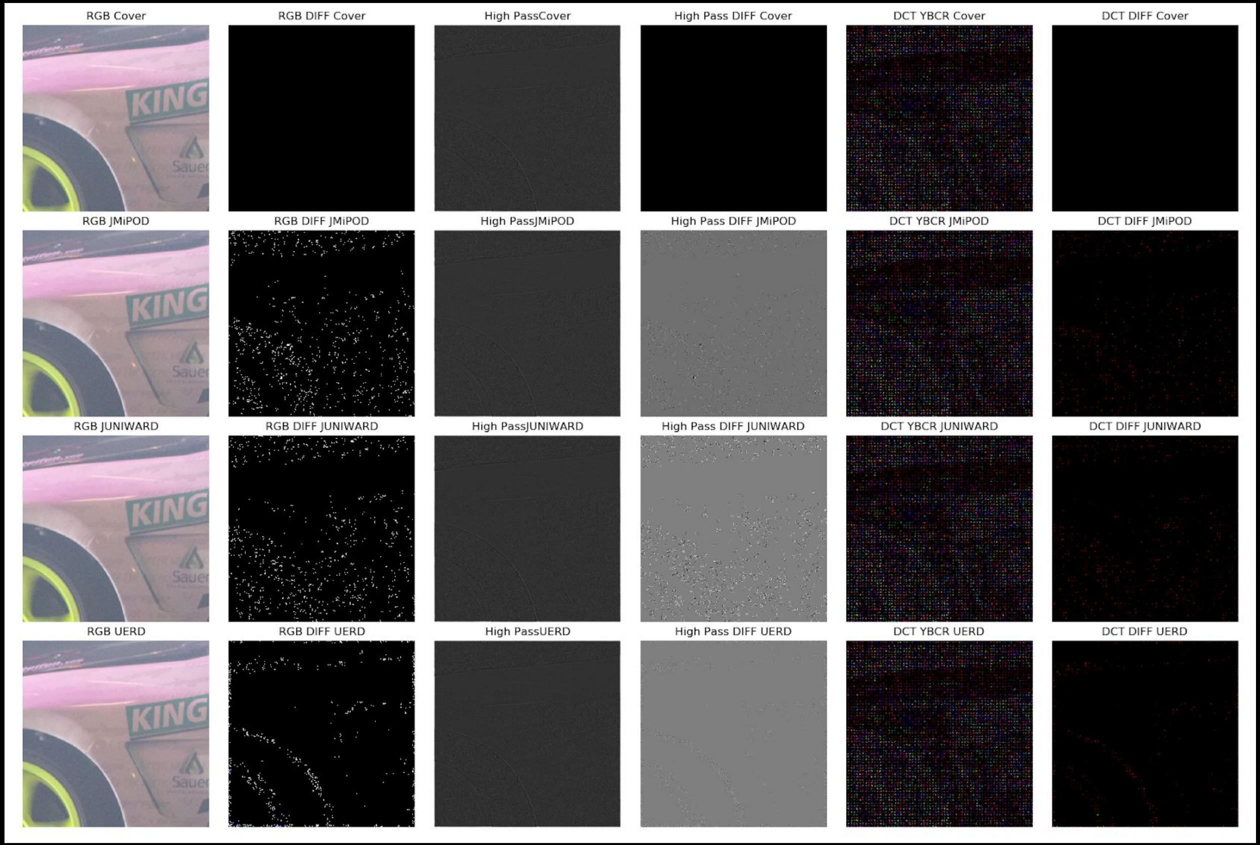
- On a diverse dataset:
    - Different acquisition settings
    - Jpeg compressions (95, 90, 75)
    - Steganography algorithms (JUNIWARD, JMiPOD, UERD)

Jpeg compression



Retrieved from: https://www.graphicsmill.com/docs/gm/working-with-jpeg.htm (21.07.20)

Visualization



Visualization

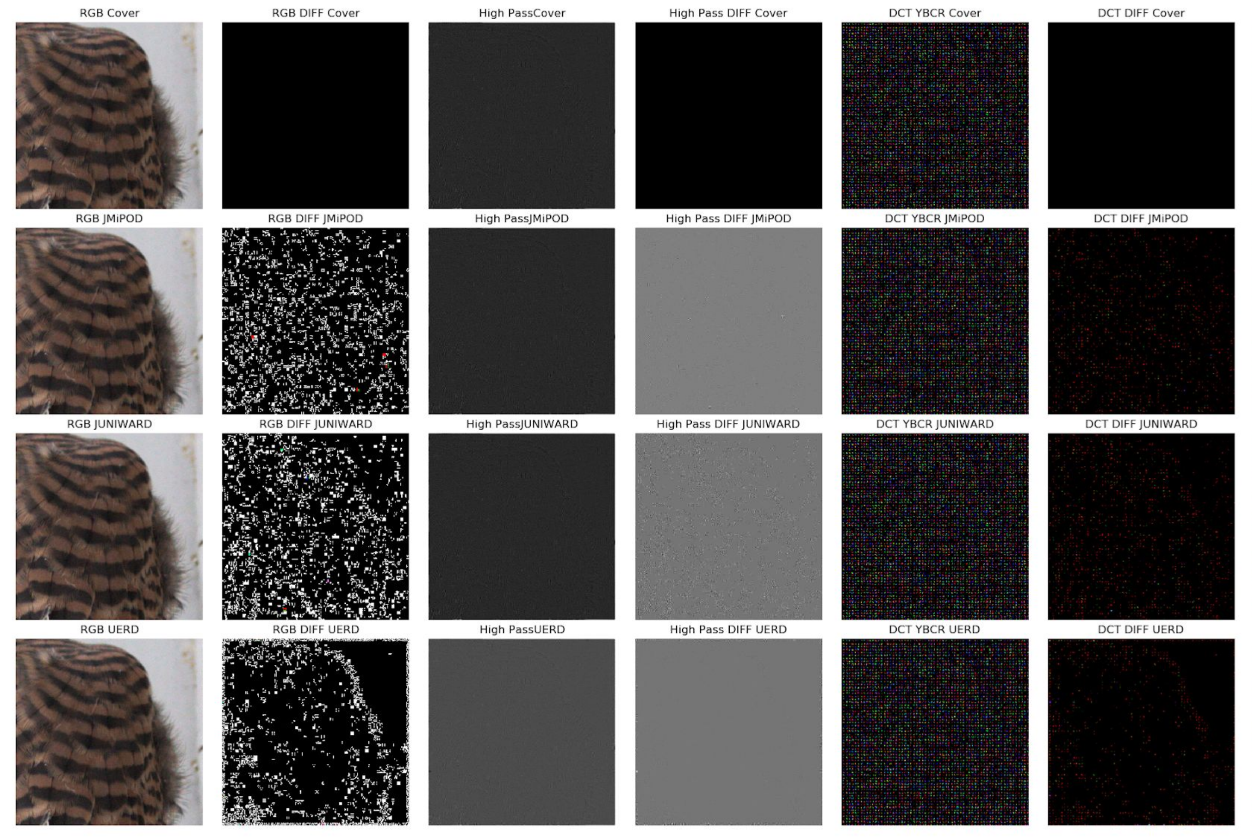


Image Difference JPEG Compression

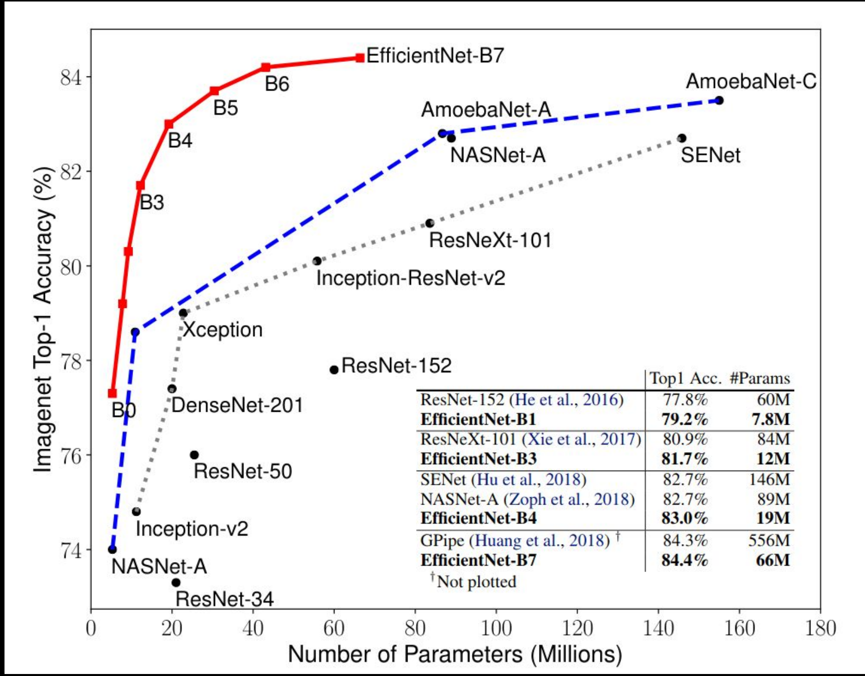
	JMiPOD	JUNIWARD	UERD
JPEG 75	45.1	29.5	26.6
JPEG 90	50.5	35.5	35.0
JPEG 95	37.2	35.7	36.3

➔ Training on 12 classes to better explain data distribution

Model Architecture

Use EfficientNet [1] CNN architecture

- Uses neural architecture search to design architecture
- 8.4x smaller and 6.1x faster than best existing network
- Transfers well to other datasets



[1] Tan et al. ICML 2019 https://arxiv.org/abs/1905.11946

Label Smoothing

- Improves generalization and learning speed [1]
- Prevents the network of being overconfident [1]
- Used with cross entropy loss:

$$H(\boldsymbol{p}, \boldsymbol{y}) = \sum_{k=1}^{K} -y_k \log(p_k)$$

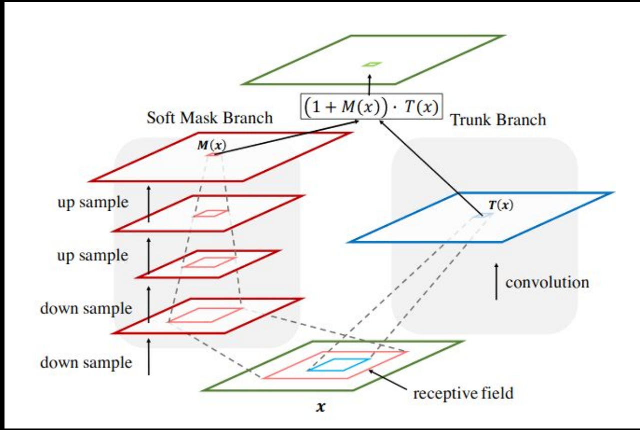
- Target class is modified such:

$$y_k^{LS} = y_k(1-\alpha) + \alpha/K$$

[1] Müller et al. Neurips 2019 https://arxiv.org/abs/1906.02629

Attention

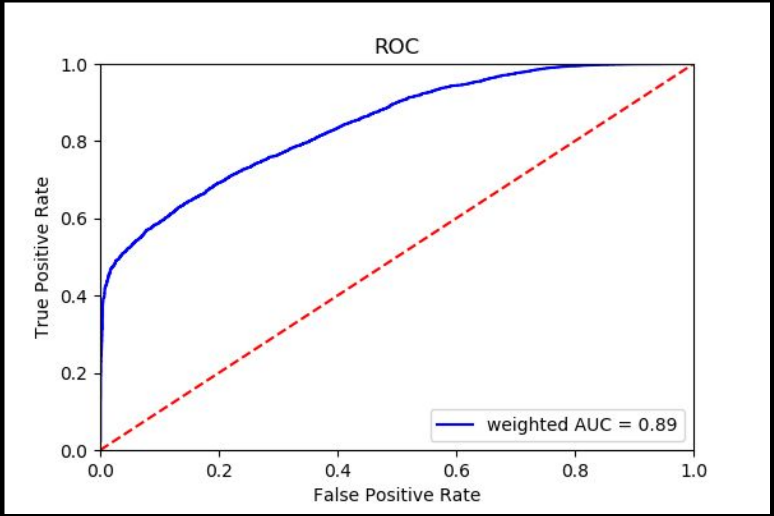
- [2] used attention to improve image classification
- Instead of learning the attention mask (left) we use high-pass filtered image



[2] Wang et al. CVPR 2017 https://arxiv.org/abs/1704.06904

Evaluation - Weighted AUC

- Performance measure for classification problem at all thresholds settings
- Defines networks capability to distinguish between classes
- Weighted AUC where false positives are stronger weighted



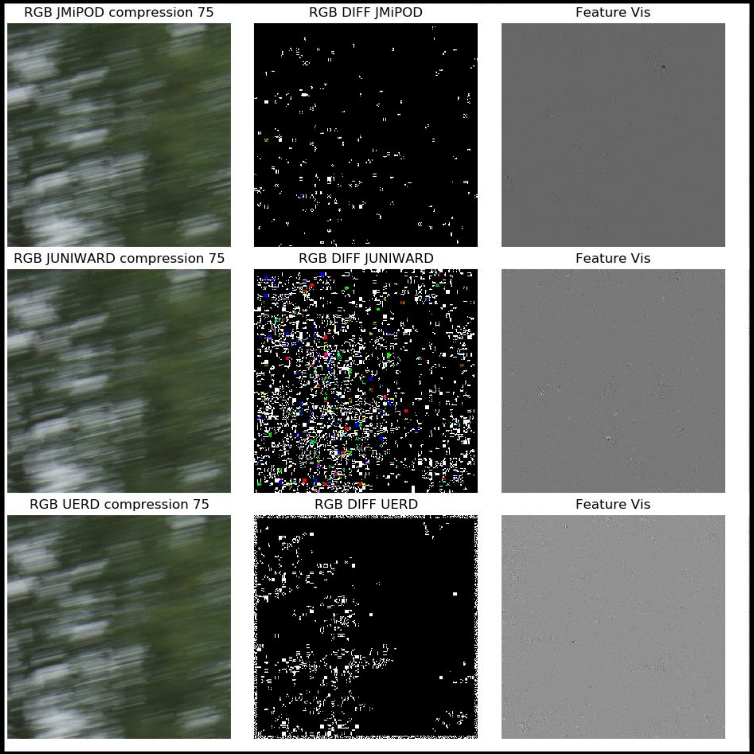
Evaluation

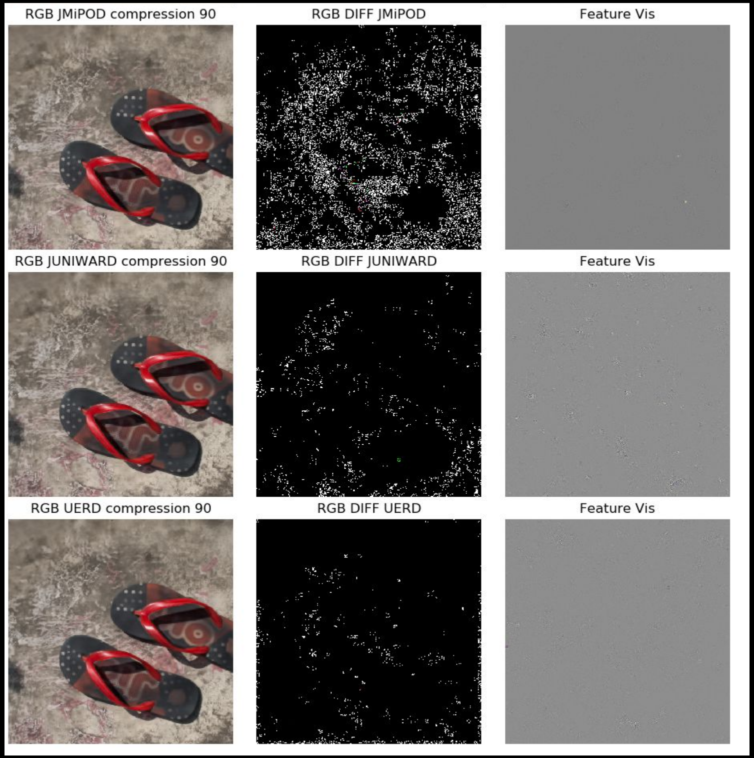
	Pretrained	Smoothing	Attention	weighted AUC		binary Accuracy	
				4 classes	12 classes	4 classes	12 classes
EfficientNet B0	✗	✗	✗	0.58	0.57	54 %	54%
EfficientNet DCT	✗	✗	✗	0.56	0.55	53 %	53%
EfficientNet B0	✓	✗	✗	0.876	0.884	72.4 %	73.1 %
EfficientNet B0	✓	0.05	✗	0.870	0.885	71.6 %	73.3 %
EfficientNet B0	✓	0.1	✗	0.865	0.887	71.1 %	73.4 %
EfficientNet B0	✓	0.2	✗	0.843	0.891	69.2 %	74.5 %
EfficientNet B0	✓	0.2	✓	---	0.877	---	72.8 %
EfficientNet B3	✓	0.2	✗	---	0.888	---	74.2 %

Evaluation 4 and 12 Classes

Accuracy in %	Cover			JMiPOD			JUNIWARD			UERD		
	75	90	95	75	90	95	75	90	95	75	90	95
Model 4 classes	93			20			61			63		
Model 12 classes	96	93	94	25	31	26	92	85	39	64	70	59

Feature Visualization via GBP [1]





[1] Springenberg et al. ICLR 15. https://arxiv.org/abs/1412.6806

Conclusion

- In depth analysis of the problem
- Effective approach to detect hidden messages in images
- Extensive ablation study with feature visualization
- Outlook:
  - Train seperate network for JMIPOD algorithm
  - Pretrain EfficientNet for training on DCT coefficients
  - Alternative preprocessing ?