LSSD: a Controlled Large JPEG Image Database for Deep-Learning-based Steganalysis

Hugo RUIZ[1], Mehdi YEDROUDJ[1], Marc CHAUMONT[1,2],
Frédéric COMBY[1], Gérard SUBSOL[1]
[1]Research-Team ICAR, LIRMM, Univ. Montpellier, CNRS, France;
[2]Univ. Nîmes, France

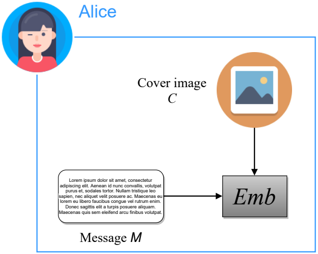
hugo.ruiz@lirmm.fr

January 11, 2021

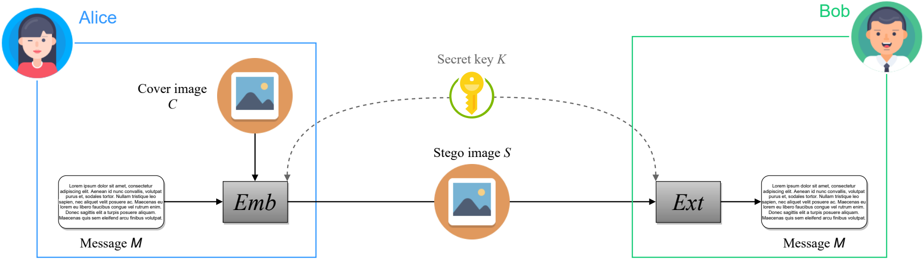
ICPR'2021, International Conference on Pattern Recognition,
MMForWILD'2021, Worshop on MultiMedia FORensics in the WILD,
Lecture Notes in Computer Science, LNCS, Springer.

January 10-15, 2021, Virtual Conference due to Covid (formerly Milan, Italy).

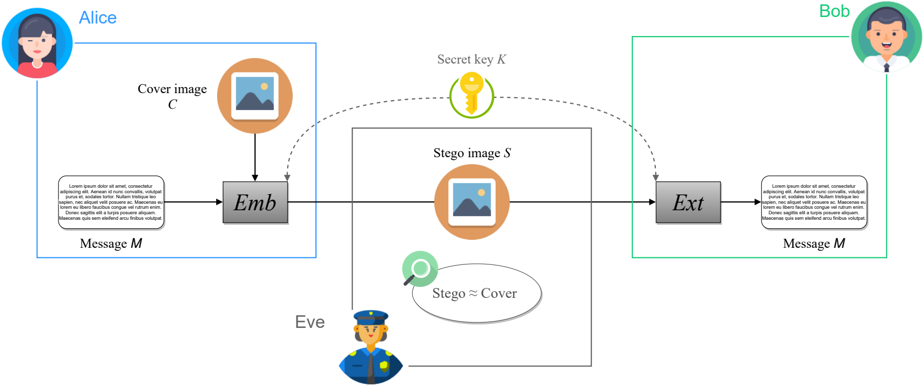
Steganography / Steganalysis



Steganography / Steganalysis



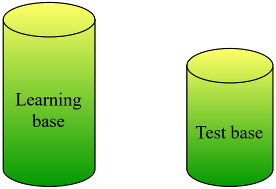
Steganography / Steganalysis



Efficient methods in steganalysis

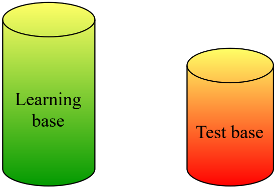
- First were based on Machine Learning [KCP13]
- Deep Learning improved performance
- But it requires a significant training database

The mismatch problem



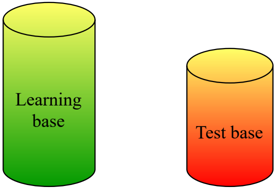
The mismatch problem

- Differences between learning base and test base
  - different image sources (Camera & ISO)
  - differences in *development* parameters (RAW $\rightarrow$ JPEG)



The mismatch problem

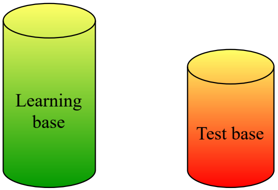
- Differences between learning base and test base
  - different image sources (Camera & ISO)
  - differences in *development* parameters (RAW $\rightarrow$ JPEG)



- Reduce performance of the method

The mismatch problem

- Differences between learning base and test base
  - different image sources (Camera & ISO)
  - differences in *development* parameters (RAW $\rightarrow$ JPEG)



- Reduce performance of the method

Objective to work with mismatch

To be closer to the real world and have a network more robust

Limits of image bases currently used in steganalysis

- ALASKA v1 (2018) & v2 (2020): big base (up to 80k images) with a lot of diversity. *link here*
- BOSS (2010): a reference base but few images (10k). *link here*
- Dresden & RAISE : very few images (∼10k both combined). *RAISE link here*

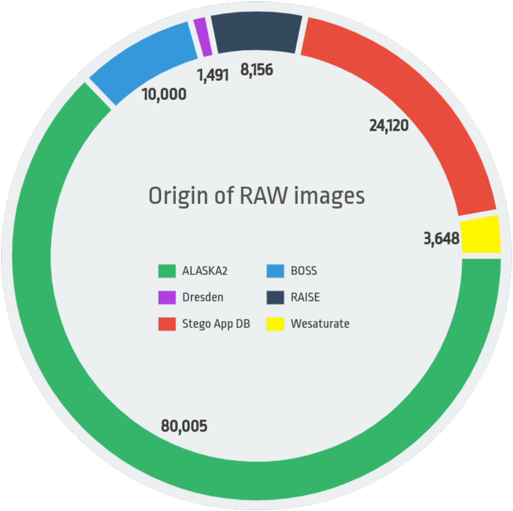
All these bases contain about 100,000 images in RAW format but some are no longer available for download.

Examples

Part of image from **ALASKA2**
256×256 JPEG image in colour

Part of image from **BOSS**
256×256 JPEG image in grayscale

Increase diversity: get as RAW images as possible



- Total: 127,420 images
- Size: $\sim 3000 \times 5000$
- Cameras: > 100

Analysis of diversity of the images

The ISO number is the value of sensitivity of the camera sensor.

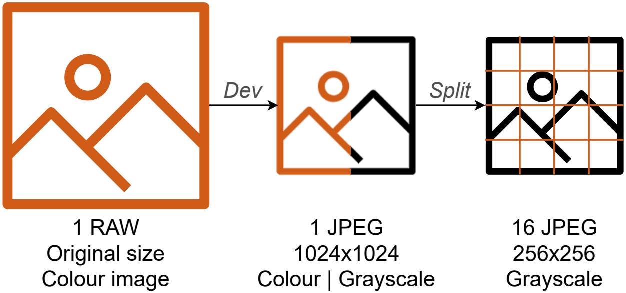
- ISO for ALASKA2 base:

	$ISO < 100$	$100 \leq ISO < 1000$	$1000 \leq ISO$
Number	12,497	55,893	11,615

- Camera models: > 100

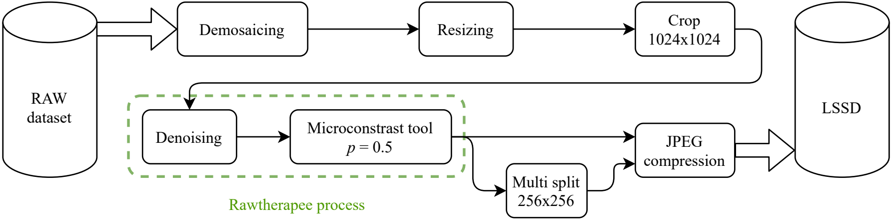
	ALASKA2	BOSS	Dresden	RAISE	Stego App
Number	40	7	25	3	26

Processing pipeline

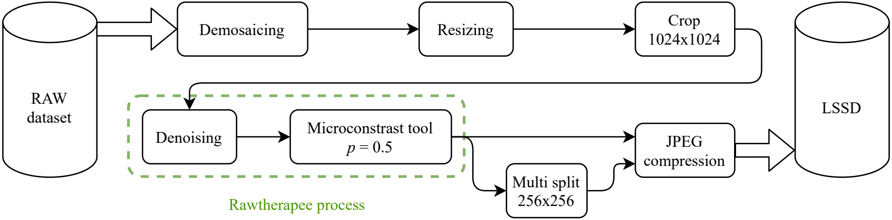


To obtain LSSD database: $127{,}420 \times 16 \simeq 2M$

Processing pipeline



Processing pipeline



With this method, it is possible to create 1024×1024 and 256×256 JPEG images.

During the Rawtherapee process, Unsharp Masking (USM) can be used and more values are available for the denoising.

Parameters for the LSSD database

Based on the script[CGB20] used for the development of the images of the ALASKA2 challenge[1].

Other parameters could be used:

- Demosaicing: AMAZE or IGV
- Resize: can be random
- Denoise: different range
- ...

Name	Value
Demosaicing	Fast or DCB
Resize & Crop	Yes
Size (resize)	1024×1024
Kernel (resize)	Nearest (0.2) Bicubic (0.5) Bilinear (0.3)
Resize factor	depends on initial size
Unsharp Masking	No
Denoise *(Pyramid Denoising)*	Yes
Intensity	[0; 60]
Detail	[0; 40]
Micro-contrast	Yes ($p = 0.5$)
Color	No
Quality factor	75

[1] `https://alaska.utt.fr/`

Quality Factor (QF), a key-parameter for JPEG images

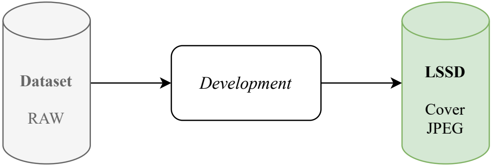
$QF \in [0, 100]$

- 0: very poor quality
- 50: minimum
- **75: our choice**
- 100: best quality available

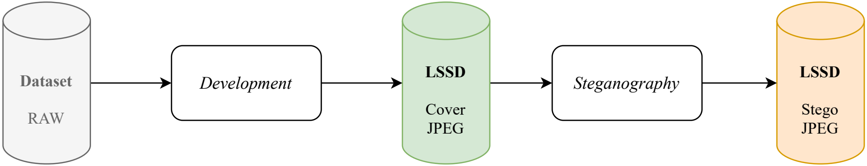
QF is linked to quantization matrices in DCT image compression.
We use only standard matrices in the first version of LSSD.

However, possible to increase diversity by changing QF.

Global process

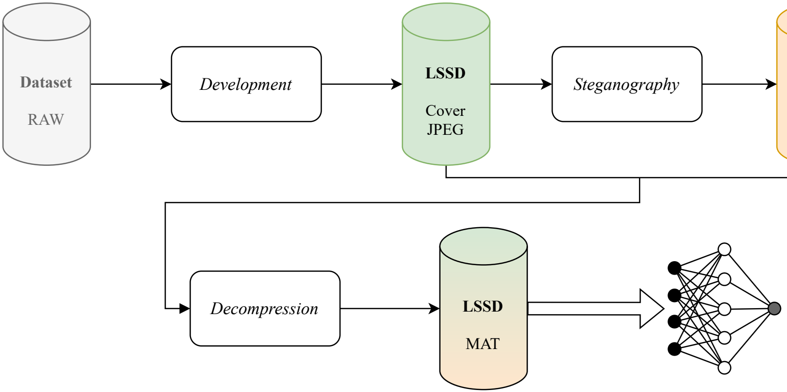


Global process



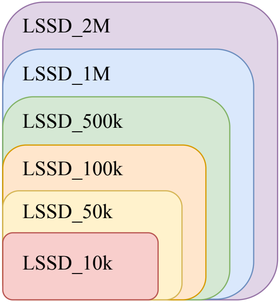
- Algorithm: J-UNIWARD [HF13]
- Payload: 0.2 bpnzacs
- Type: grayscale

Global process



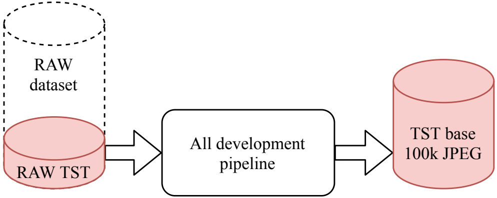
Building different bases

- With different sizes
- 6 bases: from 10k to 2M images
- The smaller ones are extracted from the larger ones
  $\Rightarrow$ same development
- Number of images = **cover images**
  Ex: LSSD_50k = 50k cover + 50k stego



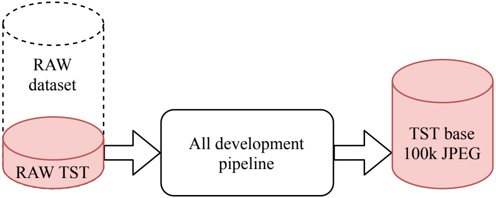
Test base

- Extract 6,250 RAW images from RAW dataset
- Same development pipeline (with same parameters)
- Test base of 100k images



Test base

- Extract 6,250 RAW images from RAW dataset
- Same development pipeline (with same parameters)
- Test base of 100k images



This test base is totally independent of the learning base

A new database: LSSD

http://www.lirmm.fr/~chaumont/LSSD.html

A new database: LSSD

A database of 2 million JPEG (color or grayscale) 256×256 images.

Different databases available for downloading separately.

http://www.lirmm.fr/~chaumont/LSSD.html

To download (access soon)

A shell script to download: `https://github.com/Yiouki/download_lssd`

Parameters:

- Base name (`-b`): LSSD_10k, 50k…or ALASKA2, BOSS…
- Type (`-t`): JPEG or MAT
- Coloring (`-c`): Color or Gray
- Nature (`-n`): Cover or Stego (Stego_P02 in our case)
- Output (`-o`): Choose the output path
- Help (`-h`): Print help

Example to download LSSD_10k gray cover images in MAT format: `sh LSSD_download_script -b LSSD_10k -t MAT -c Gray -n Cover`

Conclusions

- Steganalysis community can find multiple usages of this base:
  - scalability analysis [RCY+21]
  - learning bases exceeding millions of controlled images

- Lot of important steps to develop a complete database

- About one week to develop a complete and usable base.
  (Computer: 16 proc Intel Xeon(R) W-2145 3.70Ghz)

To be continued...

- Create and analyze the controlled mismatch between the training/learning and test base.
  With USM, denoising, demosaicing...

- Use different steganography algorithms with more payloads

- Use colour in the base

Références I

Rémi Cogranne, Quentin Giboulot, and Patrick Bas.
Challenge Academic Research on Steganalysis with Realistic Images.
In Proceedings of the IEEE International Workshop on Information Forensics and Security, WIFS'2020, Virtual Conference due to covid (Formerly New-York, NY, USA), December 2020.

Vojtech Holub and Jessica Fridrich.
Digital image steganography using universal distortion.
IH&MMSec 13 Proceedings of the first ACM workshop on Information hiding and multimedia security, 2013(1), 2013.

Sarra Kouider, Marc Chaumont, and William Puech.
Adaptive Steganography by Oracle (ASO).
In Proceeding of the IEEE International Conference on Multimedia and Expo, ICME'2013, pages 1–6, San Jose, California, USA, July 2013.

Références II

Hugo Ruiz, Marc Chaumont, Mehdi Yedroudj, Ahmed Oulad-Amara, Frédéric Comby, and Gérard Subsol.

Analysis of the Scalability of a Deep-Learning Network for Steganography "Into the Wild".

In Submited to MultiMedia FORensics in the WILD, MMForWILD'2020, in conjunction with ICPR2020 The 25th International Conference on Pattern Recognition, Virtual Conference due to covid (Formerly Milan, Italy), January 2021.