Reversible Fragile Medical Image Watermarking Scheme Resistant to Malicious Tampering Attacks

Victor Fedoseev, Anna Denisova
Samara National Research University,
Image Processing Systems Institute, RAS,
Samara, Russia,
vicanfed@gmail.com

Introduction to the Topic

Opposing Sides

- **Our global goal:** to protect medical images from tampering (unauthorized change)
- **Intruder's goal:** misdiagnosis

Basic Technique

- Fragile watermarking
- Watermark bits are distributed in a meaningful image area (region of interest, ROI)
- Correct watermark extraction confirms image authenticity
- Bit errors at watermark extraction help to localize tampering areas

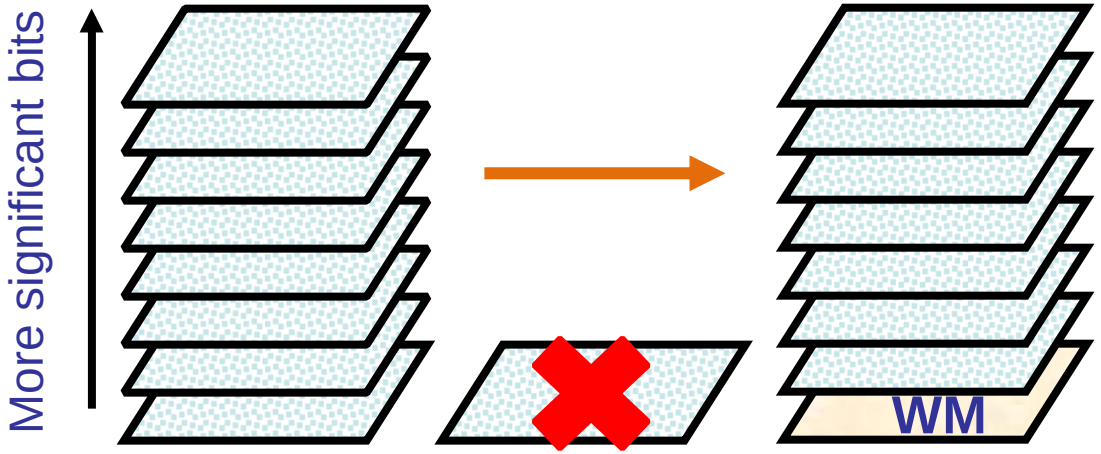
We concentrate **on the development of a specific fragile watermarking method able to eliminate shortcomings of traditional methods (based on LSB or QIM watermarking)**

Problem Statement.

🏆 LSB Watermarking vs. Blind Tampering Attack

LSB Medical Image Watermarking

Source image

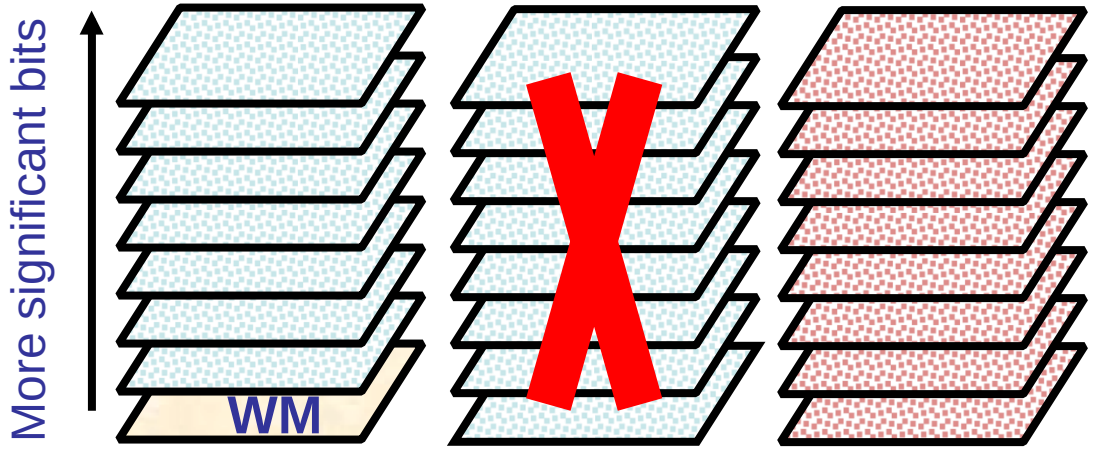


Protected image

Medical image protection by least significant bit watermarking

Blind Tampering Attack to Medical Image Watermarking

Protected image



Tampered image

"Blind" tampering attack destroys the watermark.

Hense, this tampering easily detected.

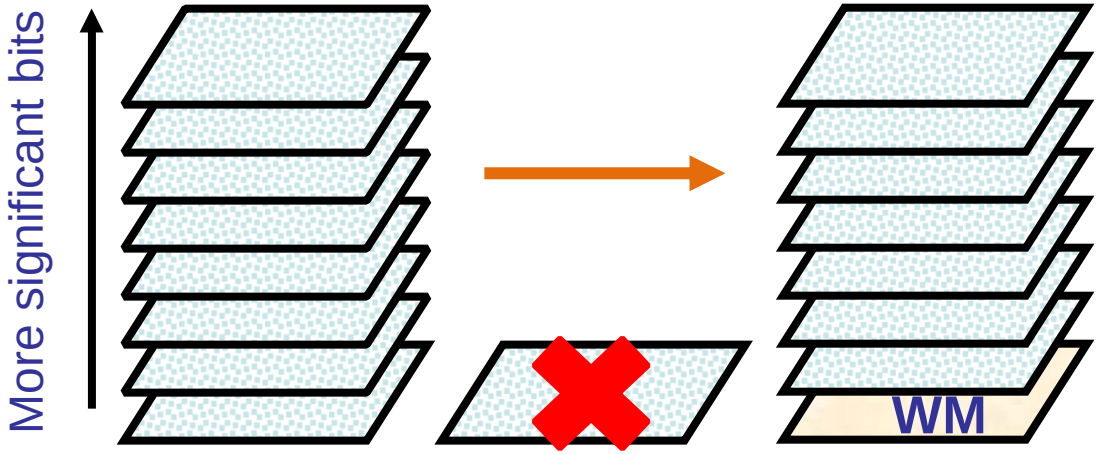
X-ray images by radiopaedia.org and by M. Galeziok et al, 2009

Problem Statement.

🏆 Malicious Tampering Attack vs. LSB Watermarking

LSB Medical Image Watermarking

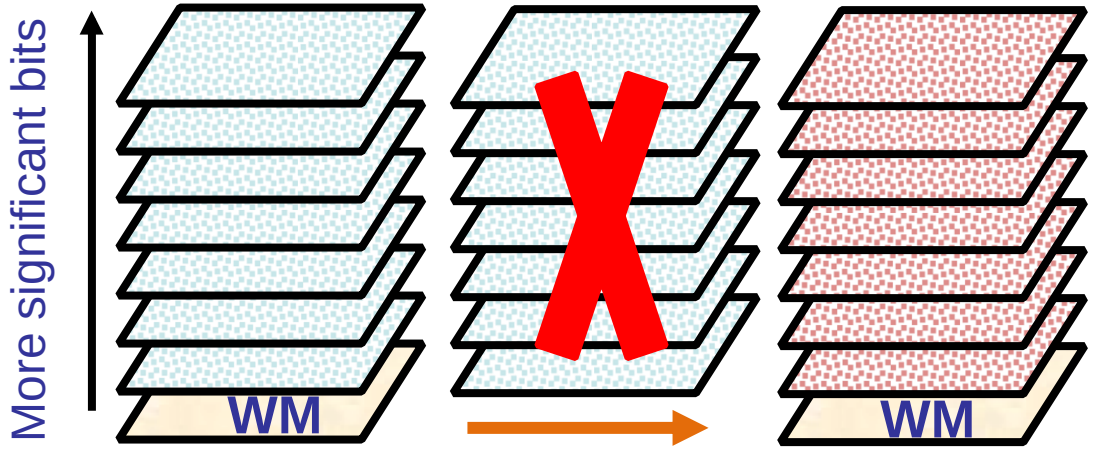
Source image



Protected image

Malicious Tampering Attack to Medical Image Watermarking

Protected image



Tampered image

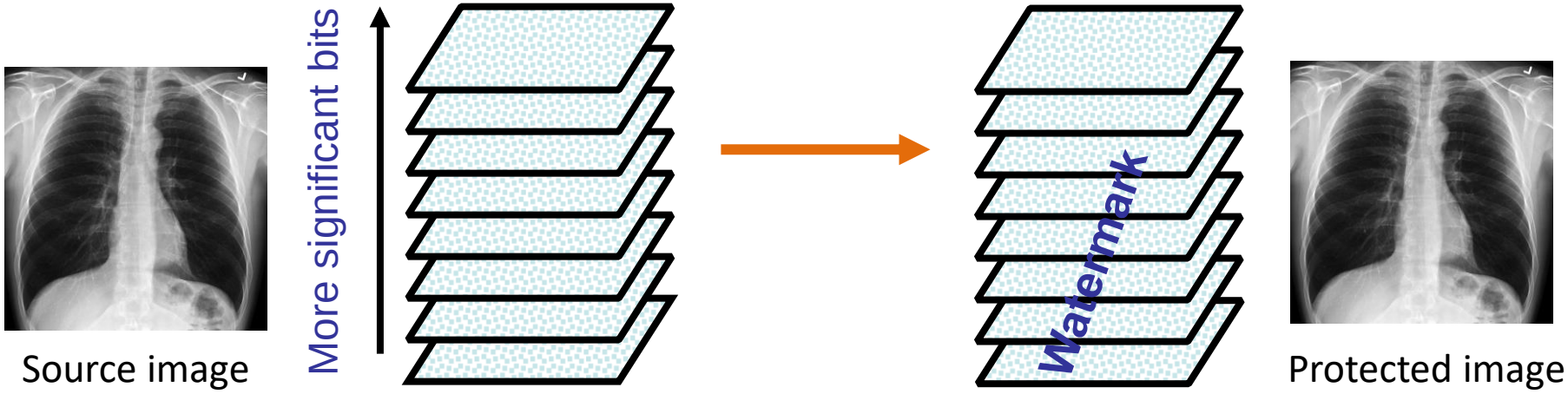
The image can be separated into the watermark and the meaningful part.

The latter can be replaced by a fake image and combined with the valid watermark

X-ray images by radiopaedia.org and by M. Galeziok et al, 2009

Problem Statement.
🏆 Our Method vs. Malicious Tampering Attack

Our Watermarking Approach

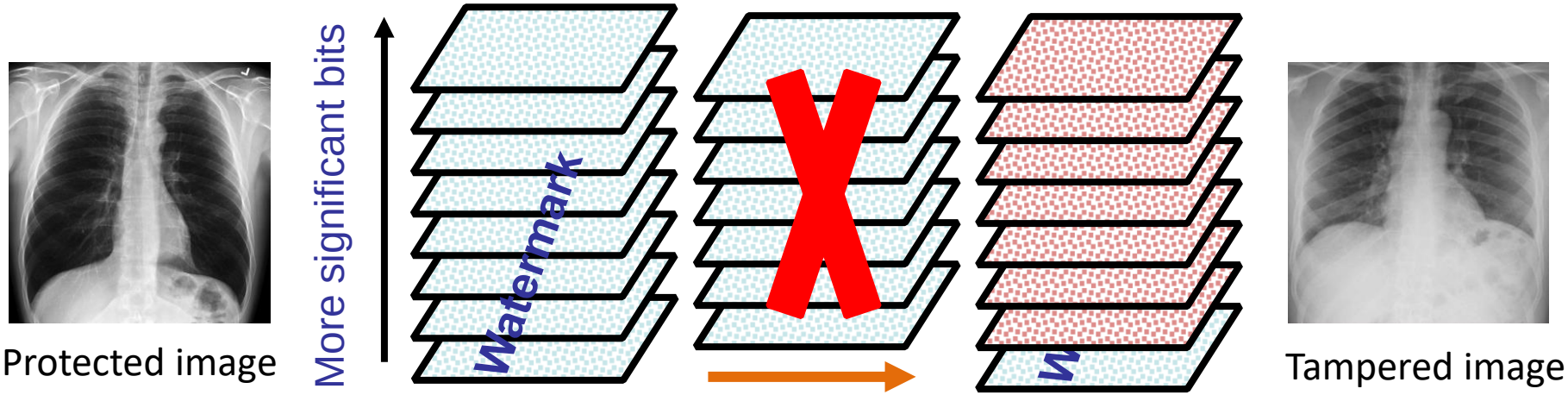


Source image

Protected image

Our watermarking method slightly influence on many bit planes

Malicious Tampering Attack to Medical Image Watermarking



Protected image

Tampered image

Due to the distribution of the watermark between various bit planes, the tampering operation causes many errors at watermark extraction

X-ray images by radiopaedia.org and by M. Galeziok et al, 2009

Known LSB- or QIM-based Watermarking Schemes

$$C^W = \left\lfloor \frac{C}{2\Delta} \right\rfloor + \text{Watermark } W \times \text{Mask of embedding positions} \times \Delta + C \pmod{\Delta}$$

Watermark W (pseudo-random binary data)

Mask of embedding positions ($E = 75\%$)

Malicious attack feasibility

- Δ is a scalar value. For LSB watermarking, $\Delta = 1$
- To attack this method, an intruder has to keep unchanged $C^W \pmod{2\Delta}$:

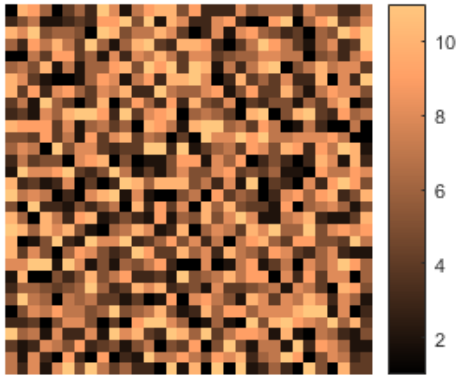
  $C^W := C^W + 2p\Delta$, where $p = \pm 1, \pm 2, \ldots$
- Second term can be very low!

Our Method

$$C^W = \left\lfloor \frac{C}{2\Delta} \right\rfloor + \text{Watermark } W \times \text{Mask of embedding positions} \times \Delta \text{ - matrix of multipliers for QIM-based watermarking} + C \pmod{\Delta}$$

Watermark W (pseudo-random binary data)

Mask of embedding positions ($E = 75\%$)



Δ - matrix of multipliers for QIM-based watermarking

Malicious attack feasibility

- Δ is a matrix of values from 1 to Δ_{max} determined by the secret key
- To attack this method, an intruder has to keep unchanged $C^W \pmod{2\Delta}$:

$$C^W := C^W + p \cdot LCM(2, \ldots, 2\Delta_{max}), \text{ where } p = \pm 1, \pm 2, \ldots$$

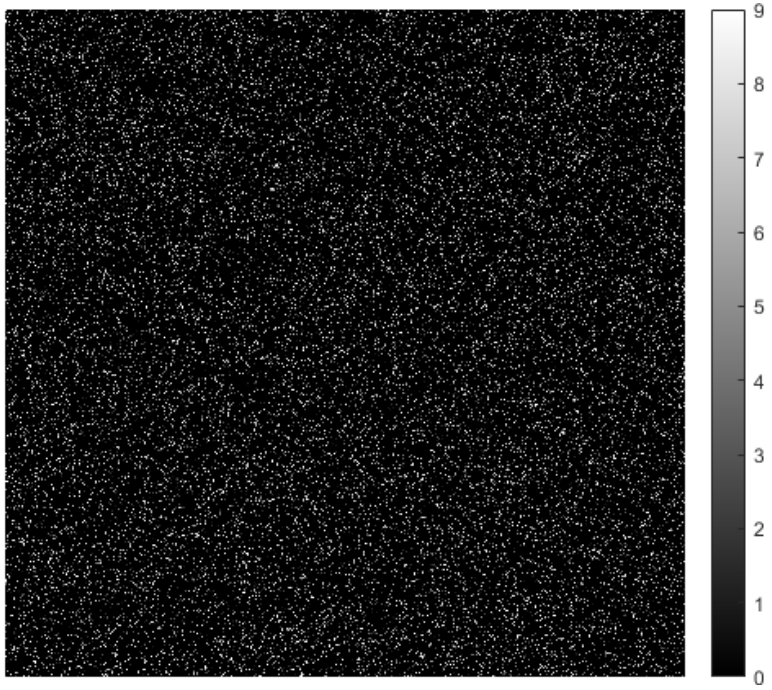
LCM is least common multiple. **It grows much faster!**

- For $\Delta_{max} = 5$, $LCM(2, \ldots, 2\Delta_{max}) = 120$ – it is enough for 8-bit images
- For $\Delta_{max} = 11$, $LCM(2, \ldots, 2\Delta_{max}) = 55440$ – it is enough for 16-bit medical images

Embedding Example

Source image

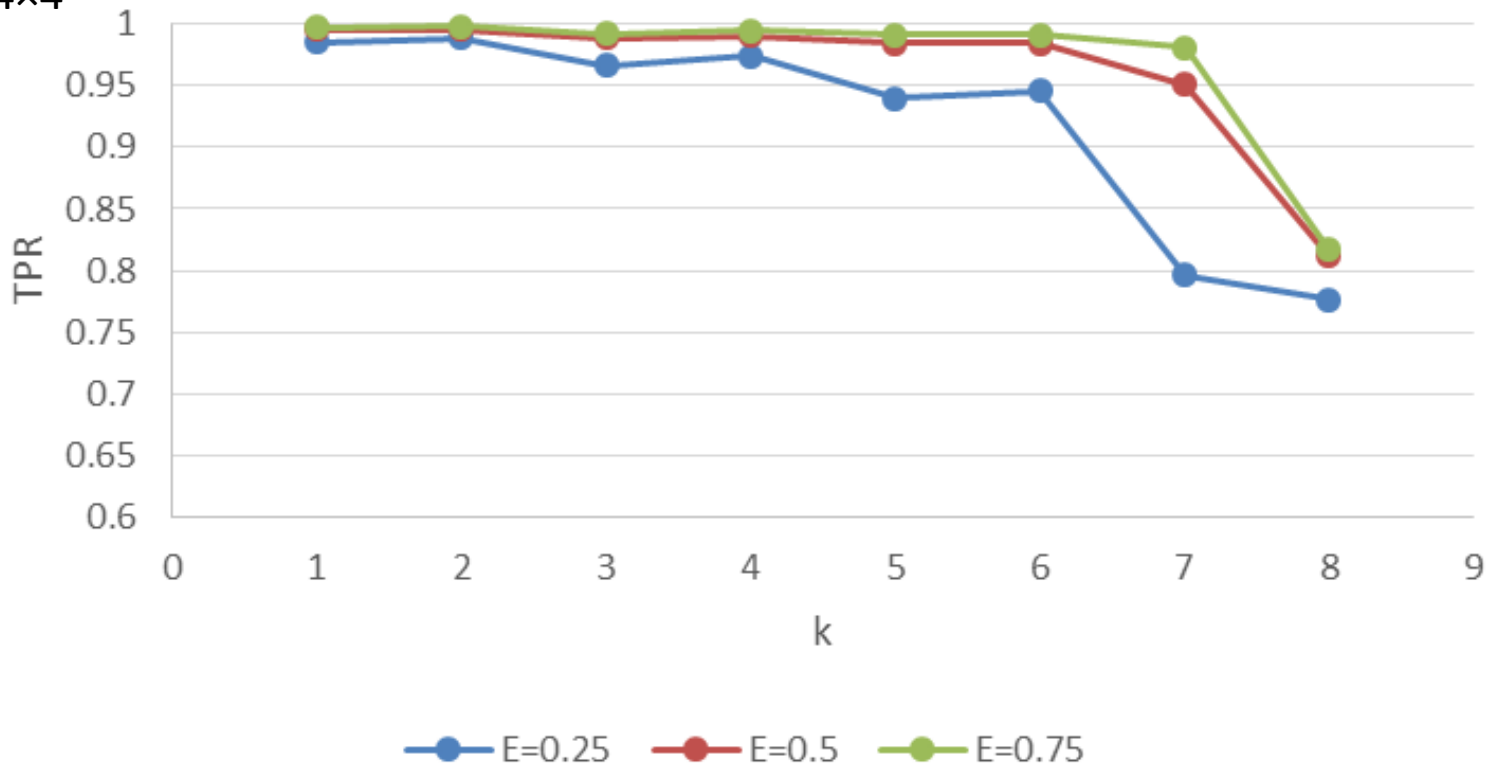
Watermarked image



Absolute value of their difference

Experiments: Localization of the Advanced Tampering

- Attacking method: $C^W := C^W + LCM(2,4,\ldots,2k)$, where $k = 1,2,\ldots,\Delta_{max}$
- 50% of pixels were tampered
- Various E values (fraction of watermarked pixels)
- **The measure is** TPR = number of blocks correctly determined as tampered / total number of tampered blocks
- Block size is 4×4
- 12-bit image



If we use $E \geq 0.5$ then $TPR \geq 0.95$ up to $k = 7$.

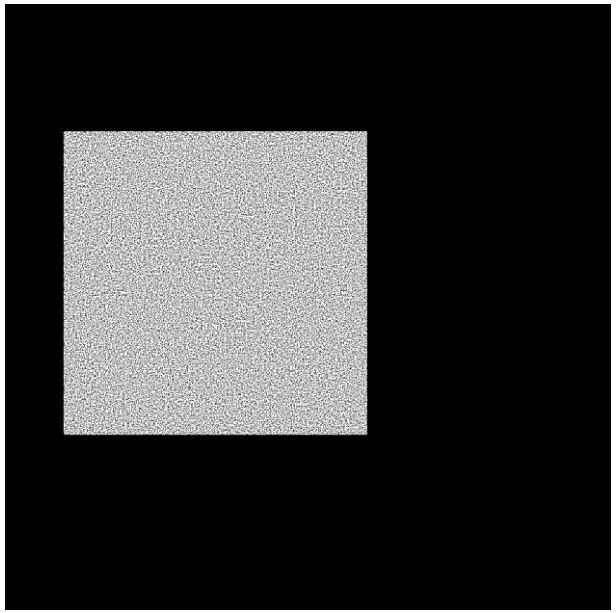
For $k = 8$ $LCM(2,4,\ldots,2k) = 1680$.

An intruder is limited to 2 possible values.

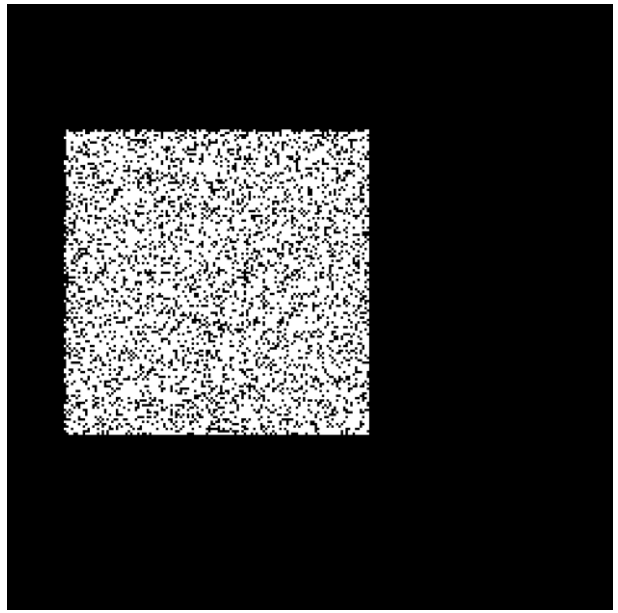
Such tampering is easy to detect visually.

Tampering Localization Example

Tampered image



Correct map of tampered pixels



Estimated map of tampered pixels

Estimated map after post-processing (morphological closing using a 9×9 window)

Conclusion

Other details not mentioned in the presentation

- Separating of an image into ROI (region of interest) and RONI (the rest area)
- Source image recovery by robust watermarking in RONI (a second watermark) to reduce the possibility of an accidental misdiagnosis
- Theoretical estimation of ROI / RONI capacity.
- Specific pixel selection approach aimed to minimize embedding distortions
- More experiments

Acknowledgments

- The work was funded by the RFBR grant 19-29-09045.

Thank you for your attention!