Recent Advances in DNN Watermarking

Hanzhou WU

h.wu.phd@ieee.org

SHANGHAI UNIVERSITY

April 2022

YouTube

Short Bio

Hanzhou Wu

PhD, Associate Professor

Department of Electronic and Information Engineering

Shanghai University, Shanghai 200444, China

Hanzhou Wu, received the B.S. and Ph.D. degrees from Southwest Jiaotong University, Chengdu, China, in June 2011 and June 2017, respectively. From October 2014 to October 2016, he was a Visiting Scholar with the New Jersey Institute of Technology, NJ, USA. He was a Research Staff with the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, from July 2017 to March 2019. He is currently an Associate Professor with Shanghai University, Shanghai, China. He has authored three book chapters and published more than 50 research papers in peer journals and conferences such as *IEEE Transactions on Dependable and Secure Computing*, *IEEE Transactions on Image Processing*, *IEEE Computational Intelligence Magazine*, *IEEE Transactions on Circuits and Systems for Video Technology*, *IEEE Signal Processing Letters* and so on. His research interests include artificial intelligence security, steganography/steganalysis and digital watermarking. He has received more than 950 citations from Google Scholar.

Background

Watermarking deep neural networks (DNNs): Protecting the intellectual property of DNNs

Computing resources

Expertise

Large-scale dataset

IBM Cloud

aws

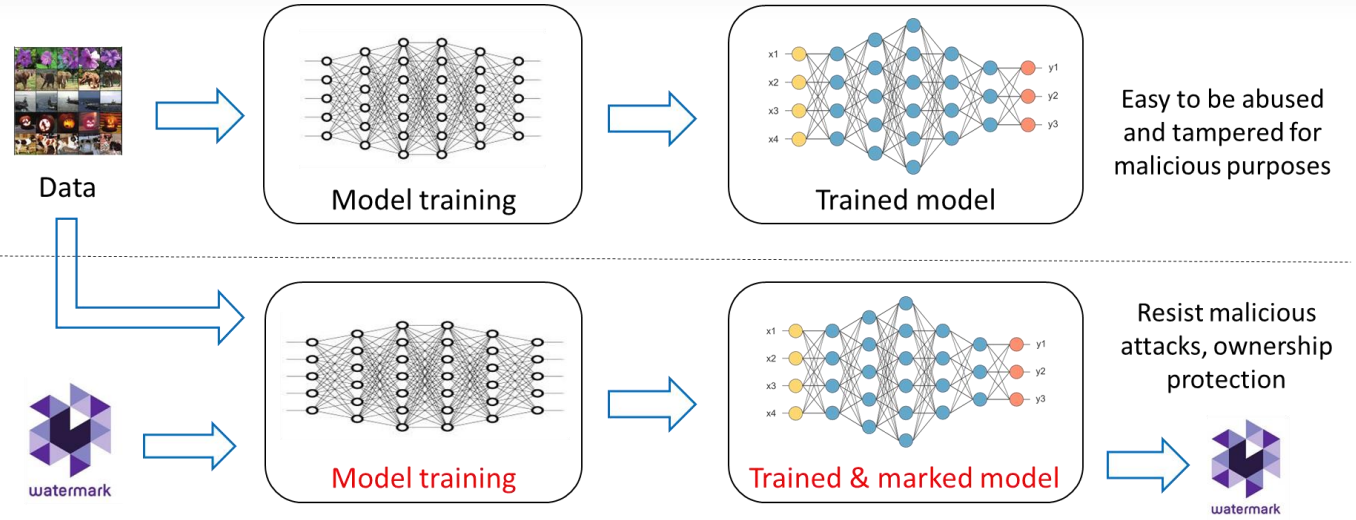
Microsoft Azure

Google Cloud Platform

Machine Learning as a Service (MLaaS)

Many threats

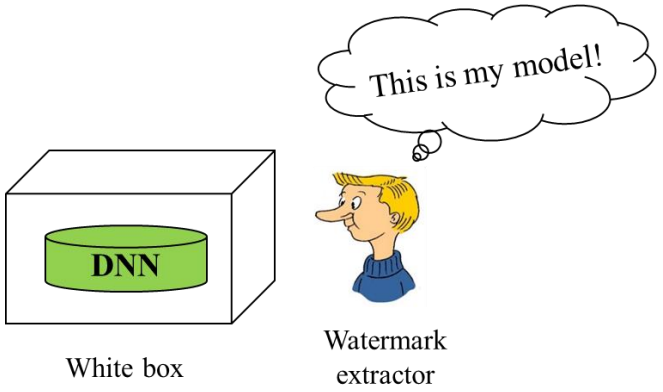
DNN Watermarking - Metrics



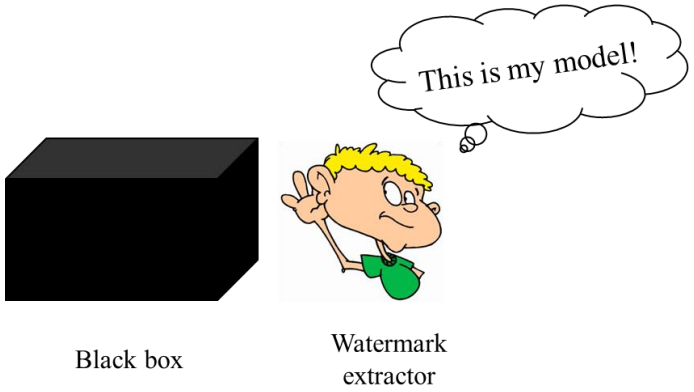
DNN watermarking, or called **model watermarking**, is to embed a *secret message (i.e., watermark)* in a given DNN model to produce a *marked* DNN model which allows us to identify the DNN ownership by extracting the embedded watermark from the target DNN marked model.

Metrics	Description
Fidelity	Task fidelity: the performance of the DNN on its original task after watermarking should not be impaired.
	Watermark fidelity: the distortion between the extracted watermark and the original one should be low.
Imperceptibility	The difference between the marked DNN model and the non-marked DNN model should be low.
Payload	It is desirable to embed as many secret bits as possible for reliable ownership verification.
Security	It should be very difficult for any unauthenticated parity to extract, tamper and forge the watermark.
Robustness	The hidden watermark can resist various attacks such as fine-tuning and model compression.
Complexity	The computational cost of embedding a watermark and verifying the ownership should be low.

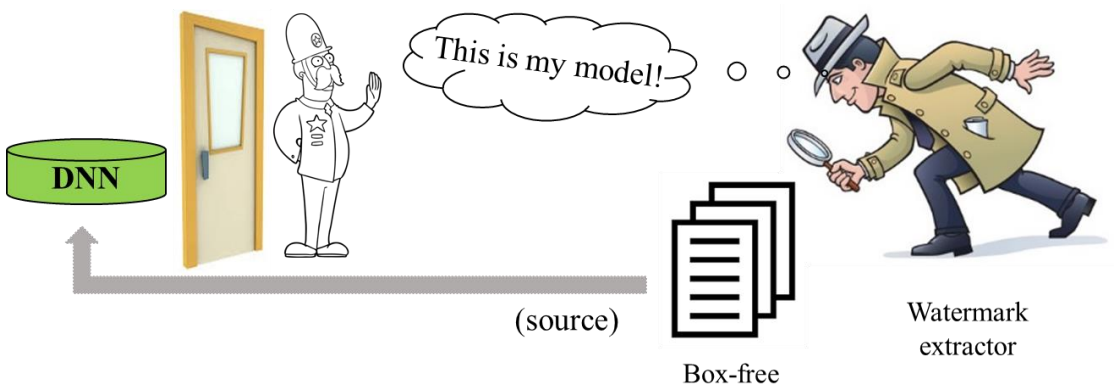
DNN Watermarking - Categories



White-box watermarking



Black-box watermarking



Box-free watermarking

Categories	Description
White-box watermarking	The watermark extractor should know the internal details of the target model.
Black-box watermarking	The watermark extractor does not know the internal details of the target model, but can query the target model with a set of special input samples
Box-free watermarking	The watermark extractor has no access to the target model, but can extract the watermark from any sample generated by the target model.

White-box DNN Watermarking - Strategies

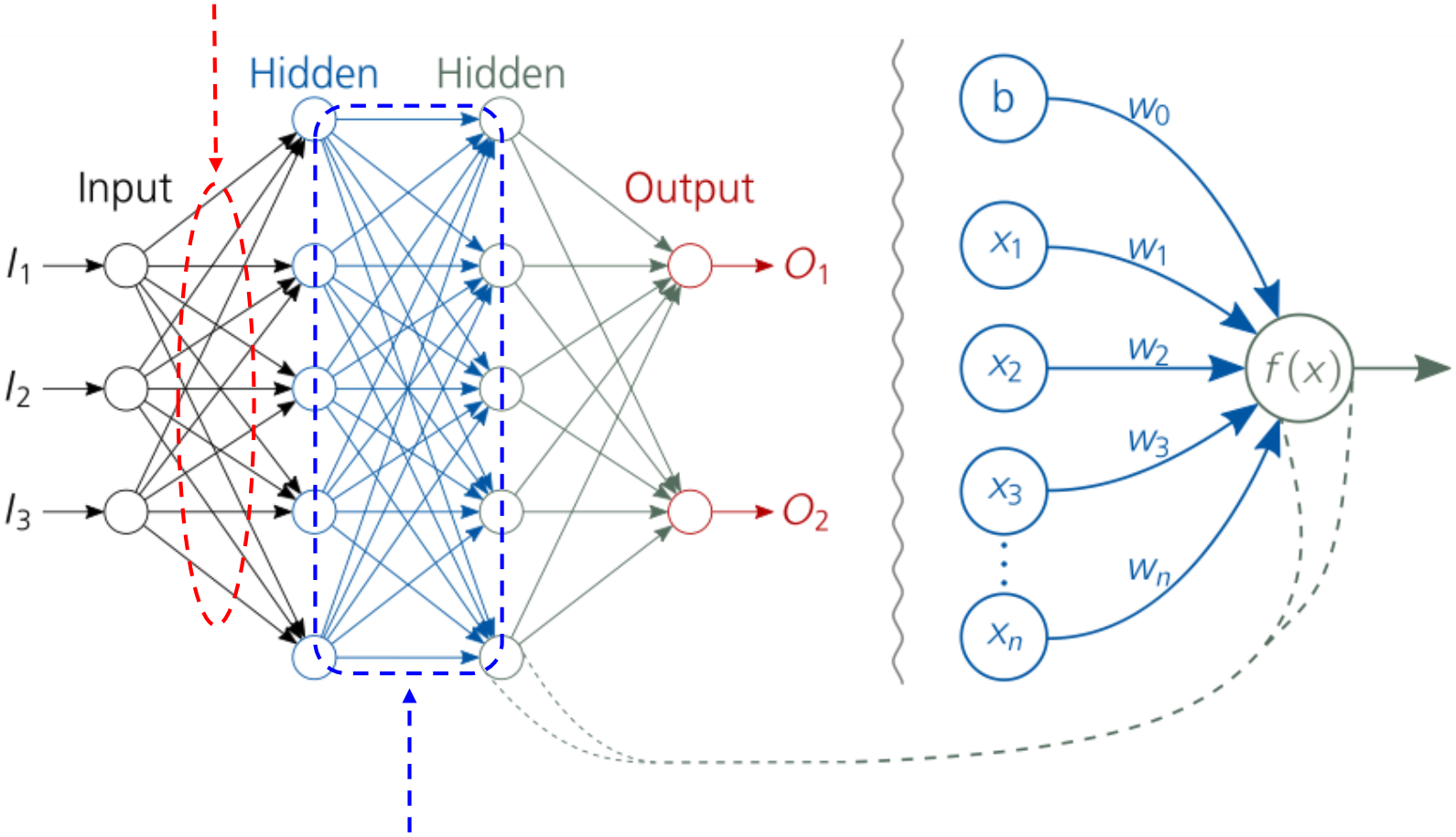
① Modifying network parameters to embed a watermark

Parameter based methods

- Optimizing watermark-involved loss
- Embedding with low-order statistics
- Adversarial training
- Dither modulation
- Spread spectrum (SS) embedding
- Others

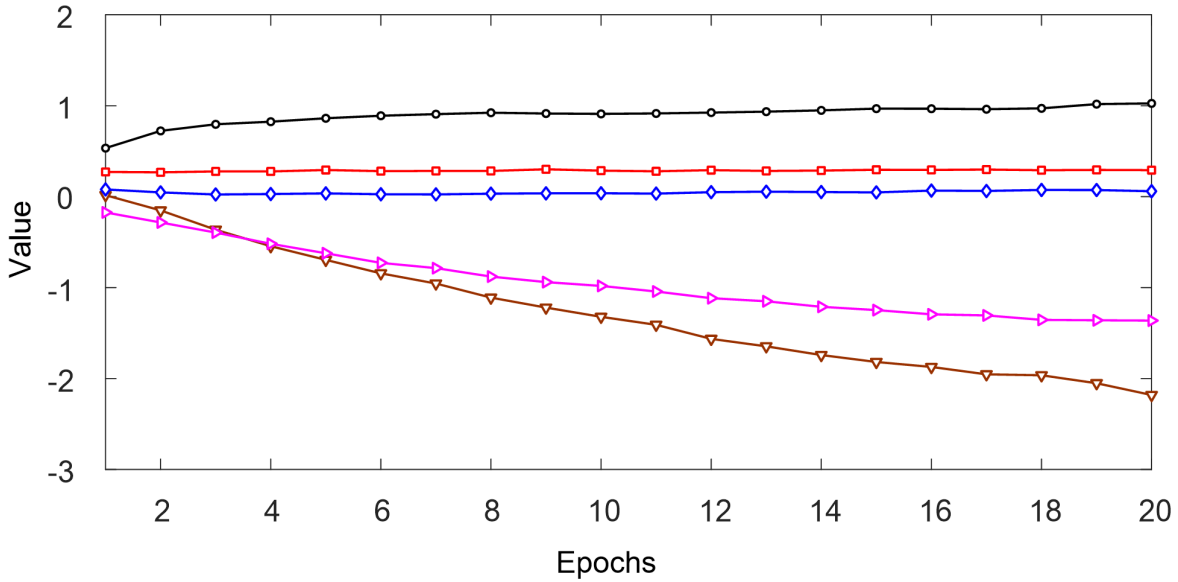
Structure based methods

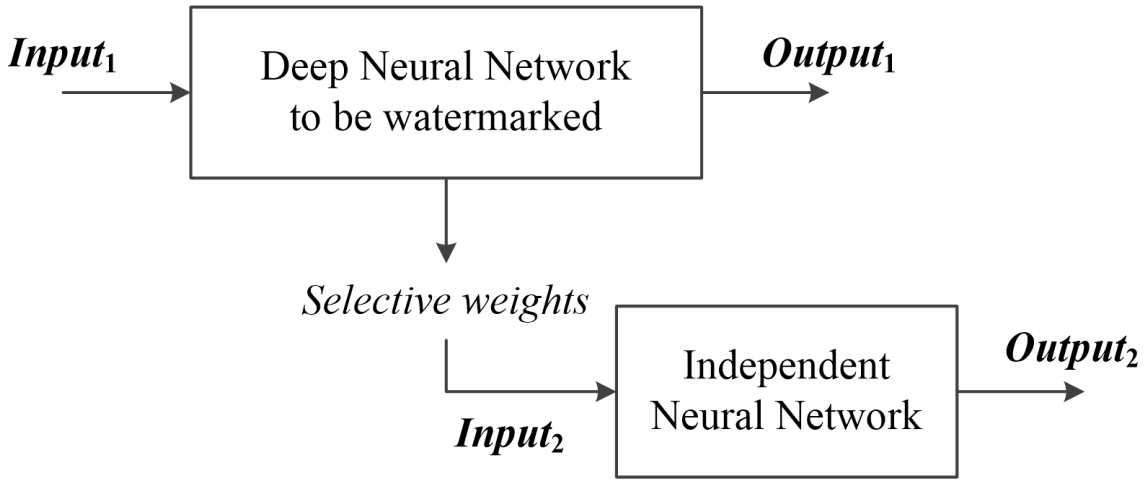
- Model compression (to reduce the size)
- Quantized index modulation (QIM)
- Others



② Modifying network structure to embed a watermark

White-box DNN Watermarking - Wang *et al.*'s algorithm



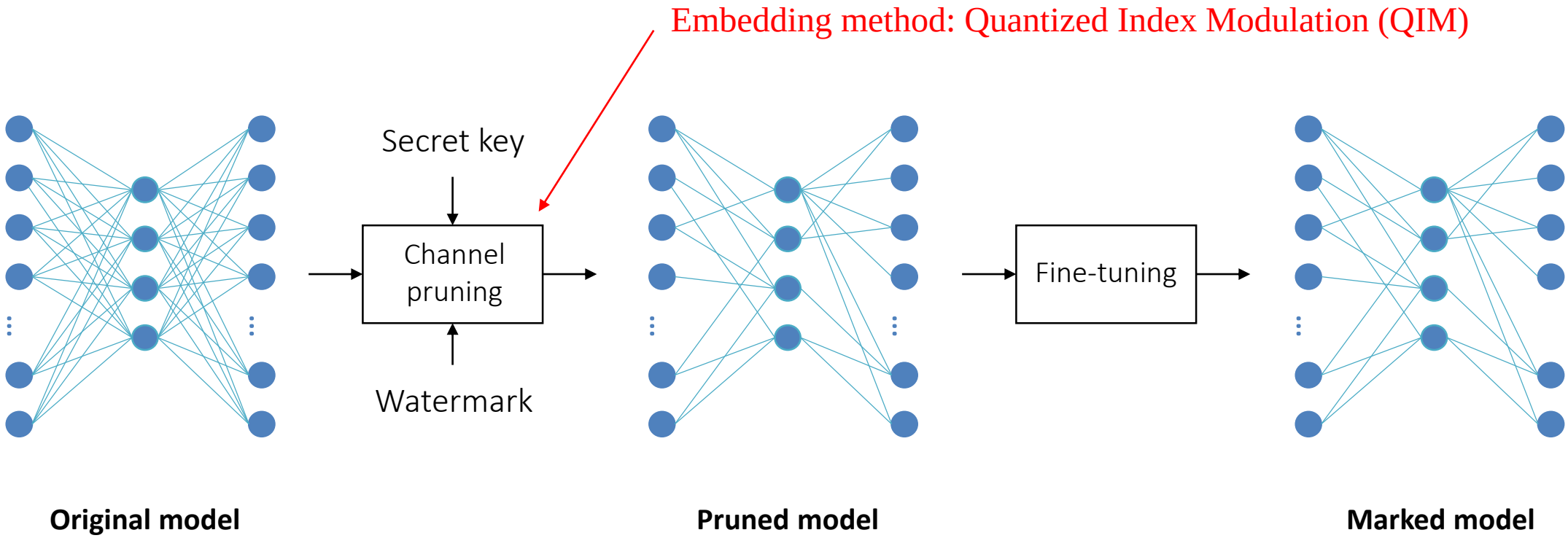


Using those earlier-convergent parameters to embed a watermark

Using an independent neural network (that can be optimized) to extract the watermark from the target model

J. Wang *et al*. **Watermarking in deep neural networks via error back-propagation**. *Proc. IS&T EI-MWSF*, 2020.

White-box DNN Watermarking - Zhao *et al.*'s algorithm



① Embedding a secret watermark

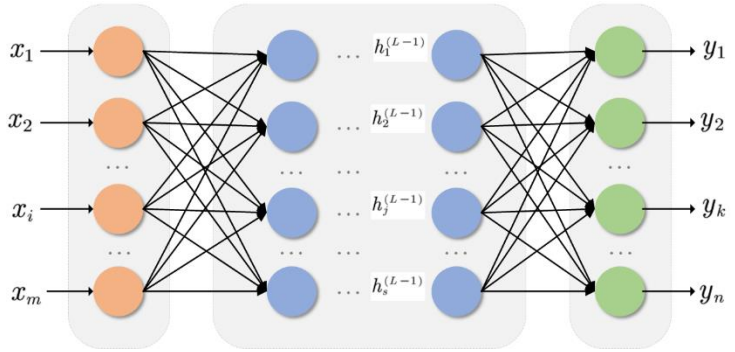
② Recovering the performance on the task

Zhao *et al*. **Structural watermarking to deep neural networks via network channel pruning**. *IEEE WIFS*, 2021.

Black-box DNN Watermarking - General framework

Clean samples
(normal data)

Trigger samples
(special data)



Marked DNN model

Put into use

To train a marked model

Leaked

Model training

- *Normal dataset*
- *Trigger dataset*

Model usage

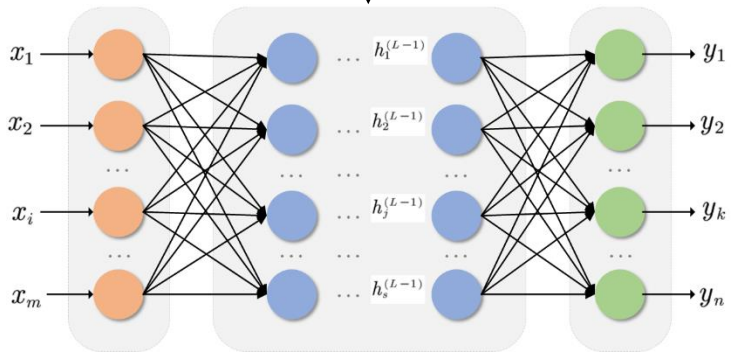
- *Normal queries by normal data*

Model verification

- *Special queries by trigger data*

Constructing the trigger sample by adding a trigger signal

Trigger samples
(special data)



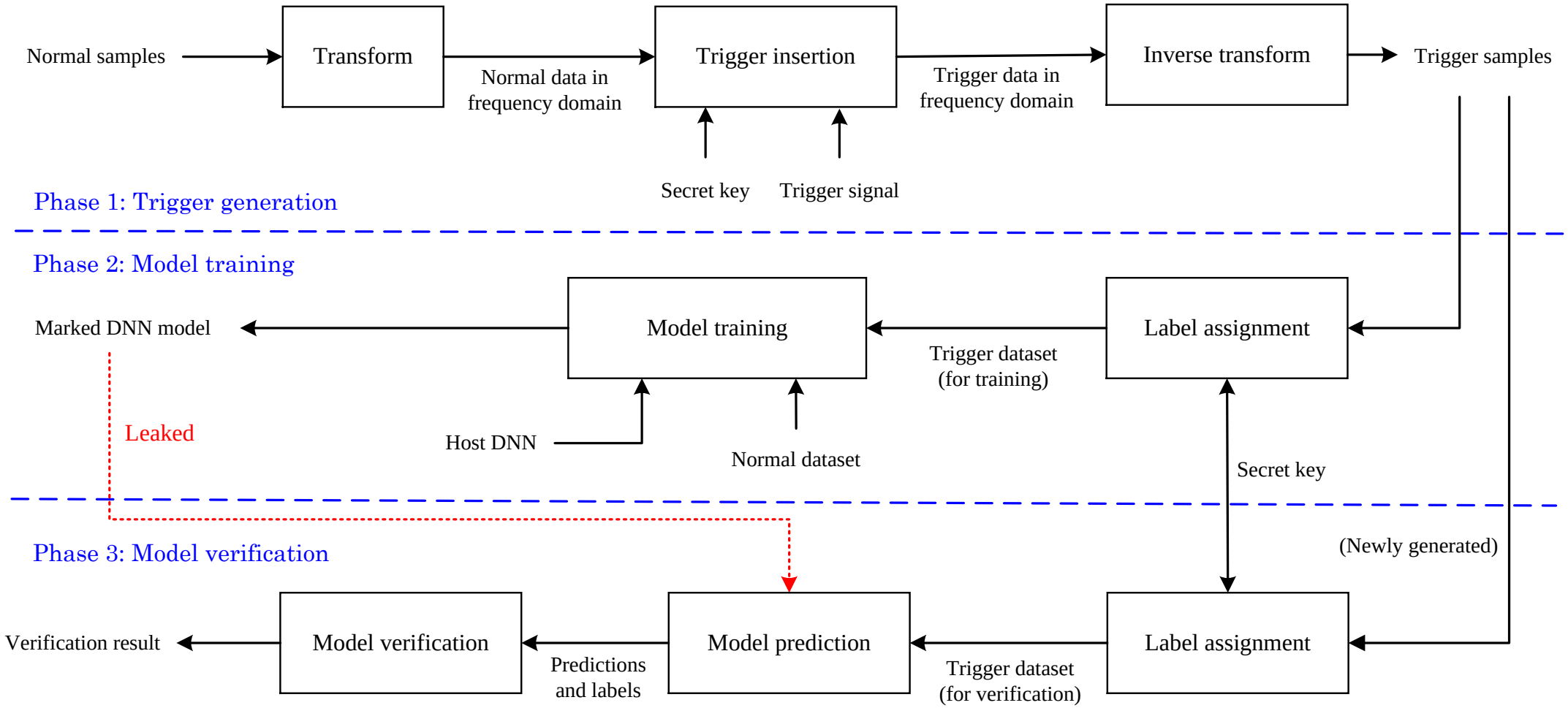
Output

Ground-truth
(pre-specified)

Equal?
Yes/No

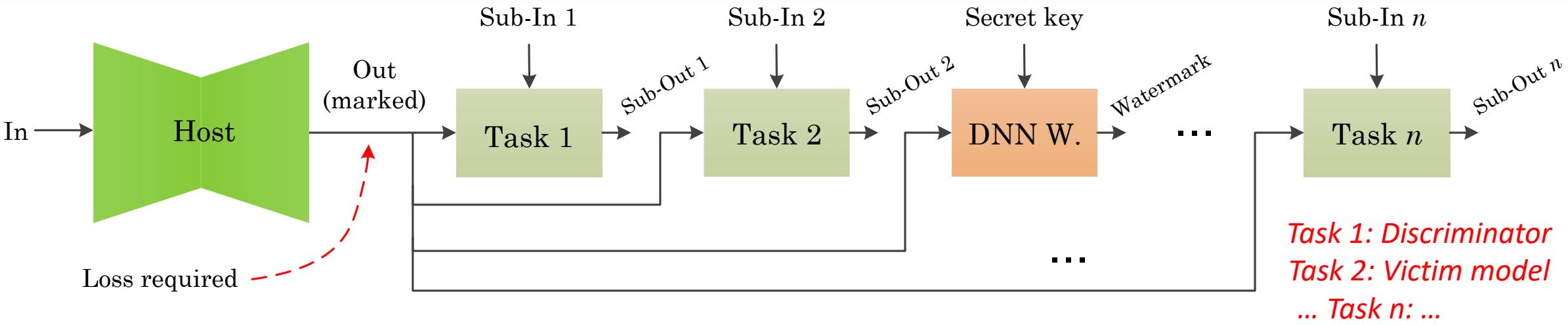
To verify the ownership by checking the output

Black-box DNN Watermarking - Wang *et al.*'s algorithm

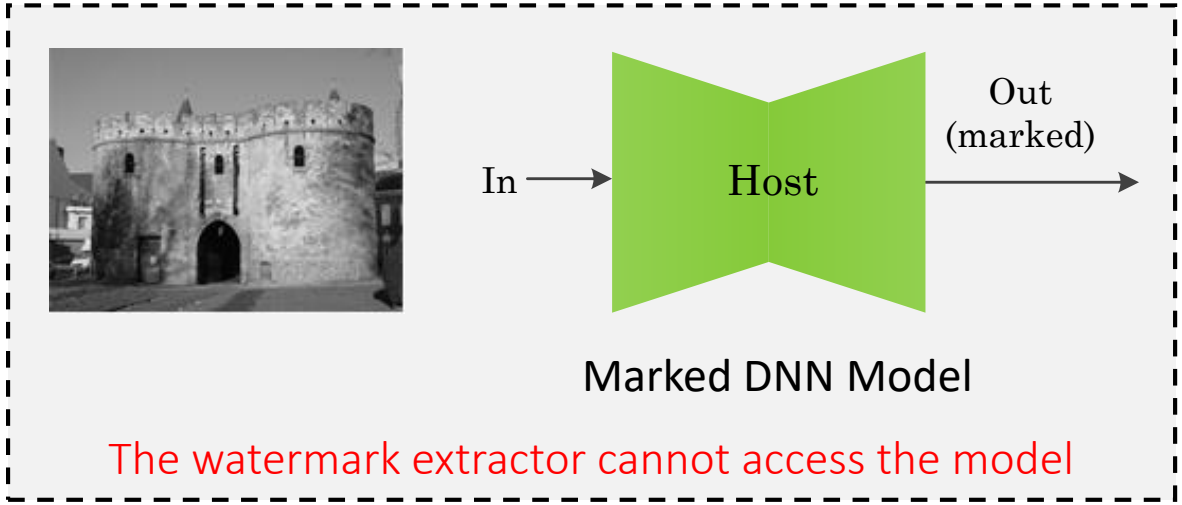


Wang *et al*. Protecting the intellectual property of speaker recognition model by black-box watermarking in the frequency domain. *Symmetry*, 2022.

Box-free DNN Watermarking - General framework



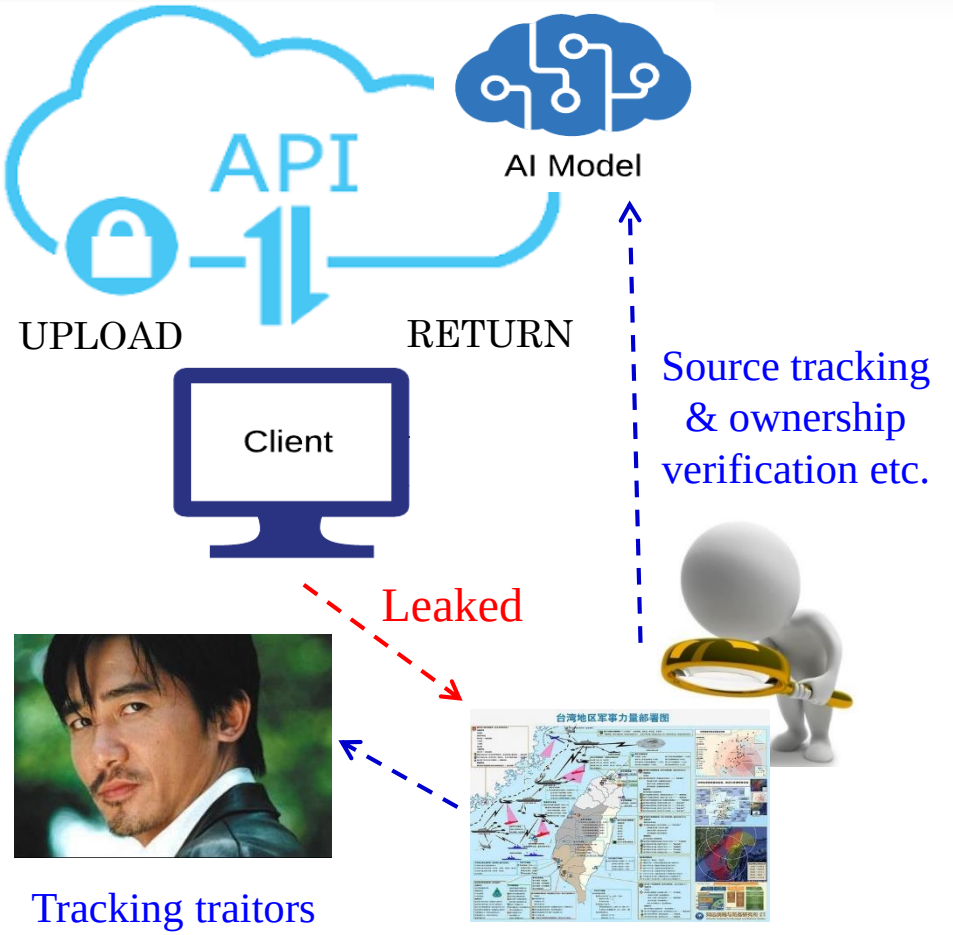
(a) DNN model training

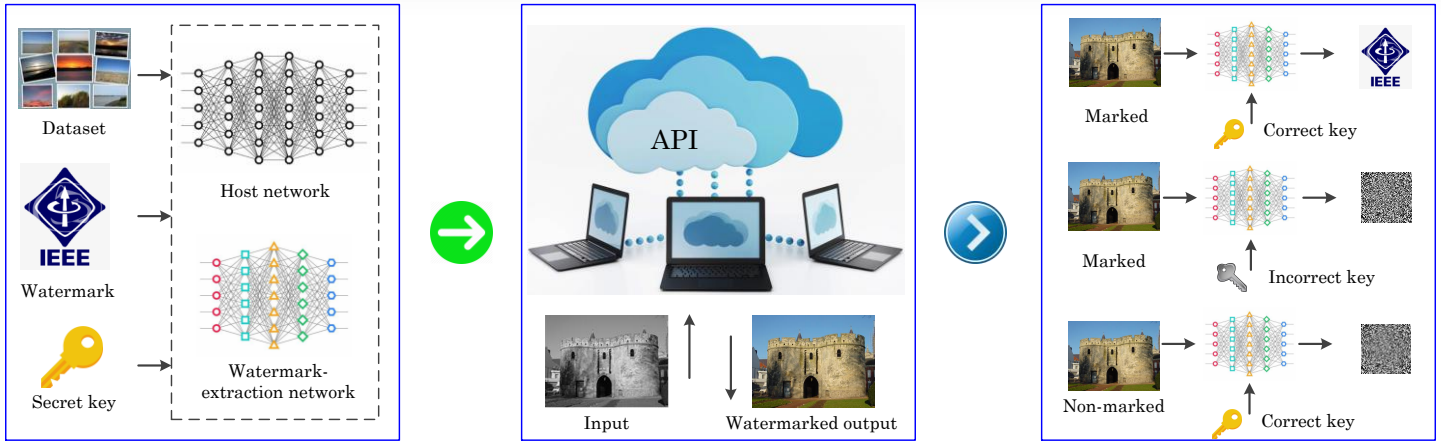


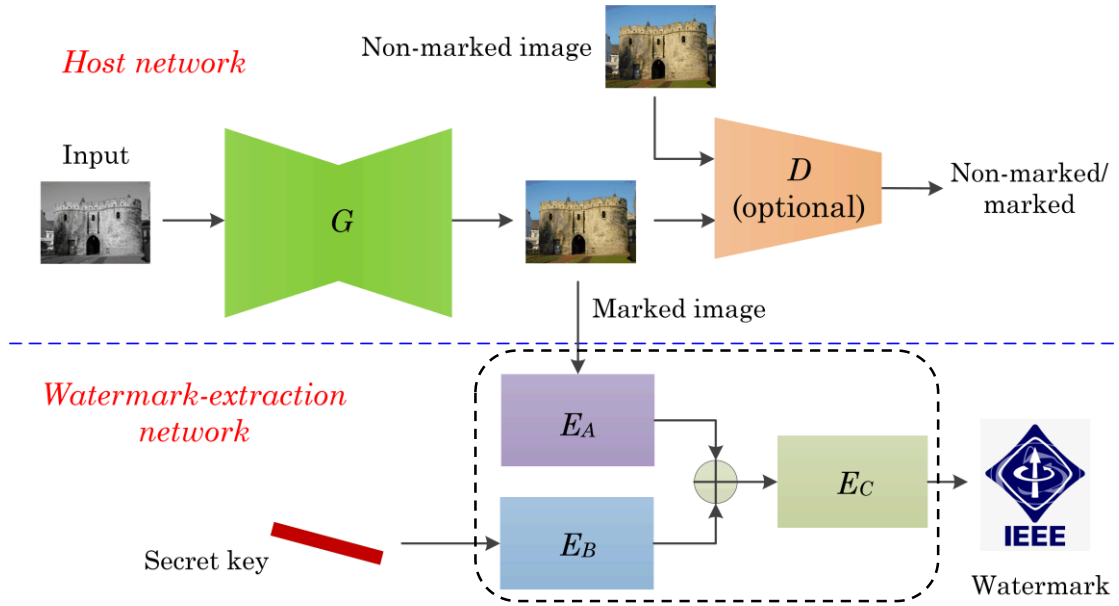
Watermark

(b) Ownership verification

Box-free DNN Watermarking - Wu *et al.*'s algorithm







Wu *et al*. **Watermarking neural networks with watermarked images**. *IEEE TCSVT*, 2021.

Conclusion and Discussion

- Conclusion
  - Mainstream works can be divided to three categories: *white-box*, *black-box*, *box-free*.
  - Some embedding strategies in media watermarking can be used for DNN watermarking.
  - Mainstream works mainly focus on designing *robust* watermark embedding operations.
- Future works
  - Fragile DNN watermarking
    - To determine whether a given DNN model was previously altered or not.
  - Theoretic research of DNN watermarking
    - Information theory
    - Game theory
    - Interpretability
    - And so on …

References

1. H. Wu, G. Liu, Y. Yao, X. Zhang. Watermarking neural networks with watermarked images. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2591-2601, 2021.
2. X. Zhao, Y. Yao, H. Wu, X. Zhang. Structural watermarking to deep neural networks via network channel pruning. *IEEE International Workshop on Information Forensics and Security*, pp. 1-6, 2021.
3. M. Li, H. Wu, X. Zhang. Generating watermarked adversarial texts. *arXiv Preprint arXiv:2110.12948*, 2021.
4. X. Zhao, H. Wu, X. Zhang. Watermarking graph neural networks by random graphs. *IEEE International Symposium on Digital Forensics and Security*, pp. 1-6, 2021.
5. H. Wu, G. Liu, X. Zhang. Hiding data hiding. *arXiv Preprint arXiv:2102.06826v3*, 2021.
6. J. Wang, H. Wu, X. Zhang, Y. Yao. Watermarking in deep neural networks via error back-propagation. *IS&T Electronic Imaging, Media Watermarking, Security and Forensics*, 2020.
7. Y. Wang, H. Wu. Protecting the intellectual property of speaker recognition model by black-box watermarking in the frequency domain. *Symmetry*, vol. 14, no. 3, pp. 619, 2022.
8. Z. Wang, G. Feng, H. Wu, X. Zhang. Data hiding in neural networks for multiple receivers. *IEEE Computational Intelligence Magazine*, vol. 16, no. 4, pp. 70-84, 2021.

Thank you so much!

Email: h.wu.phd@ieee.org