ClearStamp

A Human-Visible and Robust Model-Ownership Proof based on Transposed Model Training

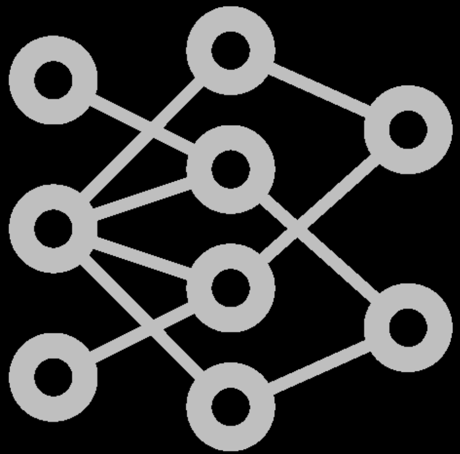
Torsten Krauß, Jasper Stang, and Alexandra Dmitrienko

University of Würzburg, Germany

33rd USENIX Security Symposium

Image

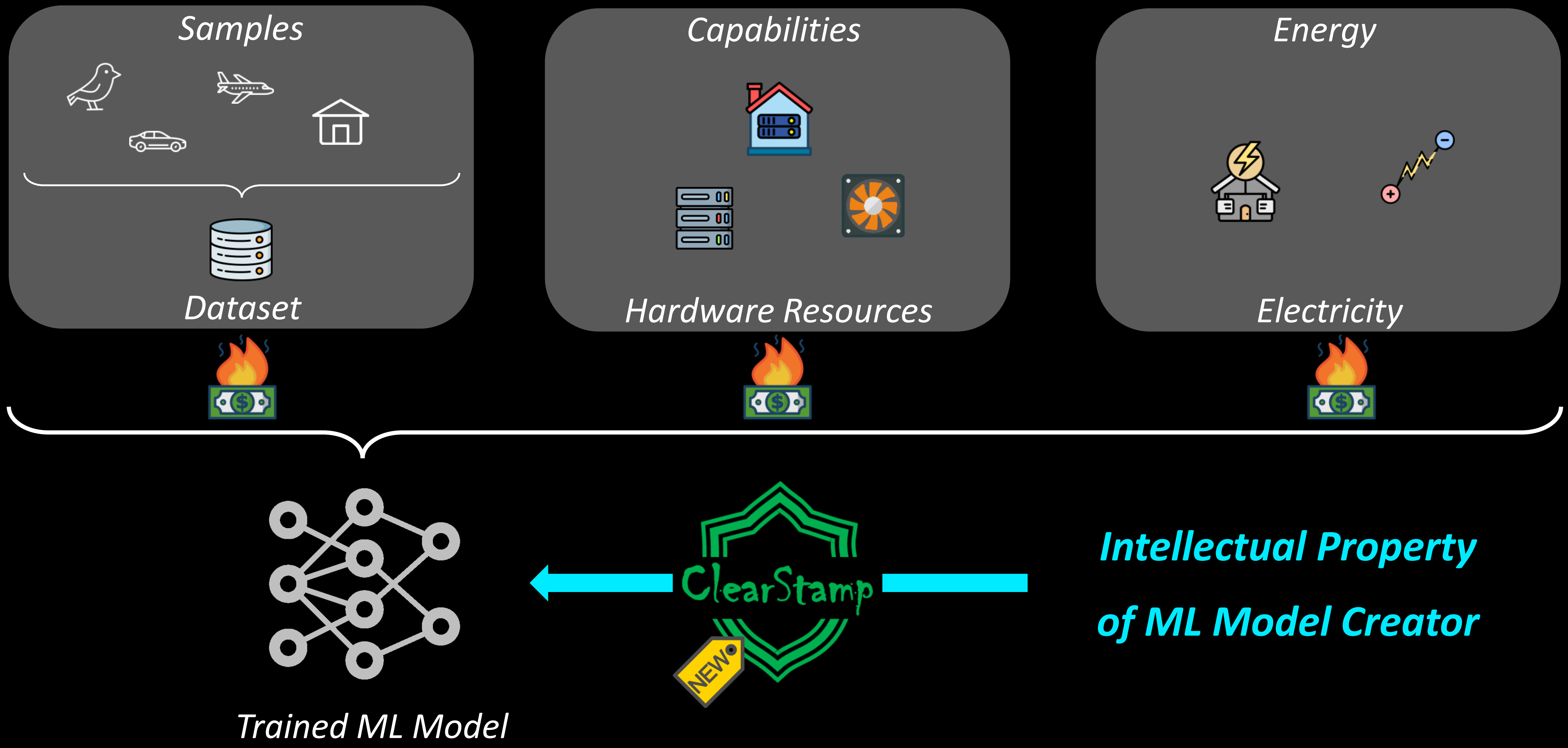
ML Model



Image

ML Model

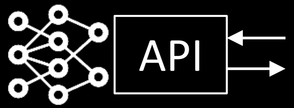
ML Models are Intellectual Property



Intellectual Property Violations

Trained ML Model

- *Theft*
- *Illicit Utilization*



- *Unauthorized Resell*
- *Unwarranted Modifications*

Intellectual Property Protection

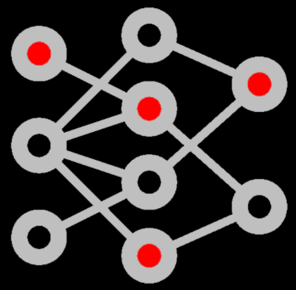
Image

Add a Visible Stamp

Same Principle

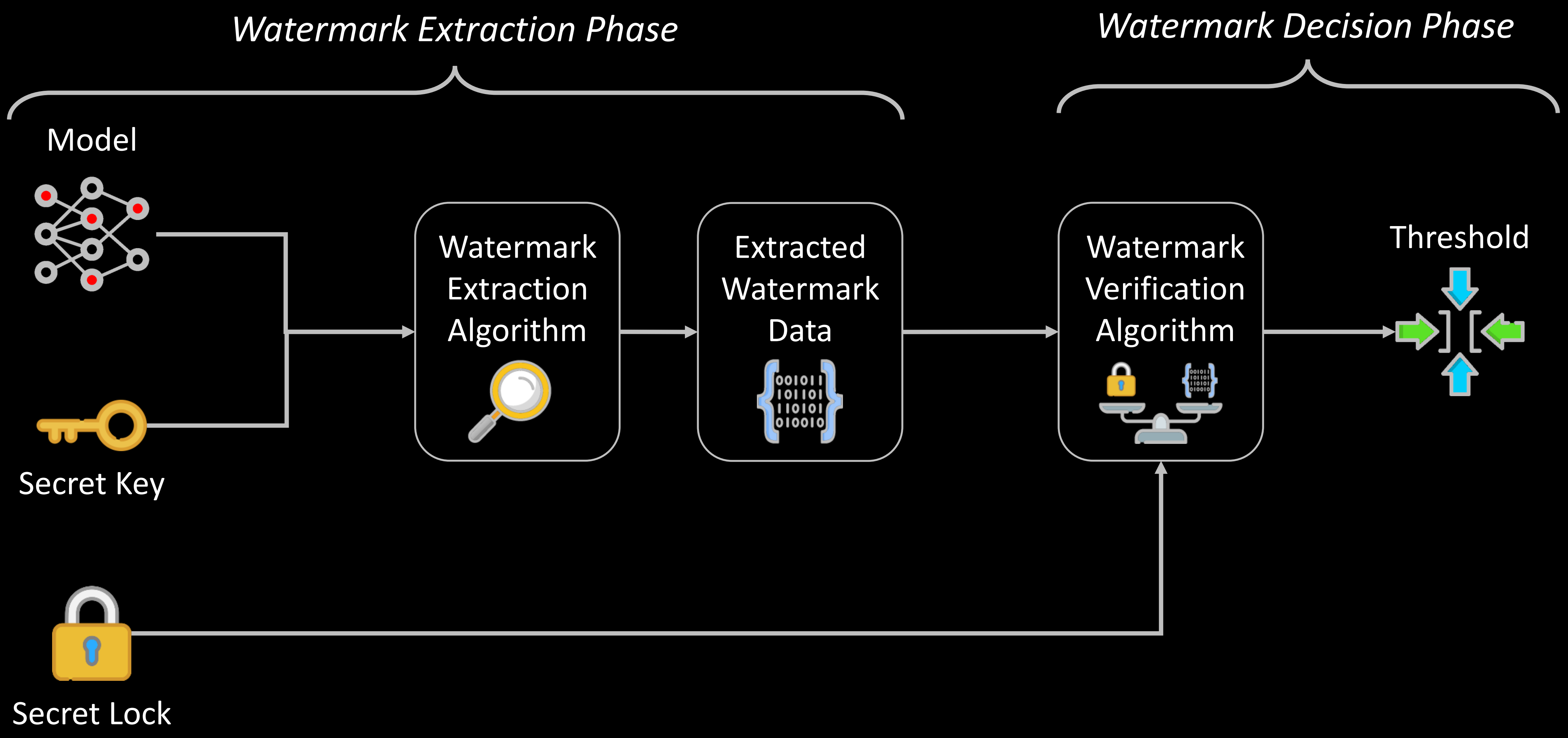
=

Trained ML Model

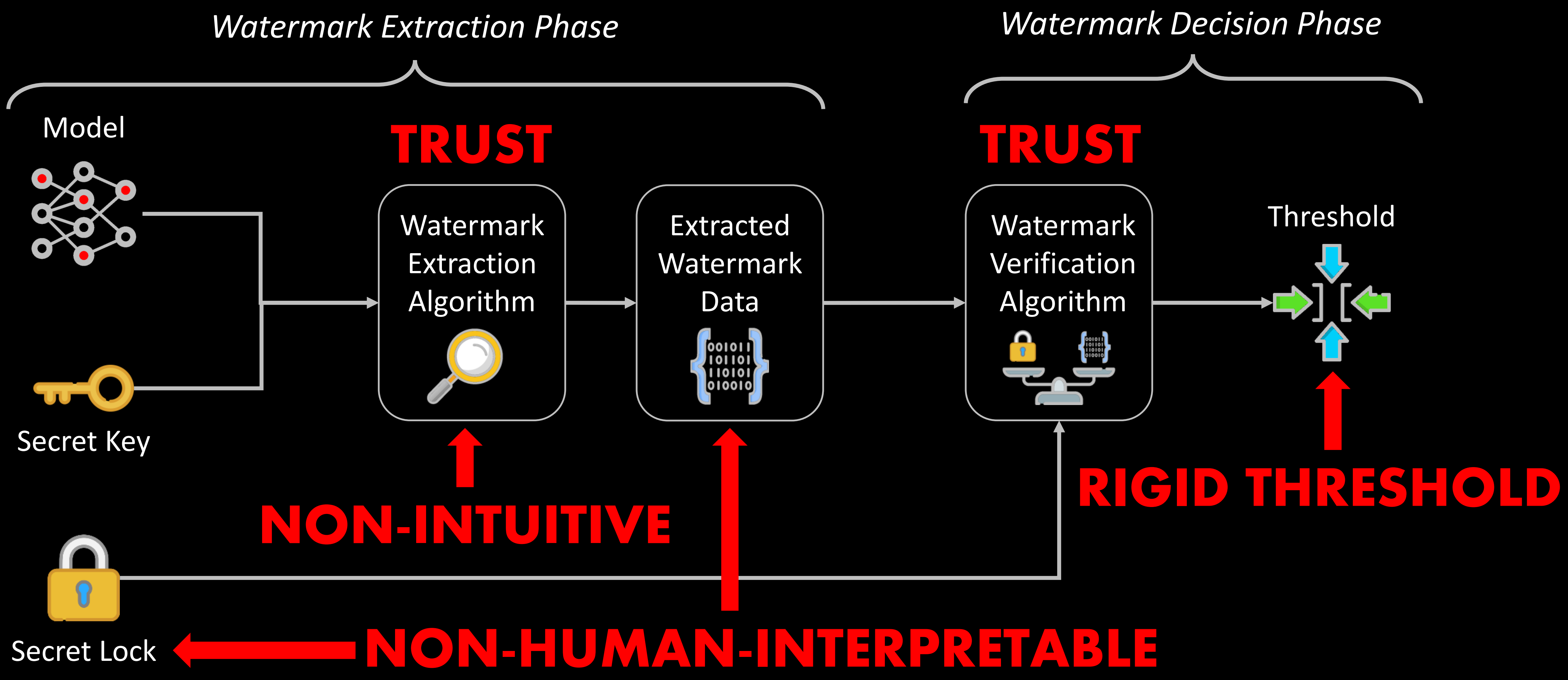


Hide Unique Information in Parameters

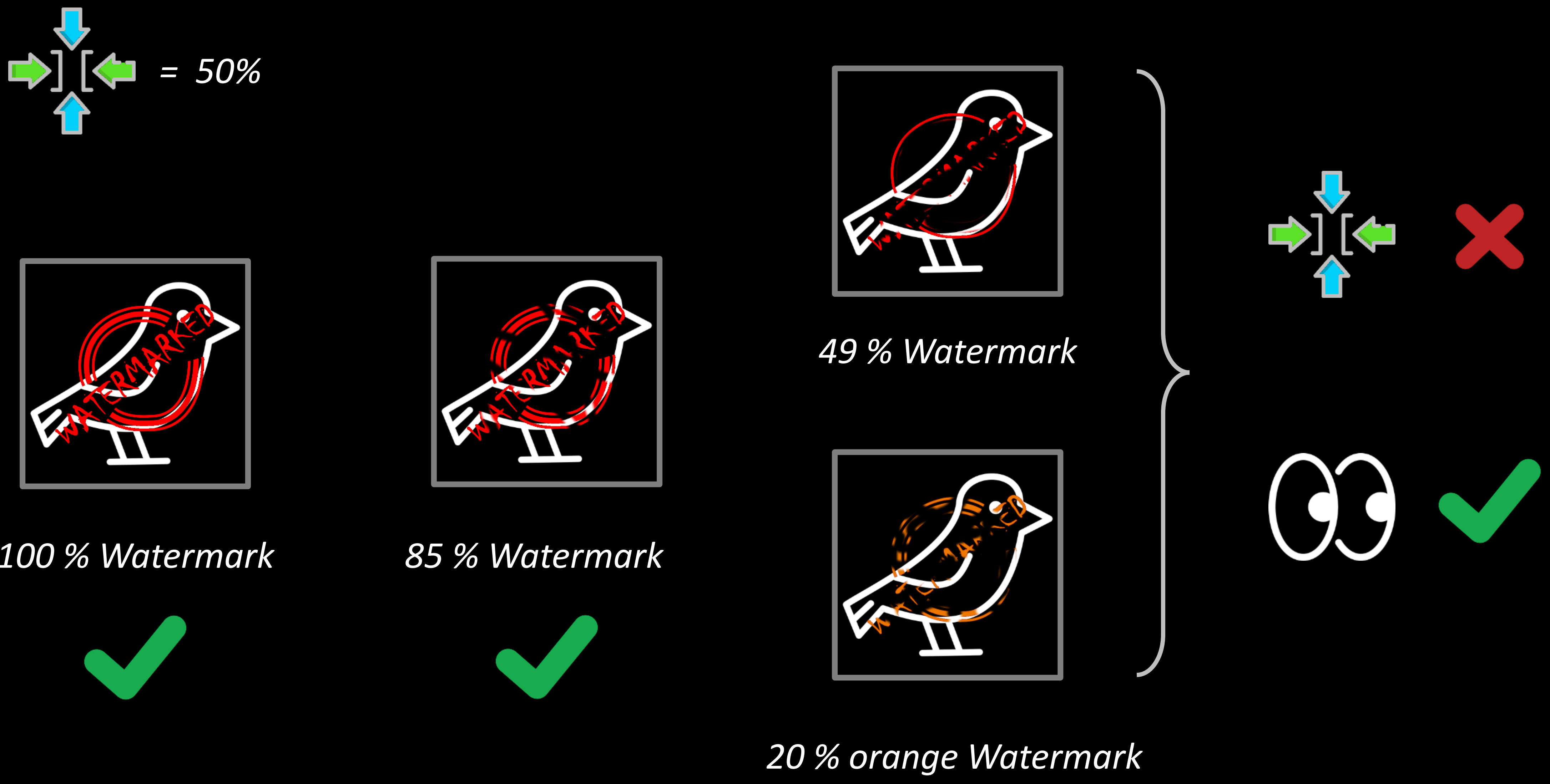
ML Model Watermark Verification



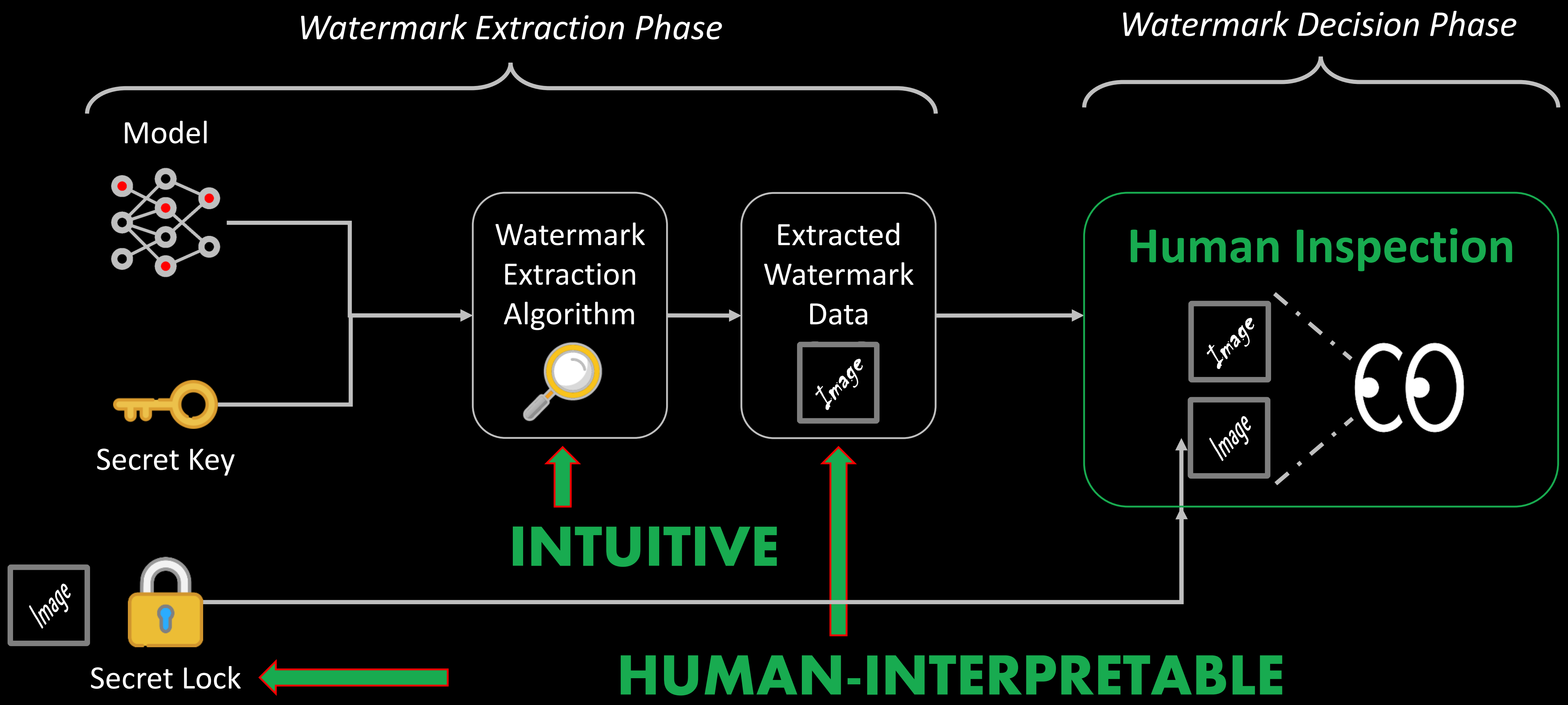
ML Model Watermark Verification



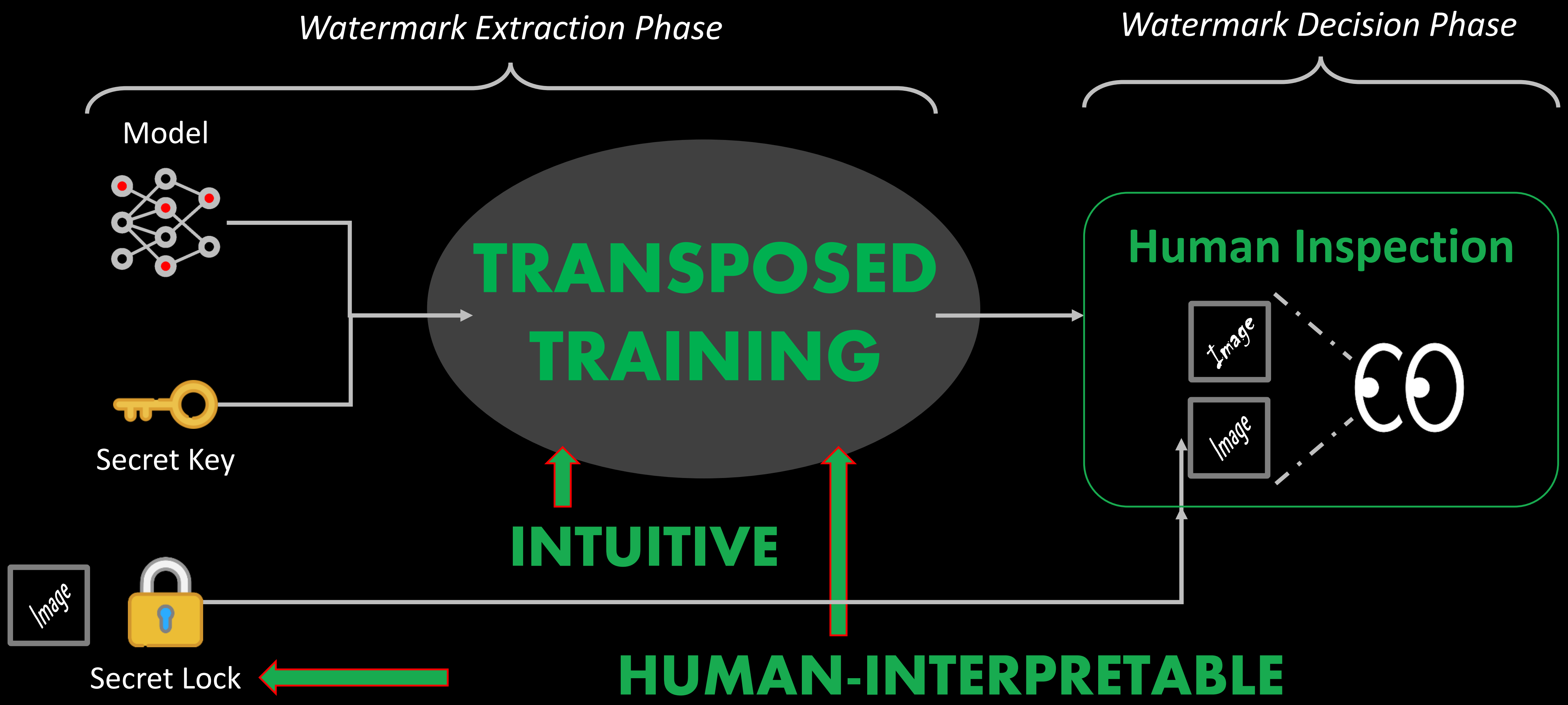
Rigid Threshold VS. Partially Removed Watermarks



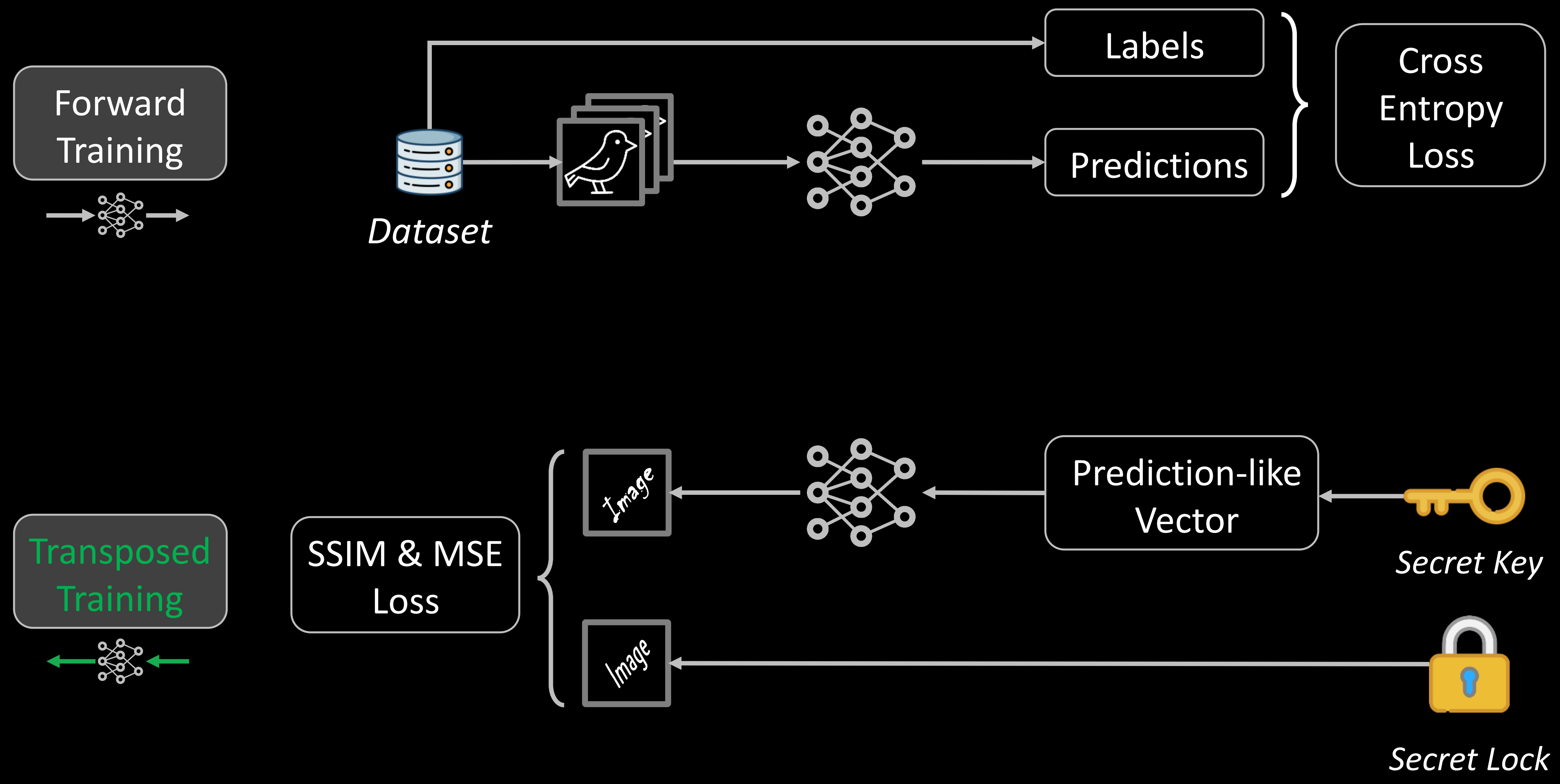
ClearStamp - Principle



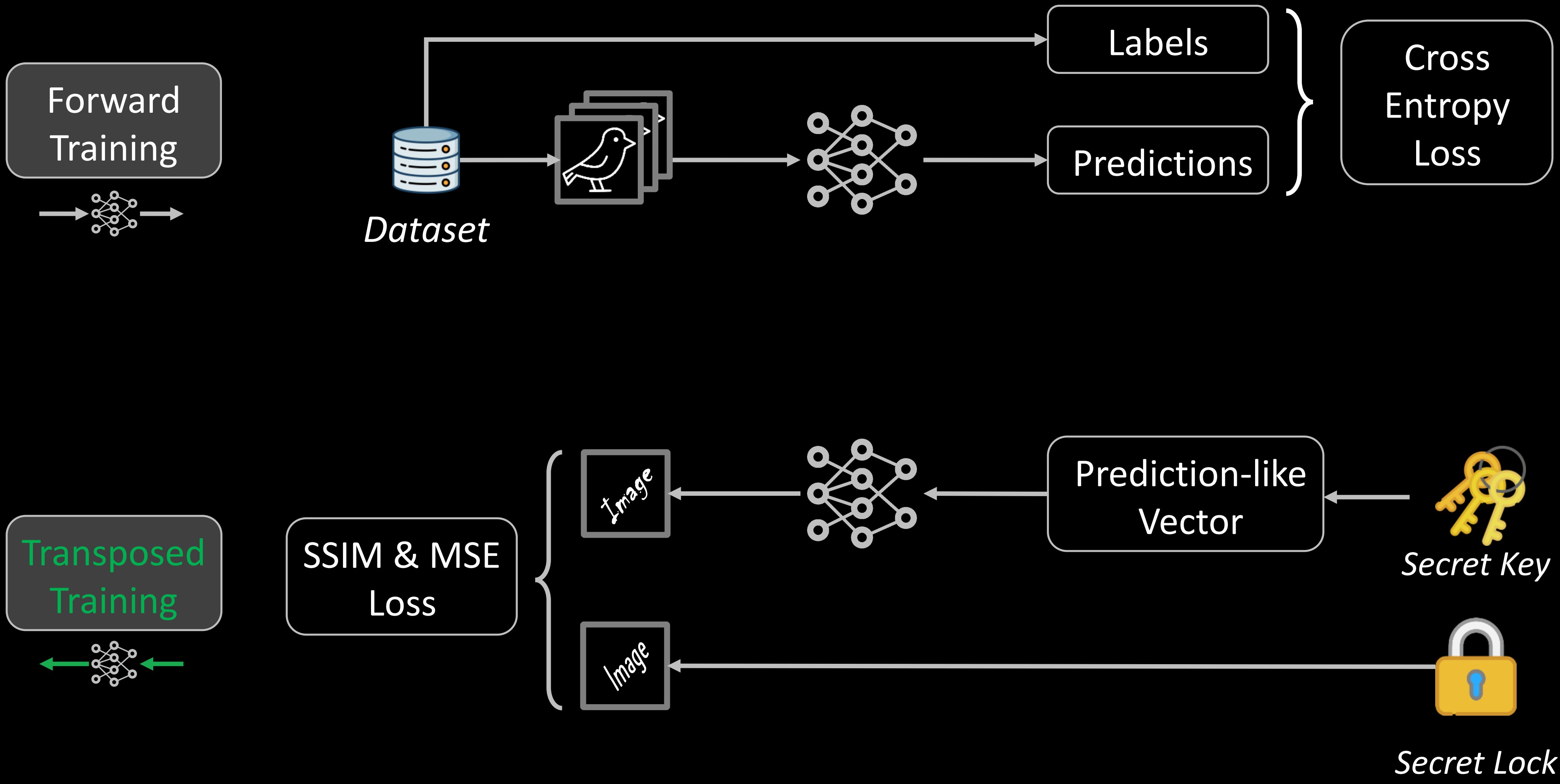
ClearStamp - Principle



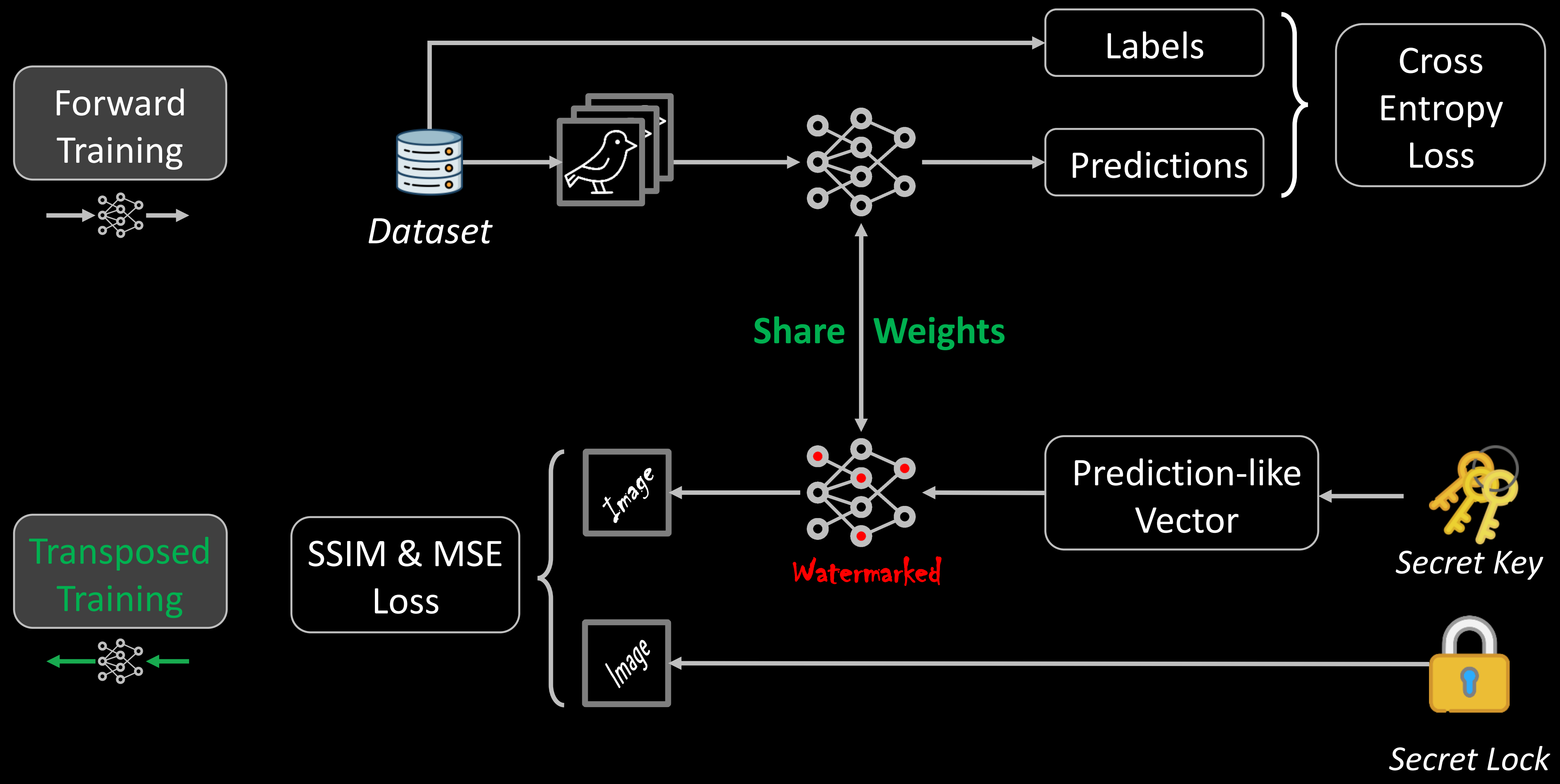
Transposed Training for Watermarking



Transposed Training for Watermarking



Transposed Training for Watermarking



Transposed Training - Details

Model Component	Forward Model	Transposed Model
Linear Layer	$y = x \cdot w^T + b$	$x = (y - b) \cdot w$
Batch Normalization	$y = \frac{x - E(x)}{\sqrt{Var(x) + \varepsilon}} \cdot \gamma + \beta$	$x = \frac{(y - \beta) \cdot \sqrt{1 + \varepsilon}}{\gamma}$ with $E(x) = 0, Var(x) = 1$
Convolutions [1]	Replace with deconvolutions [2]	
Pooling Layer [3]	Replace with Interpolations [4, 5]	
Dropout Layers [6]	Keep same dropout	
Activation Functions	Use same activation, e.g., ReLU [7]	
Skip Connections	Fixate skip connections	

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998.
[2] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. CVPR, 2010.
[3] Afia Zafar, Muhammad Aamir, Nazri Mohd Nawi, Ali Arshad, Saman Riaz, Abdulrahman Alruban, Ashit Kumar Dutta, and Sultan Almotairi. A comparison of pooling methods for convolutional neural networks. AppliedSciences, 2022.
[4] Olivier Rukundo and Hanqiang Cao. Nearest Neighbor Value Interpolation. IJACSA, 2012.
[5] Olivier Rukundo and Bodhaswar T Maharaj. Optimization of Image Interpolation based on Nearest Neighbour Algorithm. VISAPP, 2014.
[6] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014.
[7] Abien Fred Agarap. Deep Learning using Rectified Linear Units (ReLU). arXiv preprint arXiv:1803.08375, 2018.

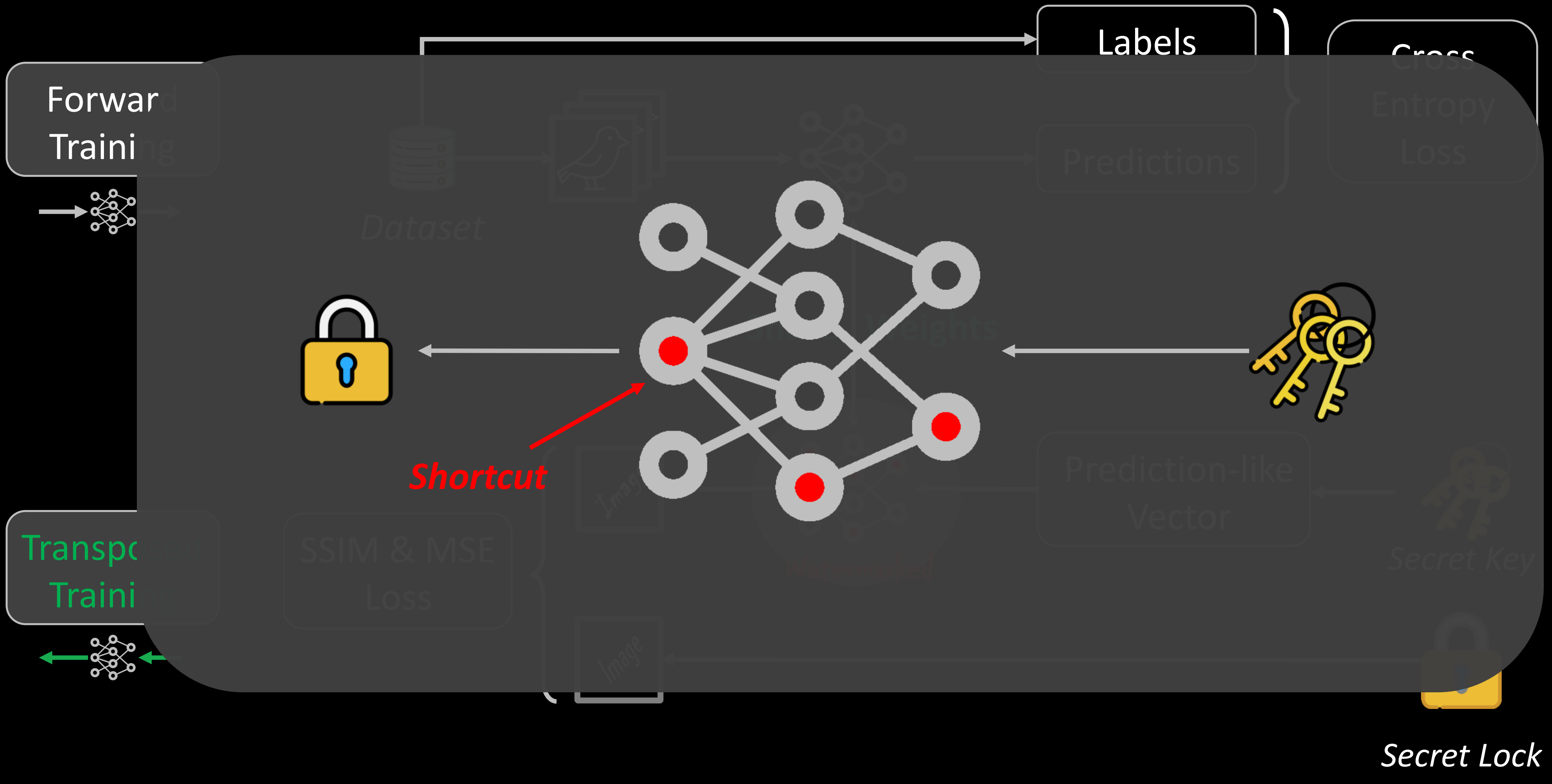
Transposed Training - Details

Model Component	Forward Model	Transposed Model
Linear Layer	$y = x \cdot w^T + b$	$x = (y - b) \cdot w$
Batch Normalization	$y = \frac{x - E(x)}{\sqrt{Var(x) + \varepsilon}} \cdot \gamma + \beta$	$x = (y - \beta) \cdot \frac{\sqrt{1+\varepsilon}}{\gamma}$ with $E(x) = 0, Var(x) = 1$
Convolutions [1]	Replace with deconvolutions [2]	
Pooling Layer [3]	Replace with Interpolations [4, 5]	
Dropout Layers [6]	Keep same dropout	
Activation Functions	Use same activation, e.g., ReLU [7]	
Skip Connections	Fixate skip connections	

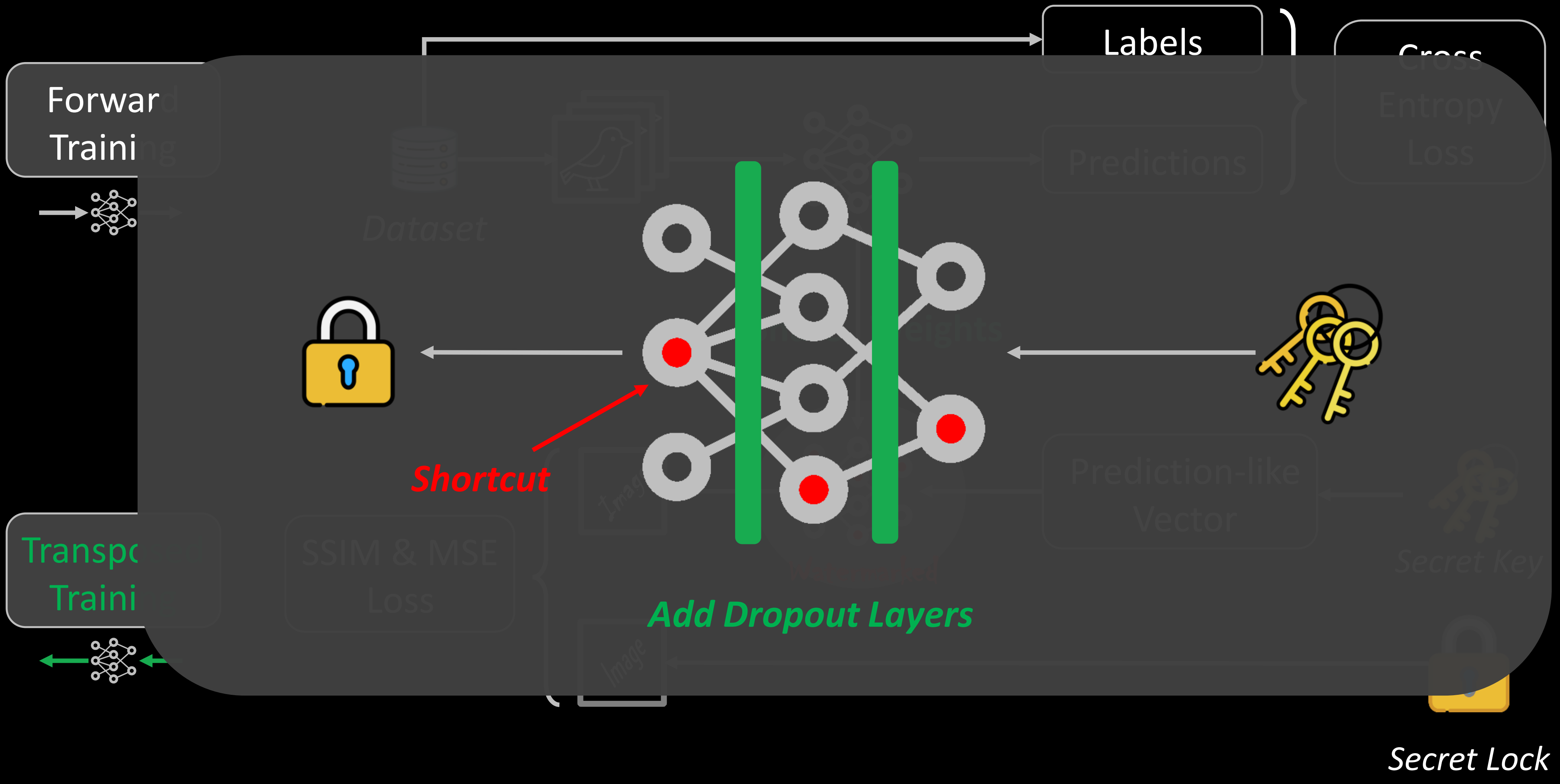
Details in Paper

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998.
[2] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. CVPR, 2010.
[3] Afia Zafar, Muhammad Aamir, Nazri Mohd Nawi, Ali Arshad, Saman Riaz, Abdulrahman Alruban, Ashit Kumar Dutta, and Sultan Almotairi. A comparison of pooling methods for convolutional neural networks. AppliedSciences, 2022.
[4] Olivier Rukundo and Hanqiang Cao. Nearest Neighbor Value Interpolation. IJACSA, 2012.
[5] Olivier Rukundo and Bodhaswar T Maharaj. Optimization of Image Interpolation based on Nearest Neighbour Algorithm. VISAPP, 2014.
[6] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014.
[7] Abien Fred Agarap. Deep Learning using Rectified Linear Units (ReLU). arXiv preprint arXiv:1803.08375, 2018.

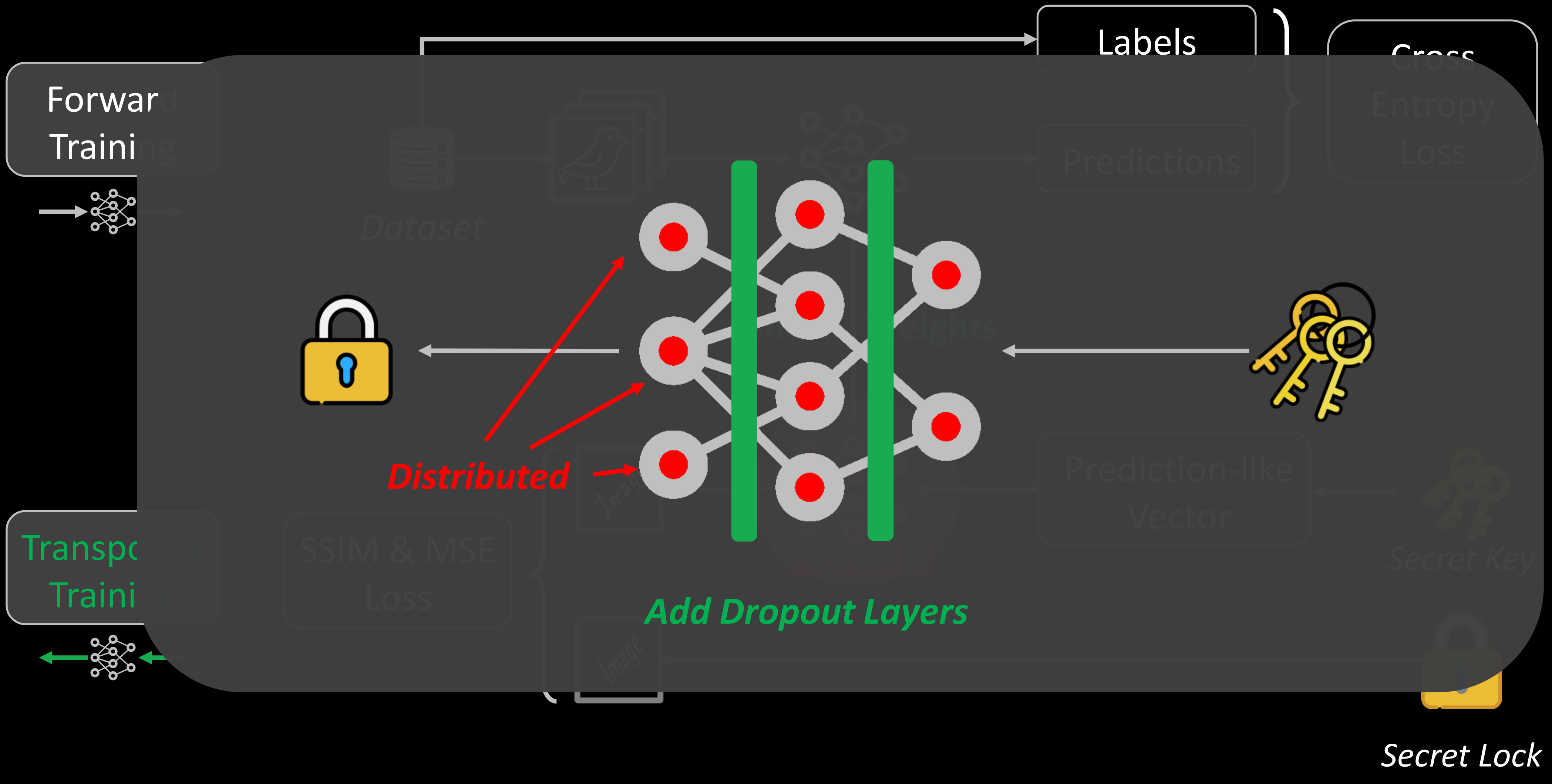
Parameter Entanglement



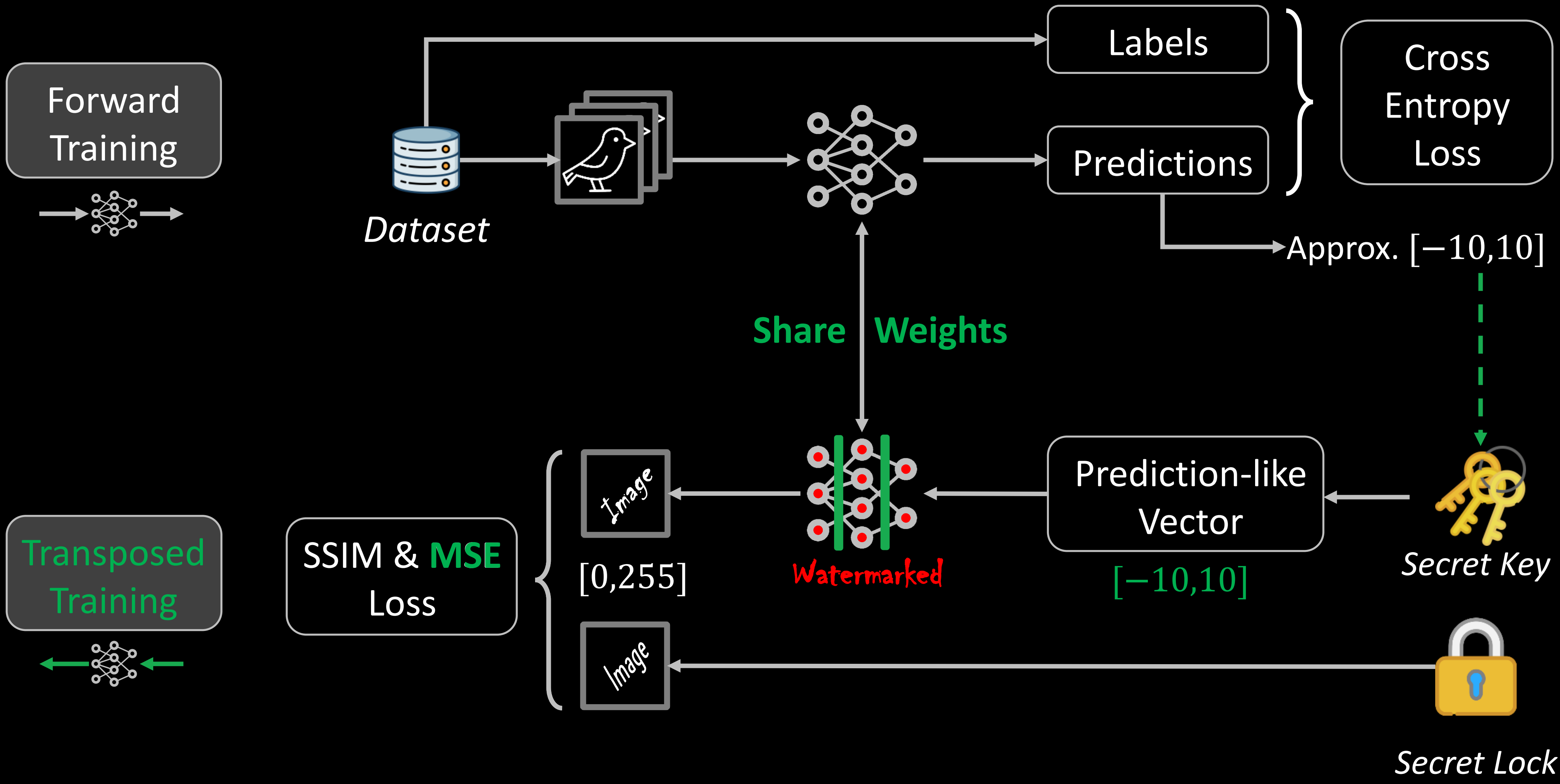
Parameter Entanglement



Parameter Entanglement



Parameter Entanglement



ClearStamp - Workflow

WATERMARKED

1. Watermark Hardening
2. Constraint Training

Predictions

Predictions

Legal / Illegal Model Distribution

3. 3rd Party Manipulation

Predictions

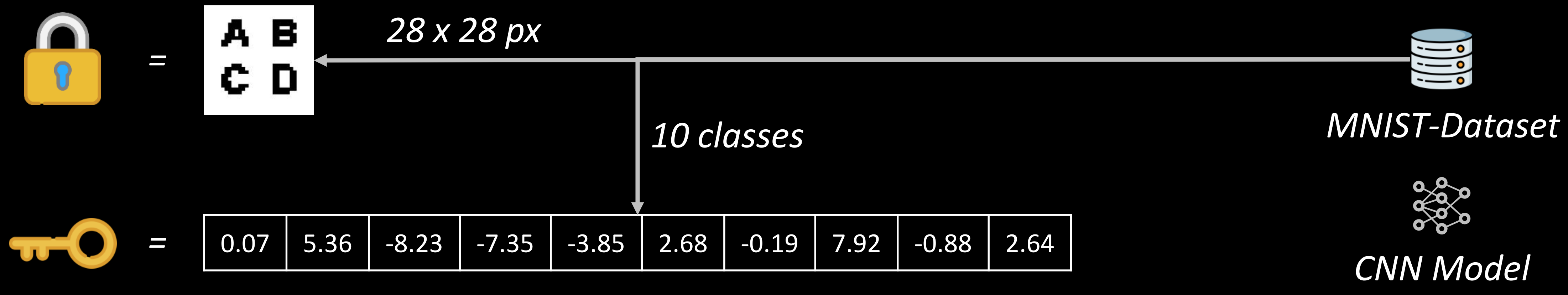
Copyright Infringement?

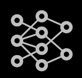
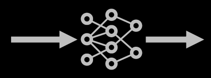
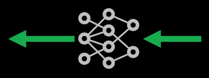
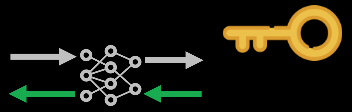
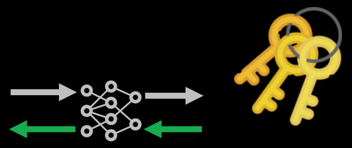
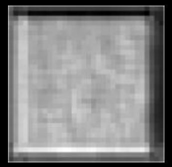
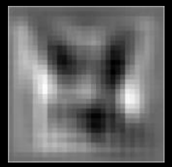
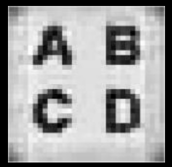
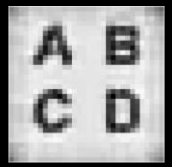
4. Watermark Testing

Image

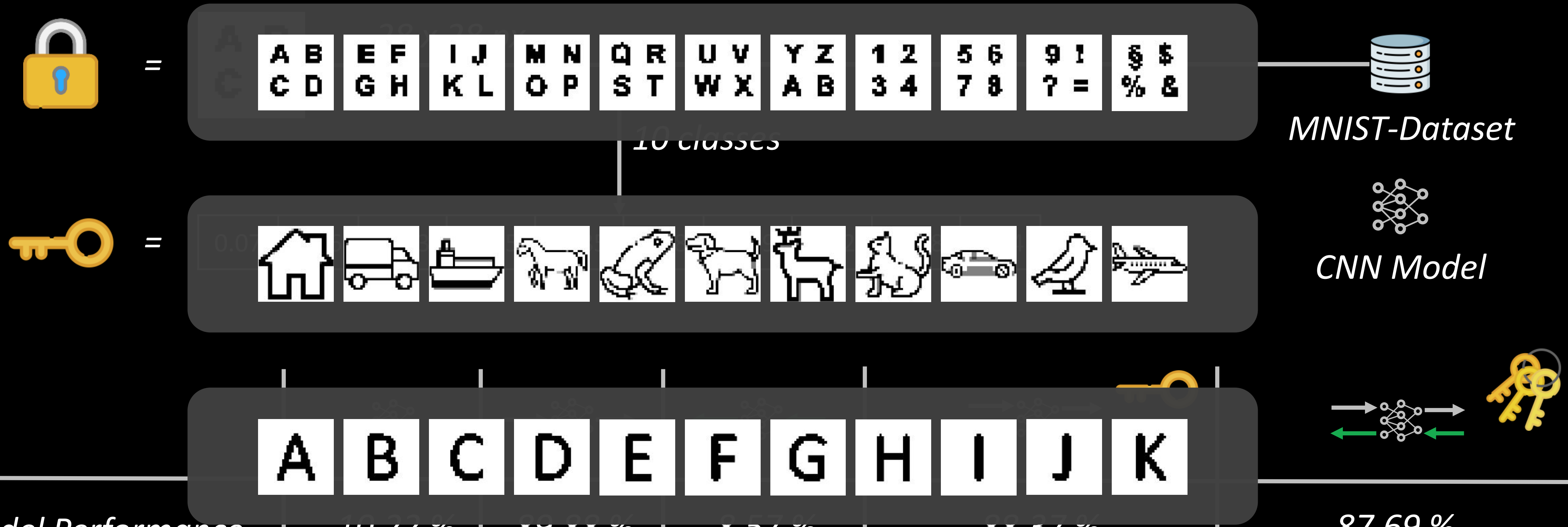
Image

ClearStamp - Evaluation



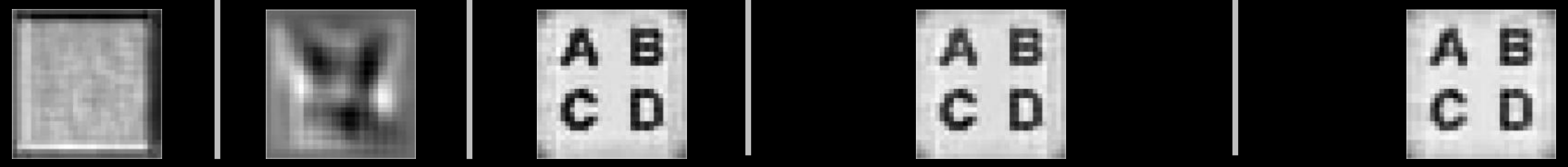
					
Model Performance	*10.22 %*	*89.88 %*	*8.57 %*	*88.37 %*	*87.69 %*
Extracted Watermark					

ClearStamp - Evaluation

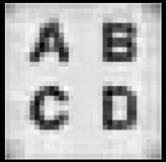
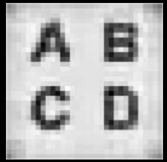
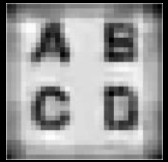
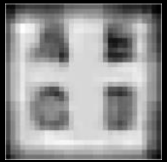
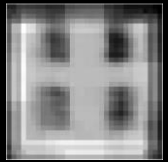
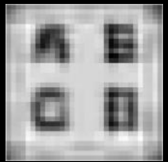


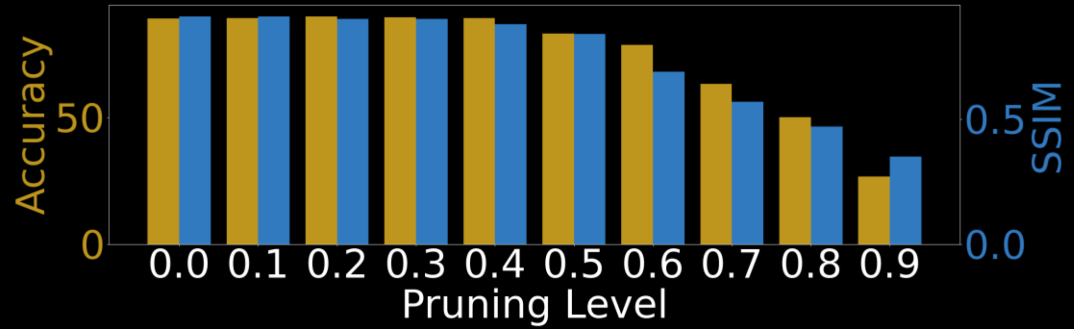
Model Performance	*10.22 %*	*89.88 %*	*8.57 %*	*88.37 %*	*87.69 %*

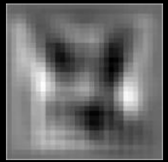
Extracted Watermark



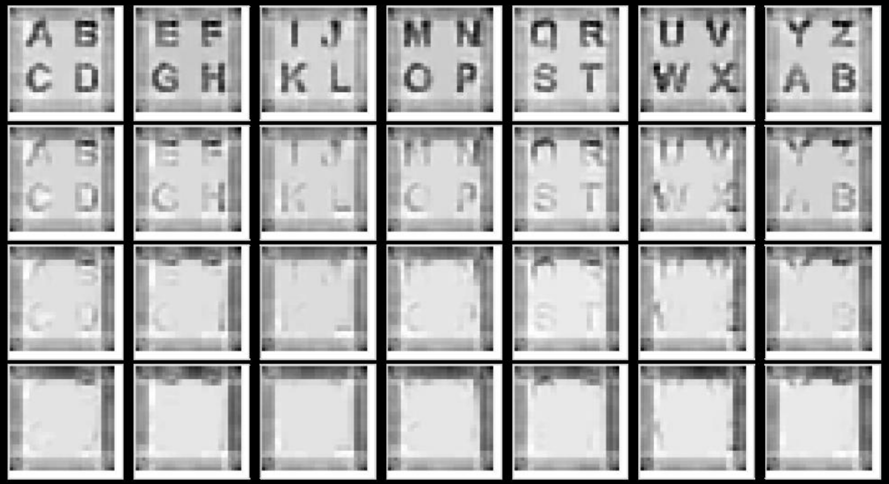
ClearStamp - Evaluation

	Fine-Tuning			*Pruning*			*Fine-Pruning*
	Same LR	$^1/_{10}$ *LR*	$^1/_{10}$ *LR & unseen data*	*60 %*	*80 %*	*90 %*	$^1/_{10}$ *LR & 40 %*
Performance	*87.42 %*	*89.86 %*	*97.02 %*	*78.56 %*	*50.10 %*	*26.76 %*	*89.79 %*
Extracted Watermark							





ClearStamp - Evaluation

Erase Watermark	*Performance*
	*83.89 %* *70.38 %* *56.39 %* *46.52 %*

Capacity

MiT License Text

8,544 bits

Bit Error Rate

5.92 %

Dot code

Runtime

Hardening

53.48 s ***(one time)***

Training

67.62 s → 94.39 s

+ 39.58 %

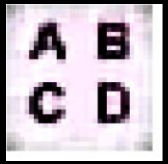
Extraction

0.02 s

ClearStamp - Evaluation

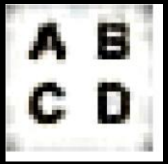
CNN

CIFAR-10



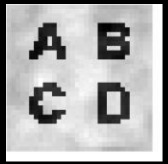
CNN

GTSRB



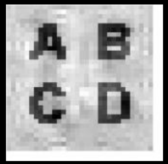
Small CNN

MNIST



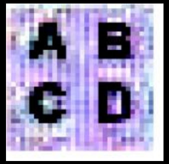
CNN + Batchnorm

MNIST



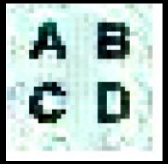
ResNet-18

CIFAR-10



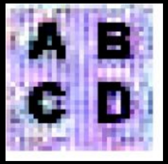
ResNet-18

GTSRB



ResNet-34

CIFAR-10

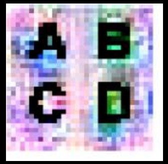


VGG11

CIFAR-100

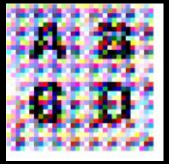
ResNet-18

CIFAR-100 → CIFAR-10

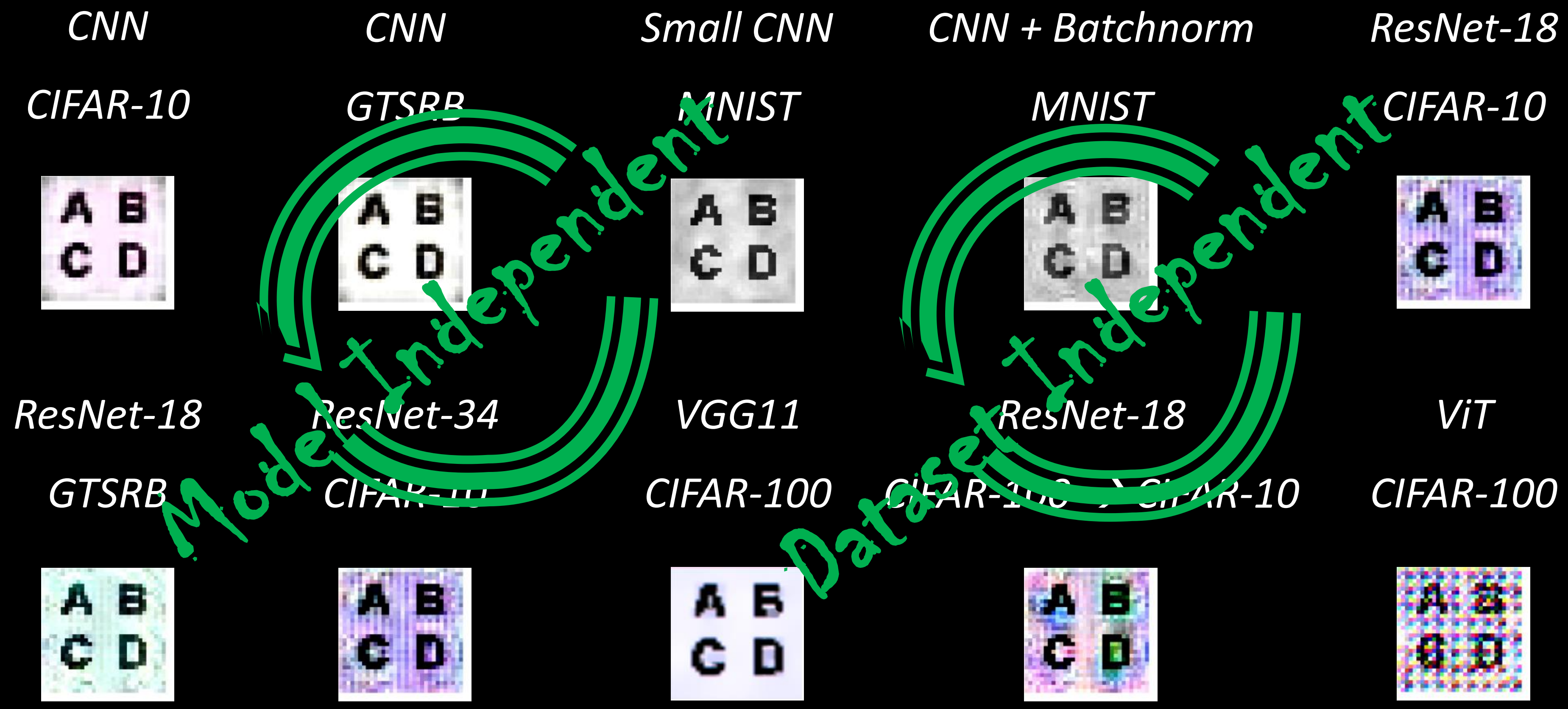


ViT

CIFAR-100



ClearStamp - Evaluation



Conclusion

Copyright Infringement of ML models

Watermarking of ML models

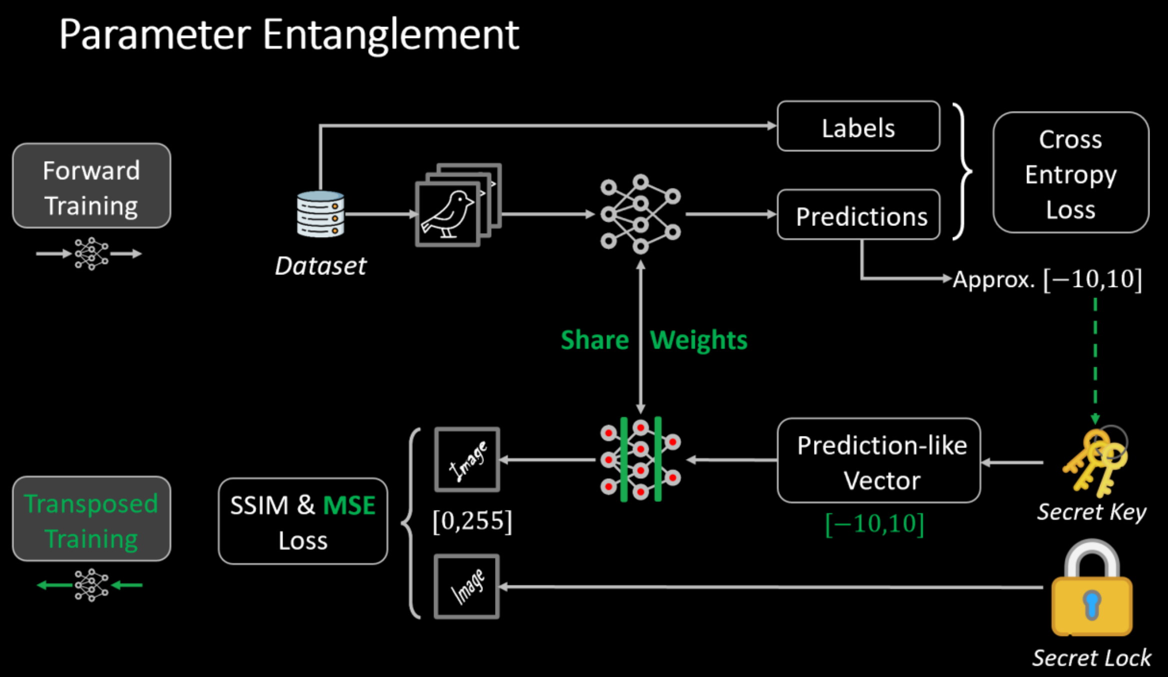
- Non-intuitive algorithms
- Non-human-interpretable
- Rigid threshold
- Partially removed watermarks

Transposed training to generate a **human-visible watermark**

ClearStamp

Thank you!!11!!1

Any Questions?





Torsten Krauß, Jasper Stang, Alexandra Dmitrienko

University of **Würzburg**