Fusion Channel Attack with POI Learning Encoder

Xiwen Ren[1,&], Xinyao Li[2,&], Ling Ning[1], Changhai Ou[1,*]

[1] Wuhan University

[2] City University of Hong Kong

Abstract: In order to challenge the security of cryptographic systems, Side-Channel Attacks exploit data leaks such as power consumption and electromagnetic emissions. Classic Side-Channel Attacks, which mainly focus on mono-channel data, fail to utilize the joint information of multi-channel data. However, previous studies of multi-channel attacks have often been limited in how they process and adapt to dynamic data. Furthermore, the different data types from various channels make it difficult to use them effectively. This study introduces the Fusion Channel Attack with POI Learning Encoder (FCA), which employs a set of POI Learning encoders that learn the inverse base transformation function family and project the data of each channel into a unified fusion latent space. Furthermore, our method introduces an optimal transport theory based metric for evaluating feature space fusion, which is used to assess the differences in feature spaces between channels. This model not only enhances the ability to process and interpret multi-source data, but also significantly improves the accuracy and applicability of SCAs in different environments.

Keywords: Multi-channel attacks, Deep Learning, Self-Attention mechanism

1 Introduction

In the contemporary digital age, Side-Channel Attacks(SCAs) exploit unintentional information leaks such as power consumption and electromagnetic(EM) emissions to avoid cryptographic security measures [1]. As digital technologies integrate more deeply into everyday life, SCAs are becoming increasingly relevant and raising risks such as identity theft [2] and unauthorised surveillance [3]. Initially, SCAs relied on single physical channels for data extraction. Nevertheless, the development of cryptographic protections has led to the employment of multi-channel approaches that enhance attack power through efficient data fusion techniques. These techniques have evolved from simple data splicing

&:These authors contributed to the work equally and should be regarded as co-first authors.
*Corresponding author.

methods [4] complex feature fusion methods [5], such as Principal Component Analysis(PCA) and Linear Discriminant Analysis(LDA) [6]. However, multi-channel attacks face significant challenges, including high computational complexity from increased input dimensions, dependency on complete data from all channels, and limited scalability after model training, which complicates the adaptation and processing of dynamic data [7]. It is therefore vital to address these issues to improve the reliability and practicality of multi-channel attacks in the context of digital security threats.

Multi-channel attacks offer advantages, but existing analytical techniques have limits. They struggle to handle the variability and high dimensionality of multi-channel data, which is crucial for developing more robust security measures. In this paper, we systematically investigate the question: Is the model really interpretable just because it effectively fuses information from different channels? We introduce a groundbreaking model, called Fusion Channel Attack with POI Learning Encoder (FCA), which employs a set of POI Learning encoders that learn the inverse base transformation function family and project the data of each channel into a unified fusion latent space. Our methodology not only enhances the ability to work with multi-channel data, but also challenges the conventional view of model interpretability in the context of integrated data sources.

Contributions. We summarize the contributions as follows:

- **A novel Fusion Channel Attack.** Our method uses a feature extractor based on the self-attention mechanism to explicitly encode points of interest (POIs) from different channels, compressing data from different channels into the same latent space. This method allows the same model to learn feature information from multiple channels, improving the efficiency and effectiveness of multi-channel data processing.

- **Interpretable integration evaluation index.** Our method introduces an optimal transport theory based metric for evaluating feature space fusion, which is used to assess the differences in feature spaces between channels. The effectiveness of the fused feature spaces is extensively demonstrated using this evaluation metric and visualisation techniques.

- **A new data set for evaluation.** In addition to proposing theoretical models and evaluation methods, we also construct a comprehensive dataset for assessment purposes. This dataset uses the new fusion degree metric to evaluate various existing multi-channel fusion attack methods. In particular, by fusing feature spaces from power and electromagnetic channels, the approach achieves more effective key recovery attacks on cryptographically protected devices with masking.

2 Methodology

By leveraging data from different channels, multi-channel attacks can significantly increase the amount of leakage information. However, utilizing multi-channel datas is still challenging. Often requiring specially designed models or complex data processing methods to implement attacks.Why is it not feasible to directly use multi-channel data to train a model? Beyond factors such as data dimensionality, we believe that the primary reason is data heterogeneity.

2.1 Heterogeneity of Multi-channel Data

Suppose the data from different channels is collected simultaneously from the same device. In this case, different channels' information can be viewed as representations of the same private variable (e.g. keys, intermediate variables). However, even though these data expose the same information, it is challenging to directly utilize them for attacks. Taking deep learning-based attack methods as an example, training a convolutional neural network (CNN) directly with data from different channels can lead to significant degradation in model performance. We believe this is because the feature spaces of data from different channels differ significantly, making it challenging to utilize this data simultaneously. The differences in feature spaces is referred as data heterogeneity. To more precisely demonstrate data heterogeneity, we employ UMAP (Uniform Manifold Approximation and Projection) to visualize the high-dimensional manifolds of the data [8]. UMAP is a dimensionality reduction technique that constructs a weighted k-nearest neighbor graph and optimizes its layout in a lower-dimensional space to preserve the topological structure. The edge weights of the graph are defined by:

$$\mu_{ij} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right) \tag{1}$$

where $d(x_i, x_j)$ is the distance between data points x_i and x_j, ho_i is the distance to the nearest neighbor of x_i, and σ_i is a scaling factor. The UMAP is based on the following optimization objective:

$$\mathcal{L} = \sum_{(i,j)\in G} \mu_{ij} \log\left(\frac{\mu_{ij}}{\nu_{ij}}\right) + (1-\mu_{ij}) \log\left(\frac{1-\mu_{ij}}{1-\nu_{ij}}\right) \tag{2}$$

where u_{ij} is the probability of an edge between points i and j in the low-dimensional space. The optimization process captures the high-dimensional manifold structure of the data in a lower-dimensional representation, making it ideal for visualizing data heterogeneity.

Taking electromagnetic side channel and power side channels as examples, data from both channels are collected from the same cryptographic device and have been standardized and aligned. The dimension of both electromagnetic and power data are reduced to 700 using Principal Component Analysis(PCA) to eliminate the impact of different channel noises. As shown in Fig 1, the shapes of the two channel datas exhibit significant differences. The UMAP visualization can be seen as a projection of the high-dimensional manifold onto a two-dimensional space, indicating that the power and electromagnetic information have significant differences in their high-dimensional manifolds. The difference

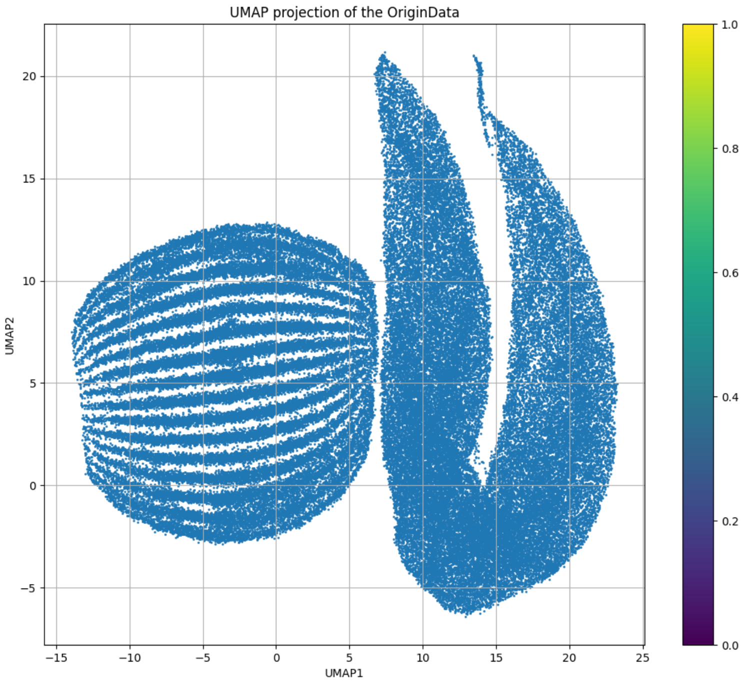

Figure 1: Visualization of Channel Feature Space

in manifolds demonstrates the disparity in the feature spaces of the two channels, as they are collected from the same device and correspond to the same private variable set.

2.2 Fusion Channel Transformation

The difficulty in directly utilizing data from different channels is due to the disparity in their feature spaces. Therefore, we aim to transform the feature spaces of the two channels into a similar latent space where they share comparable manifolds.

We refer to the unified latent space as the fusion channel space. The transformation of the feature spaces is referred as fusion channel Transformation.

The used side channels are defined as the set I.We define a base transformation functions family as $\mathcal{F} = \{f_i \mid i \in I\}$. The feature space of channel i can be viewed as projections under the base transformation function f_i. We denote the feature set of different channels as C. The transformation process can be expressed as:

$$\{c_i = f_i(a), c \in C, f \in \mathcal{F}, i \in I\} \tag{3}$$

where a is the latent variable shared by each channel, representing that the information from different channels corresponds to the same private variable.

We use a set of decoders D to learn the base transformation function family $\mathcal{F}$. The learning objective for this process is:

$$\min loss(d_i(a), f_i(a)), d \in D, f \in \mathcal{F} \tag{4}$$

A set of encoders E is adopted to learn the inverse projection process from c_i to a. The learning objective for encoders is set as:

$$\min loss(e_i(a), c_i), e \in E, c \in C \tag{5}$$

Combining these two learning processes, the final learning objective of the encoder and decoder can be expressed as:

$$\min loss(D(E(C)), \mathcal{F}(a)) \tag{6}$$

equivalently,

$$\min loss(D(E(C)), C) \tag{7}$$

However, it is evident that training a neural network based on this learning objective can easily lead to a collapsed solution where $E(x) = x$ and $D(x) = x$. To avoid such collapse, we expect the encoder and decoder to focus on the more valuable parts of the features, known as Points of Interest (POI), which refers to specific time instances in the side-channel trace that are highly correlated with the private variable.

2.3 POI Learning with Self-attention Algorithm

To make the process of finding POI learnable, we utilize the self-attention algorithm [9] to enable the encoder and decoder to automatically learn the POI of the data through training. The self-attention algorithm allows the model focus on the more valuable segment of the feature. It involves computing attention scores that determine the importance of each part of the input data. The attention matrix is calculated by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

where Q (queries), K (keys), and V (values) are linear transformations of the input data, and d_k is the dimension of the keys. Based on the self-attention algorithm, the model can better integrate information from different time slices as well. This integration process enables the model to effectively attack cryptographic devices protected by masking algorithms.

To verify the effectiveness of POI learning, we compared the performance of the POI learning encoder with the common neural network(a fully connected neural network) during the training process. We trained them using the same channel data based on the proposed training objective. The mean squared error(MSE) function is adopted as the loss function. We present the loss value changing during the training process. As shown in Fig 2 and Fig 3, the loss values of the fully connected network did not show a significant downward trend, while the loss of the POI learning encoder steadily decreased to zero. The result indicates that neural networks based on POI learning have a better feature extraction capability.

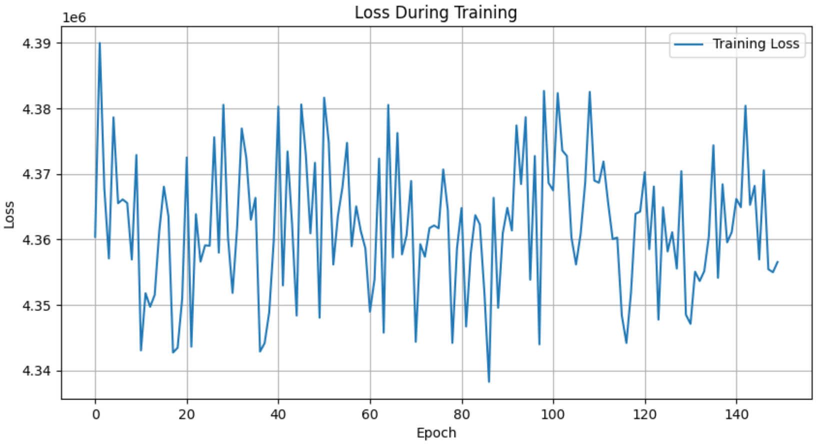

Figure 2: Loss Changing of MLP

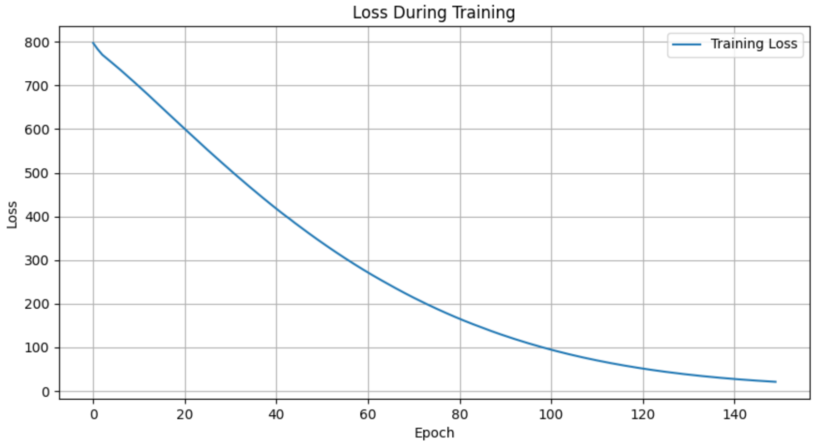

Figure 3: Loss Changing of POL Learning Network

Based on our proposed learning objective, for a set of channels I, we train a set of Encoder and Decoder pairs $\{E, D\}$. Based on Eq 5, we can obtain the latent variable a by inversely projecting the channel data c_i using Encoder e_i. In this work, we use a three-layer transformer network as the POI learning encoder with 8 self-attention heads. The Adam optimizer is employed for training. We obtain the latent variable a for power and electromagnetic channels. The UMAP visualization of the latent variables is shown in the Fig 4.

2.4 Evaluation of Fusion Channel

How do we verify that the obtained latent variable a is in the same feature space? We propose a quantitative metric called the Reduced-DW distance based on the optimal transport theory to measure the difference between two feature spaces. The DW (Distributional Wasserstein) distance is a measure of the dissimilarity between two probability distributions [10]. It quantifies the minimum cost of transporting mass to transform one distribution into another, which is particularly useful for comparing the distributions of

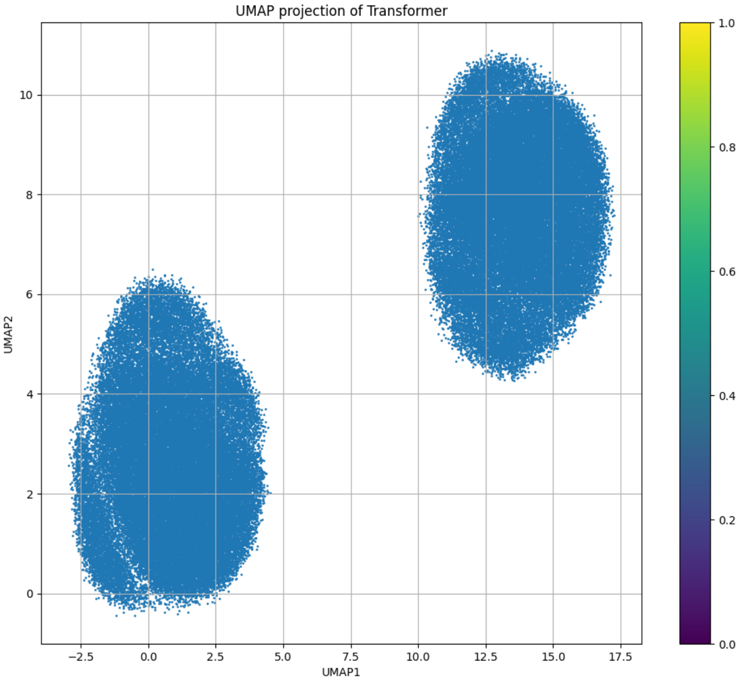

Figure 4: Visualization of Latent variables

high-dimensional data. The DW distance between two distributions P and Q is defined as:

$$DW(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|] \tag{9}$$

However, since side-channel data typically has high dimensionality, directly calculating the DW distance between two channel data sets is time-consuming and may not reflect the true distribution of the data in high-dimensional space. Therefore, we propose the Reduced-DW distance, where the side-channel data is first reduced to two dimensions using UMAP, and then the DW distance between the two channel data sets is calculated.

We calculate the Reduced-DW distance between the power and electromagnetic signals after PCA dimensionality reduction, as well as the Reduced-DW distance between the latent variables of the two channels. The former distance is 0.7234, while the latter is 0.1041. It is evident that the feature space similarity between the latent variables is higher.

.

3 Conclusion

In this work we proposed a new multi-channel fusion framework based on with POI Learning Encoder, which uses a POI Learning encoder to innovatively integrate multi-channel data into a unified latent space. In a comprehensive case study we demonstrated the application of our proposed framework through key recovery attack experiments conducted on the publicly available multi-channel dataset, the SPERO dataset [11], yielding favorable outcomes. Figs. 5- 6 compare the optimized attack results of models with and without POI Learning encoding, following comprehensive parameter adjustments. Under specific parameter settings, the POI Learning encoded model demonstrates a marked improvement in key prediction accuracy when evaluated with a limited dataset, outperforming the non-POI Learning encoded model while requiring substantially fewer training epochs.

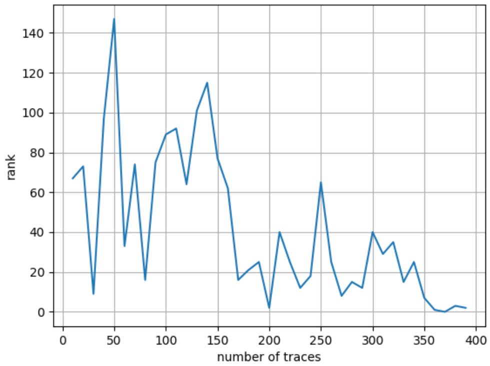

Figure 5: The key recovery results of the model trained without the POI Learning Encoder, under the conditions of 10 training epochs and a learning rate of 0.00001.

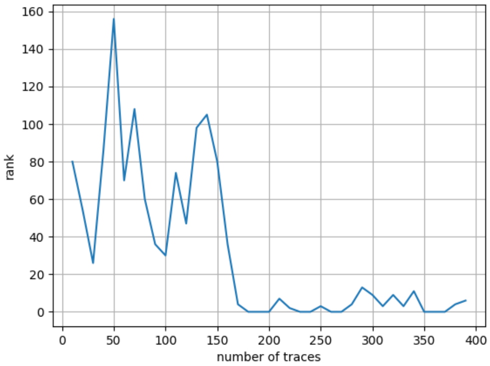

Figure 6: The key recovery results of the model trained with the POI Learning Encoder, under the conditions of 10 training epochs and a learning rate of 0.00003.

In this study, our framework not only learns and processes data from multiple channels through the same model, but also has a complete system of evaluation metrics to assess the spatial differences between different channels, making FCA interpretable. Then we not only propose the theoretical model and evaluation method, but also actually construct the dataset used for evaluation. Our work not only develops new ideas in theory, but also demonstrates its effectiveness in practical applications.

References

[1] Stefan Mangard, Elisabeth Oswald, and Thomas Popp. Power analysis attacks: Revealing the secrets of smart cards. 31, 2008.

[2] Tao Ni, Xiaokuan Zhang, and Qingchuan Zhao. Recovering fingerprints from in-display fingerprint sensors via electromagnetic side channel. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 253–267, 2023.

[3] Daniel Genkin, Noam Nissan, Roei Schuster, and Eran Tromer. Lend me your ear: Passive remote physical side channels on {PCs}. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4437–4454, 2022.

[4] Dakshi Agrawal, Josyula R Rao, and Pankaj Rohatgi. Multi-channel attacks. In *Cryptographic Hardware and Embedded Systems-CHES 2003: 5th International Workshop, Cologne, Germany, September 8–10, 2003. Proceedings 5*, pages 2–16. Springer, 2003.

[5] Wei Yang, Yongbin Zhou, Yuchen Cao, Hailong Zhang, Qian Zhang, and Huan Wang. Multi-channel fusion attacks. *IEEE Transactions on Information Forensics and Security*, 12(8):1757–1771, 2017.

[6] Yukio Tominaga. Comparative study of class data analysis with pca-lda, simca, pls, anns, and k-nn. *Chemometrics and Intelligent Laboratory Systems*, 49(1):105–115, 1999.

[7] E. De Mulder, P. Buysschaert, S.B. Ors, P. Delmotte, B. Preneel, G. Vandenbosch, and I. Verbauwhede. Electromagnetic analysis attack on an fpga implementation of an elliptic curve cryptosystem. In *EUROCON 2005 - The International Conference on "Computer as a Tool"*, volume 2, pages 1879–1882, 2005.

[8] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[10] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[11] Yunkai Bai, Rabin Yu Acharya, and Domenic Forte. Spero: Simultaneous power/em side-channel dataset using real-time and oscilloscope setups. *arXiv preprint arXiv:2405.06571*, 2024.