Shift-invariance Robustness of Convolutional Neural Networks in Side-channel Analysis

Marina Krček[1], Lichao Wu[2], Guilherme Perin[3], and Stjepan Picek[1,2]

[1] Delft University of Technology, The Netherlands
[2] Radboud University, The Netherlands
[3] Leiden University, The Netherlands

Abstract. Convolutional neural networks (CNNs) offer unrivaled performance in profiling side-channel analysis. This claim is corroborated by numerous results where CNNs break targets protected with masking and hiding countermeasures. One hiding countermeasure is commonly investigated in related works - desynchronization (misalignment). The conclusions usually state that CNNs can break desynchronization as they are shift-invariant. This paper investigates that claim in more detail and reveals that the situation is more complex. While CNNs have certain shift-invariance, it is insufficient for commonly encountered scenarios in deep learning-based side-channel analysis. We propose to use data augmentation to improve the shift-invariance and, in a more powerful version, ensembles of data augmentation. Our results show the proposed techniques work very well and improve the attack significantly, even for an order of magnitude.

Keywords: Side-channel Analysis · Deep learning · Misalignment · Countermeasures · Shift-invariance

1 Introduction

Convolutional neural networks are successfully applied in diverse domains, like image [10] and audio [13] processing. Moreover, CNNs have been successfully applied to profiling side-channel analysis (SCA). From the first paper on CNNs in SCA appearing in 2016, there have been more than 250 papers exploring the topic of deep learning-based side-channel analysis (DLSCA).[4] The primary motivation for using CNNs in DLSCA is their ability to 1) work with raw features, making feature engineering non-mandatory, and 2) deliver excellent attack performance even when the target is protected with countermeasures. While implicitly assumed by the SCA community, both points deserve more discussion.

While CNNs dramatically reduce the requirement of feature engineering, most works in the DLSCA domain consider traces with a pre-selected window of samples. More recently, results demonstrate that such selection is unnecessary:

[4] In [22], the authors reported 183 papers up to 2021. Searching papers published in 2022 and 2023 shows more than 100 additional papers.

CNNs can directly retrieve leakages from raw measurements and reach state-of-the-art performance [17,21]. Excellent attack performance, even in the presence of countermeasures, is commonly assumed but also confirmed by numerous empirical analyses. However, the rationale for how CNNs bypass these countermeasures must be clarified. Specifically, desynchronization is a commonly investigated hiding countermeasure. There, for "reasonable" desynchronization values, it is assumed that CNNs inherently handle it due to their shift-invariance property. Understanding where the shift-invariance comes from and how to improve it is a challenging problem. Although multiple works try to explain it from a theoretical perspective or improve shift-invariance by modifying CNN architecture or the dataset used for training, the mechanism is not fully understood. Since multilayer perceptron (MLP) architectures are not shift-invariant [3], we focus our attention on CNNs in this work.

While the core motivation for this work (data augmentation being the key to the shift-invariance of CNNs for DLSCA) may sound intuitive and well-known in the SCA domain, we argue this is not the case. We conducted a literature survey for DLSCA and examined around 230 related works.[5] Out of those 230 works, 70 use datasets with desynchronization. Surprisingly, only 17 out of those 70 works use data augmentation. Thus, with 24% papers using data augmentation, we cannot claim it is a well-known or accepted technique to fight desynchronization. Furthermore, we analyzed the 230 papers to see if they mention the shift-invariance property of CNNs and what is the cause of it. We found nine papers discussing how the convolutional layer is the key to shift-invariance; four papers claim it is due to the pooling layer. From a more general perspective, 18 papers claim it is due to CNNs, without going into specifics. Unfortunately, no papers in DLSCA directly connect data augmentation and shift-invariance.

To conclude, despite the numerous related works, the shift-invariance of CNNs in DLSCA remains an open question. More precisely, it needs to be clarified if CNNs have sufficient shift-invariance and, if not, how to extend it. Our main contributions are:

1. We show that commonly used CNNs in DLSCA have a limited shift-invariance and should be carefully used when attacking desynchronized datasets. Additionally, our experiments show that the pooling layer has a negligible effect on the shift-invariance of CNNs. Still, this might need a separate, more thorough investigation.
2. We show how data augmentation can successfully improve the shift-invariance of CNNs in DLSCA.
3. We show how to use ensembles of data augmentation settings to provide superior attack performance using CNNs when dealing with highly desynchronized datasets.

We emphasize that understanding what part of CNN gives shift-invariance and how to design a shift-invariant CNN are problems that have been open for several years. This work does not aim to solve those problems by providing a

[5] We do not claim this is an exhaustive search, but we are confident it is representative, as it contains more works than a recent systematization of knowledge work [22].

universal solution regardless of the evaluated dataset. Instead, we propose an efficient data augmentation approach to improve the shift-invariance of an existing model. Although one could potentially reduce the effects of desynchronization by using deeper architectures or custom architecture elements, we decided not to follow those directions as they will 1) make the hyperparameter tuning more challenging and 2) increase the chances for overfitting. On the other hand, using data augmentation and ensembles is a simpler yet very successful approach.

2 Deep Learning-based Side-channel Analysis

As commonly done in profiling attacks, we consider a setup with two phases: profiling and attack. The profiling phase corresponds to training a machine learning-based classifier, and the attack phase to test its classification performance. Let us assume a side-channel dataset $\mathcal{X}$ that consists of N measurements. $\mathcal{X}$ can be considered as a 2D array with N rows and J columns. Each point in the array $\mathcal{X}$ is an element $t_{i,j}$, where i indicates the side-channel trace index and j indicates the index of a sample (feature) inside a side-channel trace $\mathbf{x}_i$. The profiling phase aims to find the parameters $\boldsymbol{\theta}$ of a function f that minimize the empirical risk (the expected value of the loss function). In the test phase, the goal is to predict labels $\boldsymbol{Y}$ based on the traces from the attacked device and the trained model $f_{\boldsymbol{\theta}}$. More precisely, the result of predicting with a model $f_{\boldsymbol{\theta}}$ on the test set is a two-dimensional matrix P with dimensions $Q \times C$, where Q is the number of attack traces and C is the number of labels. Each value in the matrix P denotes the probability that for a specific key $\mathbf{k}$ and input $\mathbf{a}$, we obtain the label $\mathbf{y}$. Finally, the cumulative sum $S(\mathbf{k})$ for any key byte candidate $\mathbf{k}$ is a side-channel distinguisher with a common maximum log-likelihood principle: $S(\mathbf{k}) = \sum_{i=1}^{Q} \log(\mathbf{p}_{i,\mathbf{y}})$. The cumulative sums for each possible key value form a key guessing vector ordered per the probability of the key being correct (the first position in the vector is the most likely key, and the last position is the least likely key). The position of the correct key is called the key rank, and it denotes the effort the attacker requires to break the target. To reduce the effects of a random choice of test measurements, it is common to assess the average behavior over many randomly selected traces, called guessing entropy (GE) (i.e., average key rank) [23]. We consider DLSCA already well-known in the SCA community, so we only briefly overviewed it. For more details, we refer readers to, e.g., [22].

Data Normalization. The application of data normalization is crucial to obtain better deep learning performance. The most common form of data normalization adopted in previous papers is *standardization* (see Section 3 in [25] and Section 4-B in [22] for a discussion about data preprocessing). Profiling, validation, and attack sets must be aligned for consistent data normalization. Otherwise, features in validation and attack sets are distorted when normalized with μ_p and σ_p (standardization). In our case, as we aim to investigate the shift-invariance robustness of trained deep neural networks, the profiling and attack sets might contain different desynchronization levels. Therefore, we use the Horizontal MinMax Scaler [25] for data normalization in the range $[-1, 1]$. The scaler is trained on and also fits the transposed trace sets individually.

Data Augmentation. Data augmentation refers to increasing the training set's size by artificially generating additional training data with dynamic changes during a model's training. These changes must preserve the class properties of the training set. The training set represents an approximate distribution, given by a finite set $\mathcal{T}$, from a true and unknown distribution $\mathcal{R}$. By augmenting the training set $\mathcal{T}$, one expects that the $\mathcal{T}$ becomes a better representation of $\mathcal{R}$. The main idea is to improve class representation inside a dataset, so it is essential to understand what kind of effect the augmentation needs to develop. Inappropriate data augmentation settings can lead to no or even detrimental effect [6].

Metrics. As the goal of this work is to improve the attack efficiency of SCA, the shift-invariance robustness is quantified with the guessing entropy metric and the corresponding number of attack traces to reach a GE equal to 1 ($N_{\mathsf{ge=1}}$), which means successful key recovery. In all our experiments, we provide GE and corresponding $N_{\mathsf{ge=1}}$ results for a single attacked key byte from a 128-bit AES key.

3 Related Works

Cagli et al. used data augmentation to defeat datasets protected with software-based countermeasure (Random Delay Interrupt) and hardware-based countermeasure (jitter) [4]. The authors broke the targets and suggested using data augmentation to reduce overfitting. Interestingly, they did not mention data augmentation as the technique to help defeat desynchronization. B. Timon proposed a non-profiled deep learning-based SCA approach and used data augmentation to improve the learning phase [24]. Perin et al. used data augmentation to improve the performance of the deep learning-based attack on a protected ECC implementation [19]. The authors stated that data augmentation defeats overfitting, and the investigated datasets do not have desynchronization. Lu et al. used data augmentation to expand the profiling set and improve the attack performance [17]. The authors mentioned convolutional layers as relevant for shift-invariance. Perin et al. discussed various feature selection scenarios and broke several datasets [21]. The authors used data augmentation for desynchronized datasets and reported significantly fewer attack traces needed if using data augmentation in the training process. The authors did not discuss the connection between shift-invariance and data augmentation.

Since the publicly-available ASCAD datasets provide 50 and 100 samples window desynchronization [2], most papers either use synchronized datasets or desynchronizations up to 100 samples.[6] Cagli et al. used desynchronization levels up to 200 samples [4], which they managed to circumvent by using deep neural network architectures and data augmentation. Lu et al. investigated the performance of CNNs against the ASCAD dataset with desynchronization of up to

[6] Naturally, the desynchronization window should be considered relative to the length of the trace.

200 samples [17]. Zotkin et al. used data augmentation and attacked datasets with desynchronization of up to 150 samples [29]. Aligned with several DLSCA works, the authors attributed a shift-invariance property to CNNs. Hajra et al. considered scenarios where the datasets had a significant level of desynchronization, up to 400 samples and used transformer networks to handle it [9]. The authors tested scenarios where a neural network is trained on synchronized and attacked desynchronized datasets. The good performance is attributed to the shift-invariance property of CNNs.

Finally, we note that ensembles were used in various contexts in DLSCA, improving attack performance. For instance, Perin et al. used ensembles to combine predictions from multiple neural networks [20], while Zaid et al. proposed a loss function called Ensembling Loss [27].

4 CNNs and Invariance

In the context of CNNs, three general types of invariance are commonly discussed: translation (shift), rotation, and scale invariance. Of those three, shift-invariance is discussed in the context of DLSCA as it becomes relevant to defeat various desynchronization countermeasures. It is a relatively common (and long-lasting) claim that CNNs are invariant due to the convolution layer, e.g., [14,7]. More recently, it was recognized that convolutional layers could have the property of shift-equivariance and not shift-invariance [3,12]. Convolution is equivariant to a function g that translates the input, i.e., a shift of the input to a convolutional layer produces a shift in the output feature maps by the same amount [3]. The convolutional layers can lose the perfect equivariance due to downsampling [1]. In CNNs, downsampling happens when the pooling or convolutional stride is greater than one, as the intermediate representation skips samples in the input. It is a pooling layer that provides approximate shift-invariance to small translations of the input [8]. Shift-invariance means that translating the input by a small amount will not cause the values of most of the pooled outputs to change. Intuitively, pooling achieves local translation invariance [18]. Invariance is a particular case of equivariance when the transformation has no effect [15]. Combining the shift-equivariance property of convolutional layers and stability to deform pooling layers, one could expect CNNs to become shift-invariant [5]. Unfortunately, recent results showed this is not the case since small input shifts can cause drastic changes in the output [1,28]. The reason is the downsampling nature of modern convolutional and pooling layers. Some works also showed that CNNs could encode absolute spatial location in images, resulting from a lack of shift-invariance [12]. Additionally, one cannot definitively answer what part of CNNs brings the most invariance. For instance, it has been shown that the network depth and not the type of the layer contribute more to the shift-invariance [11]. This is relevant for the DLSCA perspective, as most of the results (including state-of-the-art) commonly use relatively shallow neural network architectures, e.g., [21]. Finally, larger convolution filter sizes allow for more shift-invariance [11].

5 The Shift-invariance Robustness of Deep Learning Models in SCA

In this section, we deploy several experiments to empirically verify the shift-invariance robustness of various deep neural networks in SCA. Moreover, we evaluate how data augmentation improves shift-invariance robustness and how this robustness is directly related to data augmentation settings and the shift distribution in side-channel measurements. We define four main scenarios:

1. Neural networks are trained with a synchronized profiling set.
2. Neural networks are trained with a desynchronized profiling set.
3. Neural networks are trained with synchronized profiling set and data augmentation.
4. Neural networks are trained with a desynchronized profiling set and data augmentation.

The trained networks are tested on an attack set with different levels of desynchronization to showcase the performance.

Datasets. We consider the ASCAD datasets[7], which are protected with first-order Boolean masking. In total, we consider four different datasets:

1. `ASCADf`: this dataset contains side-channel measurements consisting of 700 features representing side-channel leakages from processing the third `S-Box` output byte in the first AES encryption round. We split this dataset into profiling, validation, and attack sets with 45 000, 5 000, and 5 000 traces, respectively. All sets have the same fixed key.
2. `ASCADr`: this dataset contains side-channel measurements where each trace consists of 1 400 features. Similar to `ASCADf`, this measurement interval also represents side-channel leakages from processing the third `S-Box` output byte in the first AES encryption round. We split this dataset into profiling, validation, and attack sets with 200 000, 5 000, and 5 000 traces, respectively. The profiling set has random keys, while the validation and attack sets have the fixed key.
3. `ASCADf_desync100`: desynchronized version of `ASCADf`, and each trace is randomly shifted inside the interval $[0, 100]$. Shifts are drawn from a uniform distribution.
4. `ASCADr_desync100`: desynchronized version of `ASCADr`, and each trace is randomly shifted (also from a uniform distribution) inside the interval $[0, 100]$.

Datasets are always labeled according to the identity leakage model for the third `S-Box` output byte in the first AES encryption round.

Deep neural network architectures. The neural network architectures considered in the experiments of this section are described below. The CNN architectures always contain convolution blocks consisting of a convolution layer followed by a batch normalization layer connected to a pooling layer. Both average and max

[7] `https://github.com/ANSSI-FR/ASCAD`

pooling layers are tested. All hyperparameters were obtained with a random hyperparameter search, with the best-found architectures reported in Table 1. The architectures are denoted as `cnn_pooling_2`, `cnn_pooling_4`, `cnn_pooling_6`, and `cnn_ascadr_desync100`. We always train the models for 100 epochs. The batch size of each case will be reported in the corresponding section.

Table 1: Hyperparameter values for `cnn_pooling_2`, `cnn_pooling_4`, `cnn_pooling_6`, `cnn_ascadr_desync100`.

Hyperparameters	cnn_pooling_2	cnn_pooling_4	cnn_pooling_6	cnn_ascadr_desync100
Conv. blocks	2			5
Filters (ordered)	8, 16			12, 24, 36, 48, 60
Kernel size	30			40
Strides	2			15
Pool size	2	4	6	6
Pool strides	2	4	6	6
FC layers	1			2
FC Nb. neurons	40			50
Weight init	glorot_normal			glorot_uniform
Activation	elu			elu
Learning rate	0.001			0.00025
Optimizer	RMSprop			Adam
Epochs	100			

We analyze `cnn_pooling_2`, `cnn_pooling_4`, and `cnn_pooling_6` that contain different pooling sizes to understand if small changes in the pooling size have any effect on shift-invariance. Moreover, all models listed above contain large convolution kernel sizes (e.g., 40). Still, in this work, we skip a more detailed evaluation of the effect of filter kernel size (and its strides) on shift-invariance robustness. We evaluate architectures that provided the best results (i.e., minimum $N_{\mathtt{ge=1}}$) from our random hyperparameter search and also architectures with good performances from some related works on the evaluated datasets. Specifically, we test CNN architectures from [26] and [25] designed for the `ASCADf` and `ASCADf_desync100` datasets. Hyperparameter values are kept the same as in those papers. We denote the architectures:

– `zaid_ascad_desync_0` for architecture from [26] for `ASCADf`.
– `noConv1_ascad_desync_0` for architecture from [25]. The network is a reduced version of `zaid_ascad_desync_0`, where the convolution and batch normalization layers are removed, and the input layer is the pooling layer.
– `zaid_ascad_desync_100` for architecture from [26] for `ASCADf_desync100`.
– `noConv1_ascad_desync_100` for design from [25]. This architecture is a reduced version of `zaid_ascad_desync_100` in the same manner as `noConv1_ascad_desync_0`.

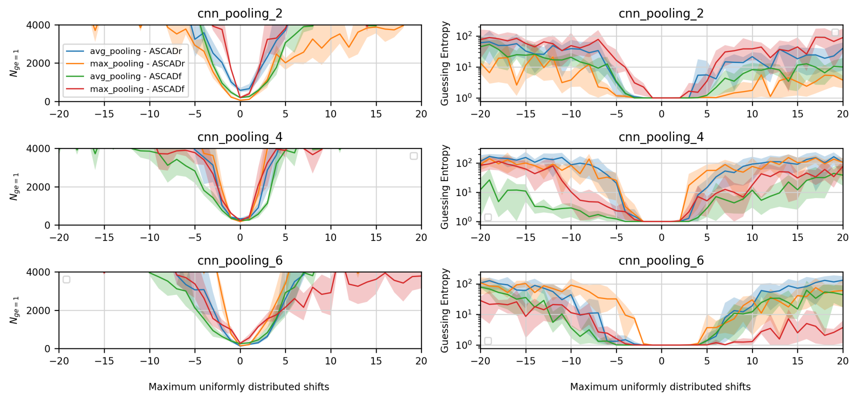
For all architectures, the loss function is always categorical cross-entropy.

Reading the graphs. The x-axis in all figures (except those illustrating the shift distributions) indicates the maximum number of shifts δ drawn from a uniform distribution and applied to the set of attack traces. For instance, when the x-axis indicates the value of 100, the set of attack traces is randomly shifted by values inside the interval $[0, 100]$. For negative x-axis values $-\delta$, the interval is $[-\delta, 0]$. To treat the trace samples at the trace boundaries, we first apply random shifts to the attack set, and after, we trim it. The shaded area around lines in the plots indicates the minimum and maximum values obtained from 10-fold experiments, while the main line is the average value. The gray region in the plots indicates the target shift-invariance region.

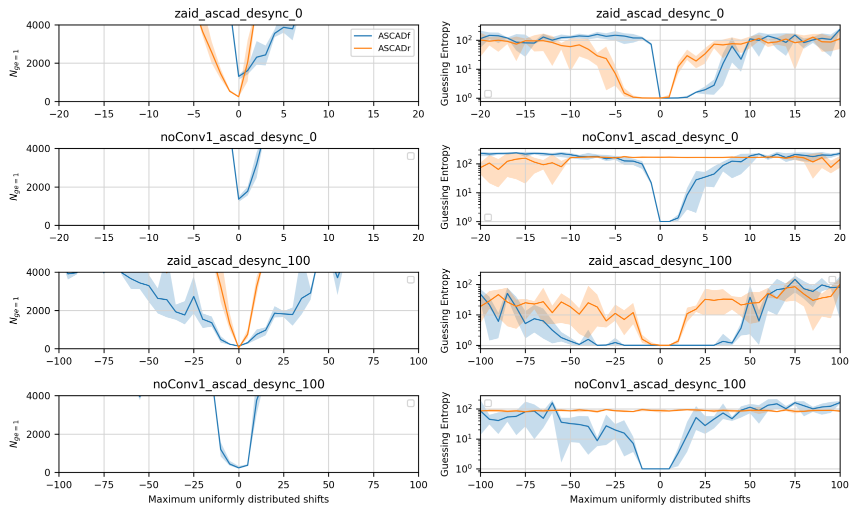
5.1 Shift-invariance Robustness of CNNs Trained with Synchronized Datasets

First, we analyze the shift-invariance robustness of `cnn_pooling_2`, `cnn_pooling_4`, and `cnn_pooling_6` architectures with the `ASCADr` and `ASCADf` datasets. Figure 1a shows the shift-invariance robustness of these three models. As we can see, the shift-invariant robustness of these networks, when trained on the synchronized datasets, is very limited in the side-channel context. Note how randomly shifting a few trace samples to the right (positive) or the left (negative) already significantly affects the model performance. With more than five sample shifts, the GE of attacked key byte substantially increases. The results show no significant difference in the shift-invariance robustness between the tested pooling size. Different pooling types (i.e., average or max pooling) show no important effects on shift-invariance for these three CNN architectures.

In the second group, we consider the CNN architectures `zaid_ascad_desync0` [26] and `noConv1_ascad_desync0` [25] that are designed for the synchronized `ASCADf` dataset. These architectures and the architectures from our first group were not designed to defeat desynchronized datasets. As shown in Figure 1b, `zaid_ascad_desync0` and `noConv1_ascad_desync0` provide a more limited shift-invariance robustness compared to CNNs from Figure 1a. The performance of `zaid_ascad_desync0` and `noConv1_ascad_desync0` models is significantly inferior compared to original results from [26,25]. The main reason comes from data normalization: in original papers, the authors considered (feature) standardization, while here, we apply horizontal MinMax normalization for the reasons mentioned in Section 2. We also show results for `zaid_ascad_desync100` and `noConv1_ascad_desync100` that were designed to break the `ASCADf_desync100` dataset. In particular, the `zaid_ascad_desync100` model provides much better shift-invariance, as randomly shifting the attack set inside the interval $[-60, 40]$ delivers successful key recovery when the attack set contains 4 000 traces. In contrast, the improved version, `noConv1_ascad_desync100` proposed in [25] provides significantly less shift-invariance robustness. Furthermore, `noConv1_ascad_desync100` cannot provide successful key recovery with the `ASCADr` dataset. This might be because the first convolution layer is absent in this architecture. The second convolution block in `zaid_ascad_desync_100` and `noConv1_ascad_desync_100` contains a pooling layer with a size and stride of 50, which is significantly higher than the pooling sizes from other CNN architectures analyzed here. As large (e.g., 50) and small (e.g., 6) pooling sizes provide similar results for synchronized datasets, we conclude that large pooling sizes in one of the hidden layers might not be the main reason for better shift-invariance robustness. Still, more research should be done to investigate the influence of pooling hyperparameters for the shift-invariance robustness and some other hyperparameters like kernel size, which are also not addressed in this work.



(a) Shift-invariance robustness of `cnn_pooling_2`, `cnn_pooling_4` and `cnn_pooling_6` architectures. Training batch size is set to 400.



(b) Shift-invariance robustness of `zaid_ascad_desync_0`, `noConv1_ascad_desync_0`, `zaid_ascad_desync_100` and `noConv1_ascad_desync_100` architectures. Training batch size is set to 50 for `zaid_ascad_desync_0`, `noConv1_ascad_desync_0` and 400 for `zaid_ascad_desync_100` and `noConv1_ascad_desync_100`.

Fig. 1: Shift-invariance robustness of profiling models trained with synchronized profiling sets.

5.2 Shift-invariance Robustness of CNNs Trained with Desynchronized Datasets

An alternative way to improve the shift-invariance robustness of CNN models is by conducting the training with desynchronized profiling traces. Figure 2 shows results for different CNN models trained with the desynchronized `ASCADr` and `ASCADf` datasets. Here, we skip results for `cnn_pooling_2`, `cnn_pooling_4`, `cnn_pooling_6`, `zaid_ascad_desync_0`, and `noConv1_ascad_desync_0` architectures as with these models, we were unable to successfully recover the key when trained and predicted with desynchronized datasets. Thus, we consider `cnn_ascadr_desync100` architecture that was found in a random hyperparameter search when `ASCADr_desync100` was used as training set and its shift-invariance robustness is provided in the top part of Figure 2. Since this model is trained with a dataset containing uniformly distributed random shifts inside the interval $[0, 100]$, its shift-invariance robustness is more salient within the same shift-interval $[0, 100]$. This is expected, as the profiling set contains traces that are shifted inside this interval. The model shows poor generalization capacity outside the interval of $[0, 100]$. For the `zaid_ascad_desync_100` and `noConv1_ascad_desync_100` architectures, we found similar shift-invariance robustness results. Both models were designed to break `ASCADf_desync100` dataset with traces containing uniform random shifts inside the interval $[0, 100]$. As shown in Figure 2, both CNN models provide satisfactory shift-invariance inside this shift interval. Outside this interval, the performance quickly deteriorates, indicating that shift-invariance is directly related to the shift distribution in the desynchronized profiling set.

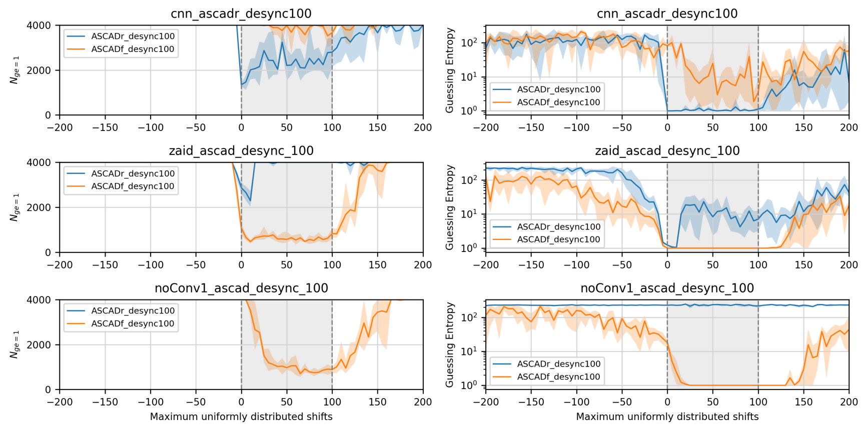


Fig. 2: The shift-invariance robustness of different neural network models trained with desynchronized datasets. For the `zaid_ascad_desync_100` and `noConv1_ascad_desync_100`, batch size is set to 200 while for `cnn_ascadr_desync100` the batch size is set to 400.

5.3 The Effect of Data Augmentation on Shift-invariance Robustness of CNNs Trained with Synchronized Datasets

When CNN models are trained with a synchronized trace set, one expects this model to generalize poorly to desynchronized attack sets, as shown by most of the results in Figure 1. One alternative to improve the model's generalization is to train with data augmentation. In this section, we only select `cnn_pooling_2` instead of also analyzing `cnn_pooling_4` and `cnn_pooling_6`, as they showed similar results in the previous section. Also, `cnn_pooling_2` provided slightly less shift-invariance compared to `cnn_pooling_4` and `cnn_pooling_6`, and we want to verify if adding data augmentation improves shift-invariance in the most critical case. The `cnn_pooling_2` architecture is configured with average pooling layers. We skip results for max pooling layers, as our preliminary analysis found no particular benefit of max pooling in shift-invariance robustness. This model is trained with the `ASCADr` and `ASCADf` datasets. In contrast, `zaid_ascad_desync_0` and `noConv1_ascad_desync_0` models are trained for the `ASCADf` dataset only (results for the `ASCADr` dataset for these two models provided no successful key recovery, so we omit these results). In terms of data augmentation, CNN models are trained with the full profiling set plus double the number of profiling traces as augmented traces (i.e., for `ASCADr`, we have 200 000 profiling traces plus 400 000 augmented traces. In comparison, for `ASCADf` we have 45 000 profiling traces plus 90 000 augmented traces). We define the number of augmented traces arbitrarily, and the analysis of the optimal number of augmented profiling traces is out-of-scope for this paper.

Figure 3 shows results for three different CNN architectures: `cnn_pooling_2` (with average pooling layer), `zaid_ascad_desync_0`, and `noConv1_ascad_desync_0`. We train these models with synchronized datasets and with data augmentation that shifts the traces for specific ranges, i.e., $[-100, 0]$, $[0, 100]$, and $[-100, 100]$. The main idea is to verify if all models can become shift-invariant inside specific shift intervals that could be present in a desynchronized attack set. Due to data augmentation, the shift-invariance robustness of all trained CNN models significantly improves compared to a scenario without data augmentation shown in Figure 1. Moreover, the shift-invariance robustness often occurs outside of the augmented shift interval. When data augmentation is implemented for random shifts inside the interval $[0, 100]$ (or $[-100, 0]$), the CNN model tends to become shift-invariant for intervals $[0, 200]$ (or $[-200, 0]$) as is visible for `cnn_pooling_2` models. This was the case also for results on `ASCADf` dataset with `cnn_pooling_2`, but we omit the figure with results due to page limit. In the case of the models designed for a specific synchronized dataset, data augmentation helped reach good attack results, at least within the data augmentation interval. Thus, training a CNN model with a synchronized dataset and data augmentation may ensure satisfactory shift-invariance robustness. This could prevent the model from presenting poor generalization for cases when the attack set is not perfectly aligned with the profiling set in the time domain.

Another finding in this analysis comes from applying data augmentation for a larger interval, which is the case for the random shifts drawn from the interval $[-100, 100]$. The shift-invariance robustness of the model improves. However, it is not as good for some cases, like with separated intervals $[0, 100]$ and $[-100, 0]$, when looking into GE for these intervals. This is more evident for the `zaid_ascad_desync_0` model: splitting the data augmentation interval from $[-100, 100]$ into two model training for intervals $[0, 100]$ and $[-100, 0]$ provides better convergence of GE inside each smaller shift interval. For `cnn_pooling_2`, the results for the interval $[0, 100]$ are superior if the model is trained with data augmentation providing random shifts inside $[0, 100]$ instead of the entire interval $[-100, 100]$. In Section 6, we propose an ensemble-based strategy to make CNN models highly shift-invariant for larger desynchronization intervals. More precisely, we propose to combine multiple data augmentation intervals into a model with better shift-invariant robustness.

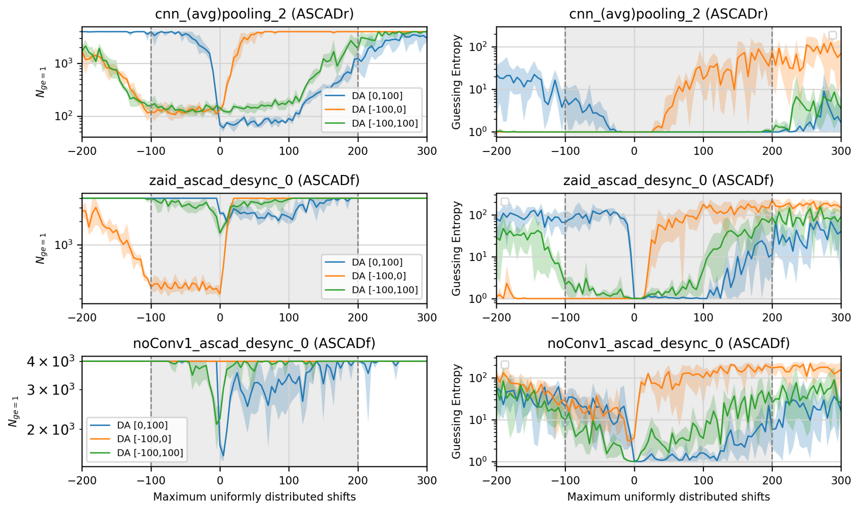


Fig. 3: The shift-invariance robustness of different neural network models trained with synchronized datasets. Expect for `noConv1_ascad_desync_0`, which has a batch size of 50. All models are trained with a batch size of 400.

5.4 The Effect of Data Augmentation on Shift-invariance Robustness of CNNs Trained with Desynchronized Datasets

A CNN architecture designed to provide an excellent generalization to desynchronized trace sets already shows improved shift-invariance robustness, as shown in Figure 2. However, shift-invariance robustness is still related to the desynchronization in the profiling set. We explore the model's generalization when the attack set contains desynchronization levels that differ from the profiling set. Data augmentation is a solution to improve the shift-invariance robustness of CNN models trained with desynchronized datasets. Our primary goal in this section is to make CNN models from Section 5.2 shift-invariant inside the trace shift interval $[-100, 200]$ when the profiling set is given by a desynchronized dataset with shifts inside the interval $[0, 100]$. This means that our trained CNN models become shift-invariant outside the shifts contained in the profiling set. Figure 4 shows the shift distribution of original desynchronized profiling traces and the shift distribution of augmented sets. Note that the shifts are applied to already desynchronized profiling traces, so it does not follow a uniform distribution. Still, we ensure the occurrence of shifts from the interval $[-100, 200]$.

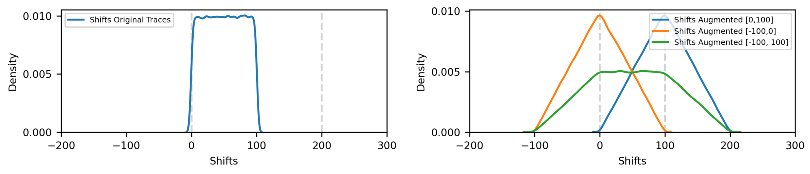


Fig. 4: Shift distribution of original (left) and augmented profiling sets (right).

Results from Figure 5 show the shift-invariance robustness of the `cnn_ascadr_desync100` and `zaid_ascad_desync_100`, while results with `noConv1_ascad_desync_100` are omitted as they are similar to `zaid_ascad_desync_100`. Models were used for predictions on desynchronized attack sets. Although these three models were designed to be shift-invariant to desynchronization of $[0, 100]$, when data augmentation is set to randomly shift the profiling set inside the shift interval $[-100, 100]$, all models become shift-invariant inside interval $[-100, 200]$. We can also see that, in some cases, the robustness goes outside of that interval.

6 Ensembles to Defeat Larger Desynchronization Levels

We evaluated the shift-invariance robustness in the previous section when trace sets are desynchronized inside the interval $[0, 100]$. This section shows that a different strategy is required for trace sets containing larger desynchronization levels to improve the shift-invariance robustness of a profiling CNN model. Specifically, we propose a solution to improve the shift-invariance robustness of a CNN model by ensembling multiple CNN models, each trained on a shorter desynchronization interval with data augmentation. We build an ensemble by averaging the output class probabilities from multiple models. The levels of desynchronization that we test are $[0, 200]$, $[0, 400]$, and $[0, 1\,000]$. To the best of our knowledge, the largest desynchronization level tested in DLSCA-related works is 400 [9].

Datasets. The experiments are conducted with the `ASCADr` and `DPAv4.2` datasets[8]. Since we want to analyze larger desynchronization levels, we select larger inter-

[8] `https://www.dpacontest.org/v4/42_traces.php`

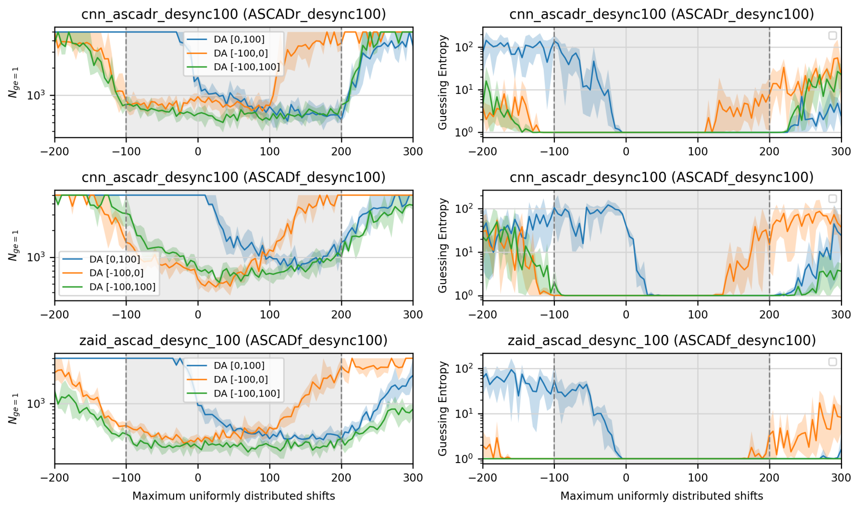


Fig. 5: The shift-invariance robustness of different neural network models trained with desynchronized datasets. All models are trained with a batch size of 400.

vals from the raw `ASCADr` and `DPAv4.2` measurements. For `ASCADr`, we select the trace interval consisting of 5 000 features, ranging from sample index 79 145 to 84 145 (the raw traces have 250 000 features). This trimmed interval includes processing the third `S-Box` byte in the first AES encryption round. This dataset is referred to as `ASCADr_5000`. For the `DPAv4.2` dataset, we select the trace interval from sample index 200 000 to 220 000, which contains the processing of the twelfth `S-Box` byte in the first AES encryption round. The resulting 20 000 interval is further resampled into 4 000 samples. This resampling process reduces the complexity of CNN models. This dataset is referred to as `DPAv4.2_4000`. To work with desynchronized datasets, before trimming each dataset into 5 000 and 4 000 samples for `ASCADr` and `DPAv4.2`, we apply random shifts from a uniform distribution and then trim the datasets. Both datasets are labeled with the identity leakage model. For `ASCADr_5000`, using data augmentation, the model process 200 000 profiling traces plus 400 000 augmented traces. For the `DPAv4.2_4000`, we process 70 000 profiling traces plus 140 000 augmented traces.

Deep neural networks. CNN architectures evaluated in the previous section cannot deliver successful key recovery results for the experiments with larger desynchronization intervals. Therefore, we deploy a new random hyperparameter search to find the best possible CNN models for the `ASCADr_5000` and `DPAv4.2_4000`. The search considers datasets with desynchronization inside the interval $[0, 100]$. As a result, we found CNN models reported in Table 2. The models are referred to as `cnn_large_desync_ascadr` and `cnn_large_desync_dpa_v42`. In `cnn_large_desync_dpa_v42`, the fully-connected layer is regularized with l1 regularization using $l = 0.00001$.

Table 2: Hyperparameter values for `cnn_large_desync_ascadr` and `cnn_large_desync_dpa_v42`.

Hyperparameters	cnn_large_desync_ascadr400	cnn_large_desync_dpa_v42
Conv. blocks	3	4
Filters (ordered)	8, 16, 24	16, 32, 48, 64
Kernel size	40	30
Strides	2	1
Pool size	2	4
Pool strides	2	4
FC layers	1	1
FC Nb. neurons	300	50
Weight init	random_normal	random_normal
Activation	selu	selu
Learning rate	0.00025	0.001
Optimizer	Adam	Adam
Batch size	400	400
Epochs	100	100

Our goal is not to promote the best CNN architectures for large desynchronization cases but to propose a robust ensemble-based data augmentation approach that is expected to be equally efficient with architectures from related works, such as [16] and [9] that evaluated desynchronization of 200 and 400 samples, respectively.

We aim to make described architectures `cnn_large_desync_ascadr` and `cnn_large_desync_dpa_v42` shift-invariant inside the shift intervals $[0, 200]$, $[0, 400]$, and $[0, 1\,000]$. For that, we train each model with profiling sets having desynchronized traces with a normal shift distribution inside these intervals, and we add data augmentation to the training process. To keep the model shift-invariant inside, e.g., the interval $[0, 200]$, data augmentation is optimized to preserve the shifts from the original profiling set inside this $[0, 200]$ interval. This means that when we add data augmentation to the profiling set, we shift already desynchronized traces. We carefully choose the minimum and maximum shift ranges for data augmentation. Figure 6 shows (on the left side) the shift distribution of the original profiling set when it is desynchronized inside the interval $[0, 200]$ and also (in the middle) the resulting shift distribution of the augmented traces. Note how our data augmentation process mainly preserves the shift distribution inside the interval $[0, 200]$, with some traces also being shifted outside these boundaries. We do the similar when desynchronization is $[0, 400]$ and $[0, 1\,000]$. Due to the page limit, we omit the figures for those intervals.

For each large desynchronization case and each dataset, we train multiple CNN models to combine them into an ensemble. We divide the data augmentation range $[-\delta, \delta]$ into smaller ranges. For instance, when desynchronization in the original profiling set is $[0, 200]$, we divide the data augmentation into the following eight separate ranges: $[-100, -75]$, $[-75, -50]$, $[-50, -25]$, $[-25, 0]$, $[0, 25]$, $[25, 50]$, $[50, 75]$, and $[75, 100]$. It means we have a *data augmentation of step 25* because we implement multiple trainings with data augmentation shift-

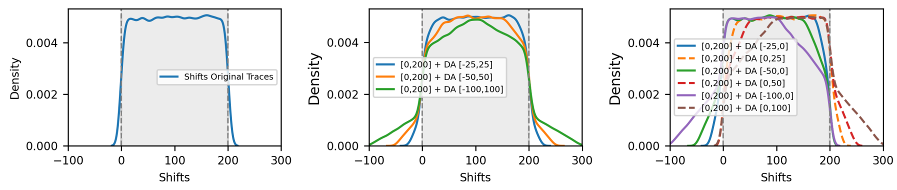


Fig. 6: Shift distributions in the original and augmented trace sets when profiling set has desynchronization in $[0, 200]$.

ing the traces with an interval of 25. We train the CNN model for each range with the original desynchronized traces and the data augmentation with the small range. Thus, we train eight separate CNN models and a CNN model with data augmentation with a $[-100, 100]$ shift range. Finally, we build an ensemble from these nine CNN models by averaging their output class probabilities. Similarly, we will test data augmentation with step 50 and 100 for dataset with desynchronization $[0, 200]$, step 50, 100 and 200 for $[0, 400]$ and step 100, 250 and 500 for $[0, 1000]$.

Figure 7 shows results for the ensembles obtained for large desynchronization cases in the `ASCADr_5000` dataset. Training a single CNN model with data augmentation inside the shift range $[-\delta, \delta]$ shows inferior results compared to ensembling multiple CNN models with specific data augmentation steps. In the top part of Figure 7, we show results for the desynchronization case of $[0, 200]$. Building an ensemble (with all tested step sizes) allows us to recover the key with approximately 10 attack traces while using data augmentation with $[-\delta, \delta]$ needs 100 attack traces. This is ten times fewer traces which indicates better performance. A similar improvement is visible in the case of desynchronization $[0, 400]$. When desynchronization is much larger, in the case of $[0, 1\,000]$, the only scenario where we can improve shift-invariance is by using an ensemble with a data augmentation step of 100. This means that the shift-invariance robustness of this CNN model is significantly improved with the usage of ensembles.

The results for the shift-invariance robustness analysis of the `cnn_large_desync_dpa_v42` model and `DPAv4.2` dataset are shown in Figure 8. When desynchronization in the profiling set is within the range $[0, 200]$, the model becomes shift-invariant for all cases, but ensembles need less traces. When the profiling set has desynchronization of $[0, 400]$, we find better results when a single model is trained with data augmentation inside the shift range $[-200, 200]$ and an ensemble of data augmentation with a step of 200. Results for the large desynchronization scenario of $[0, 1\,000]$ show that we are only able to make the `cnn_large_desync_dpa_v42` model shift-invariant inside the interval of $[0, 1\,000]$ when we build ensembles of multiple data augmentation steps. The best results are found with the data augmentation step of 100. Additionally, the ensembles provide good results outside the targeted shift intervals. Finally, in Figures 7 and 8, if data augmentation is not considered, none of the models can provide successful key recovery results.

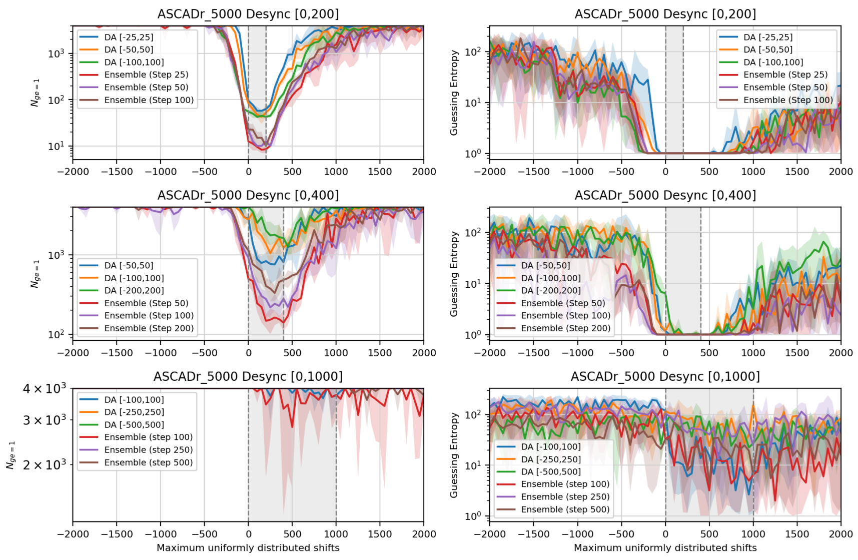


Fig. 7: The shift-invariance robustness of the `cnn_large_desync_ascadr` model trained with large desynchronization levels and the `ASCADr_5000` dataset.

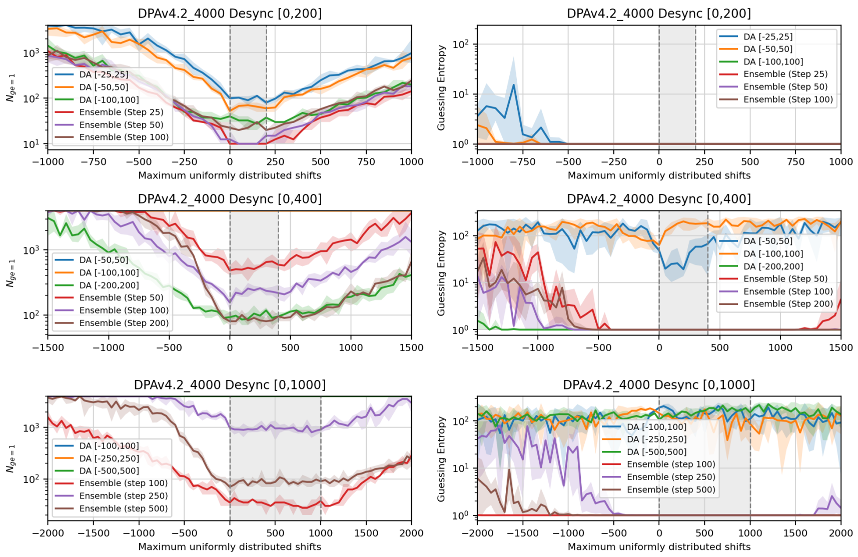


Fig. 8: The shift-invariance robustness of the `cnn_large_desync_dpa_v42` model trained with large desynchronization levels and the `DPAv4.2_4000` dataset.

7 Conclusions and Future Work

This paper investigates the shift-invariance of CNNs in DLSCA. First, we show that desynchronized datasets can easily disrupt the shift-invariance of CNNs, leading to unsuccessful attacks. Then, we discuss how shift-invariance can be improved by using data augmentation. Finally, if the desynchronization levels are large, we propose a novel method based on the ensembles of data augmentation. Our results show superior performance with data augmentation, especially with ensembles of data augmentation. Interestingly, architectural changes (e.g., modifications in the convolutional or pooling layer) provide minor improvements in shift-invariance and represent a more difficult path toward reaching it. This is especially true for more shallow neural network architectures used in DLSCA.

There are multiple future research directions. First, this paper considers the desynchronization countermeasure. It would be interesting to evaluate different types of misalignment, like jitter and random delays. Next, here, we arbitrarily set the number of augmented traces. Evaluating the optimal number of augmented traces to be used would be very relevant, especially if some guidelines based on the dataset's properties can be given.

References

1. Azulay, A., Weiss, Y.: Why do deep convolutional networks generalize so poorly to small image transformations? CoRR **abs/1805.12177** (2018), `http://arxiv.org/abs/1805.12177`
2. Benadjila, R., Prouff, E., Strullu, R., Cagli, E., Dumas, C.: Deep learning for side-channel analysis and introduction to ASCAD database. J. Cryptographic Engineering **10**(2), 163–188 (2020). `https://doi.org/10.1007/s13389-019-00220-8`, `https://doi.org/10.1007/s13389-019-00220-8`
3. Bronstein, M.M., Bruna, J., Cohen, T., Velickovic, P.: Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. CoRR **abs/2104.13478** (2021), `https://arxiv.org/abs/2104.13478`
4. Cagli, E., Dumas, C., Prouff, E.: Convolutional neural networks with data augmentation against jitter-based countermeasures. In: Fischer, W., Homma, N. (eds.) Cryptographic Hardware and Embedded Systems – CHES 2017. pp. 45–68. Springer International Publishing, Cham (2017)
5. Chaman, A., Dokmanic, I.: Truly shift-invariant convolutional neural networks. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3772–3782. IEEE Computer Society, Los Alamitos, CA, USA (jun 2021). `https://doi.org/10.1109/CVPR46437.2021.00377`, `https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00377`
6. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 113–123 (2019)
7. Gens, R., Domingos, P.M.: Deep symmetry networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), `https://proceedings.neurips.cc/paper/2014/file/f9be311e65d81a9ad8150a60844bb94c-Paper.pdf`

8. Goodfellow, I.J., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge, MA, USA (2016), http://www.deeplearningbook.org
9. Hajra, S., Saha, S., Alam, M., Mukhopadhyay, D.: Transnet: Shift invariant transformer network for side channel analysis. In: Progress in Cryptology - AFRICACRYPT 2022: 13th International Conference on Cryptology in Africa, AFRICACRYPT 2022, Fes, Morocco, July 18–20, 2022, Proceedings. p. 371–396. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-17433-9_16, https://doi.org/10.1007/978-3-031-17433-9_16
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
11. Kauderer-Abrams, E.: Quantifying translation-invariance in convolutional neural networks. CoRR **abs/1801.01450** (2018), http://arxiv.org/abs/1801.01450
12. Kayhan, O.S., van Gemert, J.C.: On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14262–14273. IEEE Computer Society, Los Alamitos, CA, USA (jun 2020). https://doi.org/10.1109/CVPR42600.2020.01428, https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01428
13. Kim, T., Lee, J., Nam, J.: Comparison and analysis of samplecnn architectures for audio classification. IEEE Journal of Selected Topics in Signal Processing **13**(2), 285–297 (2019). https://doi.org/10.1109/JSTSP.2019.2909479
14. LeCun, Y.: Generalization and network design strategies. In: Pfeifer, R., Schreter, Z., Fogelman, F., Steels, L. (eds.) Connectionism in Perspective. Elsevier, Zurich, Switzerland (1989), an extended version was published as a technical report of the University of Toronto
15. Lenc, K., Vedaldi, A.: Understanding image representations by measuring their equivariance and equivalence. International Journal of Computer Vision **127**(5), 456–476 (May 2018). https://doi.org/10.1007/s11263-018-1098-y, https://doi.org/10.1007/s11263-018-1098-y
16. Lu, X., Zhang, C., Cao, P., Gu, D., Lu, H.: Pay attention to raw traces: A deep learning architecture for end-to-end profiling attacks. IACR Transactions on Cryptographic Hardware and Embedded Systems **2021**(3), 235–274 (Jul 2021). https://doi.org/10.46586/tches.v2021.i3.235-274, https://tches.iacr.org/index.php/TCHES/article/view/8974
17. Lu, X., Zhang, C., Cao, P., Gu, D., Lu, H.: Pay attention to the raw traces: A deep learning architecture for end-to-end profiling attacks. IACR Transactions on Cryptographic Hardware and Embedded Systems (2021)
18. Marcos, D., Volpi, M., Tuia, D.: Learning rotation invariant convolutional filters for texture classification. CoRR **abs/1604.06720** (2016), http://arxiv.org/abs/1604.06720
19. Perin, G., Chmielewski, L., Batina, L., Picek, S.: Keep it unsupervised: Horizontal attacks meet deep learning. IACR Transactions on Cryptographic Hardware and Embedded Systems **2021**(1), 343–372 (Dec 2020). https://doi.org/10.46586/tches.v2021.i1.343-372, https://tches.iacr.org/index.php/TCHES/article/view/8737
20. Perin, G., Chmielewski, L., Picek, S.: Strength in numbers: Improving generalization with ensembles in machine learning-based profiled side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems **2020**(4), 337–364 (Aug 2020). https://doi.org/10.13154/tches.v2020.i4.337-364, https://tches.iacr.org/index.php/TCHES/article/view/8686

21. Perin, G., Wu, L., Picek, S.: Exploring feature selection scenarios for deep learning-based side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems **2022**(4), 828–861 (Aug 2022). `https://doi.org/10.46586/tches.v2022.i4.828-861`, `https://tches.iacr.org/index.php/TCHES/article/view/9842`
22. Picek, S., Perin, G., Mariot, L., Wu, L., Batina, L.: Sok: Deep learning-based physical side-channel analysis. ACM Comput. Surv. (oct 2022). `https://doi.org/10.1145/3569577`, `https://doi.org/10.1145/3569577`, just Accepted
23. Standaert, F.X., Malkin, T.G., Yung, M.: A unified framework for the analysis of side-channel key recovery attacks. In: Joux, A. (ed.) Advances in Cryptology - EUROCRYPT 2009. pp. 443–461. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)
24. Timon, B.: Non-profiled deep learning-based side-channel attacks. Cryptology ePrint Archive, Paper 2018/196 (2018), `https://eprint.iacr.org/2018/196`, `https://eprint.iacr.org/2018/196`
25. Wouters, L., Arribas, V., Gierlichs, B., Preneel, B.: Revisiting a methodology for efficient cnn architectures in profiling attacks. IACR Transactions on Cryptographic Hardware and Embedded Systems **2020**(3), 147–168 (Jun 2020). `https://doi.org/10.13154/tches.v2020.i3.147-168`, `https://tches.iacr.org/index.php/TCHES/article/view/8586`
26. Zaid, G., Bossuet, L., Habrard, A., Venelli, A.: Methodology for efficient cnn architectures in profiling attacks. IACR Transactions on Cryptographic Hardware and Embedded Systems **2020**(1), 1–36 (Nov 2019). `https://doi.org/10.13154/tches.v2020.i1.1-36`, `https://tches.iacr.org/index.php/TCHES/article/view/8391`
27. Zaid, G., Bossuet, L., Habrard, A., Venelli, A.: Efficiency through diversity in ensemble models applied to side-channel attacks: – a case study on public-key algorithms –. IACR Transactions on Cryptographic Hardware and Embedded Systems **2021**(3), 60–96 (Jul 2021). `https://doi.org/10.46586/tches.v2021.i3.60-96`, `https://tches.iacr.org/index.php/TCHES/article/view/8968`
28. Zhang, R.: Making convolutional networks shift-invariant again. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 7324–7334. PMLR (09–15 Jun 2019), `https://proceedings.mlr.press/v97/zhang19a.html`
29. Zotkin, Y., Olivier, F., Bourbao, E.: Deep learning vs template attacks in front of fundamental targets: experimental study. IACR Cryptol. ePrint Arch. p. 1213 (2018), `https://eprint.iacr.org/2018/1213`