Modeling attacks against device authentication using CMOS image sensor PUF

Hiroshi Yamada¹, Shunsuke Okura², Masayoshi Shirahata³, and Takeshi Fujino²

Abstract A CMOS image sensor physical unclonable function (CIS PUF), which generates a unique response extracted from manufacturing process variations, is utilized for device authentication. In this paper, we report modeling attacks on a CIS PUF, in which column fixed pattern noise is exploited in a sorting attack. When the PUF response is generated with a pairwise comparison method, unknown responses are predicted with an accuracy of over 87.8% with only 0.31% of the training sample of the entire challenge and response pairs.

Keywords: CMOS image sensor, PUF, IoT, hardware security, modeling attack

1. Introduction

A wide variety of sensors will be connected by the development of the Internet of Things (IoT) [1]. Accordingly, the security of sensor information is getting more important. Considering IoT security, it is required to know if the sensor itself is reliable [2, 3]. To meet the demand, physical unclonable function (PUF) is proposed as a lightweight solution. PUF can generate a response as a unique device identifier (ID) based on the device variation caused in the semiconductor manufacturing process [4, 5]. The unique ID obtained from PUF is used for device authentication. In this paper, we focus on the modeling attack against CIS PUF.

As the image sensor market expands [14], some PUF technologies [15, 16, 17, 18, 19, 20, 21] and applications [22, 23] using the variation inherent in the sensor devices have been proposed in order to protect the imaging data captured by cameras. Among these PUFs, the CMOS image sensors with PUF technologies (CIS PUF) [15, 16, 17] generate a response based on the RAW pixel variation in the image sensor. In the device authentication, a challenge is input to the PUF and a response is acquired, as shown in Fig. 1. If the response obtained from the CIS PUF is identical to the response enrolled in the host device, it can be confirmed that the image data is output from the correct device. Moreover, the authentication is success as long as the Hamming distance (HD) between the bit strings of the regenerated response and the enrolled response is lower than a given threshold value because some response bits will be different from the enrolled response bits due to random noise and environmental variation. However, it is vulnerable to modeling attacks, in which the attacker collects a challenge and response pair (CRP) and predicts responses to unknown challenges in order to disguise fake devices. It is reported that major PUFs such as Arbiter PUF and Ring Oscillator PUF can be cloned by a modeling attack with a machine learning attack [24, 25, 26, 27, 28] and genetic programming [29, 30]. In this paper, we focus on the modeling attack against CIS PUF.

There are two methods to generate CIS PUF responses from the RAW pixel variation: firstly, fixed pixel pairs are compared and secondly, two randomly selected pixels are compared. The latter approach is similar to the pairwise comparison method used in the Ring Oscillator PUF [9]. In this paper, the PUF using fixed pairs and random pairs are respectively called confined PUF and extensive PUF. If the number of pixels is N, the maximum number of CRPs of the confined CIS PUF [15, 16] is half the number of pixels (N/2), while that of the extensive CIS PUF [17] is N × (N − 1)/2. Hence, the maximum number of CRPs derived from the confined CIS PUF is lower than that from the extensive CIS PUF. However, the extensive CIS PUF could be vulnerable to the modeling attacks because multiple responses are generated based on the same pixels. Furthermore, a modeling attack using the column fixed pattern noise that is peculiar to the structure of the image sensor can make the response prediction easier.

This paper is organized as follows. In Sec. 2, an operation diagram of the CIS PUF and entropy of the CIS PUF response are described. In Sec. 3, simulation results of modeling attacks are presented. The modeling attack using column fixed pattern noise is then proposed in Sec. 4.
followed by the summary in Sec. 5.

2. Overview of CIS PUF

2.1 Operation of CIS PUF

A structure and a timing diagram of the CIS PUF [15] are shown in Fig. 2. The vertical scanner (V-scan) accesses pixels in a row and the pixel output signal is read out by column parallel circuits controlled by the horizontal scanner (H-scan). The RAW pixel variations are obtained by reading the difference of the clip transistor output at \( t_1 \) and the SF transistor output in the n-th row at \( t_2 \).

2.2 ID generation of CIS PUF

A PUF response bit \( b_k \) is generated by comparison of two pixels as shown in Fig. 3, where the number of pixels is 16. For example, when RAW pixel variations \( p_{0,0} \) and \( p_{0,1} \) are selected as a challenge, the response is 1 if \( p_{0,0} > p_{0,1} \), and vice versa. The confined CIS PUF compares two vertically adjacent pixels, and the total number of generated responses is 8 bits. Though the number of CRPs is small, it is expected to be resistant to the modeling attack because each pixel is used only once. On the other hand, the extensive CIS PUF utilizes two randomly selected pixels, and the number of responses is \( 16 \times 15 / 2 = 120 \) bits. Though the number of CRPs of the extensive CIS PUF is larger than that of the confined CIS PUF, the extensive CIS PUF could be vulnerable to the modeling attack because the same pixel is used multiple times. For example, if the attacker knows \( p_{0,0} > p_{1,0} \) and \( p_{1,0} > p_{2,0} \) from the used CRPs, the attacker is able to predict that \( p_{0,0} > p_{2,0} \).

2.3 Theoretical entropy of CIS PUF response

Entropy describes the unpredictability of the PUF response. If a PUF has large entropy, the number of authentication with the PUF will be large. Here, the theoretical entropy of the confined and extensive CIS PUFs are discussed, in which the sensor is supposed to be composed of \( 64 \times 64 (= 4096) \) pixels.

Confined CIS PUF

The number of total responses is 2048 bits. Since the response is generated independently, the theoretical entropy is \( \log_2 2^{2048} = 2048 \) bits.

Extensive CIS PUF

The total number of responses is 4096 \times (4096 - 1) / 2 \approx 8.4M \) bits which is 4095 times larger than that of the confined CIS PUF. However, the theoretical entropy is \( \log_2 4096! = 43.25k \) bits, because the possible combination of the pixel output is 4096!. Therefore, the theoretical entropy of the extensive CIS PUF is only 21 times larger than that of the confined CIS PUF.

Figure 4 shows distributions of the responses generated from the measurement data of the confined and the extensive CIS PUF. The horizontal and the vertical axis represent the challenge (the address of the pixel). While the responses of the confined CIS PUF looks random, some responses of the extensive CIS PUF are biased to 0 or 1 in line-by-line. This results suggest that the entropy in the extensive CIS PUF is much lower than the total number of the responses.

3. Modeling attacks against CIS PUF

In this section, we report the simulation results of a simple modeling attack on the CIS PUF in order to evaluate the predictability to unknown responses. In this simulation, the RAW pixel variation is derived from measurement data.

3.1 Overview of a simple modeling attack

In a simple modeling attack, the list of the pixel output order is generated by sorting the RAW pixel outputs based on known CRPs. The procedure of the attack is shown in
the responses, in which the maximum number is theoretical entropy. The vertical axis is the prediction accuracy of unknown responses. Figure 6(a) shows that the unknown responses of the extensive CIS PUF were predicted with an accuracy of more than 90% when the 0.52% of the CRPs were collected. On the other hand, Fig. 6(b) shows that the unknown responses of the confined CIS PUF were predicted with an accuracy of 50 ± 8%, even if more than 90% of the CRPs were collected. It is confirmed that the extensive CIS PUF is vulnerable to the modeling attack, while the confined CIS PUF is resistant to the attack.

Now let’s suppose that the threshold HD in device authentication is 15%. The attacker’s success probability of impersonation ($P_{im}$) for a 128 bit length ID is given by

$$P_{im} = \sum_{i=0}^{19} \frac{128}{109+i} (x)^{109+i} (1-x)^{19-i},$$

where $x$ is a prediction accuracy to unknown responses. Therefore, $P_{im}$ is 85% at $x = 0.878$. In other words, the modeling attack will succeed only if 0.4% of the CRPs are collected and unknown responses are predicted with an accuracy of 87.8%.

4. Modeling attack using column FPN

4.1 FPN and attack method

The CIS PUF shows column-wise fixed pattern noise called column FPN. The 12 bits RAW pixel variation of the 64 × 64 pixels is shown in Fig. 7. Vertical stripes caused by the variation in the column parallel circuits are visible, where the standard deviation among the columns is 58.9% of that among the pixels.

The column FPN decreases the entropy of the CIS PUF response. In other words, the unknown PUF response for the pixels in “small column” and “large column” can be predicted based on the collected responses. The proposed modeling attack replaces the random initialization at the first step of the modeling attack shown in Fig. 5 with a sorting of the list of column averages. The column average is calculated with “a winning rate” of the pixels in the column. The procedure for the case when the number of pixels is 4 × 4 is shown in Fig. 8. For example, let’s suppose that $p_{0,2}$ is utilized 3 times in the collected CRPs, the two responses are 1, and the other response is 0. The $p_{0,2}$ is labeled with
the winning rate of \((2 - 1)/3 = +1/3\). Following to the calculation of the winning rate of all collected CRPs, the column average is calculated. Then, the group of the pixels in the same column is sorted in the order of the column averages. The second step “pixel swap” and the third step “prediction” follow the first “column average sort”.

4.2 Simulation results

The prediction results by the modeling attack using column FPN are shown in Fig. 9. The dotted line in Fig. 9(a) represents the average of five chips by the simple modeling attack described in Sec. 3. The unknown responses of the extensive CIS PUF were predicted with the smaller number of collected CRPs when the column FPN was exploited.

In contrast, Fig. 9(b) shows that the responses of the confined CIS PUF were not predicted even if the column FPN was exploited. In addition, the prediction accuracy of the extensive CIS PUF exceeded 87.8% when 0.31% of the unknown CPRs were collected. Therefore, the attacker could impersonate the device in the authentication with collected CRPs whose number is 0.77 times smaller than that of the previous simple modeling attack. Moreover, the average training time for the modeling attack using the column FPN was increased by only about 10.3% compared to the simple modeling attack. The effective PUF response bit length of the extensive CIS PUF was only 12.7 times longer than that of the confined CIS PUF, even though the theoretical entropy of the extensive CIS PUF was 21 times larger than that of the confined CIS PUF.

4.3 Simulation results of various pixel array sizes

The modeling attack using column FPN for various pixel sizes is simulated as shown in Fig. 10, in order to verify the prediction accuracy for the larger and more popular sizes of the pixel array. It is noted here that each result is the average of five chips and the result labeled “1920 x 540” is based on the measurement data of the CIS PUF with a 2 M pixel array. The horizontal axis is the ratio of the number of collected CRPs to the theoretical entropy \(\log_2 N\) (\(N\) is the number of pixels). The vertical axis is the prediction accuracy to unknown responses. The results show that the prediction accuracy increased as the number of pixels increased and the proposed attack was more effective against the CIS PUF with a larger pixel array. Besides, the effective PUF response bit length at 87.8% prediction accuracy was around 50% of the theoretical entropy.

5. Conclusion

In this paper, the robustness of the confined CIS PUF and the vulnerability of the extensive CIS PUF to modeling attacks was confirmed. Besides, the modeling attack, in which the pixel column FPN is utilized to predict unknown PUF responses with a small number of collected CRPs, is proposed. The simulation results of the prediction using column FPN showed that the number of CRPs to predict 87.8% of unknown PUF responses was 0.77 times smaller than the simple modeling attack. Even though the total number of responses of the extensive CIS PUF was huge, the effective
response bit length in the device authentication was only 0.31\% of the total responses when the threshold HD was supposed as 15\% of a 128 bit length ID.

References


