Invertible spread-spectrum watermarking for image authentication and multilevel access to precision-critical watermarked images [*]

Josep Domingo-Ferrer and Francesc Sebé
Dept. of Computer Engineering and Mathematics
Universitat Rovira i Virgili
Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain
e-mail {jdomingo,fsebe}@etse.urv.es

Abstract

Invertible watermarking has been introduced in the literature for the purpose of image authentication. We present in this paper a spread-spectrum invertible watermarking system which can be used to authenticate images in any lossless format, i.e. *establish their integrity. A second application of invertible watermarking is multilevel access to watermarked images: depending on her clearance, the image user can "clean" the marks of more or less parts of the image, so as to gain in precision. Both applications make sense for precision-critical images (*e.g. *military, satellite, medical, quality control, reverse engineering images) whose copyright should still be protected to some extent.*

Keywords: *Invertible watermarking, copyright protection, precision-critical images, Selective release based on security clearance, Selective access to parts of a document.*

1 Introduction

In the Internet age, it has become clear that digital storage of images is far more convenient than analog storage. Indeed, the digital format is easier to manipulate with computer equipment, it is easier and safer to transmit over computer networks and, very important, it does not degrade with time. However, the ease of manipulation has its negative side as well: it is nearly straightforward to use an editing program to tamper with an image by making small changes to it.

In some precision-critical applications, it is vital for a legitimate user to be able to verify the integrity of the image before using it. The following are some examples:

- Medical images are used for disease diagnosis. Thus small changes in such images might lead to erroneous conclusions about a patient's health.
- Satellite images are used to locate military strategic targets or to produce weather forecasts. Manipulation might lead to wrong interpretation of the real situation in the photographed area.
- X-ray images are often used in quality control to trace product failures [8] and in reverse engineering to derive the schematics of complex circuits [1]. In this kind of applications precision and integrity are crucial.

When precision-critical images are sold in commercial transactions, a trade-off has to be achieved between precision and integrity requirements on one hand and copyright protection on the other hand. For example, imagine commercial satellite images. Precision is indeed important because inaccurate satellite maps are useless. However, the commercial value of those images also demands protecting them from free redistribution. Watermarking is currently the most successful approach to detecting redistribution of multimedia content, but it requires small modifications (watermarks) to be made to the content to be protected in order to embed a copyright message in it. Such modifications are subperceptual and do not diminish the perceptual value of standard artwork, but can result in some alteration of the information conveyed by precision-critical images intended for thorough human or machine processing (such as the aforementioned examples). Thus, it would be interesting to provide multilevel access to precision-critical images, in such a way that:

- Non-privileged users just see the watermarked version of the image. The copyright of the image owner is protected even if this may result in some alteration of the informational content of the precision-critical image.

[*] This work is partly funded by the Spanish CICYT under contract no. TEL98-0699-C02-02.

- The higher the clearance of the user, the more watermarks she can remove. Removing the watermark from a part of an image allows the original version of that part to be recovered and its integrity to be verified. Users with full clearance can completely invert the watermarking process so as to obtain the original precision-critical image in an authenticated way.

1.1 Our contribution

In this paper, we show how to invert under certain conditions one of the most widely used robust oblivious watermarking methods, namely the Hartung-Girod [5] spread-spectrum spatial-domain watermarking algorithm. We then use spread-spectrum invertibility to provide integrity authentication and multilevel access for precision-critical images; the latter combines integrity authentication for parts of the image with copyright protection for other parts. Section 2 discusses the principles and shortcomings of previous work on invertible watermarking. We describe in Section 3 the procedure to invert the Hartung-Girod spread-spectrum oblivious watermarking scheme. Based on this fact, an algorithm for image authentication (*i.e.*, integrity verification) and copyright protection is given in Section 4. A scheme for multilevel access to precision-critical images is discussed in Section 5 and experimental results are given. Section 6 contains some conclusions.

2 Invertible watermarking

While most watermarking schemes introduce some small amount of non-invertible distortion in the image, invertible watermarking methods are such that, if the watermarked contents are deemed authentic, the distortion due to watermarking can be removed to obtain the original contents. The first document on invertible watermarking is deemed to be the patent [7]; however, this is no public-domain know-how.

In [2] invertible watermarking methods for authentication of digital images in the JPEG format are presented. For our precision-critical application, the assumption that original images have been JPEG lossy-compressed prior to watermarking is a major shortcoming: even if the watermark can be cleanly removed, the best we can get is a lossy-compressed unwatermarked image.

Additive, non-adaptative watermarking is claimed to be invertible in [3, 4]. In the claim, the existence of an "inverse watermarking operation" is postulated for a generic additive, non-adaptative method, but details are given only on how to derive such inverse operation for the spread-spectrum, frequency-based watermarking algorithm [6].

3 Invertible spread-spectrum watermarking

In [5], a spread-spectrum technique is used to obtain an oblivious watermarking method in the spatial domain. Oblivious watermarking does not require the original image to recover the watermark embedded in the watermarked image. We will first recall the fundamentals of this method and then we will show that, under certain conditions, this kind of watermarking is invertible.

The embedding and recovery procedures of [5] are as follows:

Embedding The copyright information to be embedded is a binary sequence $a_j, \quad a_j \in \{-1, 1\}$. This discrete signal is spread by a large factor cr, called chip-rate, to obtain the sequence $b_i = a_j, \quad j \cdot cr \leq i < (j+1) \cdot cr$. The spread sequence b_i is amplified by a locally adjustable amplitude factor $\alpha_i \geq 0$ and is then modulated by a binary pseudo-noise sequence $p_i, \quad p_i \in \{-1, 1\}$. If v_i is the original signal to be marked, the resulting watermarked signal is

$$\tilde{v}_i = v_i + \alpha_i \cdot b_i \cdot p_i \qquad (1)$$

Recovery Mark recovery is performed by demodulating the watermarked signal with the same pseudo-noise signal p_i that was used for embedding, followed by summation over the window for each embedded bit, which yields the correlation sum c_j for the j'th information bit $c_j = \sum_{i=j \cdot cr}^{(j+1) \cdot cr-1} p_i \cdot \tilde{v}_i \approx \sum_{i=j \cdot cr}^{(j+1) \cdot cr-1} p_i^2 \cdot \alpha_i \cdot b_i$. The sign of $c_j \approx cr \cdot \overline{\alpha} \cdot b_i = cr \cdot \overline{\alpha} \cdot a_j$, where $\overline{\alpha} = \sum \alpha_i / cr$, is interpreted as the embedded bit $\hat{a}_j$.

With the above method, several watermarks can be superimposed (multiple marking) if different pseudo-noise sequences are used for modulation. This is due to the fact that different pseudo-noise sequences are in general orthogonal to each other and do not significantly interfere [9].

3.1 Inverting spread-spectrum watermarks

In order for the above watermarking scheme to be totally invertible, the following three conditions must be met:

1. *The seed s used to generate the pseudo-noise signal p_i must be known.* Being able to re-create p_i is needed to recover the embedded bits (see Algorithm 1 below).

2. The locally adjustable amplitude factor α_i used at each sample of the watermarked signal during the embedding phase must be known. α_i is needed to invert Equation (1) as shown in Equation (2). This requirement can be easily met by using a constant value α for all samples.

3. For every sample v_i to be modulated, its modulated value $\tilde{v}_i = v_i + \alpha_i \cdot b_i \cdot p_i$ must fall within the same range of the original values v_i (otherwise truncation would be needed, which would hamper invertibility).

Assuming that the above three conditions are met, Algorithm 1 shows how the original unwatermarked signal v_i can be recovered from $\tilde{v}_i$:

Algorithm 1 (Spread-spectrum watermark inversion)

1. *Recover all embedded bits, where the j-th embedded bit $\hat{a}_j \in \{-1, 1\}$ is obtained as the sign of the correlation sum $c_j = \sum_{i=j \cdot cr}^{(j+1) \cdot cr - 1} p_i \cdot \tilde{v}_i$.*
2. *Spread the recovered sequence $\hat{a}_j$ by the chip-rate cr value, to obtain the sequence $\hat{b}_i = \hat{a}_j, \quad j \cdot cr \le i < (j+1) \cdot cr$.*
3. *Recover the original $\hat{v}_i$ sequence by computing*

$$\hat{v}_i = \tilde{v}_i - \alpha_i \cdot \hat{b}_i \cdot p_i \qquad (2)$$

Note that $\hat{v}_i = v_i, \forall i$ if $\hat{a}_j = a_j, \forall j$, *i.e.* if the embedded bits are correctly recovered, the unwatermarked image will match the original one. This can be readily seen by comparing Equations (1) and (2).

4 Image authentication and copyright protection using invertible spatial-domain watermarking

Based on the spatial-domain spread-spectrum watermarking described above, we next adapt the ideas of [3] to give a construction that, given an image, allows the hash of the image and a copyright message to be embedded in its pixels; the hash is used as a MAC (Message Authentication Code). Anyone knowing the embedding key and the amplitude factor α is able to recover the embedded MAC and copyright message, undo the watermark, get the original image and check for MAC validity.

Without loss of generality, we will assume a monochrome image in what follows. Let the original image be $\mathbf{X} = \{x_i : 1 \le i \le n\}$, where x_i is the color level of the i-th pixel and n is the number of pixels in the image. Let x_i be the grayscale level of the pixel, which is assumed to take integer values between 0 and $MAXCOLOR$.

4.1 Invertible addition

One of the three conditions stated above for a watermark to be invertible is that the value of a modulated pixel must fall into the grayscale range of original pixels. In [7], modular addition modulo $MAXCOLOR$ is proposed as another way to keep modulated pixel values within $[0, MAXCOLOR]$. In [3], this operation is criticized because of possible visual artifacts in the watermarked image resulting from grayscale values close to 0 being flipped to grayscale values close to $MAXCOLOR$, and grayscale values close to $MAXCOLOR$ being flipped to values close to 0 (nearly white pixels become nearly black and conversely). We claim that, in addition to visual artifacts, modular addition may lead to incorrect watermark reconstruction when inverting the Hartung-Girod watermarking. This is illustrated by the following example.

Example 1 *In the watermarking procedure described in Section 3 above, assume that $MAXCOLOR = 255$, $\alpha = 3$ and that we want to embed $a = 1$ in the first four pixels of the original image. If the values of those four pixels are $(v_0, v_1, v_2, v_3) = (1, 2, 3, 2)$ and the first four bits of the pseudorandom sequence are $(p_0, p_1, p_2, p_3) = (-1, 1, 1, -1)$, we spread a over four pixels to obtain $(b_0, b_1, b_2, b_3) = (1, 1, 1, 1)$ and compute*

$$\tilde{v}_i = v_i + \alpha \cdot b_i \cdot p_i, \ \ \text{for } i = 0 \text{ to } 3$$

This yields $(\tilde{v}_0, \tilde{v}_1, \tilde{v}_2, \tilde{v}_3) = (254, 5, 6, 255)$. Values 254 and 255 result from modular reduction of -2 and -1 (which were out of range). Now, when trying to recover the embedded bit, we compute

$$c = \sum_{i=0}^{3} p_i \cdot \tilde{v}_i = -254 + 5 + 6 - 255 = -498$$

Since the sign of c is negative, we reach the erroneous conclusion that the embedded bit was $\hat{a} = -1$.

A better way to keep modulated pixels within range is to pre-process the image in the following simple way:

Algorithm 2 (Gray-level pre-processing(α))

```
For $i = 1$ to $n$ do:
    1. If $x_i < \alpha$ then $x_i := \alpha$
    2. If $x_i > MAXCOLOR - \alpha$ then $x_i := MAXCOLOR - \alpha$
```

Algorithm 2 does indeed result in some non-invertible distortion, so when watermarking is inverted, there be may some slight difference between the grayscale values of some pixels of the original and the watermarked images. However, the advantages over modular addition are clear:

- There are no visual artifacts in the watermarked image, because the magnitude of grayscale changes is at most α levels.

- Erroneous bit recovery illustrated in Example 1 is avoided.

4.2 Hash embedding and verification

According to [3], invertible watermarking for image authentication consists of computing a hash of the original image and embedding the hash bits in the image. Our embedding algorithm takes as input the pre-processed image resulting from Algorithm 2 and depends on two parameters: a seed s for pseudo-random number generation and the amplitude factor α.

Algorithm 3 (Hash embedding(s,α))

1. *Compute the hash H of the pre-processed image $\mathbf{X}$. Concatenate a copyright message to H, to obtain the sequence H'.*

2. *Construct the sequence a_i to be embedded by doing, for $i = 1$ to $|H'|$:*

   - *If $H'_i = 0$ then $a_i := -1$*
   - *If $H'_i = 1$ then $a_i := 1$*

3. *Using the spread-spectrum Hartung-Girod technique with parameter α and seed s, embed the sequence $\{a_i : i = 1, \cdots, |H'|\}$ into $\mathbf{X}$. Let $\mathbf{X}'$ be the resulting watermarked image.*

The algorithm for image authentication, *i.e.* for verification of image integrity, is now straightforward:

Algorithm 4 (Integrity verification(s,α))

1. *Use Algorithm 1 to recover the embedded sequence $\hat{H}'$ and the unwatermarked image $\hat{\mathbf{X}}$ from $\mathbf{X}'$.*

2. *Compute the hash H of $\hat{\mathbf{X}}$*

3. *Compare H with the hash $\hat{H}$ contained in $\hat{H}'$. If they agree, then $\hat{\mathbf{X}} = \mathbf{X}$ and image is deemed authentic. If they do not, $\hat{\mathbf{X}}$ is deemed non-authentic.*

5 Multilevel access to precision-critical images

As mentioned in the introduction, our next goal is to devise a mechanism for user access to precision-critical images whereby the user can invert the embedded watermarks (and thus the distortion they cause) to an extent proportional to her clearance. Users with no clearance at all only see the watermarked image, which is copyright-protected and perceptually good but not suitable for high-precision processing. At the other end, users with full clearance can completely invert the watermarking process so as to obtain the original precision-critical image from the watermarked one. Between both extreme user types, users with intermediate clearances can invert the watermarking for some parts of the image. When unwatermarking a portion of the image, Algorithm 4 can be used to check its integrity.

Assumption 1 *We assume that the image to be protected can be divided into r semantically significant disjoint subimages. The i-th subimage is watermarked using an invertible method seeded with a different seed s_i. Formally speaking, if we denote the original image by $\mathbf{X}$, the watermarked image by $\mathbf{X}'$ and the watermarking transformation by F, we have*

$$\mathbf{X}' = F(\mathbf{X}, \{s_1, \cdots, s_r\})$$

Under the above assumption, knowledge of s_i allows the original i-th subimage to be retrieved from its watermarked version. More formally, given a subset $S \subset \{s_1, \cdots, s_r\}$, we can compute $\mathbf{X}'(S) = F^{-1}(\mathbf{X}', S)$, where $\mathbf{X}'(S)$ is a partially unwatermarked image resulting from inverting the watermarks in the subimages of $\mathbf{X}'$ that were watermarked with seeds in S.

From the previous discussion, we can see that, the larger the subset S of seeds known by a user, the more watermarks the user can invert, *i.e.* the closer is the unwatermarked image $\mathbf{X}'(S)$ to the original $\mathbf{X}$. This suggests the following algorithm to implement multilevel access to the watermarked file $\mathbf{X}'$:

Algorithm 5

1. *Let CH be a clearance hierarchy comprising u user categories (for example, for medical images, we could think of "doctor", "nurse", "other users"). For each category j, let k_j be a secret key known only to users in that category (the user does not actually need to know k_j, which can reside in her smart card).*

2. *For $i = 1, \cdots, r$ and $j = 1, \cdots, u$, encrypt s_i with some redundancy R_i under k_j to get $E_{k_j}(s_i||R_i)$ if s_i should be revealed to user category j. Note that different seeds may be used for each image, whereas the key k_j corresponding to category j is assumed to stay stable.*

3. *Assuming that the invertible watermarking algorithm used allows multiple marking without significant increase of the non-invertible distortion, embed $E_{k_j}(s_i||R_i)$, for $i = 1, \cdots, r$ and $j = 1, \cdots, u$, into*

the watermarked file $\mathbf{X}'$ to get a rewatermarked file $\mathbf{X}''$. A public seed is used for this second watermarking round.

From $\mathbf{X}''$, a user can recover and decrypt the subset S of seeds her category is entitled to know, and thus retrieve $\mathbf{X}'(S)$. Redundancy R_i encrypted with s_i allows the user to check that s_i was correctly decrypted.

As pointed out in Section 3, the Hartung-Girod method is an example of invertible watermarking which supports multiple marking. The only shortcoming of using n marking rounds (as required by Step 3 of Algorithm 5) is that α in the pre-processing algorithm (Algorithm 2) must be replaced with $n\alpha$. Thus, two marking rounds result in an increase of the non-invertible distortion introduced at the pre-processing stage. An alternative to avoid embedding $E_{k_j}(s_i||R_i)$ in the image is to keep those encrypted values in a freely accessible public repository.

Figure 1 shows the original image "Chips" [1] and its division into 12 subimages. The Hartung-Girod method has been used to embed the same watermark in each subimage. Figure 2 shows the completely watermarked version of the image and the partially unwatermarked version once the watermarks in the four subimages in the top row have been inverted using Algorithm 1. The partially unwatermarked version offers maximum precision and integrity verification to a user wishing to inspect the chip contact area depicted in the upper four subimages; the eight remaining subimages showing the rest of the chip still carry watermarks whose copyright messages can be used to prove ownership in case of unlawful redistribution. Table 1 gives the variation in the PSNR (Peak Signal-to-Noise Ratio) as the amplitude factor α and the number of unwatermarked subimages increase (the order of subimage unwatermarking is row-wise from top to bottom; within a row, subimages are unwatermarked from left to right); the top left column of the table corresponds to the completely watermarked image and the bottom right column is ∞ because the completely unwatermarked image exactly matches the pre-processed original image resulting from Algorithm 2. PSNRs compare the one-round watermarked images $\mathbf{X}'(S)$ with the pre-processed original $\mathbf{X}$. It would not make sense to use the two-round watermarked image $\mathbf{X}''$ in the comparison, because the second watermarking round can be undone by every user, as it uses a public seed.

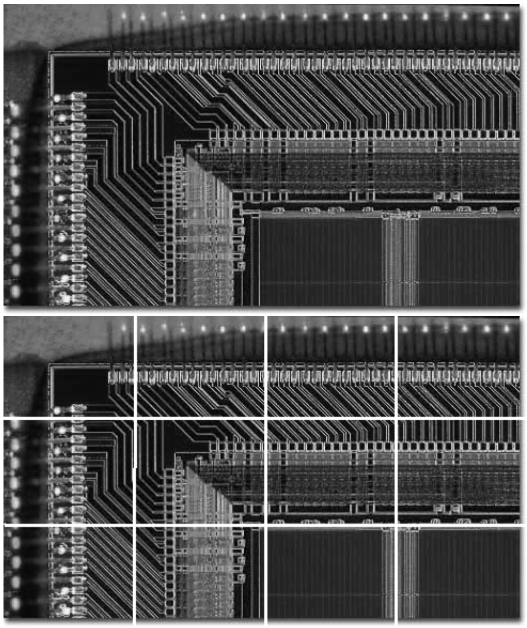

Figure 1. Top, original "Chips" image. Bottom, subimage division of "Chips" (12 tiles)

Note 1 (On the pre-processing distortion) *The PSNR between the original image "Chips" and its original pre-processed version (that is, between the input and the output of Algorithm 1) decreases as α increases and ranges between 57.69 dB when $\alpha = 4$ and 47.71 dB when $\alpha = 10$. Thus, for the values of α used, the distortion introduced by the pre-processing algorithm is very small.*

6 Conclusions

Invertible watermarking methods must be oblivious (*i.e.* not require the original image to recover watermarks) and the Hartung-Girod method is one of the simplest and most widely used robust oblivious watermarking techniques. We have presented in this paper a construction to invert this watermarking algorithm.

Image authentication is the most straightforward application of invertible watermarking and is especially interesting for precision-critical images. We have described how to combine image authentication and copyright protection, by constructing a scheme for multilevel access: a user can invert the watermarks of as many parts of the image as her clearance permits. In this way, privileged users can get more precise and authenticated images, whereas unprivileged users just see the watermarked image. This is a way to achieve a trade-off between the need of copyright protection (unprivileged users are a broad group not especially trusted) and the need for precision in some applications (privileged users requiring such precision will typically be a small group which the image owner trusts to some extent).

It is our belief that the results reported here will facilitate the use of copy detection techniques in domains where

[1] http://www.microscopy.fsu.edu/micro/gallery/chips/chipshots.html

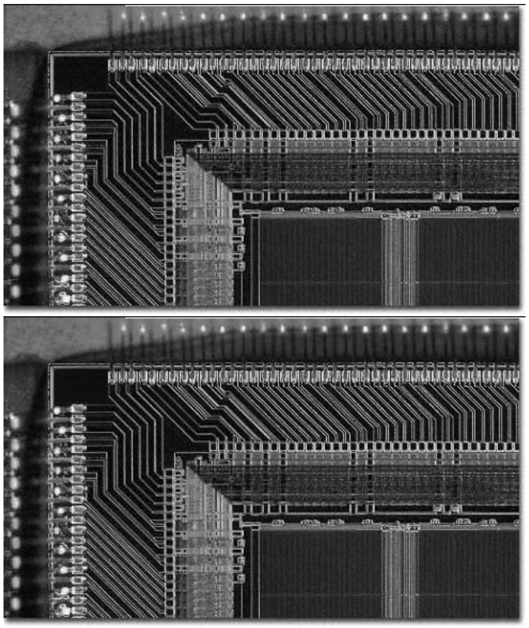

Figure 2. Top, completely watermarked "Chips" image. Bottom, "Chips" with the top four images unwatermarked

Table 1. Perceptual quality PSNR (dB) as a function of α and the number of unwatermarked subimages

	No. of unwatermarked subimages						
α	0	1	2	3	4	5	6
4	35.99	36.33	36.74	37.20	37.71	38.29	38.96
5	33.98	34.36	34.77	35.23	35.74	36.3	36.99
6	32.36	32.74	33.15	33.61	34.12	34.70	35.37
7	30.98	31.36	31.78	32.23	32.75	33.33	34.00
8	29.79	30.17	30.58	31.04	31.55	32.13	32.80
9	28.73	29.11	29.52	29.98	30.49	31.07	31.74
10	27.78	28.16	28.57	29.03	29.54	30.12	30.79

	No. of unwatermarked subimages					
α	7	8	9	10	11	12
4	39.75	40.72	41.97	43.73	46.74	∞
5	37.78	38.75	40.00	41.76	44.77	∞
6	36.16	37.13	38.38	40.14	43.15	∞
7	34.79	35.76	37.01	38.77	41.78	∞
8	33.59	34.56	35.81	37.57	40.58	∞
9	32.53	33.50	34.75	36.51	39.52	∞
10	31.58	32.55	33.80	35.56	38.57	∞

they are not currently being used because of the distortion they normally introduce. Examples are medical, military, quality-control and reverse-engineering images.

References

[1] DigiRay, `http://www.digiray.com/reverse-engineering/index.html`

[2] J. Fridrich, M. Goljan and R. Du, "Invertible authentication watermark for JPEG images", in *IEEE International Conference on Information Technology: Coding and Computing - ITCC'2001*, IEEE Computer Society, 2001, pp. 223-227.

[3] J. Fridrich, M. Goljan and R. Du, "Invertible authentication", in *Proc. SPIE Security and Watermarking of Multimedia Contents*, San José CA, Jan. 23-26, 2001.

[4] M. Goljan, J. Fridrich and R. Du, "Distortion-free data embedding", in *4th Information Hiding Workshop*, Pittsburgh PA, April 2001.

[5] F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video", *Signal Processing*, 66(3): 283-301, May 1998.

[6] A. Herrigel, J. Ó Ruanaidh, H. Petersen, S. Pereira and T. Pun, "Secure copyright protection techniques for digital images", in *2nd Information Hiding Workshop*, Portland, Oregon, 1998.

[7] C. W. Honsinger, P. Jones, M. Rabbani and J. C. Stoffel, "Lossless recovery of an original image containing embedded data", US Patent Application, Docket No: 77102/E-D, 1999.

[8] J. Kumagai, "Chip detectives", *IEEE Spectrum*, 37(11): 43-48, Nov. 2000.

[9] D. Nicholson, *Spread Spectrum Signal Design - Low Probability of Exploitation and Anti-Jam Systems*, Computer Science Press, 1988.

[10] J. Domingo-Ferrer, J. M. Mateo-Sanz and F. Sebé, "Watermarking for Multilevel Access to Statistical Databases", in *IEEE International Conference on Information Technology: Coding and Computing - ITCC'2001*, IEEE Computer Society, 2001, pp. 243-247.