Stéganographie par Deep Learning

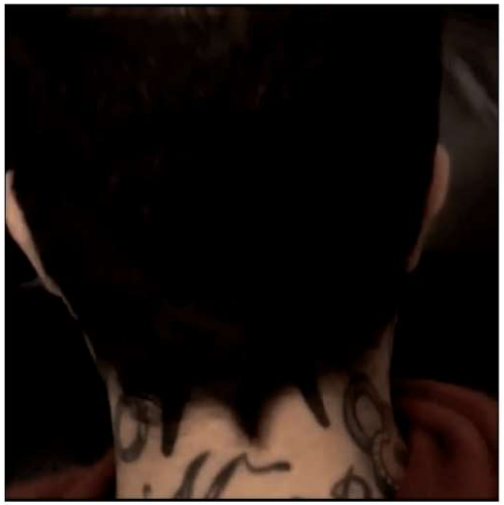
Marc CHAUMONT[1,2]
LIRMM[1], Univ Nîmes[2], Montpellier, France

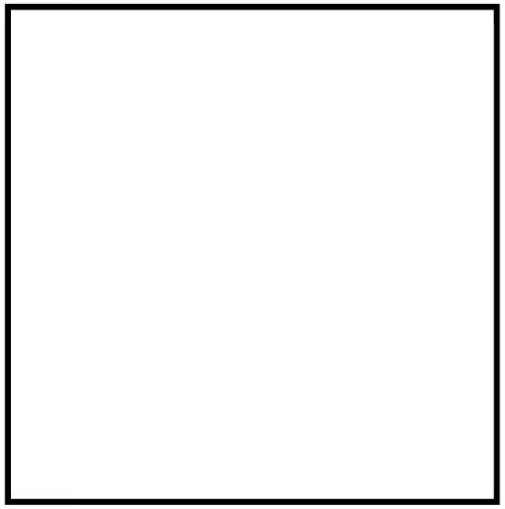
March 8, 2020

Présentation de laboratoire. LIRMM le 9 mars 2020.

Outline

La stéganographie (une histoire ancienne)





" 499 avant J-C. Histiée sélectionne son plus fidèle esclave, lui fait raser la tête et tatouer un ordre de révolte sur le crâne; Dès que les cheveux eurent repoussé, il l'envoie à Aristagoras qui lui rase de nouveau la tête afin de lire le message. " Wikipedia.

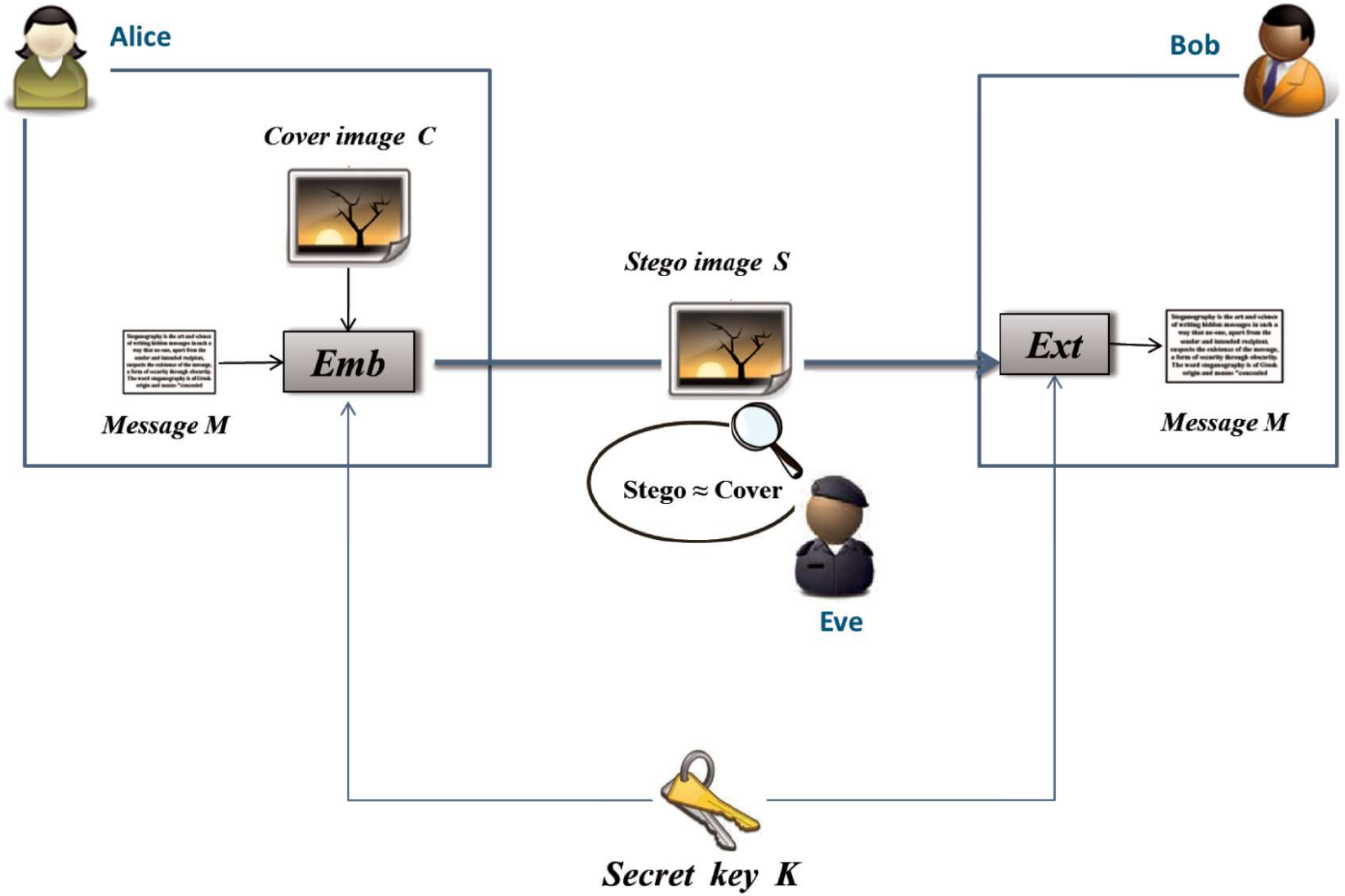
Une toute petite communauté de chercheurs

Stéganographie moderne (utilisation de codes) depuis début 2000.

- Petit groupe au LIRMM.
- Quatre spécialistes en France.
- Une vingtaine d'équipes dans le monde.

Adossé à la société savante
IEEE Signal Processing Information Forensics & Security
https://signalprocessingsociety.org/get-involved/information-forensics-and-security

Stéganographie / Stéganalyse



Un exemple simpliste d'insertion (1)

EXEMPLE : INSERTION LSB

- Prenons l'exemple d'une image en niveau de gris valeur ou les valeurs des pixels appartiennent à {0,…,255} :

Un exemple simpliste d'insertion (2)

EXEMPLE : INSERTION LSB

Zone de l'œil
Agrandie (15x15)

52	59	64	56	54	57	53	50	55	49	50	53	51	53	53
51	54	52	51	52	52	49	50	47	49	48	46	50	59	48
53	52	52	58	51	47	50	51	46	47	46	47	44	51	55
53	51	53	55	51	53	45	43	42	46	45	47	48	79	66
48	47	55	47	51	48	46	44	45	47	44	50	56	70	90
53	53	47	43	54	49	50	40	46	47	75	62	69	94	128
77	67	43	47	62	60	45	39	35	40	87	87	65	86	121
90	81	50	60	80	79	48	37	35	38	55	90	65	50	72
103	90	50	57	94	93	76	47	40	42	68	112	77	56	66
115	107	71	52	92	98	90	72	66	66	90	108	74	53	87
117	121	99	67	58	84	98	84	84	91	83	72	57	66	126
108	127	115	88	59	60	79	86	90	80	76	55	65	113	173
105	114	125	113	89	62	57	54	60	57	64	77	107	160	198
113	112	117	120	113	95	82	63	65	75	88	127	158	188	202
115	119	119	124	132	120	108	101	102	116	137	163	186	197	198

Valeur des pixels appartenant à {0, ..., 255}

Un exemple simpliste d'insertion (3)

EXEMPLE : INSERTION LSB

- Représentation en binaire des entiers de {0, ..., 255} :

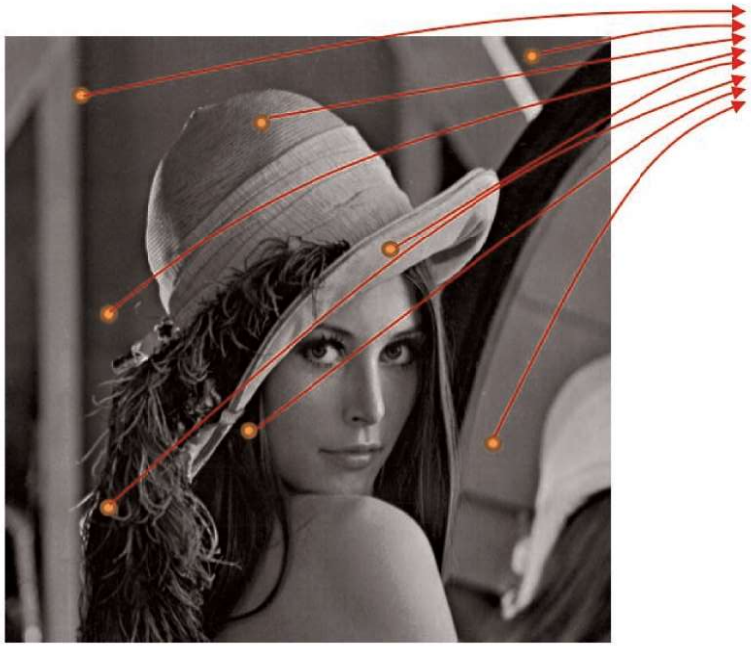
octet

ENTIER (BASE 10)	BIT 7	BIT 6	BIT 5	BIT 4	BIT 3	BIT 2	BIT 1	BIT 0
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	1	0
3	0	0	0	0	0	0	1	1
4	0	0	0	0	0	1	0	0
...								
254	1	1	1	1	1	1	1	0
255	1	1	1	1	1	1	1	1

Un exemple simpliste d'insertion (4)

EXEMPLE : INSERTION LSB

- L'insertion par substitution du bit de poids faible (LSB : Low Significant Bit Data-Hiding) :



Choix de quelques pixels

Un exemple simpliste d'insertion (5)

EXEMPLE : INSERTION LSB

- L'insertion par substitution du bit de poids faible (LSB : Low Significant Bit Data-Hiding) :

Choix de quelques pixels

Chacun des pixels va être modifié pour « embarquer » un bit

Un exemple simpliste d'insertion (6)

EXEMPLE : INSERTION LSB

- L'insertion par substitution du bit de poids faible (LSB : Low Significant Bit Data-Hiding) :

Exemple :

- Supposons que le pixel vaut 42
- Supposons que le bit à insérer vaut 1

- 42 en binaire s'écrit 101 010
- Après insertion on a 101 011
  c'est-à-dire la valeur 43

... modification invisible à l'œil

La stéganographie moderne (adaptative)

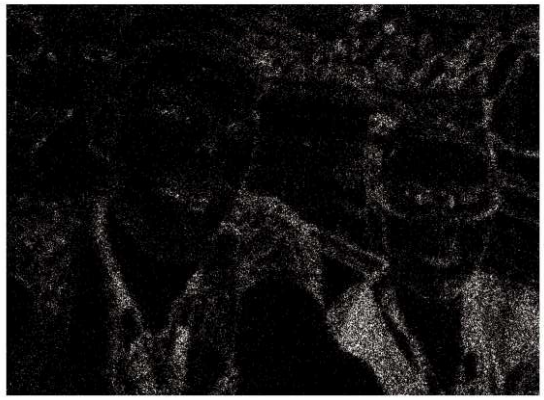
Stéganographie non adaptative



https://stylesuxx.github.io/steganography/

Stéganographie adaptative



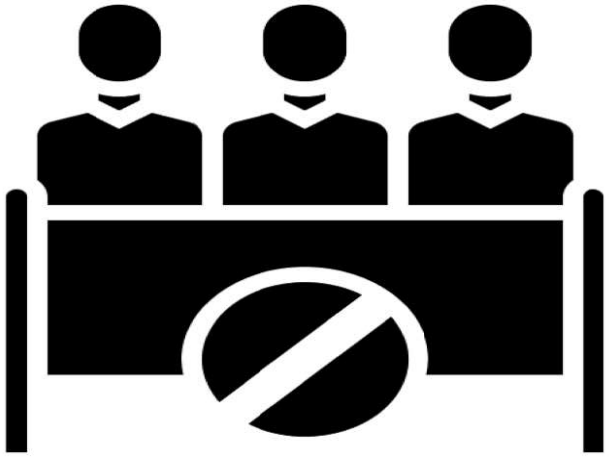


V. Holub, J. Fridrich, and T. Denemark
Universal Distortion Function for Steganography in an Arbitrary Domain

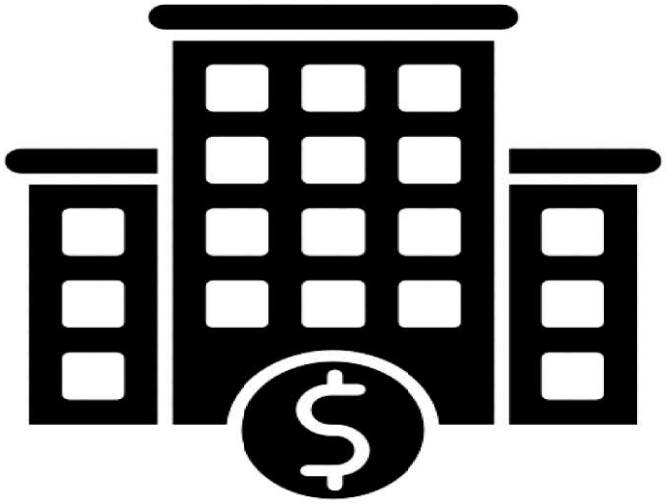
Les acteurs / utilisateurs

Agences gouvernementales

Journalistes / Dissidents



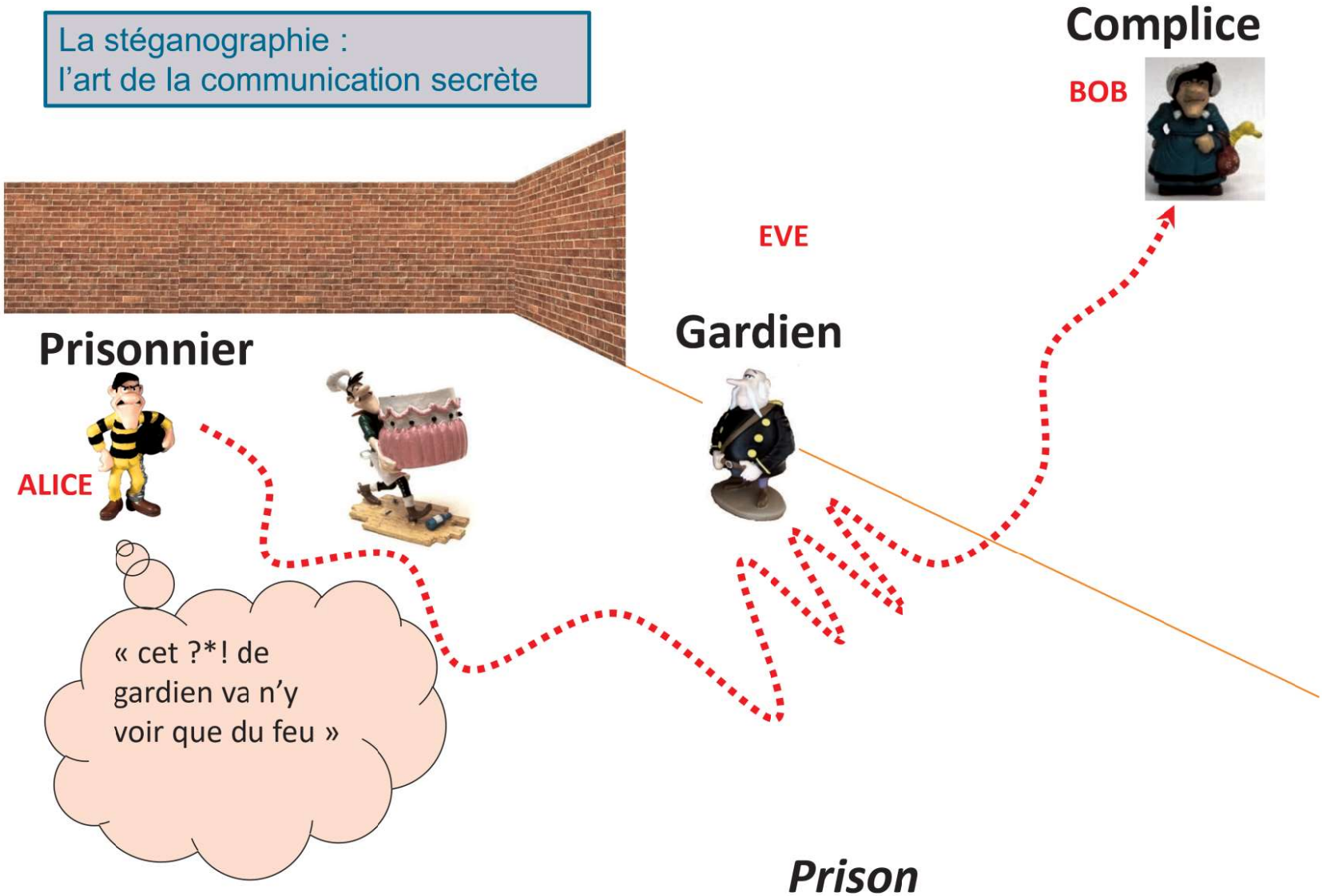
Entreprises



Terroristes

La notion de sécurité...

La stéganographie :
l'art de la communication secrète



[Simmons, 1983]

Modèle de gardien passif

Le scénario de stéganalyse "à clairvoyance" (pire des cas)

1. Une **unique (jetable) clef d'insertion**(*) pour chaque insertion.
2. Scénario similaire au principe de Kerckhoffs (cryptographie). Eve connaît/dipose :
   - de l'algorithme d'insertion et d'extraction du message,
   - de l'ensemble des paramètres publics de ces deux algorithmes,
   - d'une base d'images similaires (connait le developement) à celles utilisées par Alice et Bob,
   - on suppose souvent que les images sont de même taille et que le nombre de bits insérés est fixe.

(*) L. Pibre, J. Pasquet, D. Ienco, and M. Chaumont
Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover source-mismatch

Les seuls paramètres qu'Eve ne connaît pas sont les clefs secrètes utilisées par Alice et Bob pour insérer/extraire les messages dans les images.

Le scénario de stéganalyse "à clairvoyance" (remarque)

$\rightarrow$ On fait rarement des attaques sur la clef.

Exemple d'attaque sur la clef quand il y a une erreur de déploiement (instantiation) de l'algorithme d'insertion/extraction :

T. Pevný and A. D. Ker.
Steganographic key leakage through payload metadata

Eve "détecte" (elle classe) si "cover" ou "stego"

Deux hypothèses :

$$\begin{aligned} \mathcal{H}_0 &= \mathbf{x} \sim P_c, \\ \mathcal{H}_1 &= \mathbf{x} \sim P_s. \end{aligned}$$

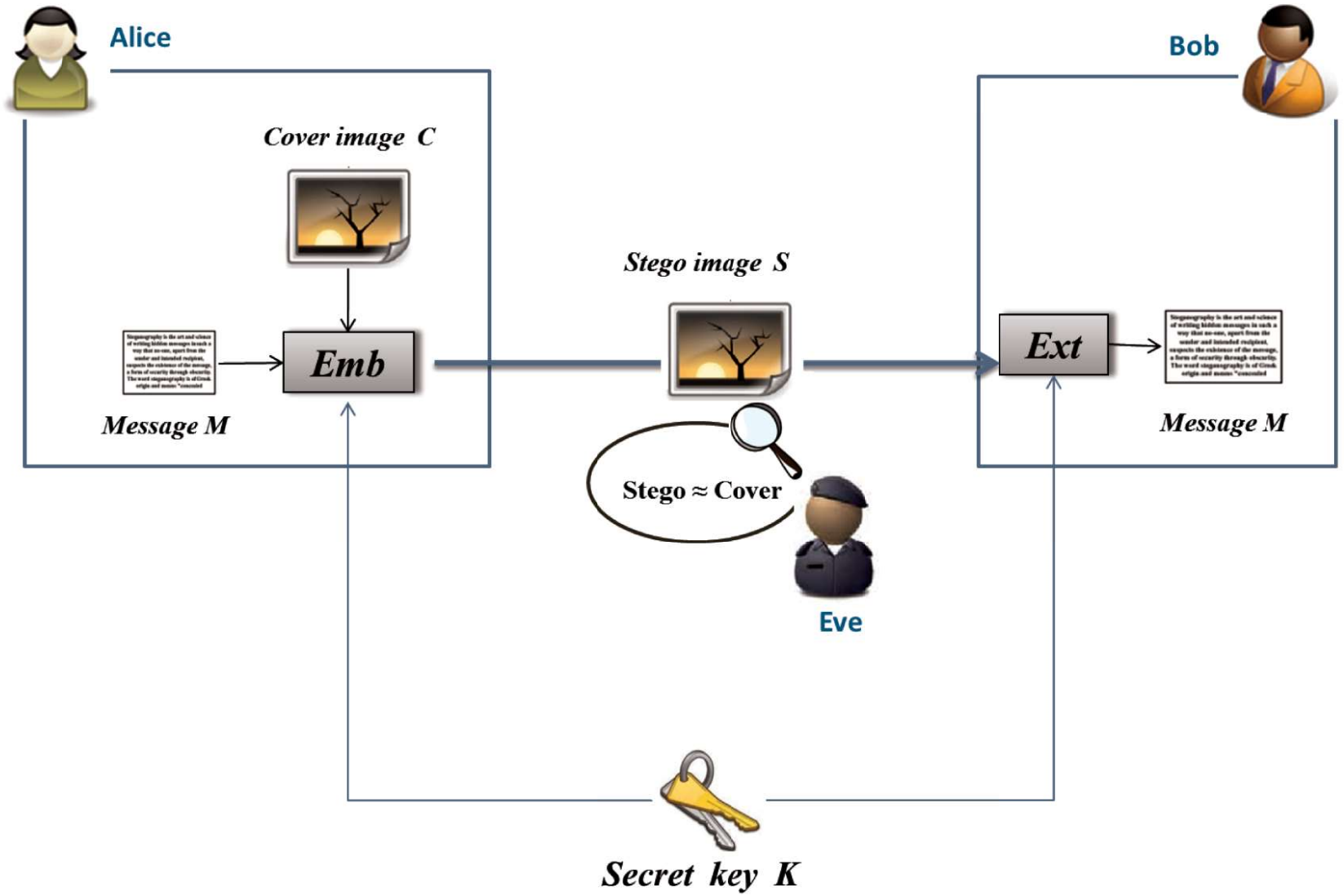
On construit un détecteur $\delta(.)$:

$$\begin{aligned} \delta : \mathbb{N}^{h \times w} &\rightarrow \{0, 1\} \\ \mathbf{x} &\rightarrow \delta(\mathbf{x}) \end{aligned}$$

- par test d'hypothèses,
- par Machine-Learning ou Deep-Learning.

Outline

Steganography / Steganalysis



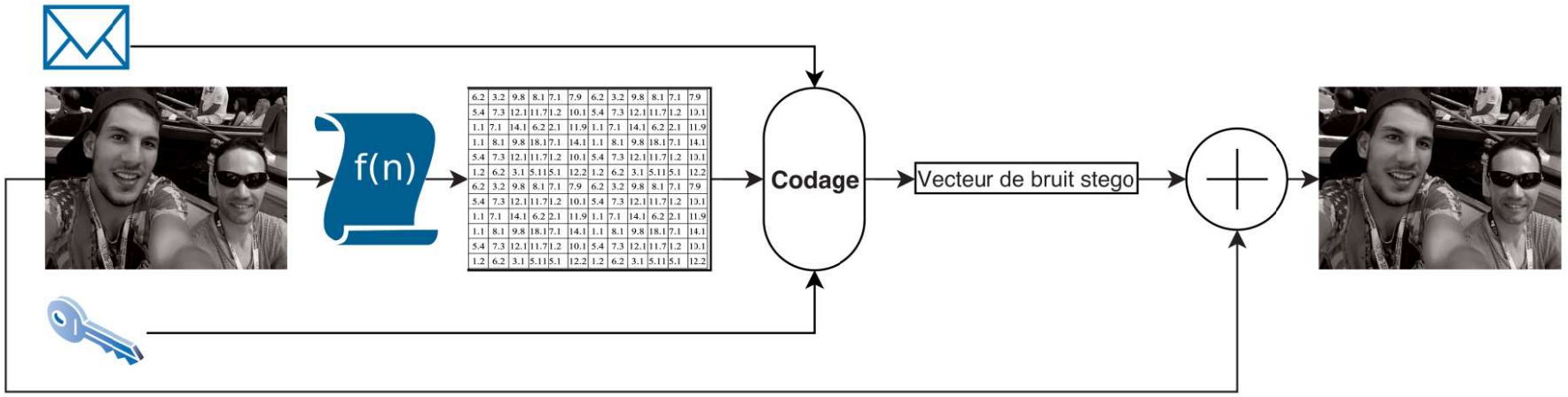
Stéganographie adaptative

Protocole opérationnel

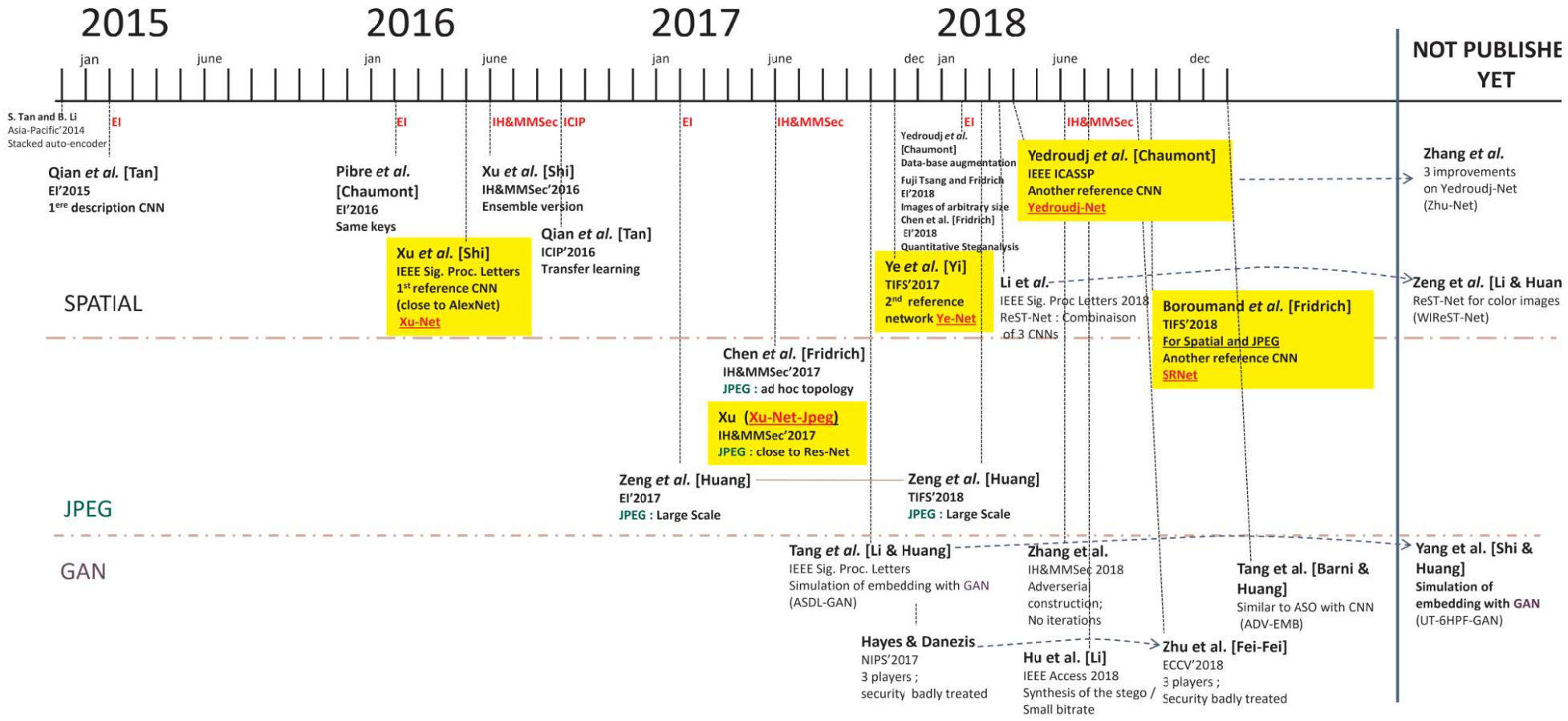
1. Génération d'une carte des coûts
2. Encodage du message
3. Insertion du message

Exemples

- WOW
- S-UNIWARD
- MIPOD
- HILL

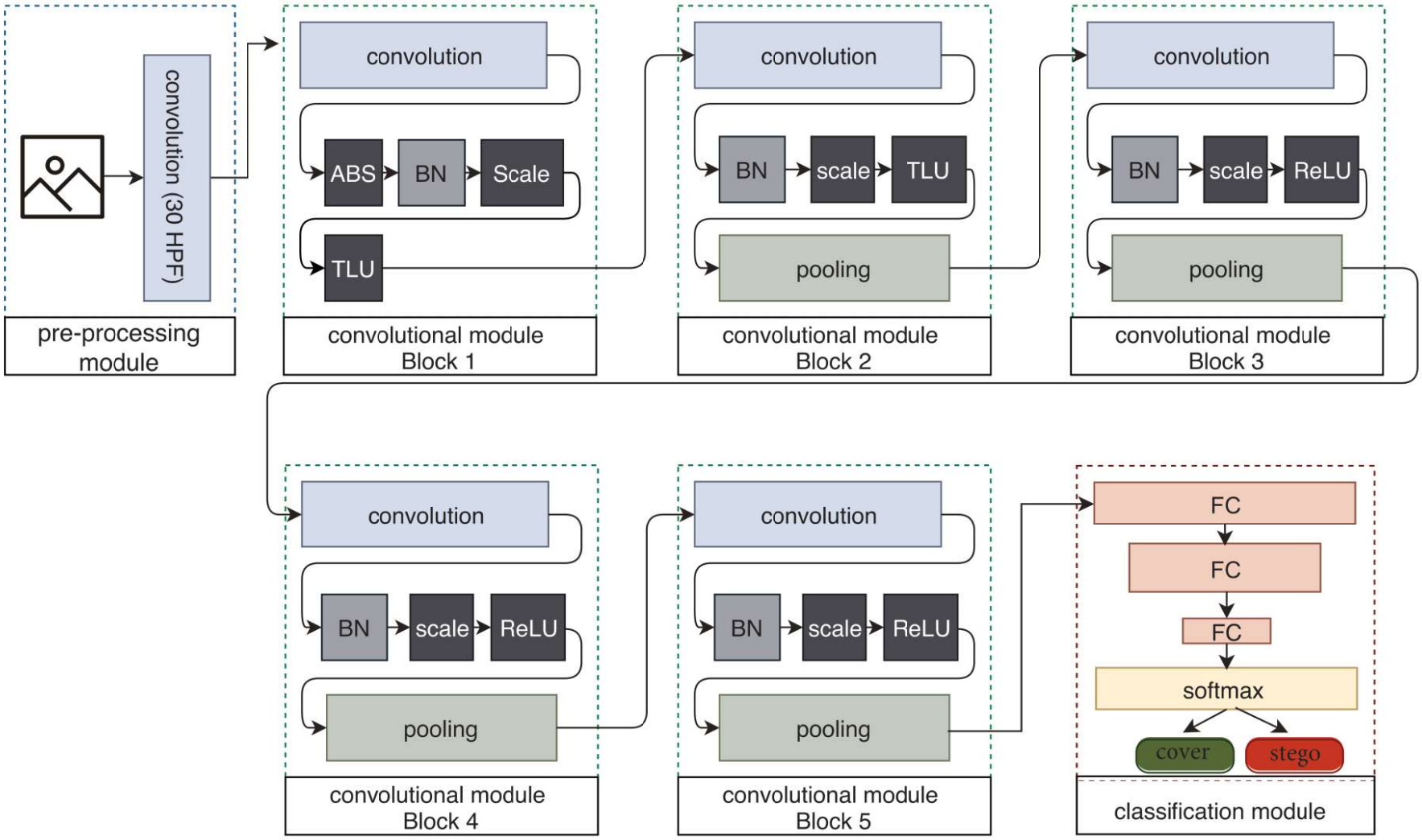


La stéganalyse par deep



Chronology of the main CNNs for steganography and steganalysis from 2015 to 2018

Yedroudj-Net



M. Yedroudj, F. Comby, and M. Chaumont

Yedrouj-Net: An efficient CNN for spatial steganalysis

Outline

Les quatre familles :

Famille :

- **1) par synthèse (pas de mofications):**

  Principe : Synthétise une image via un générateur (GAN) et extrait la "graine"/message par un "extracteur"

- **2) par insertion adverserielle itérée :**

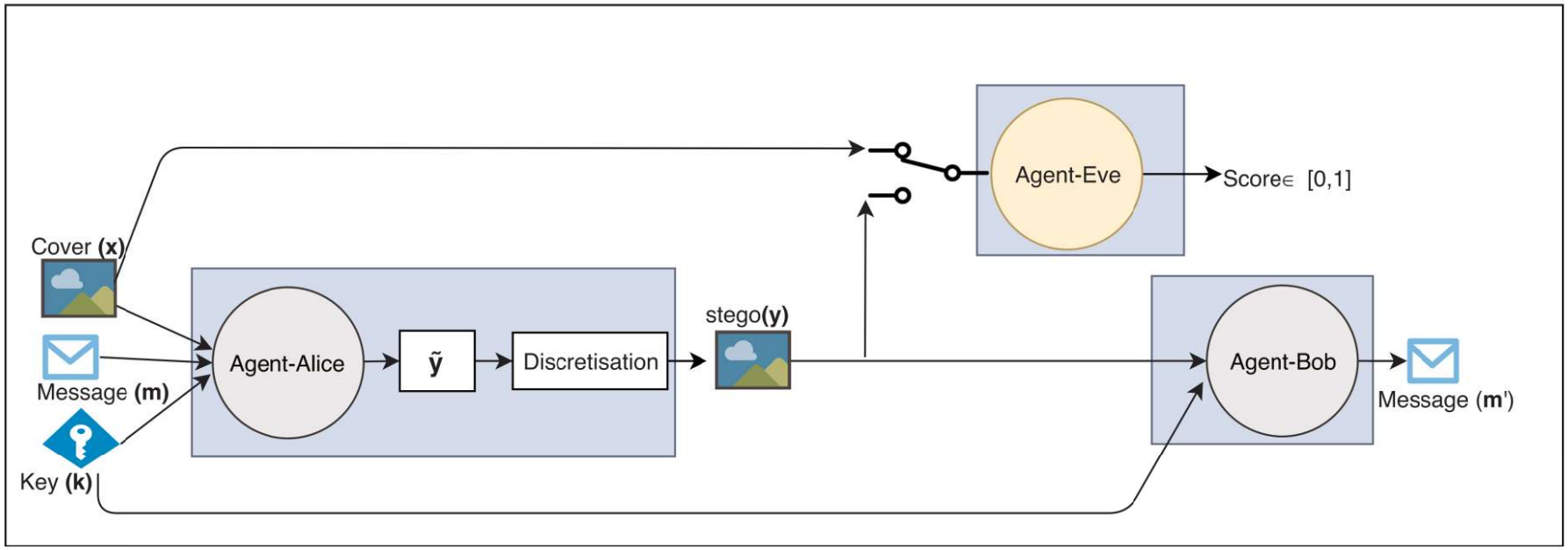
  Principe : On utilise un algoritme d'insertion pour produire des exemples "adverserielles" et un reseau "discriminateur".

- **3) par génération de la carte de probabilités de modifications :**

  Principe : Un réseau générant une carte puis simulant l'insertion et un reseau "discriminateur".

- **4) par jeu à 3 joueurs :**

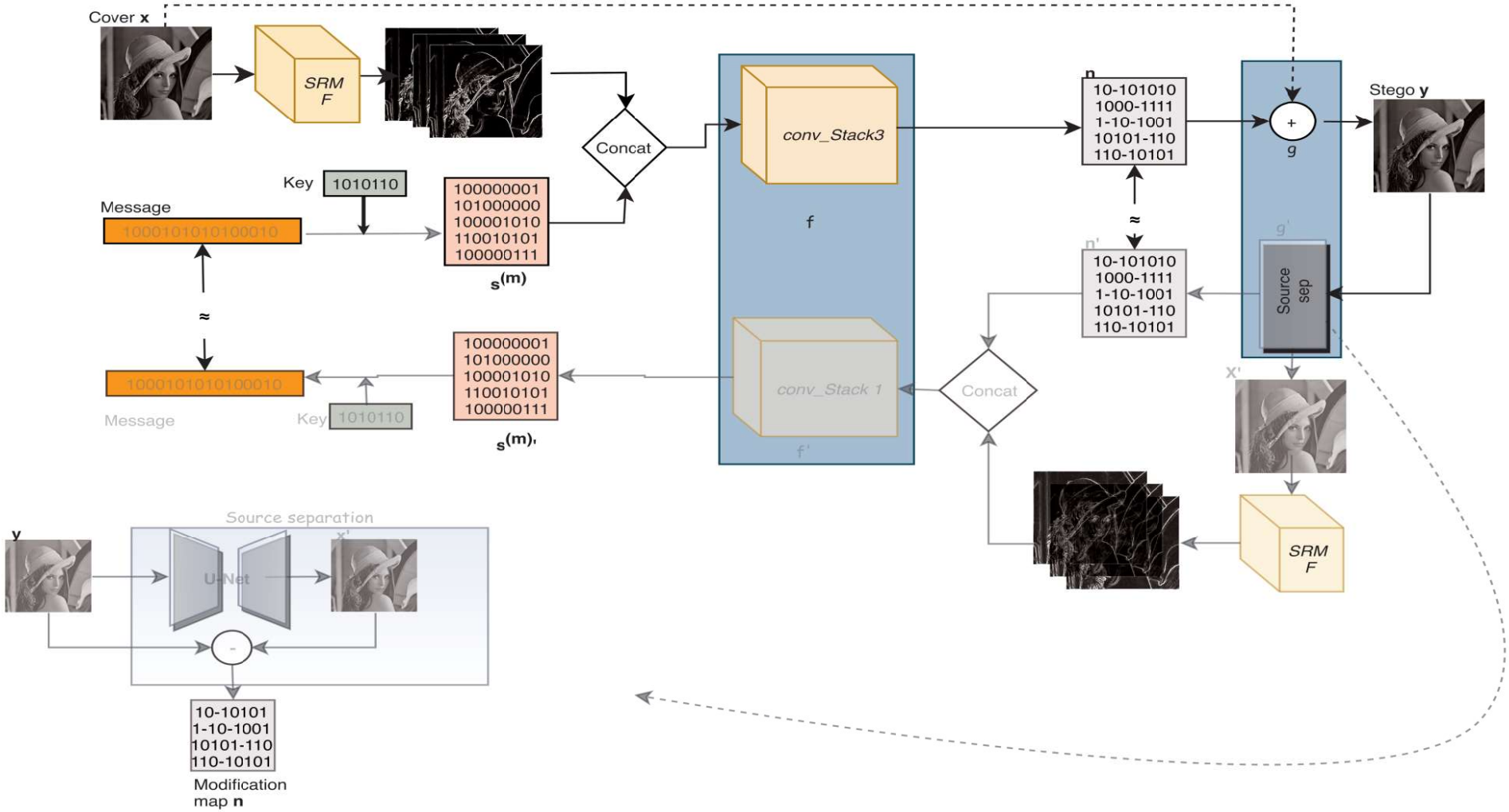
Approche à 3 joueurs



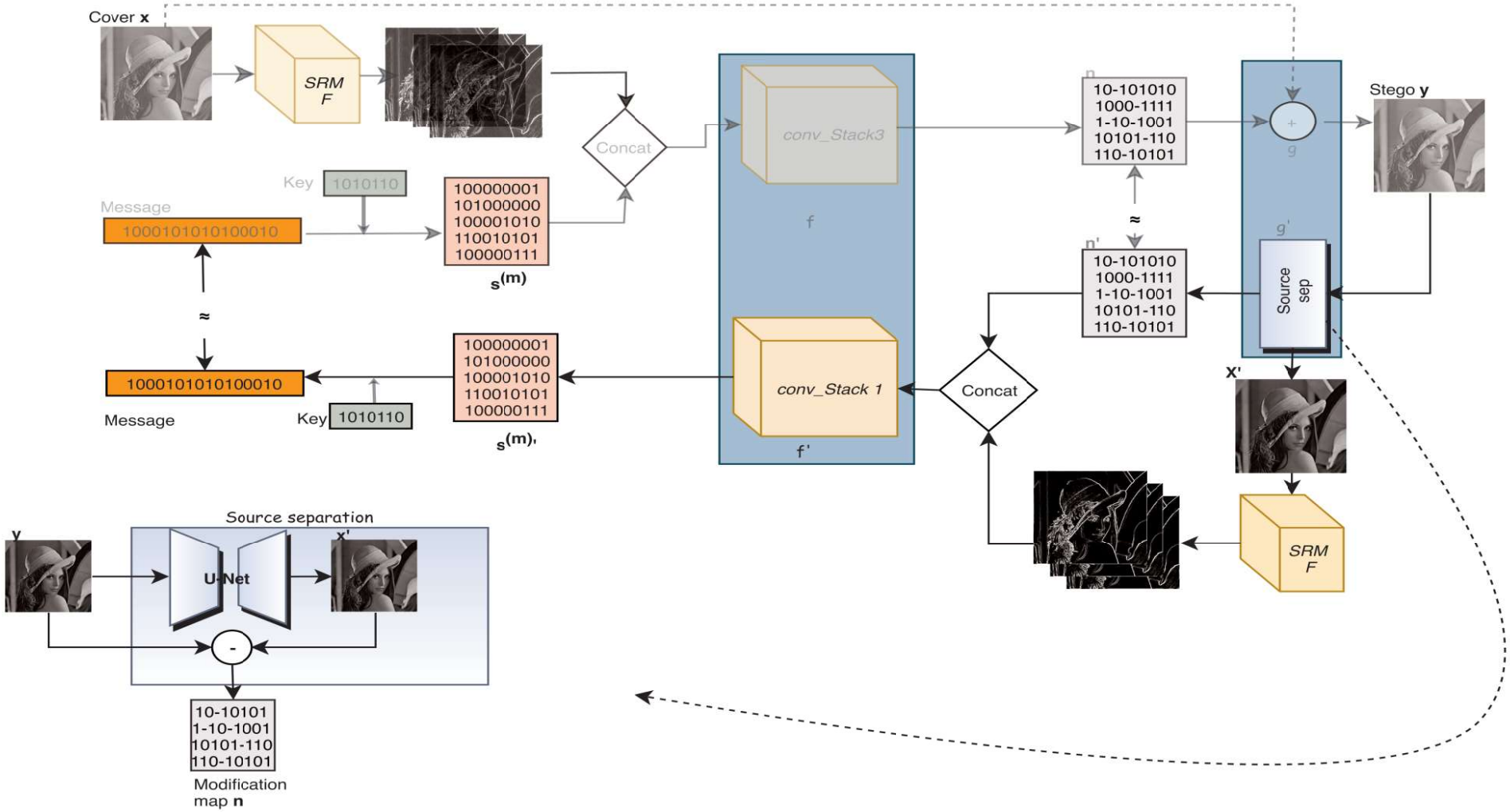
Algorithm 1: *3-players game* TRAINING

while not converge OR $loop \leq max\text{-}iter$ **do**
  // Agent-Alice et Agent-Bob
  for $iter\_team1 \leq it1$ **do**
    forward-propagation (*cover*, *message*, *key*);
    // Minimize $\mathcal{L}_{Bob} = \left(\sum_{i=1}^{m}(m_i - m'_i)^2\right)/m$
    update_Agent-Bob ($\mathcal{L}_{Bob}$);
    // Minimize $\mathcal{L}_{Alice} = \lambda_A \cdot (dist(\mathbf{x},\mathbf{y}) - \beta) + \lambda_B \cdot \mathcal{L}_{Bob} - \lambda_E \cdot \mathcal{L}_{Eve}$,
    update_Agent-Alice ($\mathcal{L}_{Alice}$);
  end
  // Agent-Eve
  for $iter\_team2 \leq it2$ **do**
    forward-propagation (*cover*, *stego*);
    // Minimize $\mathcal{L}_{Eve} = dist(l - Agent\text{-}Eve(\mathbf{z}))$
    update_Agent-Eve ($\mathcal{L}_{Eve}$);
  end
end

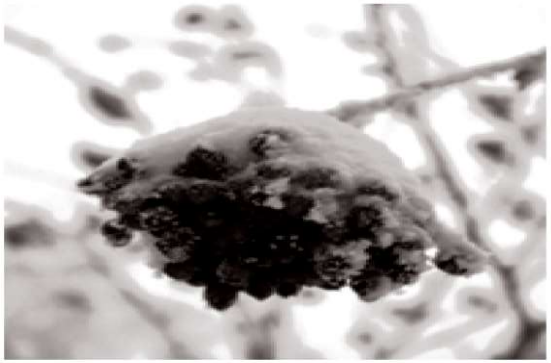
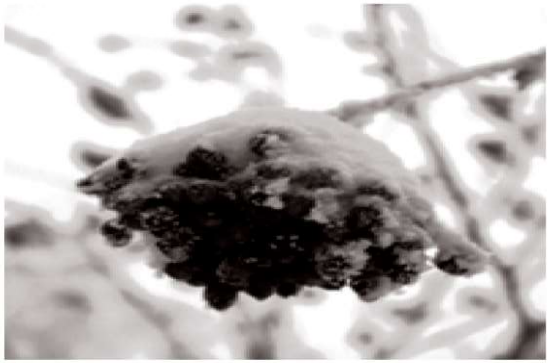
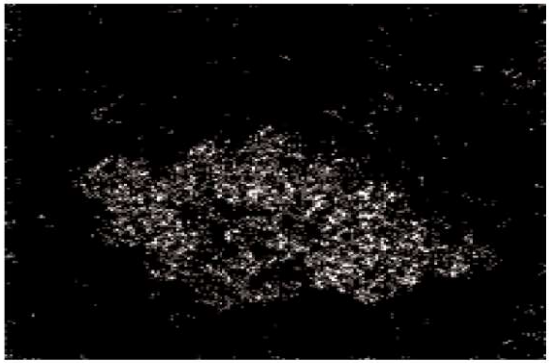
Agent-Alice et Agent-Bob



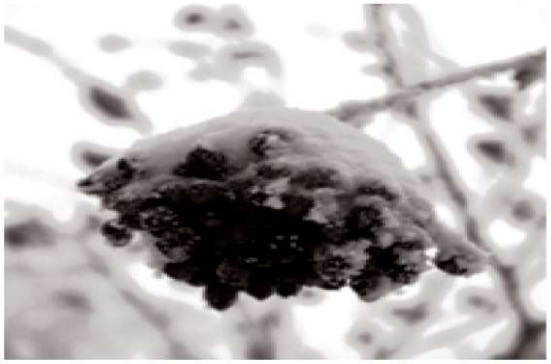
Agent-Alice et Agent-Bob

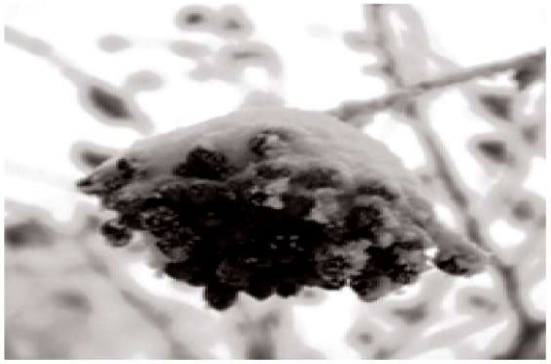


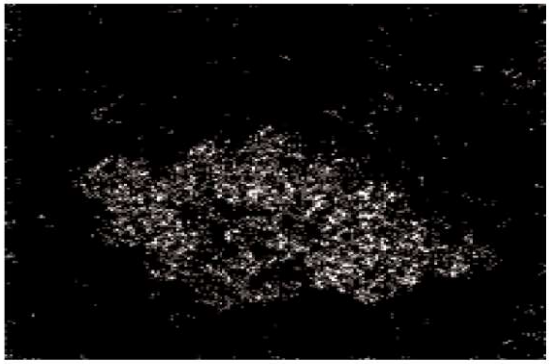
Résultats

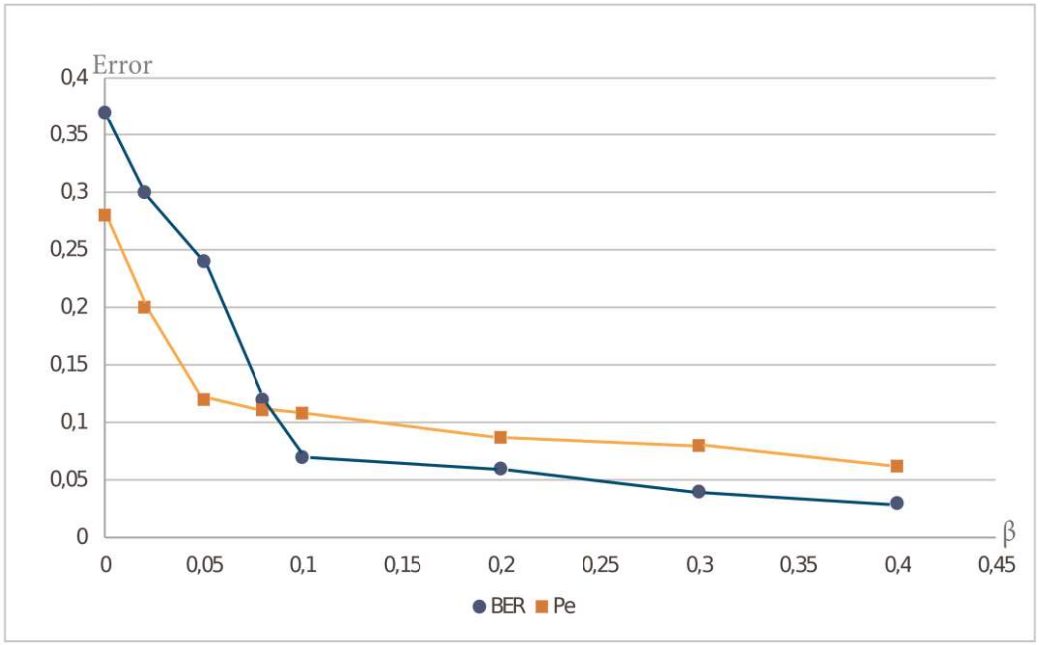




Résultats



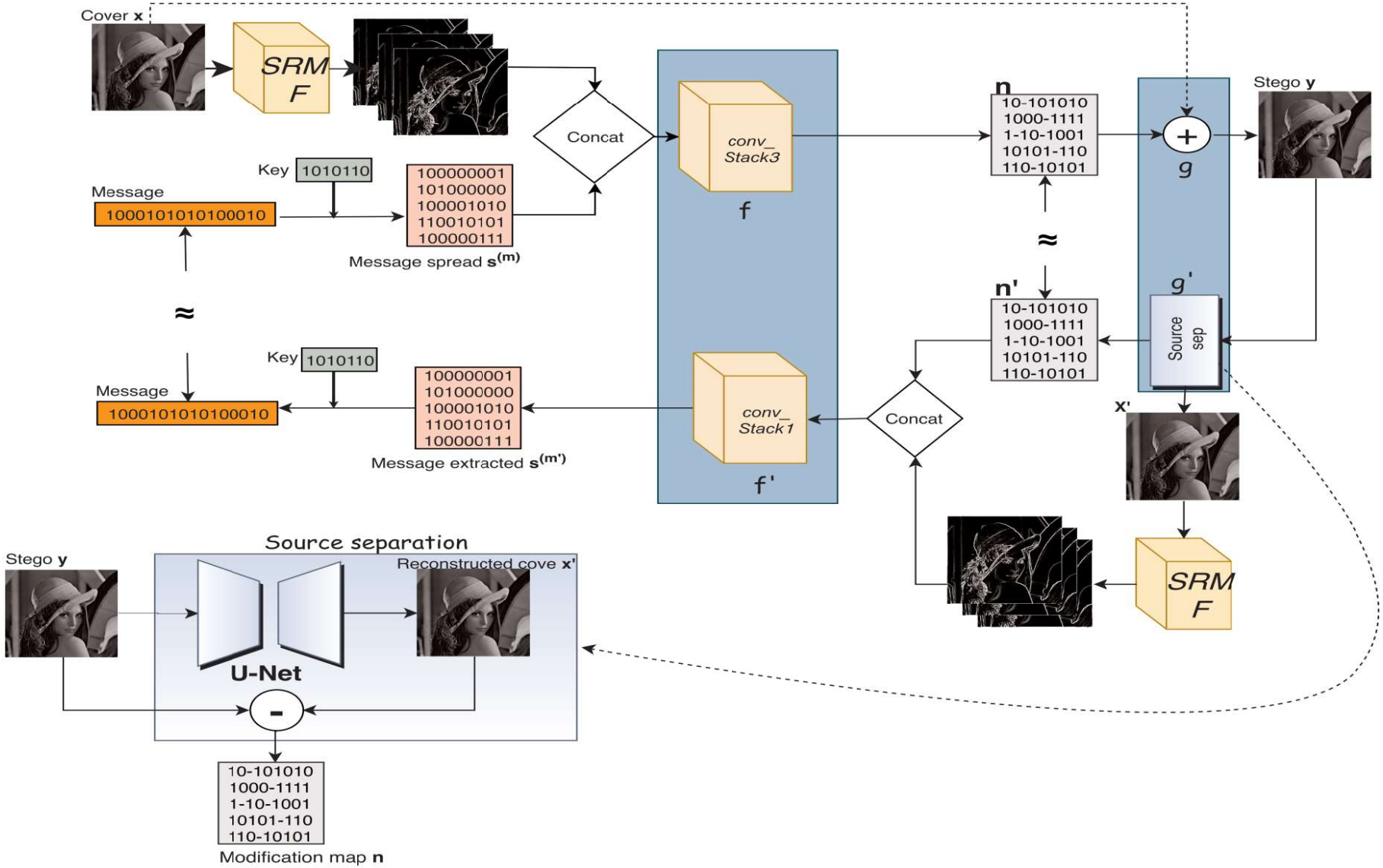






- Pour $\beta = 0.1$
  - $BER = 6\%$
  - Payload réel= 0.3 bpp
  - $Pe = 11.2\%$
- WOW
  - Payload réel= 0.3 bpp
  - $Pe = 20.9\%$

Reflexion sur l'abscence d'un bloc de codage/decodage



Points forts de la méthode

Avantage de la méthode proposée

- Clé secrète
- Stegos avec des valeurs entières
- Adaptative / la quantité d'information à insérer
- Images de taille scalable (256×256)
- Stéganalyseur efficace
- Pas d'hypothèse sur l'additivité de la fonction de distortion
- Simultation d'un équilibre de Nash (Agent-Eve est dans le processus)