Combining Similarity and Adversarial Learning to Generate Visual Explanation: Application to Medical Image Classification

Martin Charachon[1,2], Céline Hudelot[2], Paul-Henry Cournède[2], Camille Ruppli[1], Roberto Ardon[1]

[1]Incepto Medical
[2]Université Paris-Saclay, CentraleSupélec, MICS

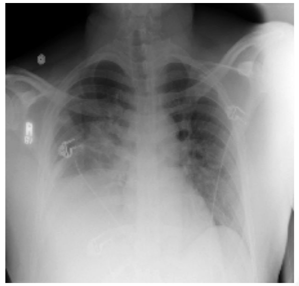
Introduction - Context

université PARIS-SACLAY

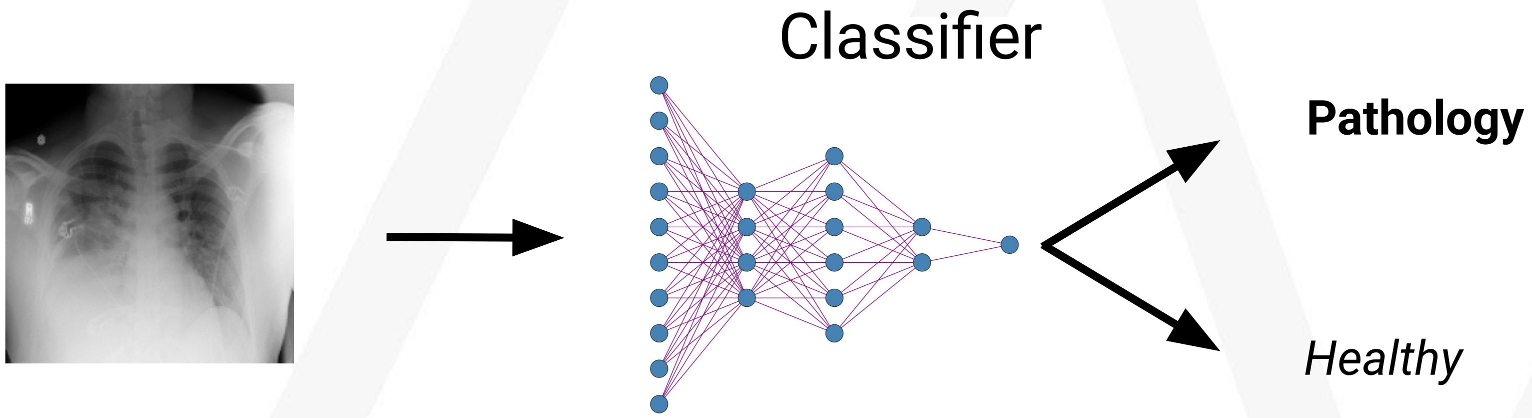
CentraleSupélec

INCEPTO

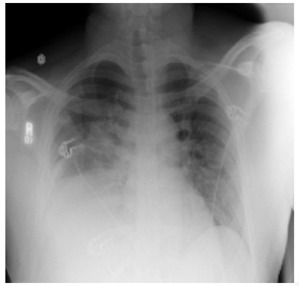
Introduction - Context

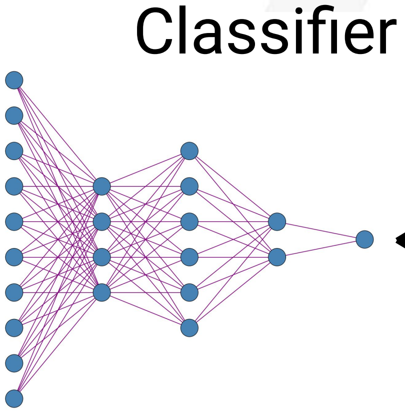


Introduction - Context



Introduction - Context

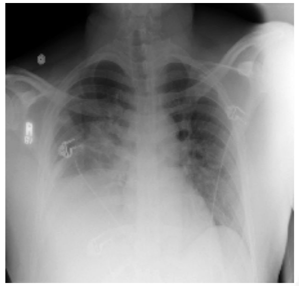




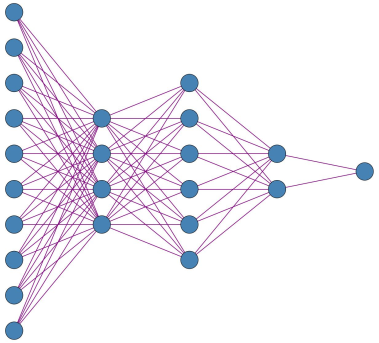
Pathology

Healthy

Introduction - Context



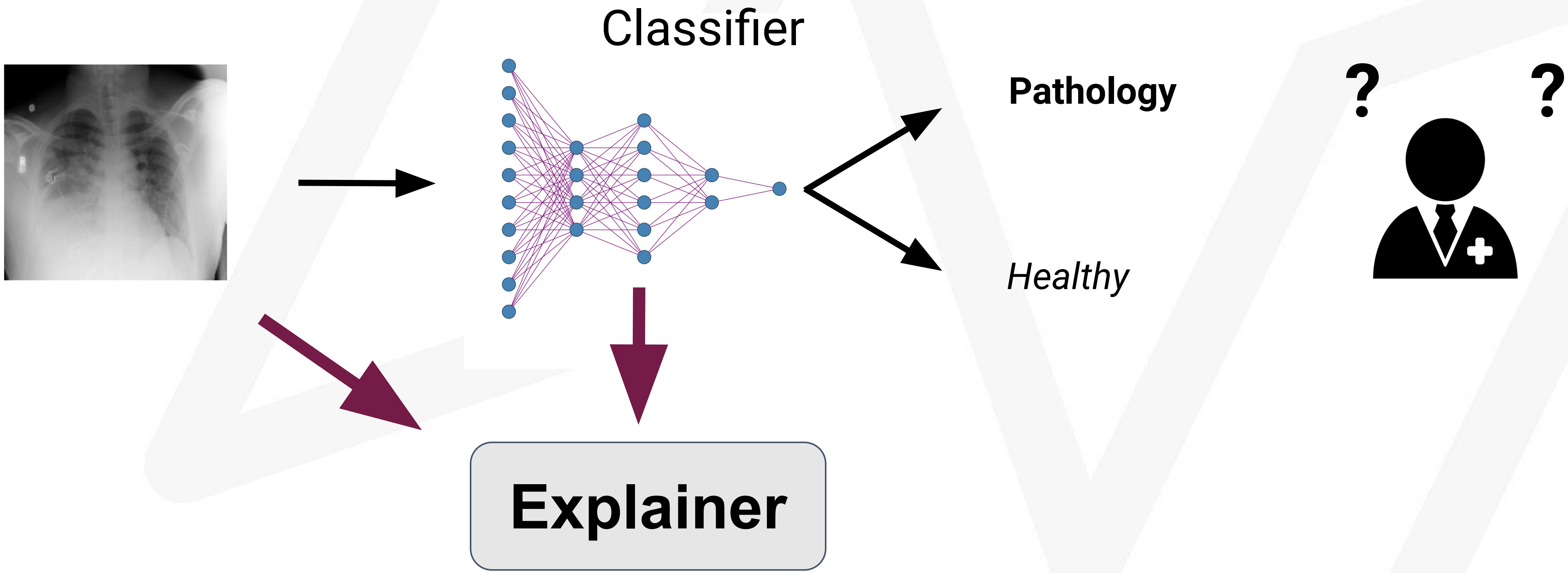
Classifier



Pathology

Healthy

Introduction - Context

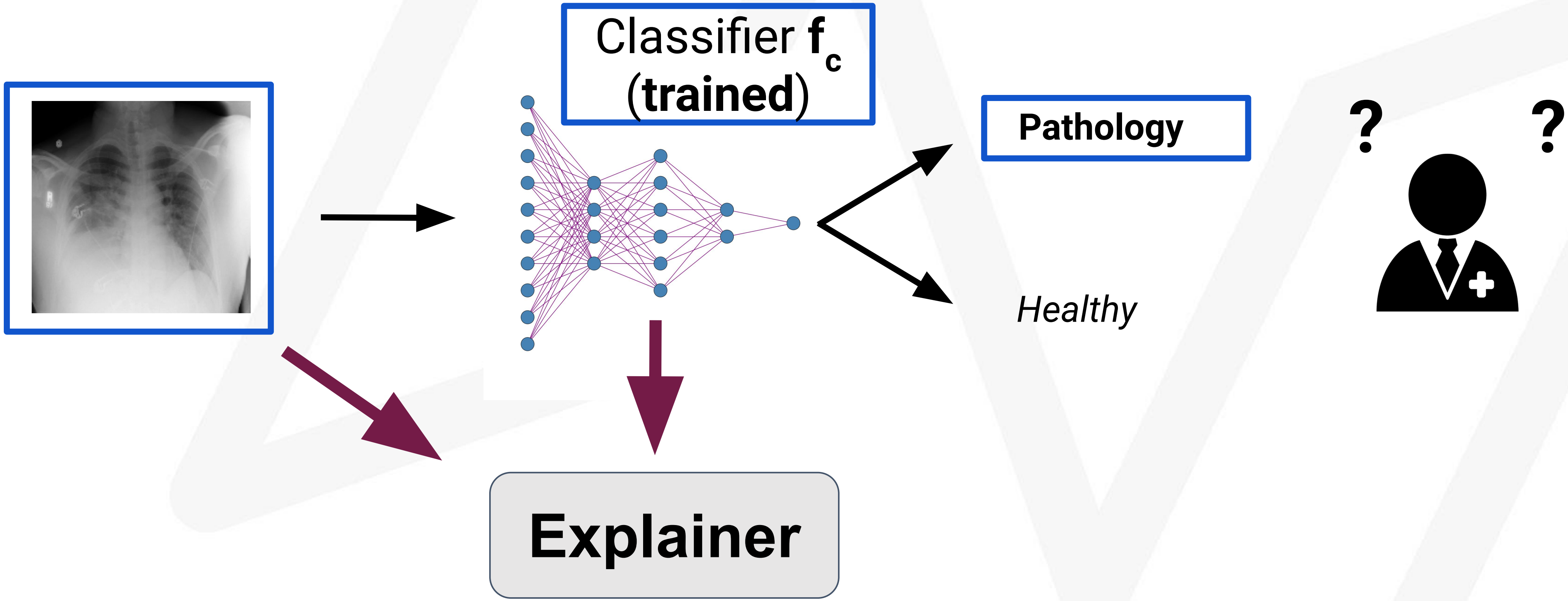


université PARIS-SACLAY

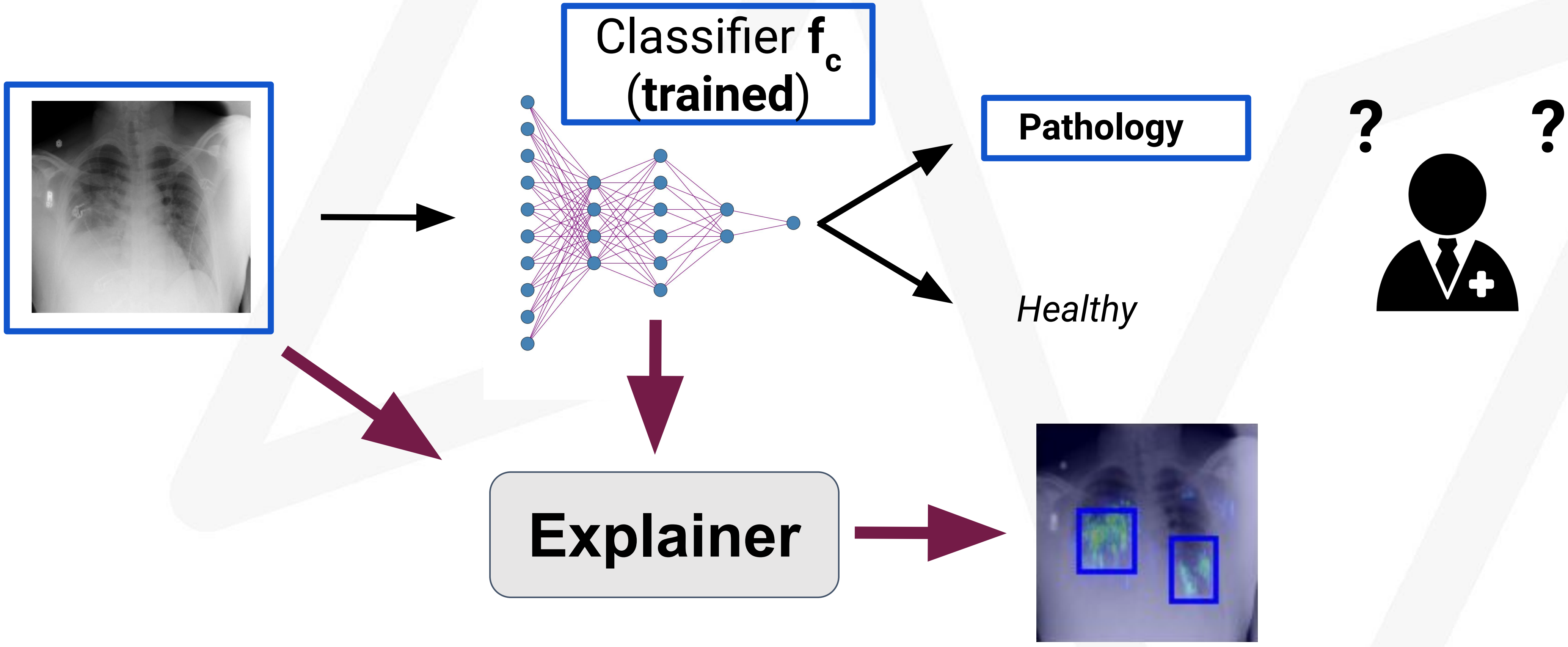
CentraleSupélec

INCEPTO

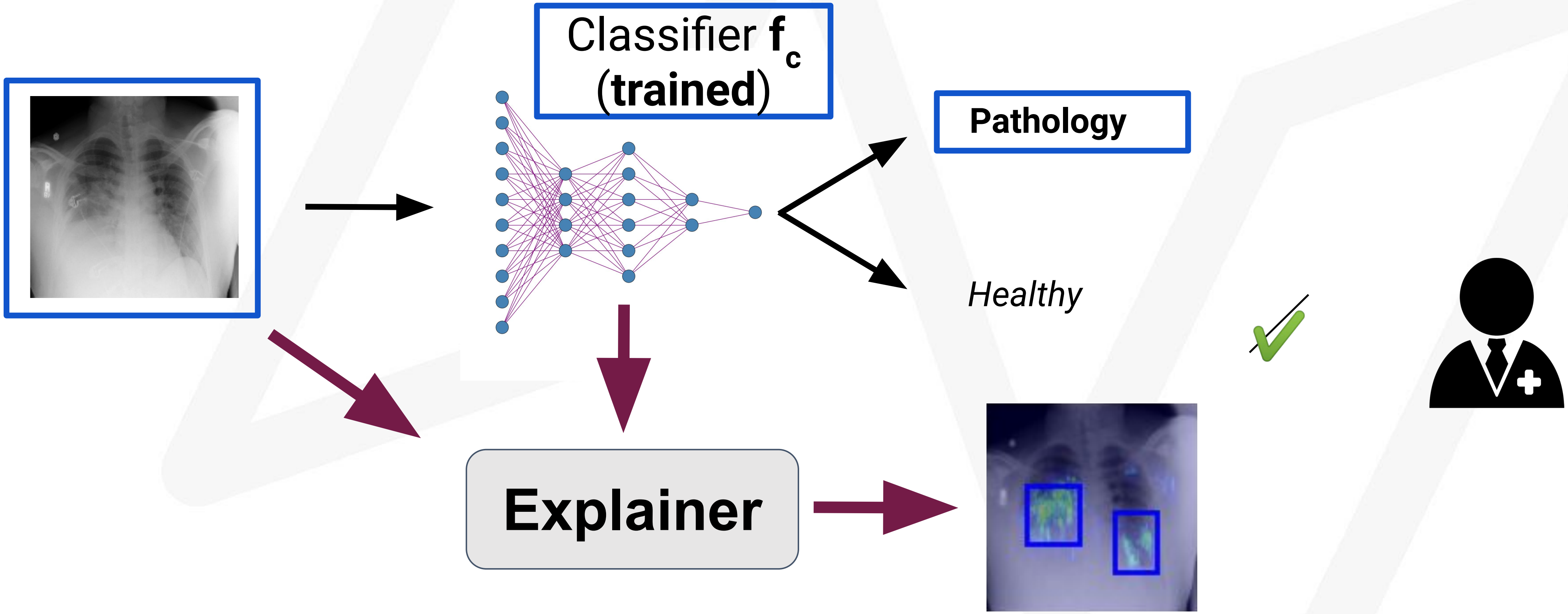
Introduction - Context



Introduction - Context

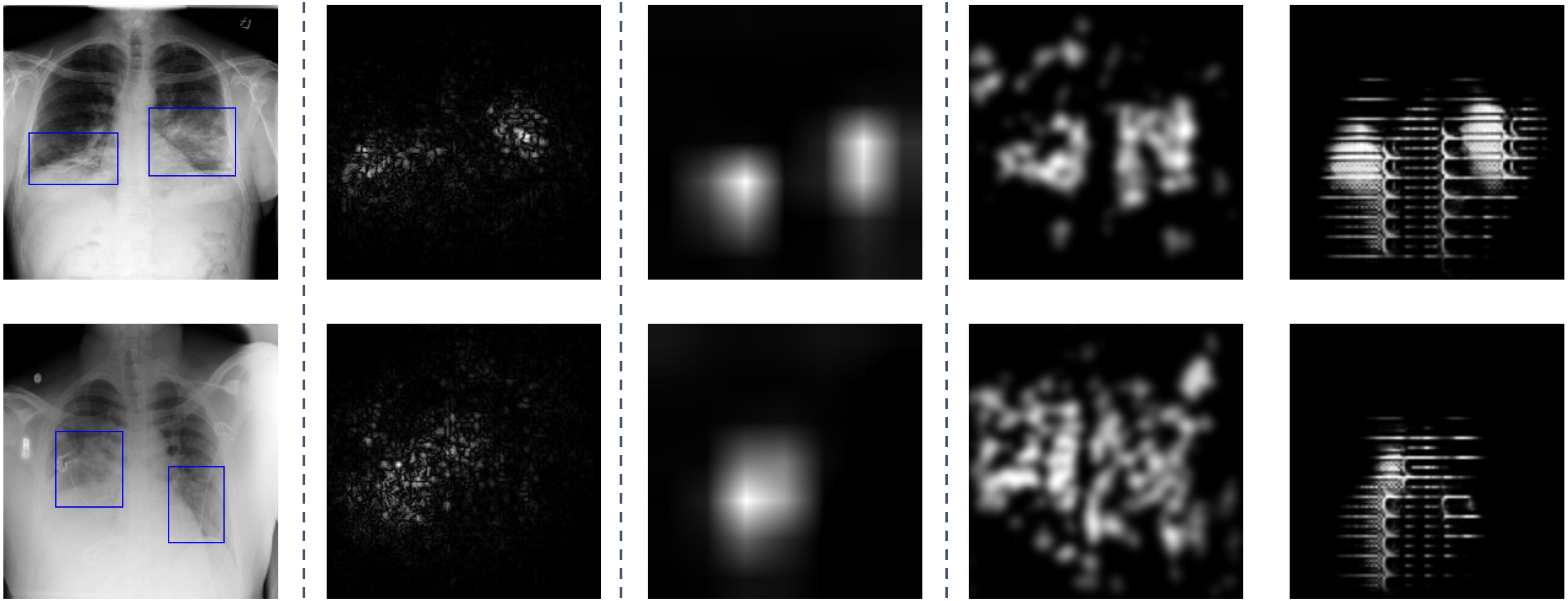


Introduction - Context



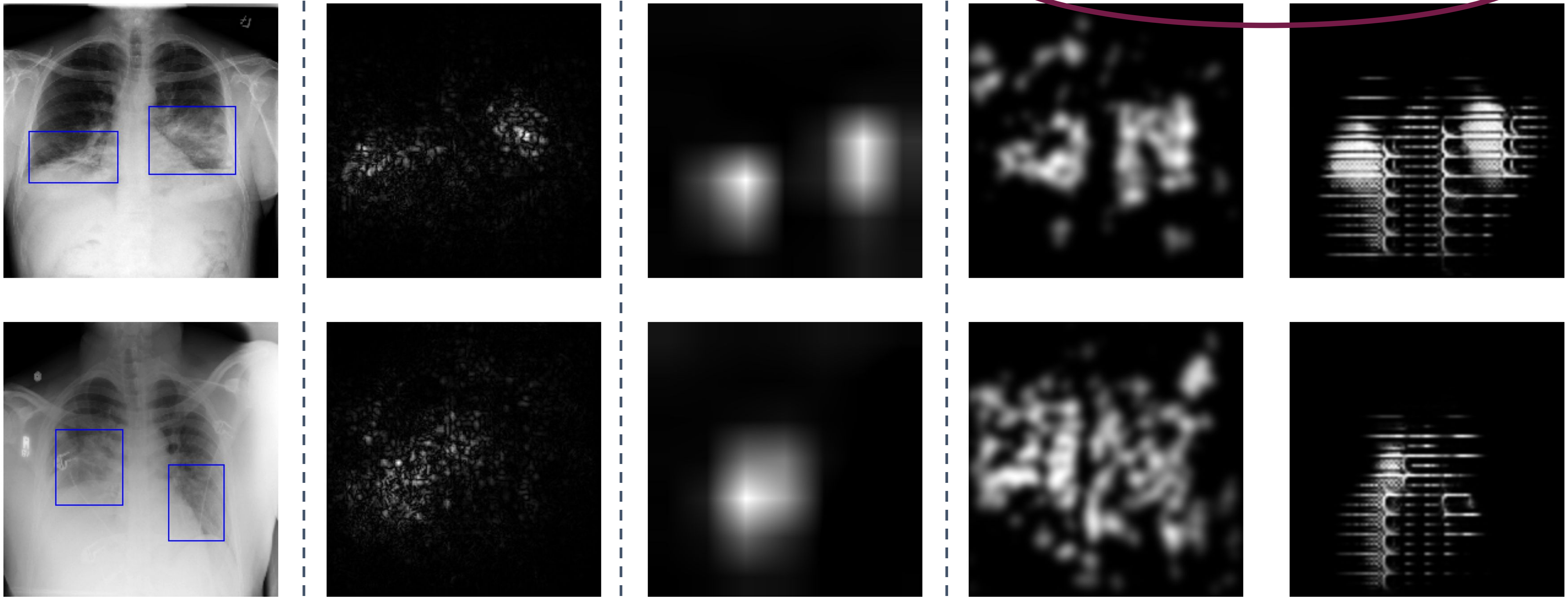
Prior Work

	Gradient [1]	CAM [2]	Perturbation based [3, 4, 5]



Prior Work

Gradient [1] | CAM [2] | Perturbation based [3, 4, 5]



Prior Work: Perturbation-based

Explanation method	Generation	Optimization	Regularization	$x_p \in D$	Indep. p	Real-time Situation
BBMP [3]	Perturbation Mask	Unique x	+++	✗	✗	$\sim$
Mask Generator [4]	Perturbation Mask	Database D	++	✗	✗	✓
Perturbation-ball [5]	Adversarial Image	Unique x	+++	✓	✓	$\sim$

Prior Work: Perturbation-based

Explanation method	Generation	Optimization	Regularization	$x_p \in D$	Indep. p	Real-time Situation
BBMP [3]	Perturbation Mask	Unique x	+++	✗	✗	$\sim$
Mask Generator [4]	Perturbation Mask	Database D	++	✗	✗	✓
Perturbation-ball [5]	Adversarial Image	Unique x	+++	✓	✓	$\sim$

Heuristic regularization

Ad-hoc
Perturbation

Computation cost

université
PARIS-SACLAY

CentraleSupélec

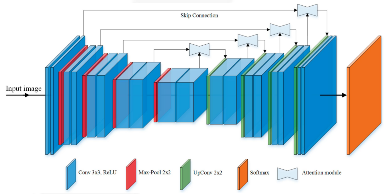
INCEPTO

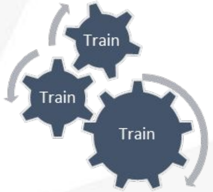
“Naive” Approach

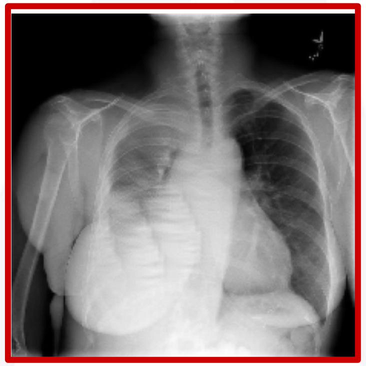
x

Original image

g_a







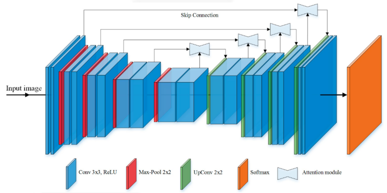
$g_a(x)$

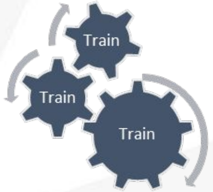
“Naive” Approach

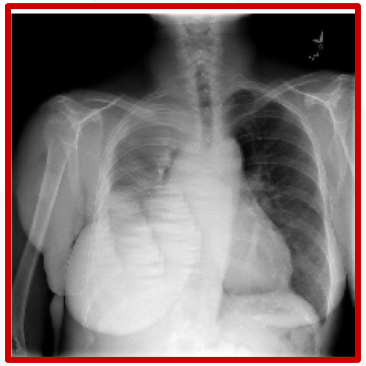
x

Original image

g_a



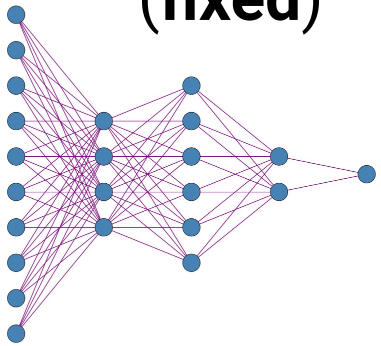




$g_a(x)$

Adversarial example

Classifier f_c
(fixed)



$f_c(g_a(x)) \neq f_c(x)$

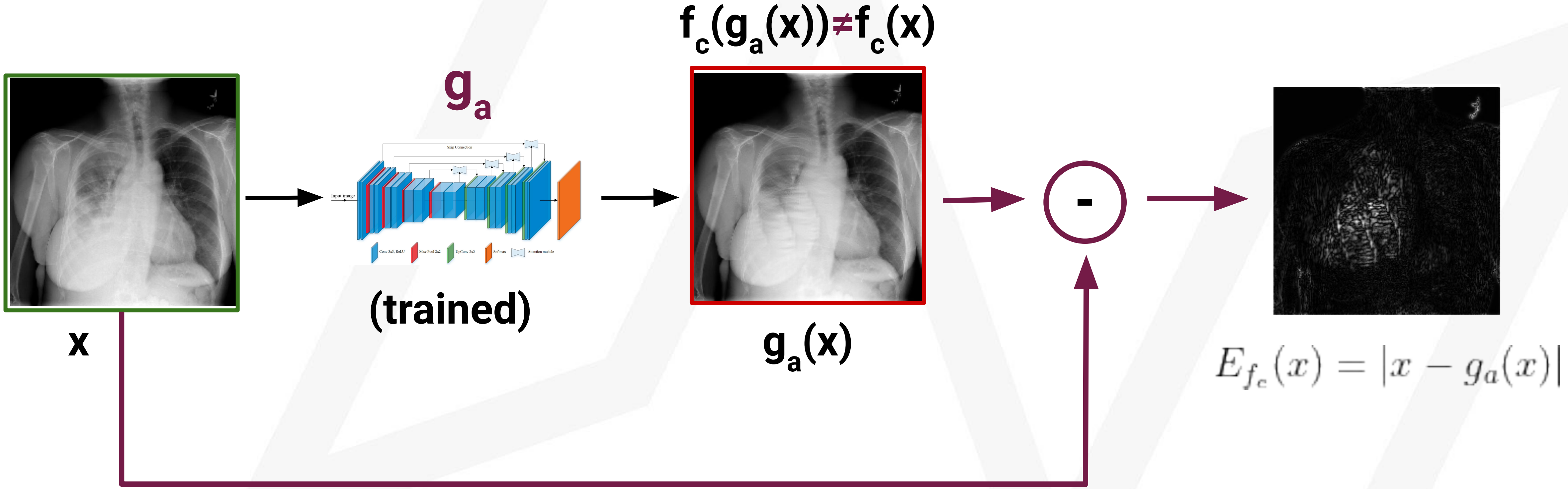
≠ Prediction

université
PARIS-SACLAY

CentraleSupélec

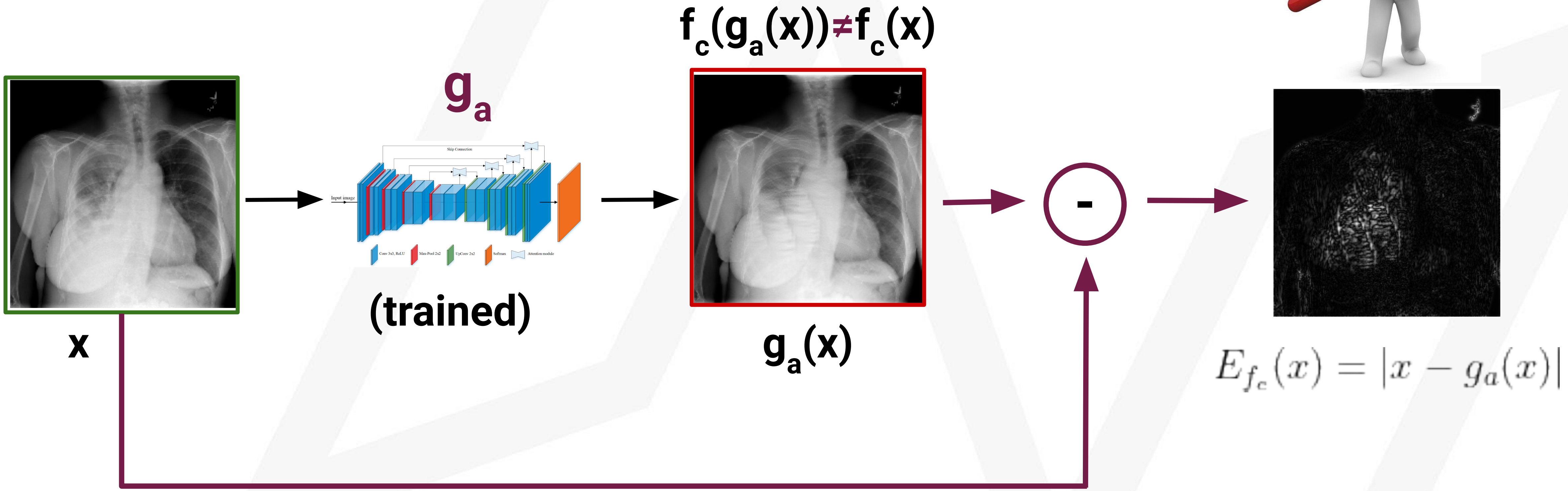
INCEPTO

“Naive” Approach



université PARIS-SACLAY
CentraleSupélec
INCEPTO

“Naive” Approach



~~Heuristic regularization~~

~~Ad-hoc Perturbation~~

Computation cost

université PARIS-SACLAY

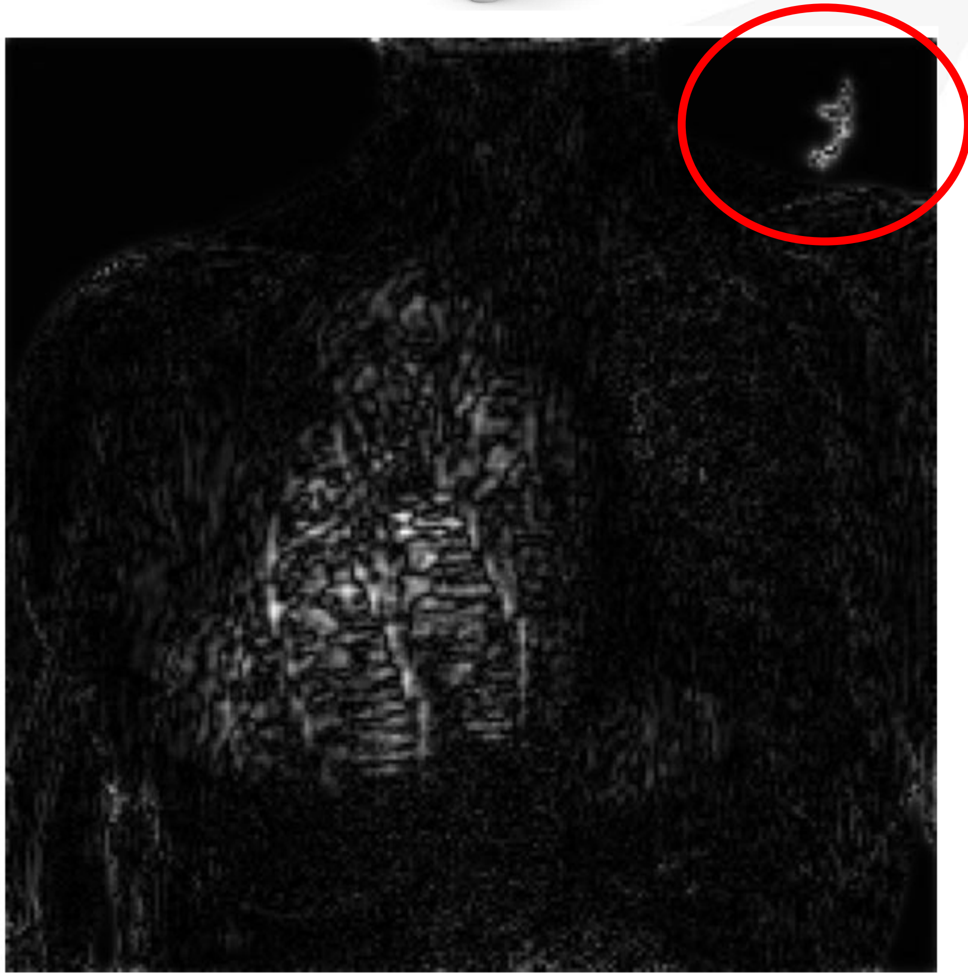
CentraleSupélec

INCEPTO

“Naive” Approach

Issues:

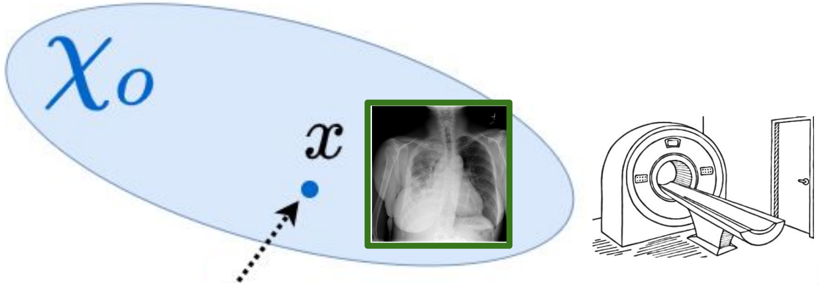
→ **Non discriminative** differences in **$|x - g_a(x)|$**



“Naive” Approach

Issues:

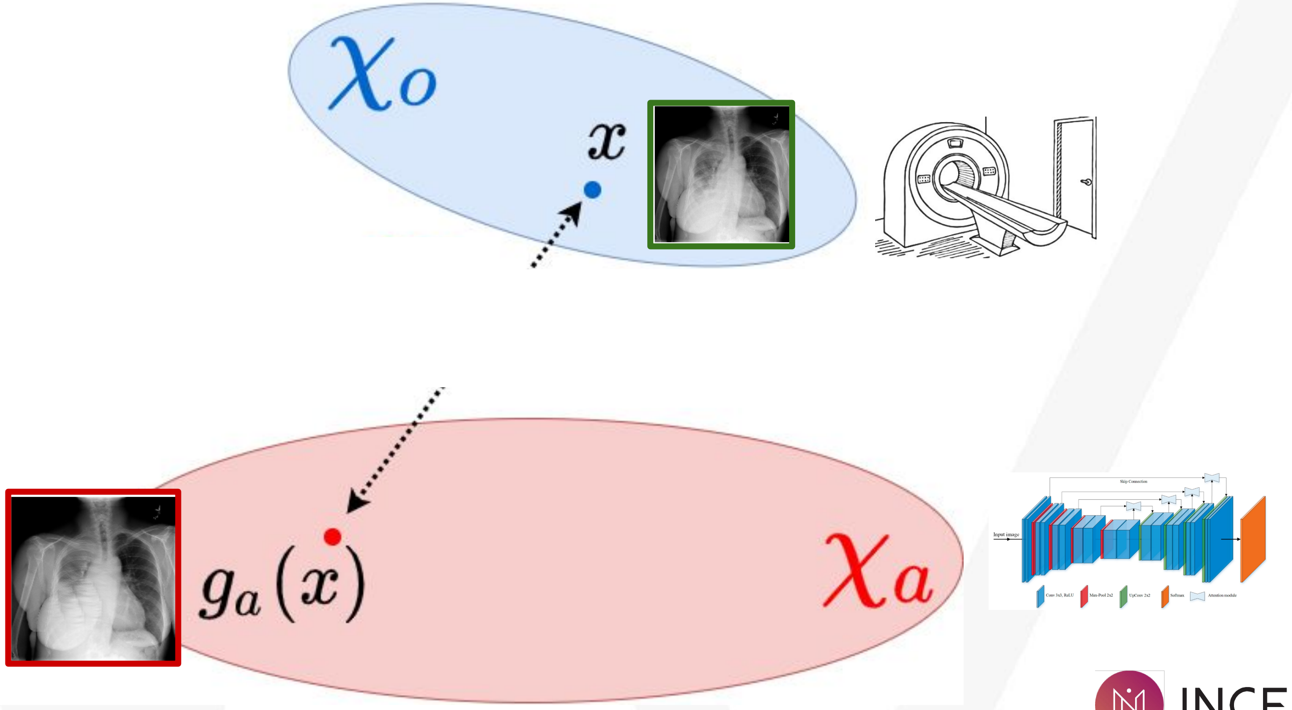
→ **Non discriminative** differences in $|\mathbf{x} - \mathbf{g_a}(\mathbf{x})|$
→ medical device space χ_o



“Naive” Approach

Issues:

→ **Non discriminative** differences in $|\mathbf{x} - \mathbf{g_a}(\mathbf{x})|$
→ medical device space χ_o
→ model generation space χ_a



université PARIS-SACLAY

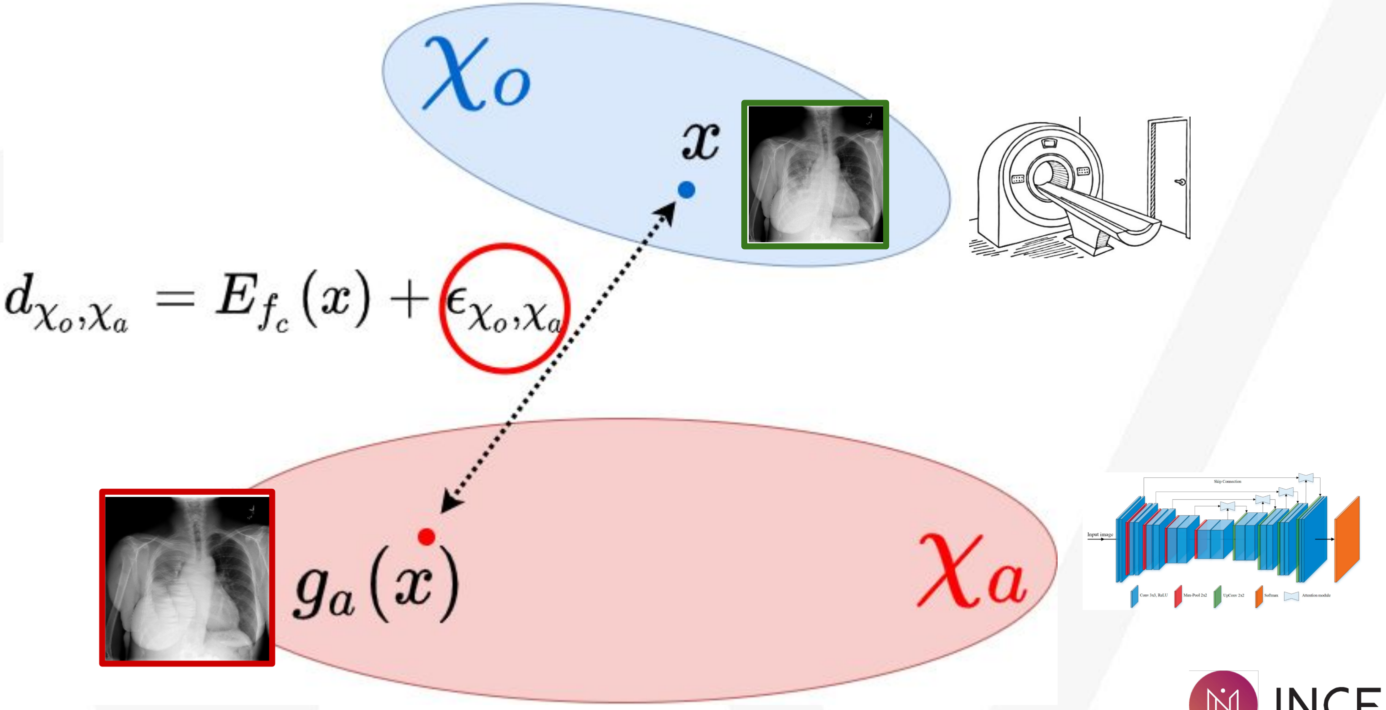
CentraleSupélec

INCEPTO

“Naive” Approach

Issues:

→ **Non discriminative** differences in |**x - g$_a$(x)**|
→ medical device space χ_o
→ model generation space χ_a

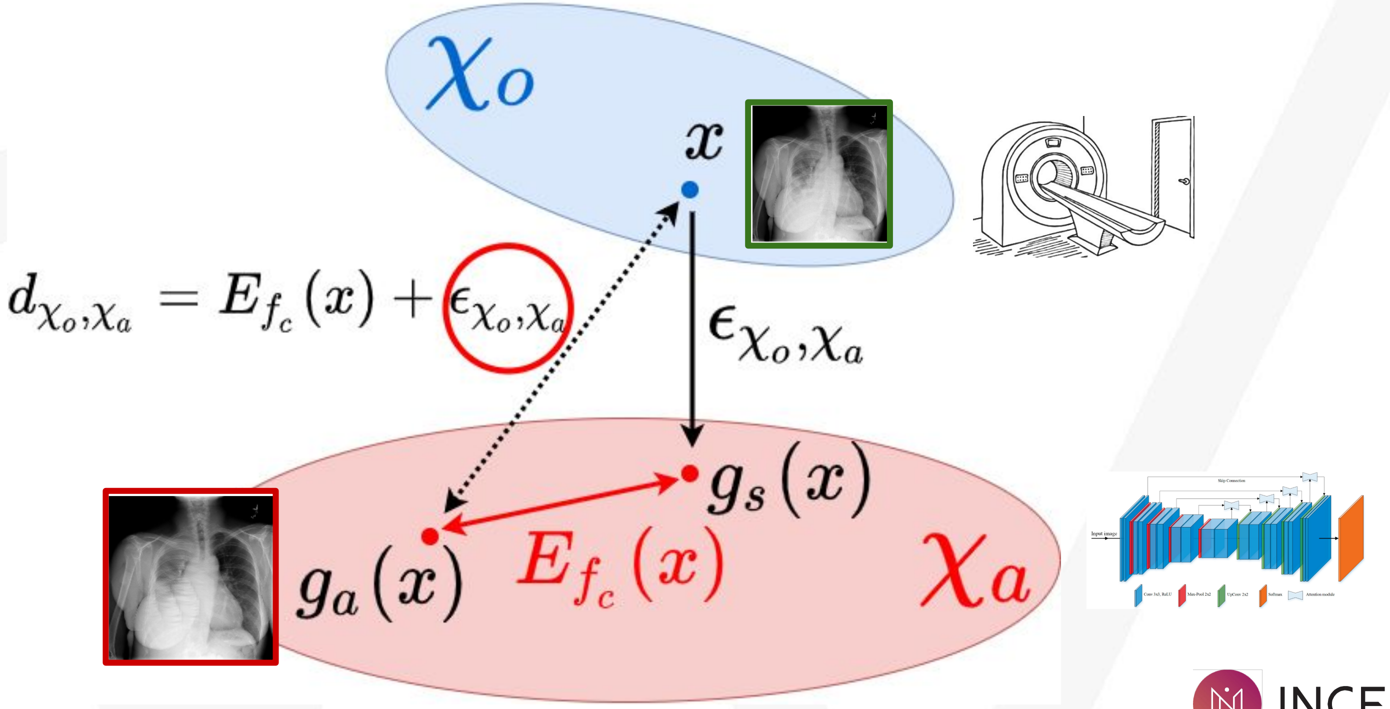


Proposed Method

Approach:

⟶ Learn to generate an adversarial example **$g_a(x) \in$** χ_a

⟶ Learn to **project** x in space $\chi_a \rightarrow$ **$g_s(x)$**



université PARIS-SACLAY

CentraleSupélec

INCEPTO

Proposed Method

Approach:

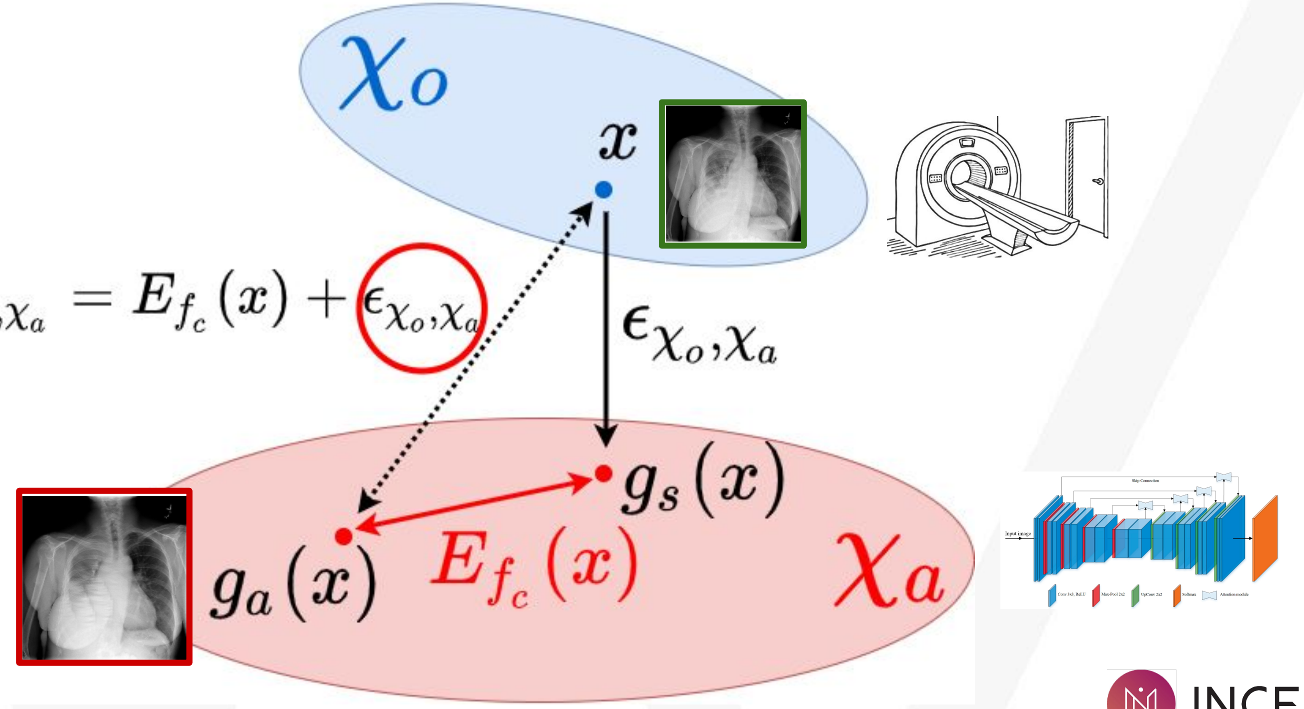
⟶ Learn to generate an adversarial example $\mathbf{g_a(x)} \in \chi_a$

⟶ Learn to **project** x in space $\chi_a \rightarrow \mathbf{g_s(x)}$

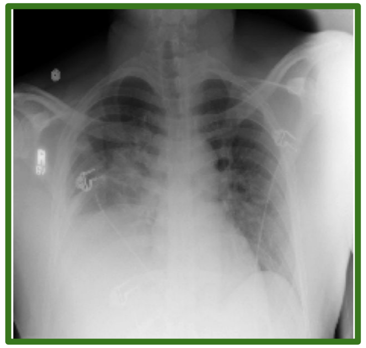
Explanation definition:

$$E_{f_c}(x) = |g_s(x) - g_a(x)|$$

$$d_{\chi_o,\chi_a} = E_{f_c}(x) + \epsilon_{\chi_o,\chi_a}$$



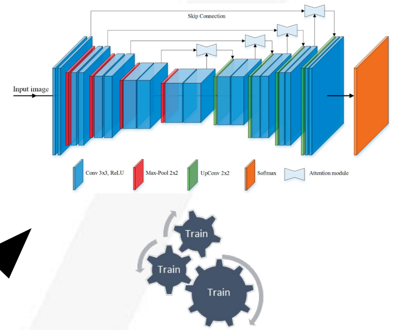
Proposed Method



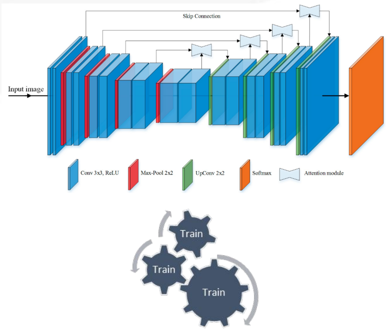
x

Original image

g_a



g_s

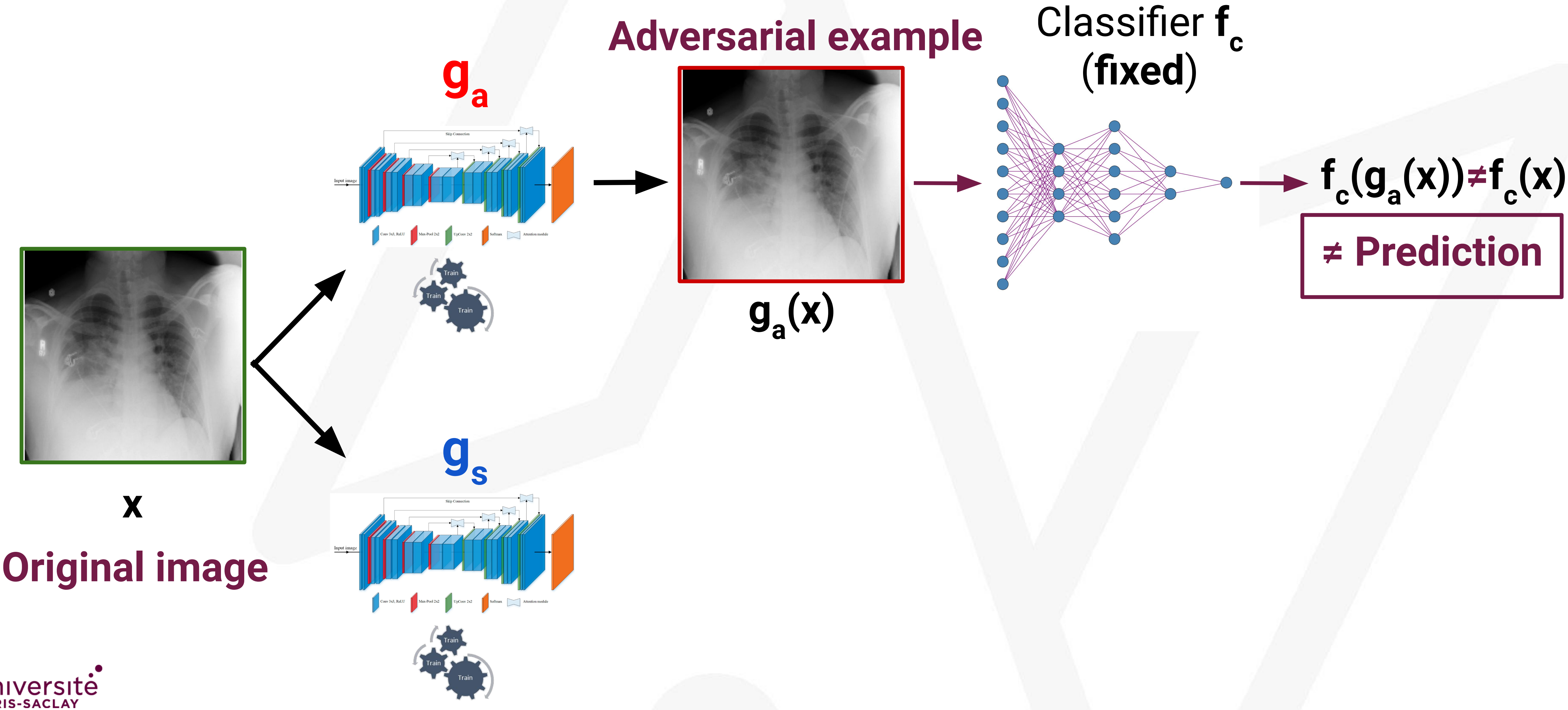


université PARIS-SACLAY

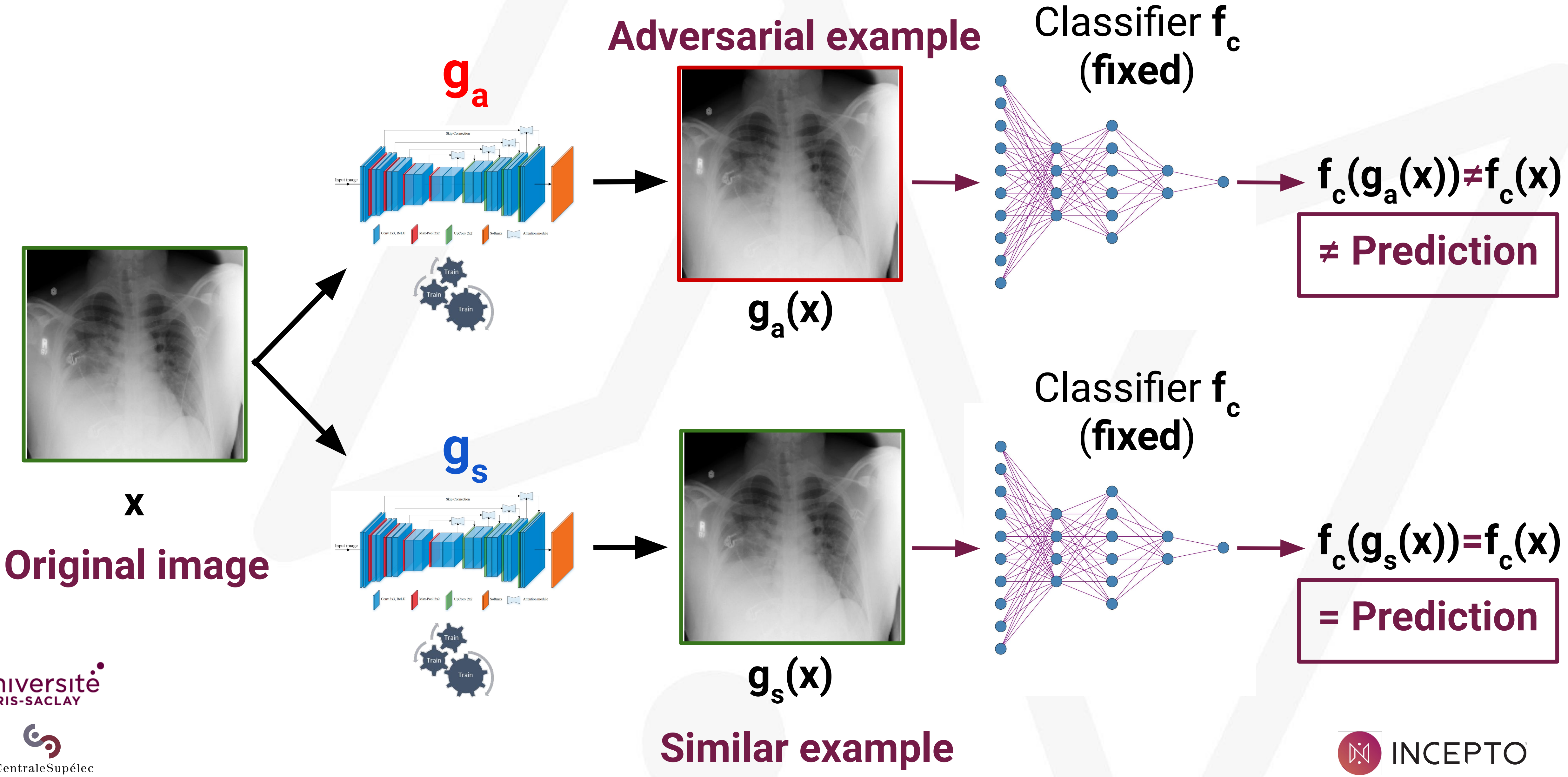
CentraleSupélec

INCEPTO

Proposed Method



Proposed Method

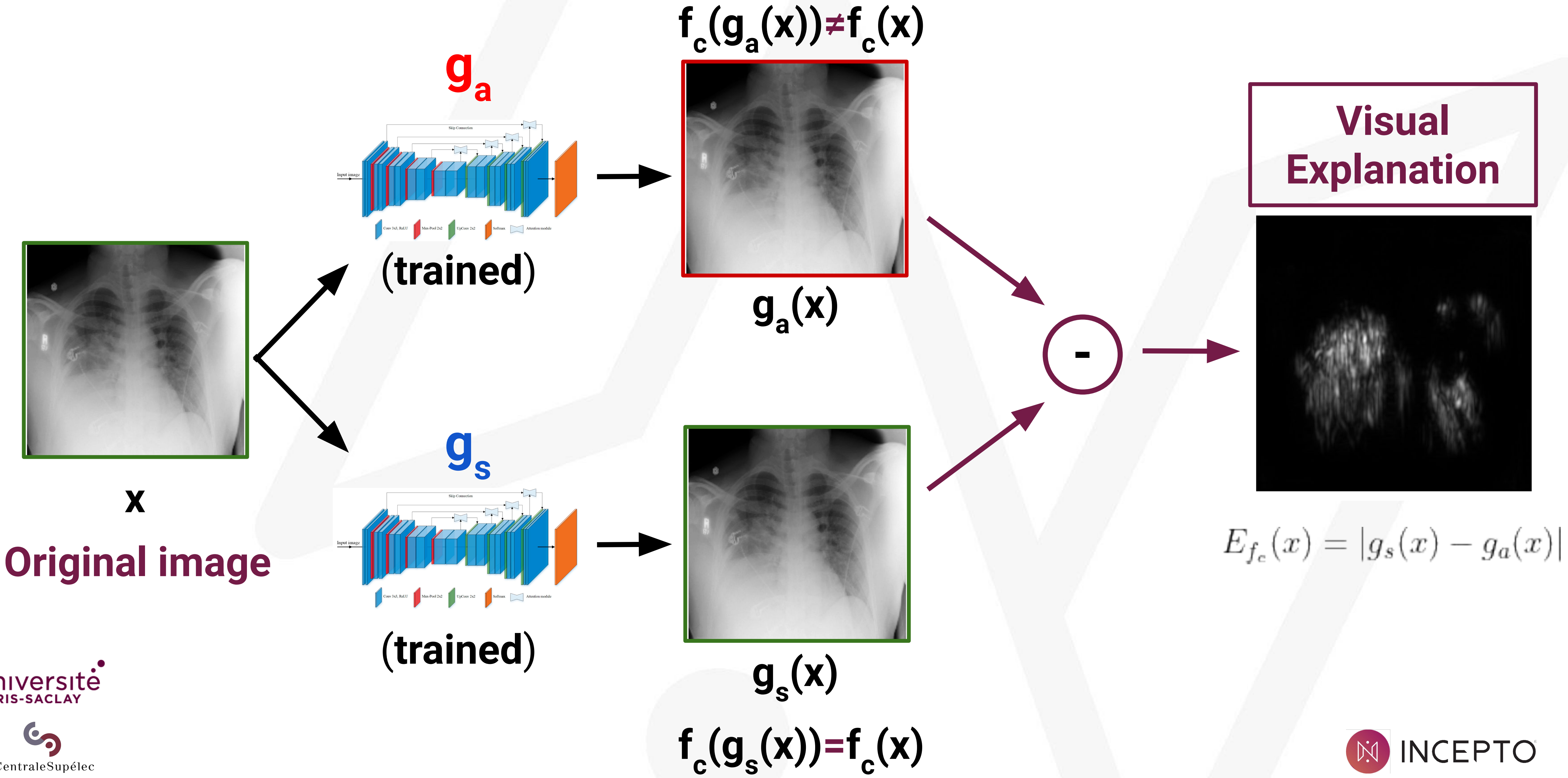


université PARIS-SACLAY

CentraleSupélec

INCEPTO

Proposed Method

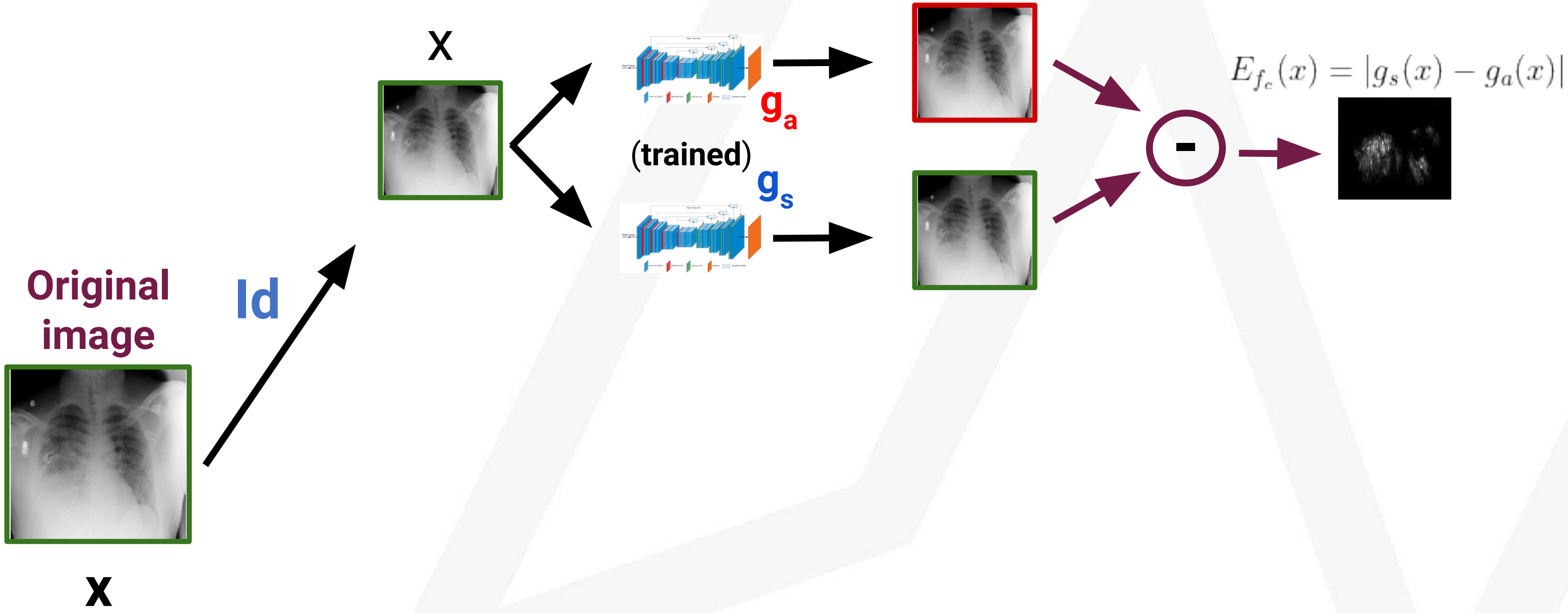


université PARIS-SACLAY

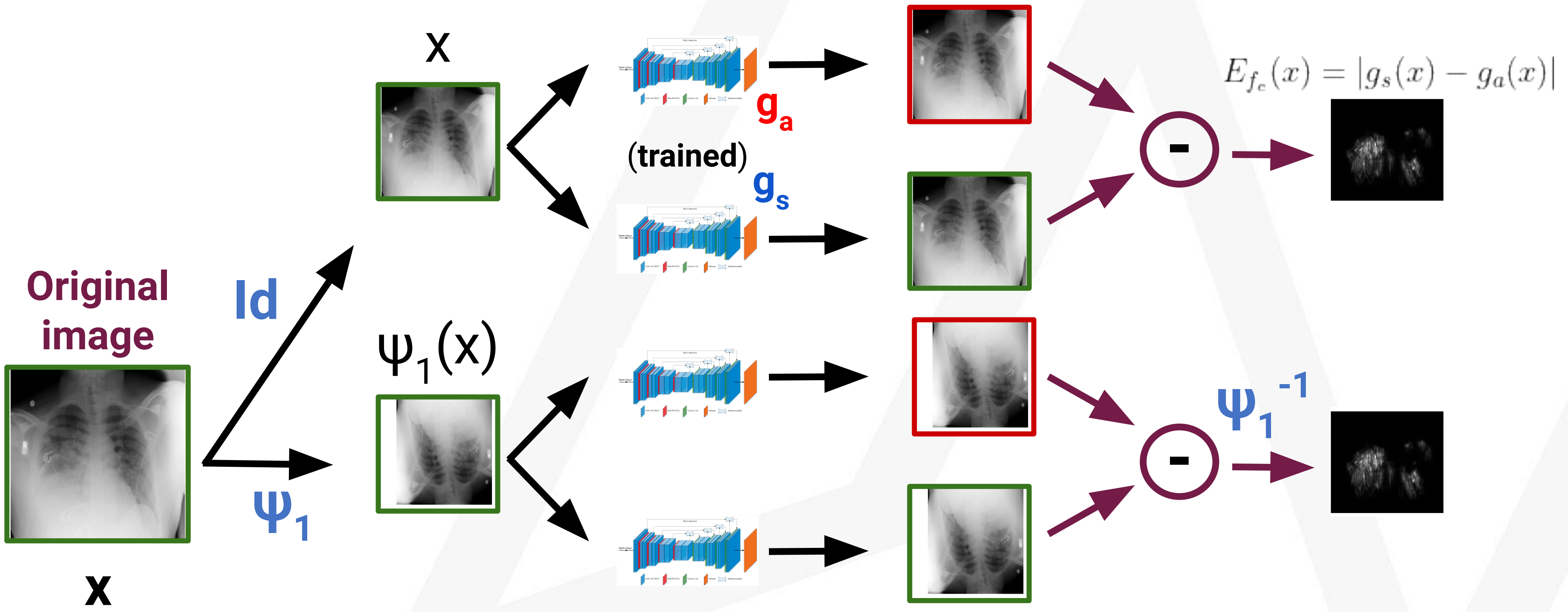
CentraleSupélec

INCEPTO

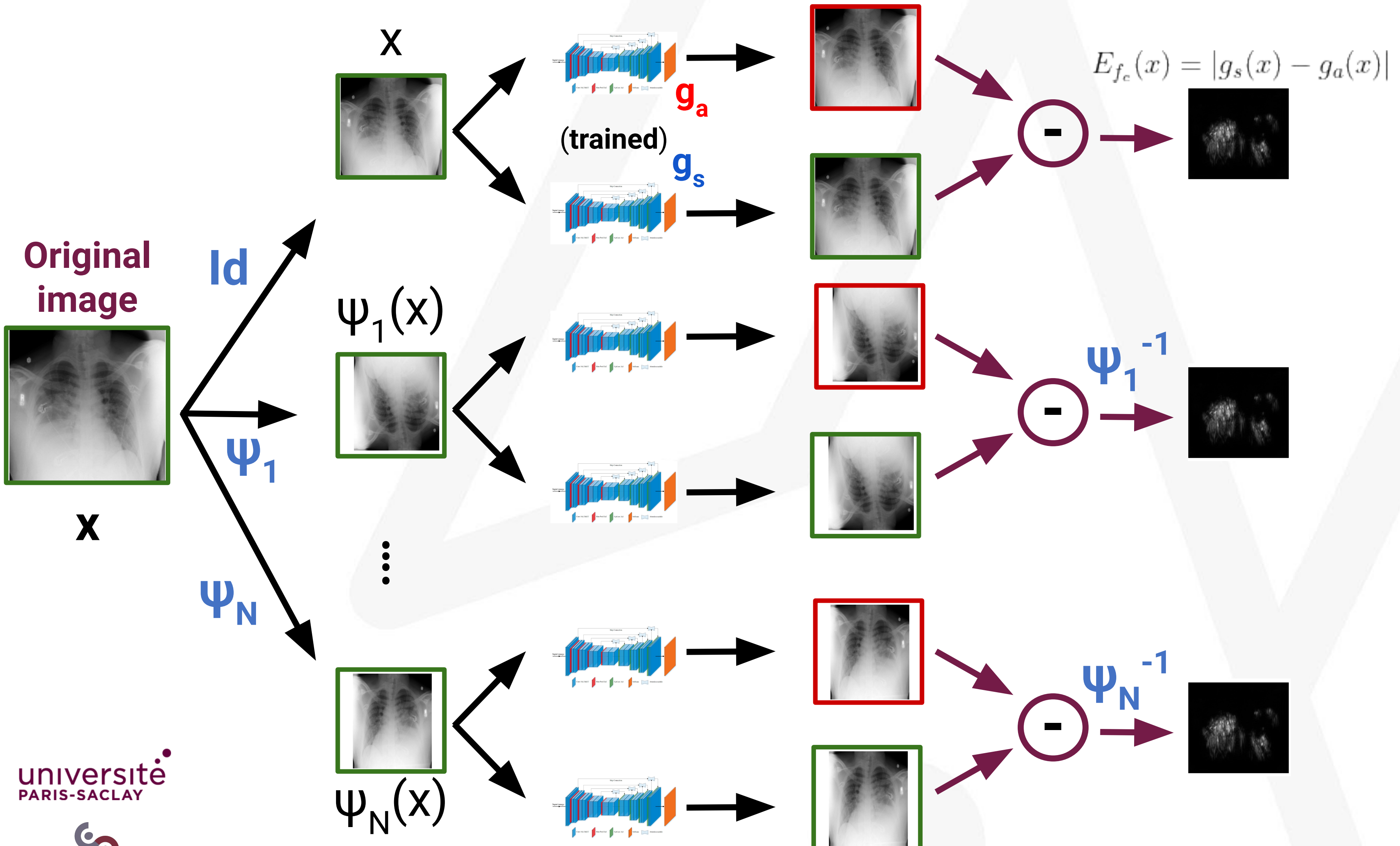
Proposed Method - Additional regularization



Proposed Method - Additional regularization

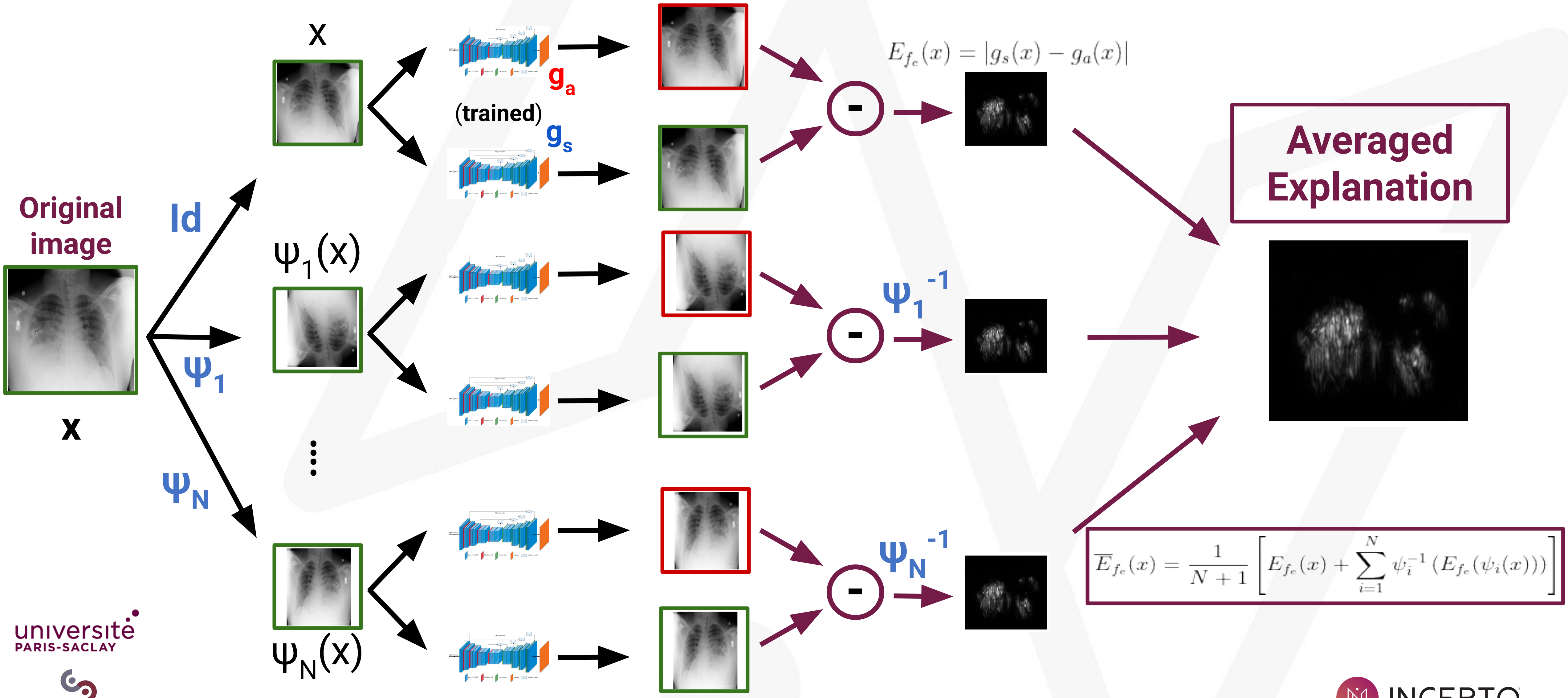


Proposed Method - Additional regularization



université PARIS-SACLAY
CentraleSupélec
INCEPTO

Proposed Method - Additional regularization



université PARIS-SACLAY

CentraleSupélec

INCEPTO

Experimental Results

Weak Localization

$$IoU_i = \frac{M_{GT} \cap M_{Ei}}{M_{GT} \cup M_{Ei}}$$

$$AUC_{Loc} = \sum_i P_i(R_i - R_{i-1})$$

IOU SCORES AT DIFFERENT THRESHOLDS OF BINARIZATION - COMPARISON TO STATE OF THE ART METHODS WITHOUT (TOP) AND WITH (BOTTOM) AUGMENTATIONS

Explanation method	IOU				
Percentile	80	85	90	95	98
Gradient [1]	0.203	0.199	0.187	0.152	0.097
GradCAM [2]	0.237	0.225	0.195	0.138	0.070
BBMP [3]	0.233	0.226	0.204	0.154	0.087
Mask Generator [4]	0.222	0.219	0.208	0.169	0.103
"Naive"	0.177	0.173	0.158	0.118	0.064
Ours	0.248 **0.292**	0.250 **0.292**	0.232 **0.272**	0.173 ***0.206***	0.097 ***0.115***

ESTIMATED AUC SCORES FOR PRECISION-RECALL AND COMPUTATION TIME - COMPARISON TO STATE OF THE ART METHODS WITHOUT (TOP) AND WITH (BOTTOM) AUGMENTATIONS

Explanation method	Total AUC	Partial AUC	Time (s)
Gradient [1]	0.287	0.189	2.04
GradCAM [2]	0.326	0.235	0.78
BBMP [3]	0.326	0.229	17.14
Mask Generator [4]	0.327	0.226	0.09
"Naive"	0.238	0.145	0.10
Ours	0.339 **0.412**	0.256 **0.328**	0.05 0.63

université PARIS-SACLAY

CentraleSupélec

INCEPTO

Experimental Results

Weak Localization

$$IoU_i = \frac{M_{GT} \cap M_{Ei}}{M_{GT} \cup M_{Ei}}$$

$$AUC_{Loc} = \sum_i P_i(R_i - R_{i-1})$$

IOU SCORES AT DIFFERENT THRESHOLDS OF BINARIZATION - COMPARISON TO STATE OF THE ART METHODS WITHOUT (TOP) AND WITH (BOTTOM) AUGMENTATIONS

Explanation method *Percentile*	**IOU** 80	85	90	95	98
Gradient [1]	0.203	0.199	0.187	0.152	0.097
GradCAM [2]	0.237	0.225	0.195	0.138	0.070
BBMP [3]	0.233	0.226	0.204	0.154	0.087
Mask Generator [4]	0.222	0.219	0.208	0.169	0.103
"Naive"	0.177	0.173	0.158	0.118	0.064
Ours	0.248	0.250	0.232	0.173	0.097
	0.292	0.292	0.272	*0.206*	*0.115*

ESTIMATED AUC SCORES FOR PRECISION-RECALL AND COMPUTATION TIME - COMPARISON TO STATE OF THE ART METHODS WITHOUT (TOP) AND WITH (BOTTOM) AUGMENTATIONS

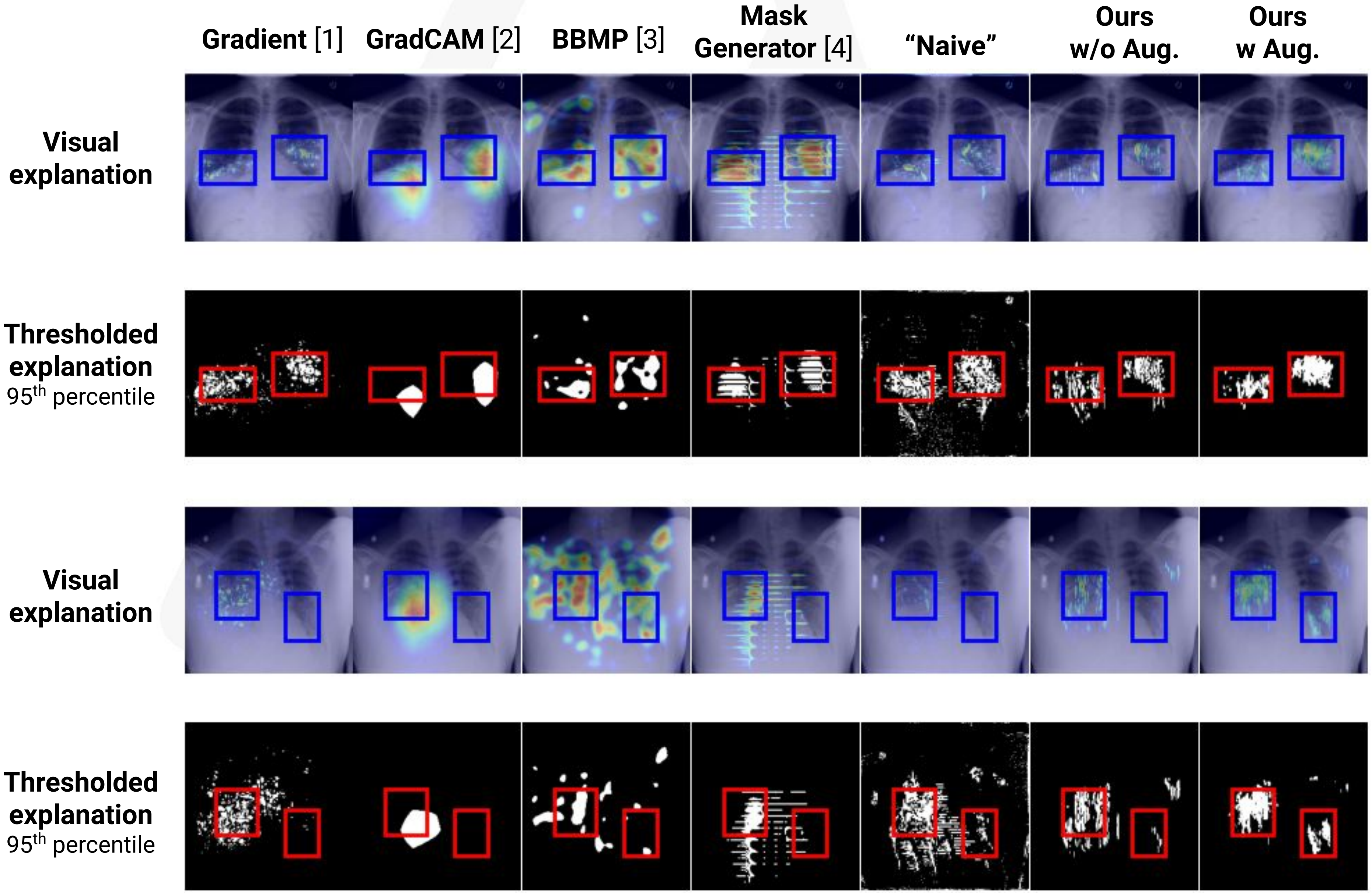
Explanation method	**Total AUC**	**Partial AUC**	**Time (s)**
Gradient [1]	0.287	0.189	2.04
GradCAM [2]	0.326	0.235	0.78
BBMP [3]	0.326	0.229	17.14
Mask Generator [4]	0.327	0.226	0.09
"Naive"	0.238	0.145	0.10
Ours	0.339	0.256	0.05
	0.412	0.328	0.63

université PARIS-SACLAY

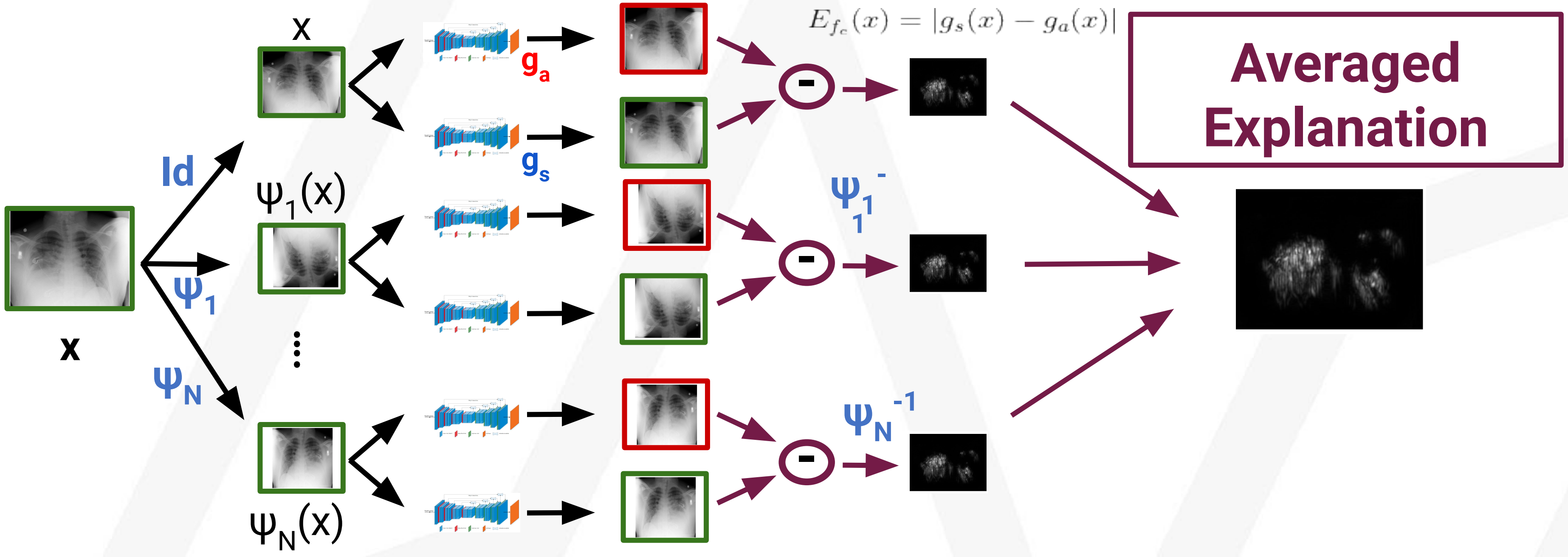
CentraleSupélec

INCEPTO

Experimental Results



Summary of Contribution



$$\overline{E}_{f_c}(x) = \frac{1}{N+1}\left[E_{f_c}(x) + \sum_{i=1}^{N} \psi_i^{-1}\left(E_{f_c}(\psi_i(x))\right)\right]$$

~~Heuristic regularization~~

~~Ad-hoc Perturbation~~

Computation cost

université PARIS-SACLAY

CentraleSupélec

INCEPTO

References

[1] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,” in ICLR, 2014
[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in ICCV, 2017
[3] R. C. Fong and A. Vedaldi, “Interpretable explanations of black boxes by meaningful perturbation,” in ICCV, 2017
[4] P. Dabkowski and Y. Gal, “Real time image saliency for black box classifiers,” in NIPS, 2017
[5] . Elliott, S. Law, and C. Russell, “Adversarial perturbations on the perceptual ball,” ArXiv, 2019

Thank you for your attention

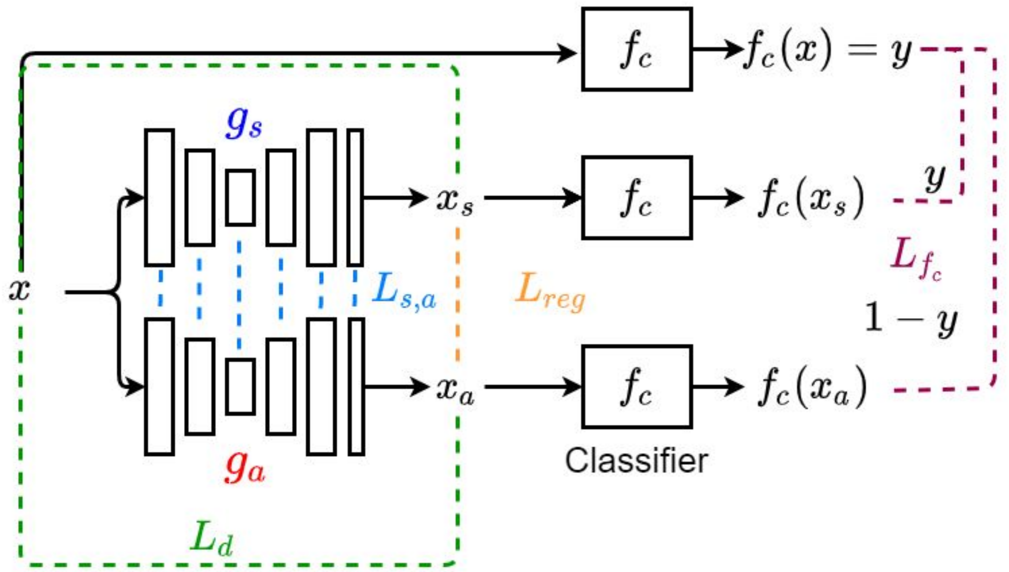
Any Question ?

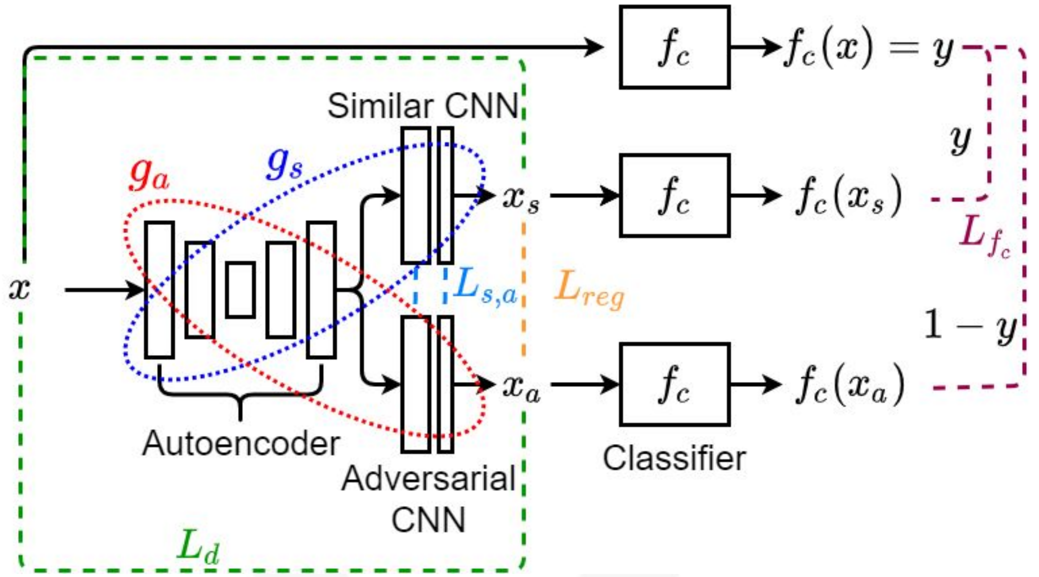
INCEPTO

Appendices

Joint Optimization

$$(\bar{g}_s, \bar{g}_a) = \underset{g_s, g_a}{\operatorname{argmin}} \left\{ \begin{array}{l} \mathbb{E}_x \left(\begin{array}{ll} L_d(x, g_s(x), g_a(x)) & + \\ L_{f_c}(x, g_s(x), g_a(x)) & + \\ L_{reg}(x, g_s(x), g_a(x)) & \end{array} \right) \\ + \quad L_{s,a}(g_s, g_a) \end{array} \right\}$$

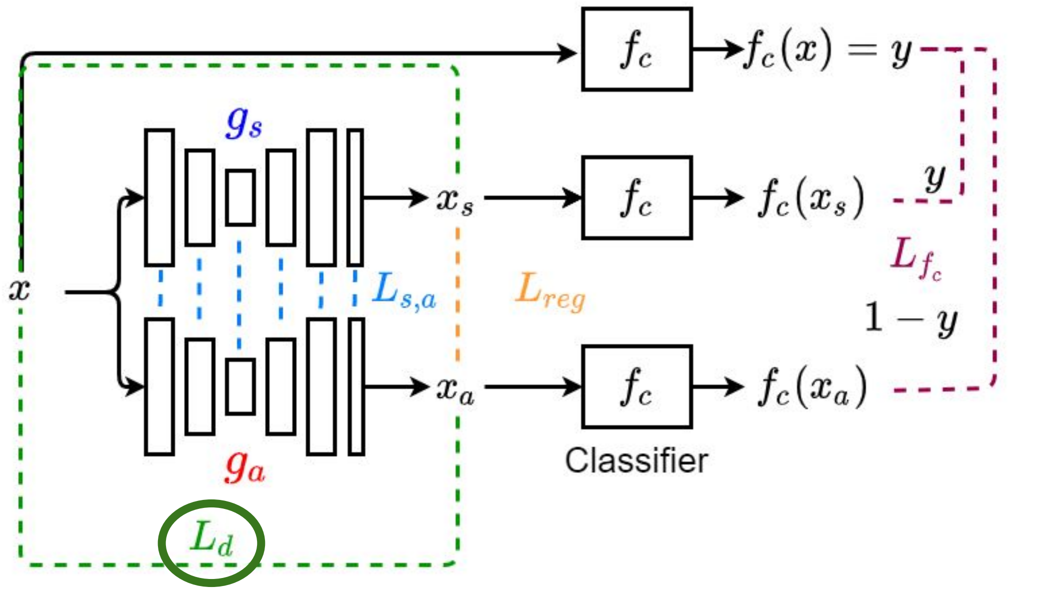


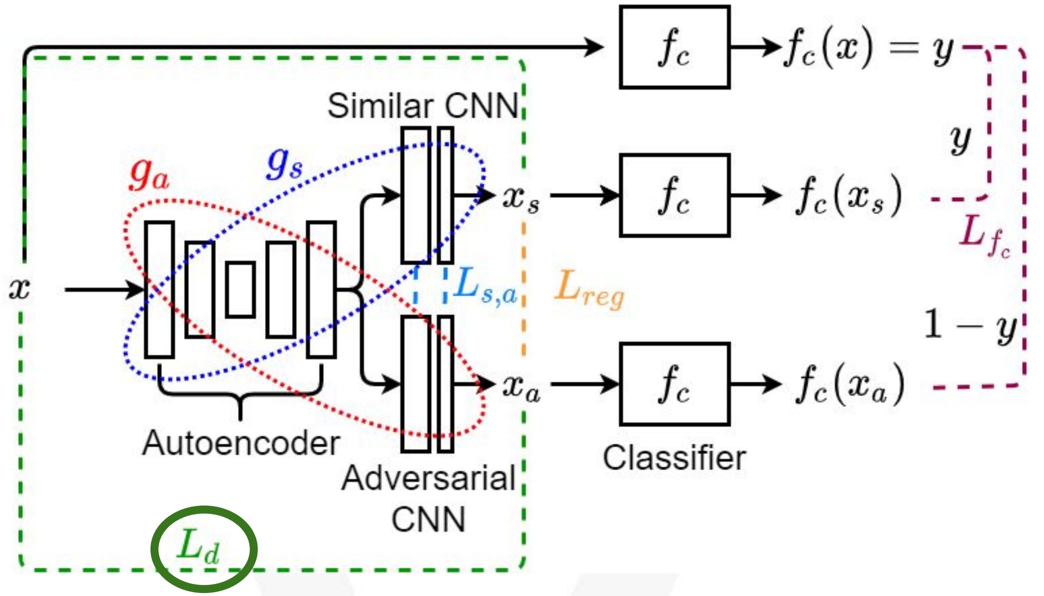


Joint Optimization

x, $g_s(x)$ and $g_a(x)$ should be similar

$$(\bar{g}_s, \bar{g}_a) = \operatorname*{argmin}_{g_s, g_a} \left\{ \begin{array}{l} \mathbb{E}_x \left(\begin{array}{ll} \boxed{L_d(x, g_s(x), g_a(x))} & + \\ L_{f_c}(x, g_s(x), g_a(x)) & + \\ L_{reg}(x, g_s(x), g_a(x)) & \end{array} \right) \\ \\ + \quad L_{s,a}(g_s, g_a) \end{array} \right\}$$

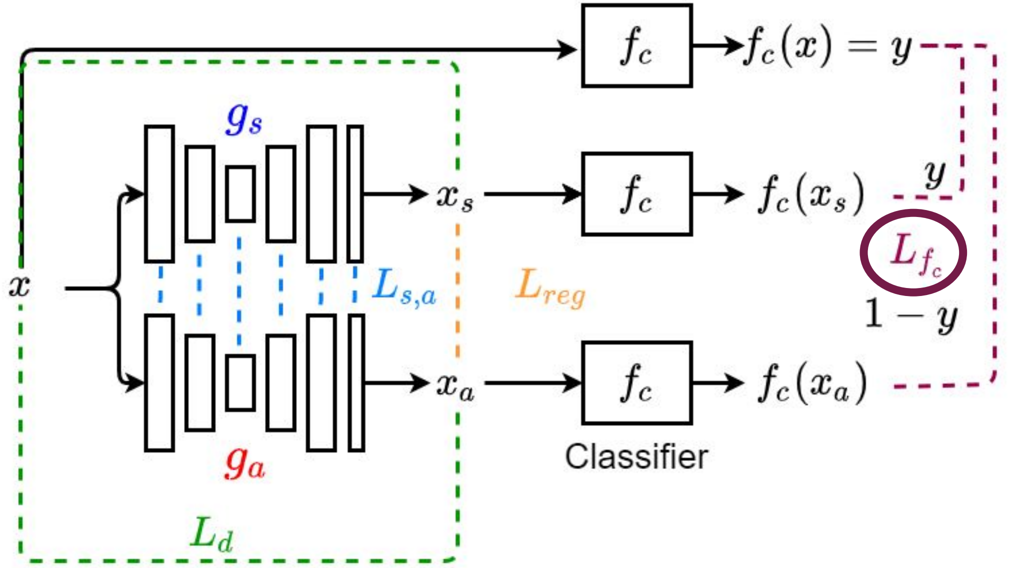


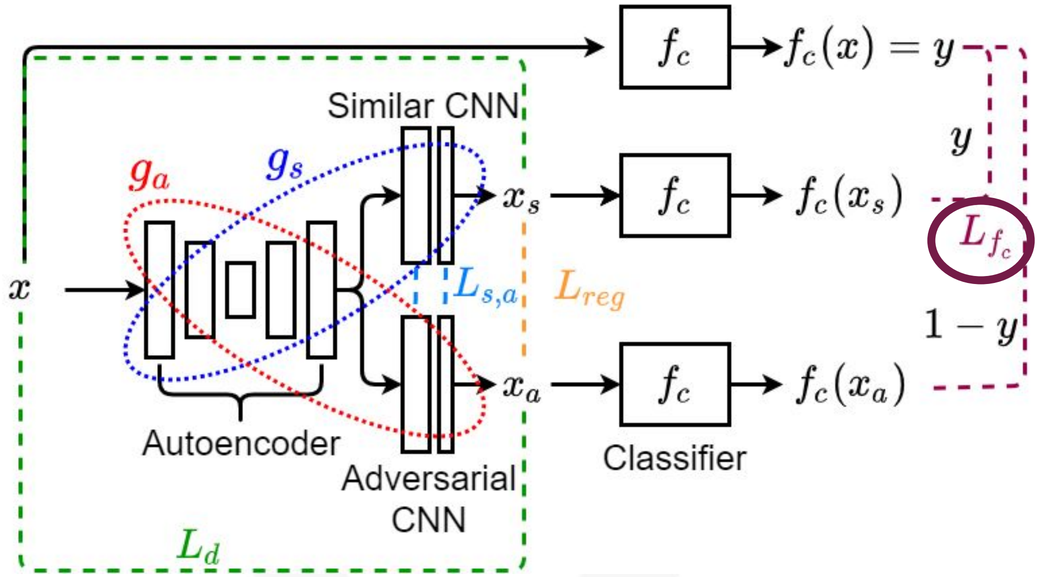


Joint Optimization

$f_c(g_s(x)) = f_c(x)$
$f_c(g_s(x)) \neq f_c(x)$

$$(\bar{g}_s, \bar{g}_a) = \underset{g_s, g_a}{\operatorname{argmin}} \left\{ \begin{array}{l} \mathbb{E}_x \left(\begin{array}{l} L_d(x, g_s(x), g_a(x)) + \\ \boxed{L_{f_c}(x, g_s(x), g_a(x))} + \\ L_{reg}(x, g_s(x), g_a(x)) \end{array} \right) \\ + \quad L_{s,a}(g_s, g_a) \end{array} \right\}$$



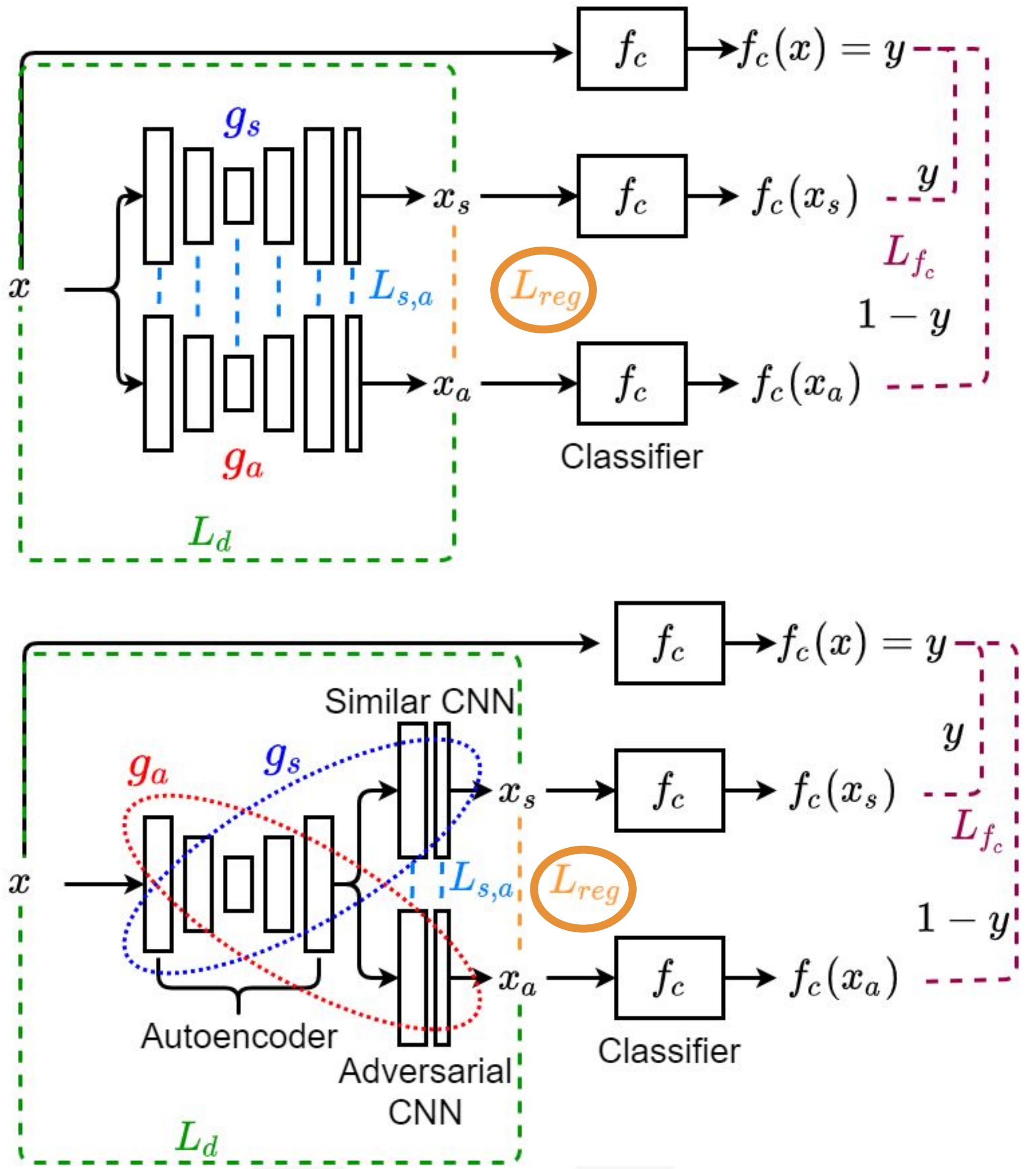


université PARIS-SACLAY
CentraleSupélec
INCEPTO

Joint Optimization

$$(\bar{g}_s, \bar{g}_a) = \underset{g_s, g_a}{\operatorname{argmin}} \left\{ \begin{array}{l} \mathbb{E}_x \left(\begin{array}{ll} L_d(x, g_s(x), g_a(x)) & + \\ L_{f_c}(x, g_s(x), g_a(x)) & + \\ \boxed{L_{reg}(x, g_s(x), g_a(x))} & \end{array} \right) \\ + \quad L_{s,a}(g_s, g_a) \end{array} \right\}$$

$g_s(x)$ close to $g_a(x)$
Smooth differences



universite PARIS-SACLAY

CentraleSupélec

INCEPTO

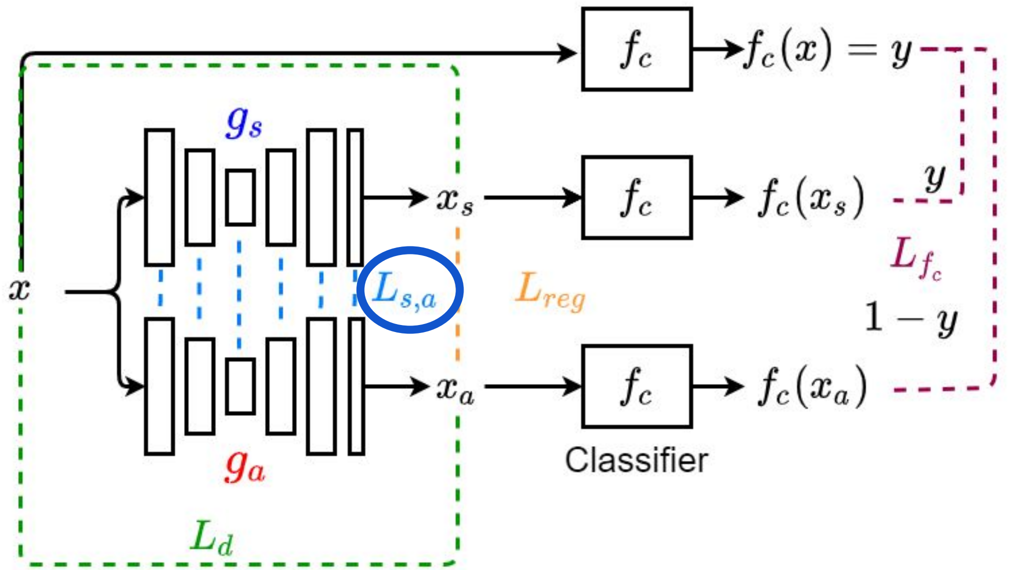
Joint Optimization

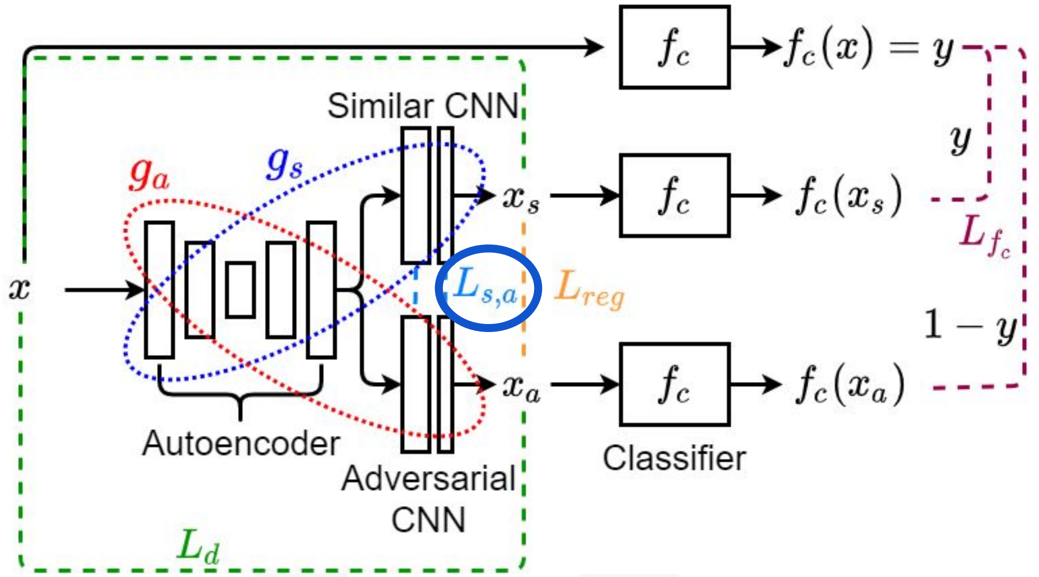
$$(\bar{g}_s, \bar{g}_a) = \underset{g_s, g_a}{\operatorname{argmin}} \left\{ \begin{array}{l} \mathbb{E}_x \left(\begin{array}{ll} L_d(x, g_s(x), g_a(x)) & + \\ L_{f_c}(x, g_s(x), g_a(x)) & + \\ L_{reg}(x, g_s(x), g_a(x)) & \end{array} \right) \\ + \boxed{L_{s,a}(g_s, g_a)} \end{array} \right\}$$

g_s and g_a parameters close

$x \longrightarrow g_s(x) \in \chi_s$

$x \longrightarrow g_a(x) \in \chi_a$ $\longrightarrow \chi_s \sim \chi_a$





université PARIS-SACLAY

CentraleSupélec

INCEPTO

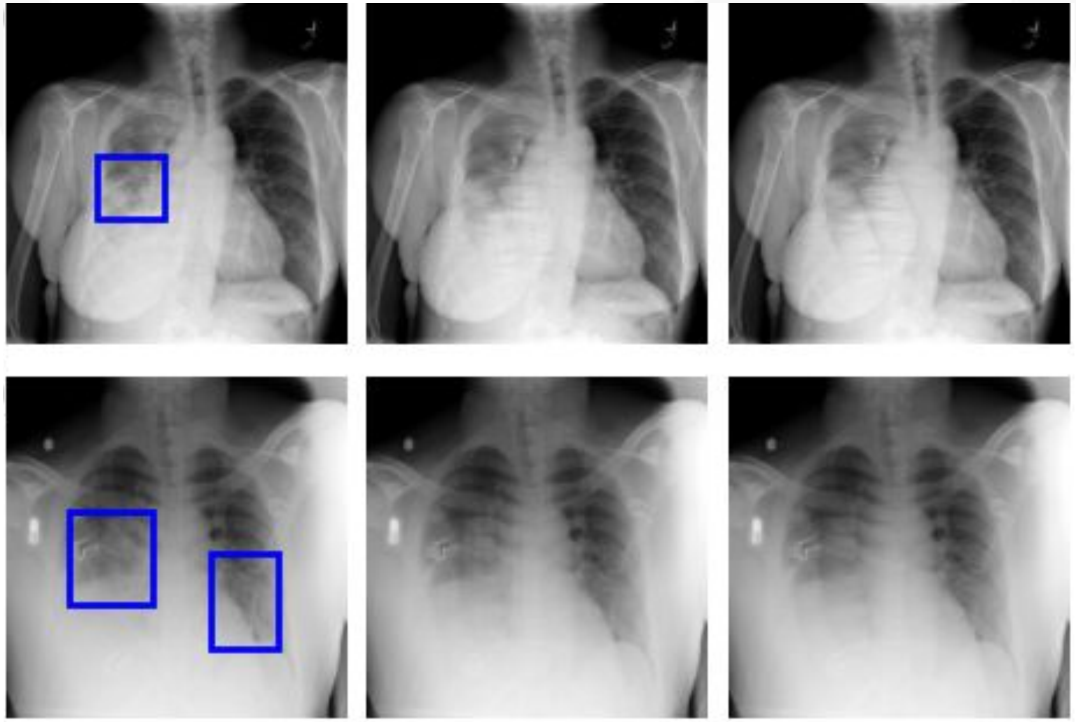
Experimental Results

Adversarial and Similar Generation

SUMMARY: SIMILAR AND ADVERSARIAL GENERATION

Explanation method	L_{reg}	$L_{s,a}$	AUC_{os}	$AUC_{\bar{o}a}$	$x \leftrightarrow x_s$		$x \leftrightarrow x_a$		$x_s \leftrightarrow x_a$	
"Naive"	✓	-	-	0.939	-	-	0.994	41.92	-	-
Duo AE (TV)	✓	✗	1.0	0.905	0.996	44.07	0.987	39.47	0.994	43.89
Duo AE (W,TV)	✓	✓	1.0	0.958	0.995	41.99	0.987	39.08	0.995	44.26
Single AE (TV)	✓	✗	1.0	0.961	**0.997**	**44.57**	0.989	40.67	0.996	45.25
Single AE (W)	✗	✓	0.998	0.949	0.995	43.61	0.994	42.42	0.999	52.26
Single AE (W, TV)	✓	✓	0.998	0.952	0.995	43.88	**0.994**	**42.63**	0.999	51.93

Original image x | Similar image $g_s(x)$ | Adversarial image $g_a(x)$



Experimental Results

Weak Localization

$$IoU_i = \frac{M_{GT} \cap M_{Ei}}{M_{GT} \cup M_{Ei}}$$

$$AUC_{Loc} = \sum_i P_i(R_i - R_{i-1})$$

IOU SCORES AT DIFFERENT THRESHOLDS OF BINARIZATION - COMPARISON TO STATE OF THE ART METHODS WITHOUT (TOP) AND WITH (BOTTOM) AUGMENTATIONS

Explanation method			**IOU**		
Percentile	80	85	90	95	98
Gradient [1]	0.203	0.199	0.187	0.152	0.097
	0.256	*0.252*	*0.236*	*0.190*	*0.117*
GradCAM [2]	0.237	0.225	0.195	0.138	0.070
	0.271	*0.263*	*0.244*	*0.190*	*0.105*
BBMP [3]	0.233	0.226	0.204	0.154	0.087
Mask Generator [4]	0.222	0.219	0.208	0.169	0.103
	0.259	*0.264*	*0.259*	**0.221**	**0.137**
"Naive"	0.177	0.173	0.158	0.118	0.064
	0.239	*0.230*	*0.208*	*0.156*	*0.087*
Ours	0.248	0.250	0.232	0.173	0.097
	0.292	**0.292**	**0.272**	*0.206*	*0.115*

ESTIMATED AUC SCORES FOR PRECISION-RECALL AND COMPUTATION TIME - COMPARISON TO STATE OF THE ART METHODS WITHOUT (TOP) AND WITH (BOTTOM) AUGMENTATIONS

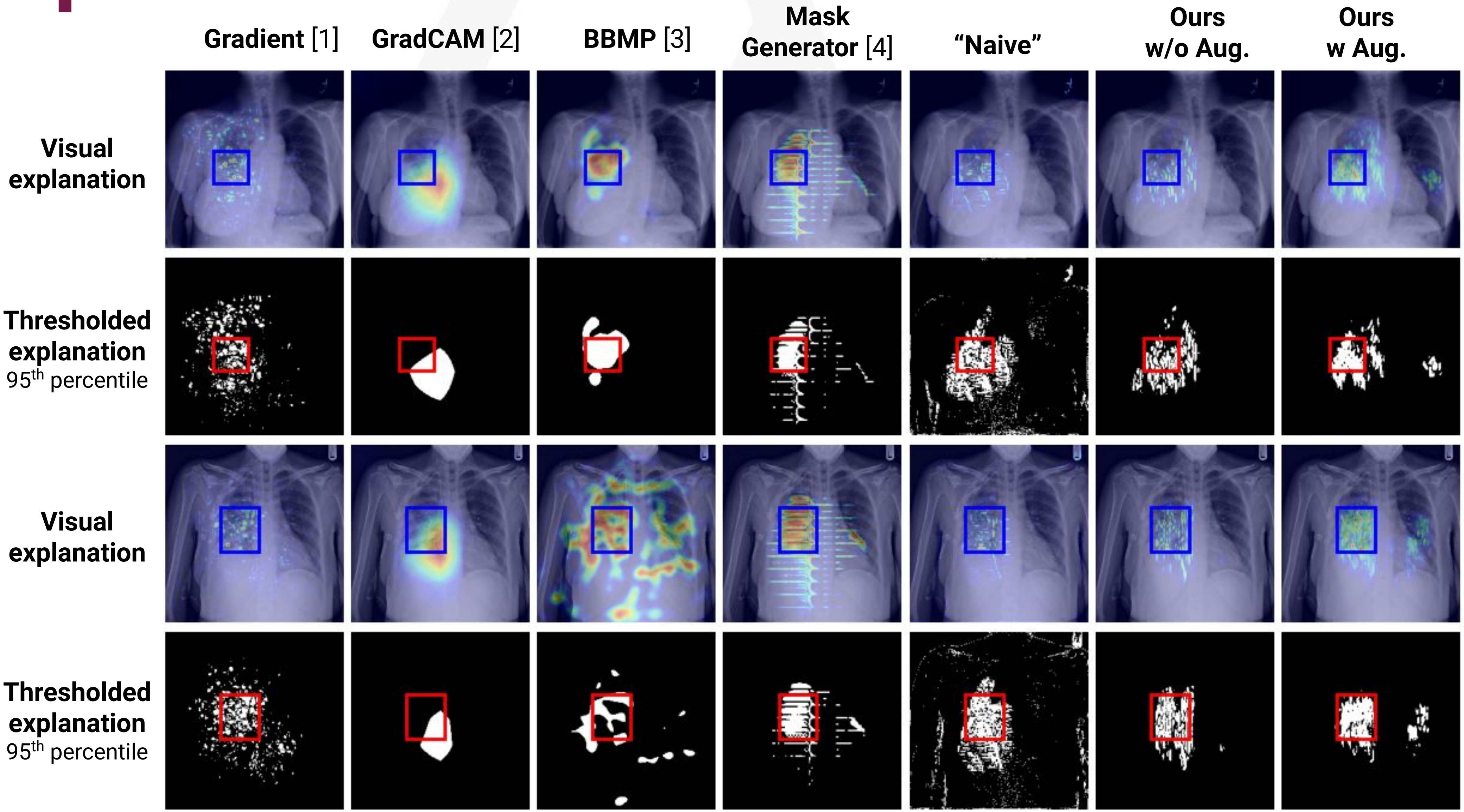
Explanation method	**Total AUC**	**Partial AUC**	**Time (s)**
Gradient [1]	0.287	0.189	2.04
	0.374	*0.274*	2.83
GradCAM [2]	0.326	0.235	0.78
	0.397	*0.302*	5.09
BBMP [3]	0.326	0.229	17.14
Mask Generator [4]	0.327	0.226	0.09
	0.404	*0.308*	0.68
"Naive"	0.238	0.145	0.10
	0.325	*0.232*	0.75
Ours	0.339	0.256	0.05
	0.412	**0.328**	0.63

université PARIS-SACLAY

CentraleSupélec

INCEPTO

Experimental Results



Université Paris-Saclay

CentraleSupélec

INCEPTO