Reconstruction d'images avec les machines à inférences récurentielles

par

Alexandre Adam

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)

Département de physique
Université de Montréal

Résumé

Abstract

Table des matières

Liste des tableaux

Liste des figures

Acronymes

RIM Recurrent Inference Machine — Machine à inférence récurrentielles.

VAE Variational AutoEncoder — Auto-encodeur variationnel de Bayes.

GRU Gated Recurrent Unit— Unité récurrentielle à porte.

BPTT BackPropagation Through Time — Rétropropagation temporelle des gradients.

LSTM Long Short Term Memory unit — Unité à mémoire longue et courte.

ADAM ADAptive Momentum estimation — Estimation adaptive de l'impulsion.

RMSProp Root Mean Squared Propagation — Propagation de la moyenne quadratique.

MAP Maximum A Posteriori.

MLE Maximum Likelihood Estimate — Maximum de la vraisemblance.

ELBO Evidence Lower BOund — Limite inférieur sur l'évidence.

HST Hubble Space Telescope.

WFC3 Wide Field Camera 3.

KL Kullback-Leibler.

Liste des symboles

$\mathbb{1}$ Matrice identité.

$\mathbf{1}$ Vecteur dont chaque élément correspond à la valeur 1.

$\mathbb{R}$ Ensemble des nombres réels.

π Pi.

$\boldsymbol{\nabla}$ Gradient.

$\boldsymbol{\nabla}^2$ Laplacien.

κ Convergence — densité surfacique de masse projeté sur l'axe de visée.

$\boldsymbol{\alpha}$ Angles de déflections.

$\boldsymbol{\beta}$ Coordonnées angulaires du plan de la source.

$\boldsymbol{\theta}$ Coordonnées angulaires du plan de la lentille.

$\boldsymbol{\xi}$ Coordnnée comobiles sur le plan de la lentille.

$\boldsymbol{\eta}$ Coordonnées comobiles sur le plan de la source.

D_s Distance du diamètre angulaire entre l'observateur et la source.

D_ℓ Distance du diamètre angulaire entre l'observateur et la lentille.

$D_{\ell s}$ Distance du diamètre angulaire entre la lentille et la source.

$g_{\mu\nu}$ Un élément de la métrique.

$\eta_{\mu\nu}$ Un élément de la métrique de Minkwosky.

$\mathcal{L}$ Lagrangien.

c Vitesse de la lumière.

G Constante universelle de la gravitation.

ρ Densité.

Σ Densité de surface.

Σ_c Densité de surface critique.

Φ Potentiel.

φ Liste des paramètres pour l'algorithme d'inférence d'un problème inverse.

ϕ Liste des paramètres pour un processus d'inférence.

θ Liste des paramètres pour un processus génératif.

$\hat{\mathbf{x}}^{(t)}$ Estimé de vecteur des paramètres physiques après t itérations de la relation de récurrence.

$\mathbf{y}$ Vecteur des quantités observées.

F Modèle physique.

$\mathcal{X}$ Espace vectoriel des paramètres physiques.

$\mathcal{Y}$ Espace vectoriel des quantités observées.

$\mathbf{z}$ Variable latente.

$\mathbf{h}^{(t)}$ État latent d'une céllule mémoire après t itérations de la relation de récurrence.

t Paramètre du temps (continu) ou indice d'une relation de récurrence (discret).

T Nombre total d'itérations de la relation de récurrence.

$\mathcal{D}$ Ensemble de données d'entraînement.

$\mathcal{T}$ Ensemble de données d'essai.

$\mathcal{I}$ Information de Fisher.

$\mathbf{H}$ Hessienne.

$D_{\mathrm{KL}}(\cdot \parallel \cdot)$ Distance de Kullback-Leibler.

$\mathbb{E}_{P(X)}\big[\,\cdot\,\big]$ Opérateur de l'espérance mathématique par rapport à la variable aléatoire X distribué selon $P(X)$.

$\|\cdot\|_2$ Norme euclidienne.

$I(X;Y)$ Information mutuelle entre les variables aléatoire X et Y.

$\mathcal{L}_\varphi$ Fonction objective d'entraînement pour les paramètres φ.

$\mathcal{N}$ Loi normale.

$\mathcal{TN}$ Loi normale tronquée.

$\mathcal{U}$ Loi uniforme.

$\boldsymbol{\mu}$ Moyenne.

$\boldsymbol{\Sigma}$ Covariance.

σ^2 Variance.

σ Déviation standard.

$\oplus$ Concaténation.

$\odot$ Produit d'Hadamard.

À Maman et Julia

Remerciements

Chapitre 1

Introduction

1.1 Lentilles gravitationnelles de type galaxie-galaxie

Fritz Zwicky (1937), suivant les calculs publiés par Einstein (1936), est le premier à observer correctement que l'anneau d'Einstein provenant d'une lentille gravitationnelle est une observable particulièrement riche en information. Ce phénomène est produit lorsque la lumière d'une source lointaine est déviée par le champ gravitationnel d'une galaxie-lentille (ou de tout autre objets massif comme un trou noir, une étoile, etc.) exceptionnellement bien aligné avec cette source selon le point de vue d'un observateur sur Terre. L'article de Zwicky (1937) articule précisément les idées qui nous motivent encore aujourd'hui (plus de 85 ans plus tard) à étudier ces objets, soit

1. d'imager des galaxies trop lointaine pour que l'on puisse les résoudre avec nos télescopes ;

2. de mesurer directement la masse gravitationnelle de ces galaxies.

La première lentille gravitationnelle est découverte par Walsh et al. (1979), suivant l'identification de deux spectres radios de quasars identiques, QSO 0957+561 A et B, séparés par seulement 5.7 secondes d'arcs et capturés avec le téléscope Mark II à l'observatioire de Jodrell Bank. Les spectres partagent la même magnitude, $m = 17$, le même décalage vers le rouge, $z = 1.405$, et possèdent des détails chimiques suspicieusement semblables, ce qui suggère fortement que les spectres sont deux copies du même noyau actif d'une galaxie en arrière plan produit par l'effet de lentille gravitationnelle d'une galaxie en avant plan, invisible dans le domaine radio à une fréquence de 966 Mhz. Cette hypothèse est rapidement confirmée par l'observation optique de la galaxie-lentille ($z = 0.39$) avec l'observatoire Palomar (Young et al., 1980), simulténament observé et confirmé par le téléscope de $2.2\,m$ de l'Université d'Hawaii au mont Mauna Kea (Stockton, 1980), ainsi que la modélisation de sa distribution de masse, de son environnement et des angles de déflection qui causerait l'apparition d'une image double du quasar par des modèles d'une complexité rapidement croissante suivant les observations subséquentes du système (Young et al., 1981).

FIGURE 1.1 – Le quasar double (QSO 0957+561 A et B) et la galaxie-lentille (G). Crédit : ESA/Hubble et NASA

1.1.1 Les angles de déflections

Dans les paragraphes qui suivent, je dérive les équations centrales qui nous permettent d'étudier les lentilles gravitationnelles de type galaxie-galaxie. Des traitements similaires peuvent être trouvé dans les manuels de références de Meneghetti (2013) et Carroll (2003).

Supposons qu'un photon est sur une trajectoire parallèle à l'axe de visée $\mathbf{e}_{\parallel}$ d'un observateur sur Terre. Supposons de plus que la source d'un champ gravitationnel Φ est situé sur l'axe de visée, ce qui a pour effet de courber la trajectoire de ce photon entre son point d'origine A et son point d'arrivé B. On définit l'angle de déviation comme la déviation totale de cette trajectoire dans la direction perpendiculaire à l'axe de visée de l'observateur. De façon générale, cette déviation s'écrit

$$\boldsymbol{\alpha} = -\int_{\lambda_A}^{\lambda_B} \ddot{\mathbf{x}} \times \mathbf{e}_{\parallel} d\lambda \,, \tag{1.1}$$

où λ paramétrise la trajectoire du photon $\mathbf{x}(\lambda)$. Le signe négatif nous indique qu'on prend la perspective de l'observateur.

La trajectoire d'un photon est sujette au principe de Fermat, qui stipule que la lumière suit une trajectoire qui extrémise la durée du parcours entre deux points. Dans le language du calcul des

variations, la variation de la durée s'écrit

$$\delta T = \delta \int_A^B n(\mathbf{x}(\ell)) \frac{d\ell}{c} = 0\,, \tag{1.2}$$

où ℓ est un élément de longueur sur la trajectoire et n est un indice de réfraction. Pour déterminer l'indice de réfraction du champ gravitationnel d'une galaxie, on doit utiliser le formalisme de la relativité générale. Selon le principe d'équivalence (fort), l'effet d'un champ gravitationnel est localement indistinguable d'une accélération causée par la courbure d'un espace-temps décrit par une métrique $g_{\mu\nu}$. La trajectoire d'un photon se trouve alors en cherchant les géodésiques de cet espace-temps. On fait l'approximation que le potentiel Φ d'une galaxie est celui d'un gas parfait, c'est-à-dire qu'il satisfait une équation de Poisson

$$\boldsymbol{\nabla}^2 \Phi = 4\pi G \rho. \tag{1.3}$$

Dans la limite où ce potentiel est faible $\dfrac{2\Phi}{c^2} \ll 1$, la métrique $g_{\mu\nu}$ est décrite par une expansion au premier ordre autour de la métrique de Minkowsky

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \approx \left(1 + \frac{2\Phi}{c^2}\right) c^2 dt^2 - \left(1 - \frac{2\Phi}{c^2}\right) d\mathbf{x}^2. \tag{1.4}$$

Puisqu'un photon suit une géodésique de l'espace-temps $ds^2 = 0$, on peut déterminer l'indice de réfraction en réarrangeant l'équation (1.4)

$$n \equiv c \left(\frac{\|d\mathbf{x}\|}{dt}\right)^{-1} \approx 1 - \frac{2\Phi}{c^2}. \tag{1.5}$$

En réécrivant l'élément de longueur $d\ell$ en terme du paramètre de la trajectoire $d\ell = \left\| \frac{d\mathbf{x}}{d\lambda} \right\| d\lambda$, on peut réécrire l'équation (1.2) sous la forme

$$\delta \int_{\lambda_A}^{\lambda_B} n(\mathbf{x}) \|\dot{\mathbf{x}}\| d\lambda = 0. \tag{1.6}$$

Par correspondance avec la fonctionnelle de l'action $J(x) = \int_{\lambda_0}^{\lambda_1} \mathcal{L}(\lambda, x, \dot{x}) d\lambda$ on trouve que le lagrangien de la trajectoire s'écrit $\mathcal{L} = n(\mathbf{x})\sqrt{\dot{x}^2}$. La trajectoire qui satisfait (1.2) est une solution des équations d'Euler-Lagrange

$$\frac{d}{d\lambda} \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{x}}} - \frac{\partial \mathcal{L}}{\partial \mathbf{x}} = 0. \tag{1.7}$$

On a donc

$$\frac{d}{d\lambda} n \frac{\dot{\mathbf{x}}}{\|\dot{\mathbf{x}}\|} - \|\dot{\mathbf{x}}\| \boldsymbol{\nabla} n = 0, \tag{1.8}$$

Puisque le choix du paramètres λ est libre, on peut le choisir tel que $\|\dot{\mathbf{x}}\| = 1$ en tout point de la

trajectoire. Ainsi,

$$\begin{aligned} \frac{d}{d\lambda} n\dot{\mathbf{x}} - \boldsymbol{\nabla} n = 0 \\ \implies n\ddot{\mathbf{x}} + (\boldsymbol{\nabla} n \cdot \dot{\mathbf{x}})\dot{\mathbf{x}} - \boldsymbol{\nabla} n = 0 \end{aligned} \tag{1.9}$$

À ce point de la dérivation, on utilise l'approximation de Born. C'est-à-dire qu'on approxime la trajectoire du photon comme une ligne droite sur l'axe de visée $\mathbf{e}_\parallel$. Cette approximation est justfiée dans le contexte des lentilles gravitationnelles de type galaxie-galaxie, puisque les angles de déviation sont généralement de l'ordre de l'arcseconde ou plus petit. Comme, le vecteur $\dot{\mathbf{x}}$ est tangent à la trajectoire du photon, on obtient

$$\ddot{\mathbf{x}} \times \mathbf{e}_\parallel = \frac{1}{n} \boldsymbol{\nabla}_\perp n = \boldsymbol{\nabla}_\perp \log n \approx -\frac{2}{c^2} \boldsymbol{\nabla}_\perp \Phi \,, \tag{1.10}$$

où $\boldsymbol{\nabla}_\perp$ est un gradient selon les coordonnées perpendiculaires à $\mathbf{e}_\parallel$. On note que le facteur 2 qui apparaît dans l'équation (1.10) est un effet qui vient de la relativité générale. Ce facteur corrige la solution que l'on aurait obtenu avec une dérivation classique (newtonienne).

On est maintenant en mesure de calculer l'angle de déviation. J'introduit le paramètre d'impact $\boldsymbol{\xi}$ qui est la distance perpendiculaire entre la position d'origine du photon sur le plan de la lentille et l'axe de visé (voir Figure 1.2). Dans le cas où le potentiel est généré par une masse M ponctuelle, ç.-à-d. qu'on suppose $\rho = M\delta^3(\mathbf{x})$, où δ est la fonction delta de Dirac, alors le potentiel qui satisfait l'équation de Poisson (1.3) est la fonction de Green $\Phi = -\dfrac{GM}{\sqrt{\xi^2 + z^2}}$, où z est la coordonné sur l'axe de visée. L'équation (1.1) se réécrit finalement comme

$$\begin{aligned} \boldsymbol{\alpha}(\boldsymbol{\xi}) &= -\frac{2GM}{c^2} \int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\xi}} \frac{1}{(\xi^2 + z^2)^{1/2}} dz \\ \implies \boldsymbol{\alpha}(\boldsymbol{\xi}) &= \frac{4GM}{c^2 \xi^2} \boldsymbol{\xi} \end{aligned} \tag{1.11}$$

Cette solution se généralise naturellement à un profil de masse quelconque en assumant qu'il s'exprime comme une somme d'élément de masses $dm = \Sigma d^2\boldsymbol{\xi}'$, où $\Sigma = \int \rho dz$ est un densité surfacique de masse. L'angle de déviation total mesuré à un point $\boldsymbol{\xi}$ est alors une convolution sur tout le plan de la lentille (mince) puisque l'équation (1.11) dépend linairement de la masse M :

$$\boldsymbol{\alpha}(\boldsymbol{\xi}) = \frac{4G}{c^2} \int_{\mathbb{R}^2} \Sigma(\boldsymbol{\xi}') \frac{\boldsymbol{\xi} - \boldsymbol{\xi}'}{\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|^2} d^2\boldsymbol{\xi}' \tag{1.12}$$

L'angle de déviation est une quantité cruciale pour résoudre une lentille gravitationnelle puisqu'il décrit une transformation des coordonnées angulaires du plan de la lentille ($\boldsymbol{\theta}$) vers les coordonnées angulaires du plan de la source ($\boldsymbol{\beta}$). On assume que les distances entre l'observateur et la lentille D_ℓ, entre l'observateur et la source D_s et entre la lentille et la source $D_{\ell s}$, sont beaucoup plus grandes que les distances perpendiculaires à l'axe de visée $\boldsymbol{\xi}$ ou $\boldsymbol{\eta}$ (voir figure 1.2). Cette approximation est justifiée pour les objets qui nous intéresse, pour lesquels les distances parallèles à l'axe de visée sont

généralement de l'ordre du Gpc, alors que les distances perpendiculaire sont généralement de l'ordre du kpc ; soit 6 ordres de grandeurs de différences. Ainsi, on peut faire un argument géométrique (euclidien)

$$\begin{aligned} D_s\boldsymbol{\theta} &= \boldsymbol{\eta}' \\ D_s\boldsymbol{\beta} &= \boldsymbol{\eta} \\ D_{\ell s}\boldsymbol{\alpha} &= \boldsymbol{\eta}' - \boldsymbol{\eta} \\ \implies D_s\boldsymbol{\beta} &= D_s\boldsymbol{\theta} - D_{\ell s}\boldsymbol{\alpha} \end{aligned} \tag{1.13}$$

La dernière relation est l'équation maîtresse qui nous permet de tracer les rayons lumineux d'une source vers un détecteur fictif dans nos simulations. On notera que cette relation reste valide pour un univers courbe et/ou en expansion (ç.-à-d. décrit par une géométrie non-euclidienne), à condition qu'on utilise une notion de distance qui satisfait, par définition, la relation trigonométrique euclidienne

$$D \equiv \frac{\xi}{\theta} \tag{1.14}$$

Il est généralement pratique de travailler avec la forme adimensionnelle de l'équation (1.13). On introduit la densité critique

$$\Sigma_c = \frac{c^2}{4\pi G}\frac{D_s}{D_{\ell s}D_\ell}\,, \tag{1.15}$$

qui nous permet de définir la quantité qu'on nomme convergence $\kappa(\boldsymbol{\theta}) \equiv \dfrac{\Sigma(\boldsymbol{\theta})}{\Sigma_c}$. On définit ainsi l'angle réduit

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{1}{\pi}\int_{\mathbb{R}^2} \kappa(\boldsymbol{\theta})\frac{\boldsymbol{\theta}-\boldsymbol{\theta}'}{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|}d^2\boldsymbol{\theta}'\,, \tag{1.16}$$

qui satisfait l'équation de la lentille adimensionnelle

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})\,. \tag{1.17}$$

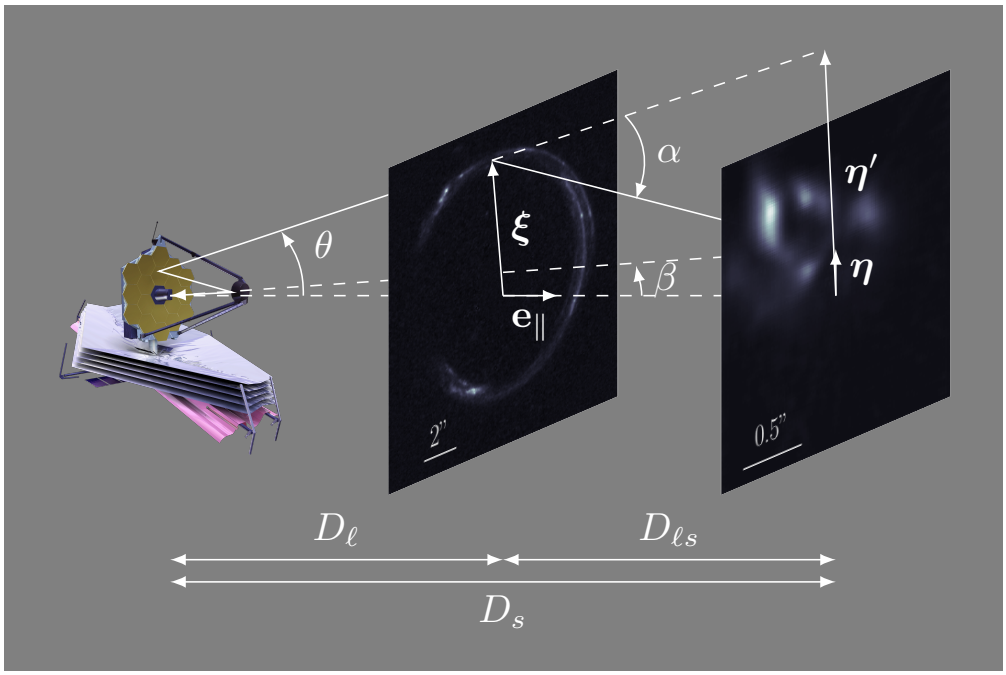

FIGURE 1.2 – Schéma d'une lentille gravitationnelle.

1.1.2 Applications

1.2 Interférométrie par masque non-régulier

1.2.1 Les angles de fermeture

1.2.2 Applications

1.3 Auto-encodeur variationnel

1.3.1 Description du modèle

Les auto-encodeurs variationnels (VAE) ont été introduits par Kingma and Welling (2013) comme une approche pour inférer approximativement les variables latentes (ou cachées) qui modélisent une distribution a posteriori définie implicitement via un échantillon de données. Dans cette section, j'introduis les concepts principaux relié à ce type de modélisation. Le lecteur peut aussi se référer au livre blanc de Kingma and Welling (2019).

On définit $\mathbf{z} \sim q(\mathbf{z})$ comme une variable latente et $\mathbf{x}$ comme un example d'un échantillon de donnée $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$. Notre objectif est de modéliser la distribution $p(\mathbf{x})$, implicitement décrite par notre échantillon. On suppose, sans perte de généralité, que la distribution de $\mathbf{x}$ fait partie d'une famille de distribution , caractérisé par θ, conditionnelle à la variable cachée : $p_\theta(\mathbf{x} \mid \mathbf{z})$. Déterminer

p_θ est généralement difficile, voir intraitable, si la dimensionalité de $\mathbf{x}$ est grande, ce qui est le cas pour des images pour lesquelles on trouve facilement $\dim(\mathbf{x}) > 10^4$. Pour résoudre cette difficulté, on introduit un modèle paramétrique d'inférence $q_\phi(\mathbf{z} \mid \mathbf{x})$ dont le rôle est de modéliser la distribution a posteriori de la variable latente pour la distribution qui nous intéresse

$$q_\phi(\mathbf{z} \mid \mathbf{x}) \approx p_\theta(\mathbf{z} \mid \mathbf{x})\,. \tag{1.18}$$

La notion de distance entre ces deux distributions est mesurée par la divergence de Kullback-Leibler $D_{\mathrm{KL}}(\cdot \parallel \cdot) \geq 0$:

$$\begin{aligned} D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log q_\phi(\mathbf{z} \mid \mathbf{x}) - \log p_\theta(\mathbf{z} \mid \mathbf{x})\Big] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log q_\phi(\mathbf{z} \mid \mathbf{x}) - \log \frac{p_\theta(\mathbf{z}, \mathbf{x})}{p_\theta(\mathbf{x})}\Big] \\ &= \log p_\theta(\mathbf{x}) - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\Big]}_{\equiv \mathcal{L}_{\phi,\theta}(\mathbf{x})}\,. \end{aligned} \tag{1.19}$$

On remarque par cette manipulation que la distance D_{KL}, en plus de mesurer la distance entre les deux distributions a posteriori (par définition), mesure aussi la différence entre le terme $\mathcal{L}_{\phi,\theta}(\mathbf{x})$, qu'on nomme limite inférieure sur l'évidence (de l'anglais *evidence lower bound* : ELBO), et la distribution qui nous intéresse $p_\theta(\mathbf{x})$. L'objectif d'un modèle VAE est de maximiser la ELBO, $\mathcal{L}_{\phi,\theta}$. En observant l'équation (1.19), on réalise que que ceci accomplit deux objectifs simultanément qui suivent du fait que la divergence KL est une quantité positive :

1. Améliorer le processus génératif $p_\theta(\mathbf{x})$ puisque $\log p_\theta(\mathbf{x}) \geq \mathcal{L}_{\phi,\theta}(\mathbf{x})$;

2. Améliorer le processus d'inférence puisque $D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) = \log p_\theta(\mathbf{x}) - \mathcal{L}_{\phi,\theta}(\mathbf{x})$ est simultanément minimisé.

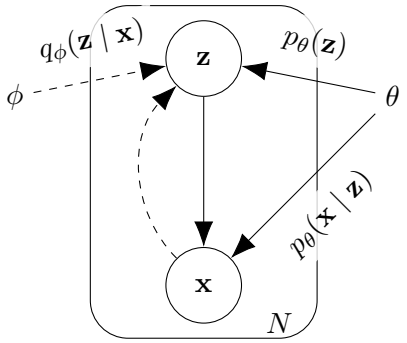

FIGURE 1.3 – Modèle graphique d'un VAE. Les flèches pleines indiquent le processus génératif, alors que les flèches pointillées indiquent le processus d'inférence.

1.3.2 Le truc de reparamétrisation

Le gradient de la ELBO par rapport aux paramètres variationnels, $\boldsymbol{\nabla}_{\phi,\theta}\mathcal{L}_{\phi,\theta}(\mathbf{x})$, est une quantité qu'on doit calculer pour faire usage d'algorithmes comme la grimpe de gradient stochastique pour maximiser la ELBO en terme de ϕ et θ. Or, la liste de paramètres ϕ apparait dans la distribution de prélevement pour calculer l'espérance mathématique $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}$ dans la ELBO (1.19). Cette opération n'a pas de dérivée formelle en terme de ϕ.

Pour résoudre ce problème, on utilise le truc de reparamétrisation (Kingma and Welling, 2013), qui consiste à restreindre la forme fonctionnelle de $q_\phi(\mathbf{z} \mid \mathbf{x})$ à une famille paramétrique qui s'exprime comme la transformation différentiable d'une variable aléatoire auxiliaire $\boldsymbol{\epsilon}$. On considère le cas où $q_\phi(\mathbf{z} \mid \mathbf{x})$ et $p(\boldsymbol{\epsilon})$ font partie de la famille gaussienne isotropique

$$p(\boldsymbol{\epsilon}) \equiv \mathcal{N}(0, \mathbb{1})\,; \tag{1.20}$$

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}),\, \mathbb{1}e^{\log \boldsymbol{\sigma}^2_\phi(\mathbf{x})})\,; \tag{1.21}$$

$$\mathbf{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\epsilon}\,. \tag{1.22}$$

$\odot$ symbolise le produit d'Hadamard, ou encore le produit élément-par-élément de vecteurs. La reparamétrisation fait en sortes que les paramètres variationnelles ne participent plus au procsessus de prélevement, maintenant pris en charge par $\boldsymbol{\epsilon}$. Cette propriété est cruciale dans le but de prendre le gradient de la ELBO (1.19). En effet, on peut maintenant échanger les opérateurs $\boldsymbol{\nabla}_{\phi,\theta}$ et $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} = \mathbb{E}_{p(\boldsymbol{\epsilon})}$, ce qui nous permet d'appliquer le gradient à l'intérieur de l'espérance mathématique. De plus, ϕ décrit maintenant une fonction générique dont le rôle est d'inférer les paramètres d'une distribution gaussienne isotropique (1.21), $f_\phi(\mathbf{x}) = (\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$, étant donné la valeure d'un échantillon $\mathbf{x}$. En pratique, on peut construire une approximation de cette fonction avec un réseau de neuronnes convolutionnelles.

Pour déterminer la forme fonctionnelle de la ELBO, on stipule a priori que la distribution marginale des variables latentes devrait correspondre à une distribution normale isotropique

$$p_\theta(\mathbf{z}) = \mathcal{N}(0, \mathbb{1}) \tag{1.23}$$

On est libre de faire ce choix sans pour autant limiter les formes possibles de la distribution qui nous intéresse $p_\theta(\mathbf{x})$. On peut alors exprimer la ELBO comme

$$\mathcal{L}_{\phi,\theta}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\Big]\,; \tag{1.24}$$

$$\implies \mathcal{L}_{\phi,\theta}(\mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{x} \mid \mathbf{z})\Big]}_{\text{terme de reconstruction}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\Big]}_{\equiv -D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}))}\,. \tag{1.25}$$

La divergence de KL obtenue au second terme du membre droit de l'équation (1.25) admet une

solution fermée étant donné les familles paramétriques stipulées pour $p_\theta(\mathbf{z})$ (1.23) et $q_\phi(\mathbf{z} \mid \mathbf{x})$ (1.21)

$$-D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z})) = \frac{1}{2} \sum_{j=1}^{\dim(\mathbf{z})} (1 + [\log \boldsymbol{\sigma}_\phi^2]_j - [\boldsymbol{\mu}_\phi]_j - [\boldsymbol{\sigma}_\phi^2]_j) \tag{1.26}$$

Une dérivation de ce terme est donnée dans l'appendice B de Kingma and Welling (2013). Le premier terme du membre droit de l'équation (1.25) est nommé *terme de reconstruction* puisqu'il connecte avec l'objectif des fonctions de type auto-encodeurs d'apprendre une représentation latente d'un échantillon de données. La reconstruction s'accomplit en utilisant d'abord le modèle d'inférence $\mathbf{z}^{(1:L)} \overset{\text{i.i.d}}{\sim} q_\phi(\mathbf{z} \mid \mathbf{x})$ [1] pour obtenir un échantillon de représentations latentes à partir des équations (1.20) à (1.22), puis en utilisant le modèle génératif $\hat{\mathbf{x}}^{(i)} \sim p_\theta(\mathbf{x} \mid \mathbf{z}^{(i)})$ pour obtenir un échantillon de reconstructions $\hat{\mathbf{x}}^{(1:L)}$ similaire à l'exemple originel $\mathbf{x}$. Comme on a déjà une variable auxiliaire $\boldsymbol{\epsilon}$ qui se charge de l'aspect génératif du modèle, on peut construire une approximation du modèle génératif avec une fonction générique des variables latentes $g_\theta(\mathbf{z}^{(i)}) = \hat{\mathbf{x}}^{(i)}$. Encore une fois, un réseau de neuronnes convolutionnelles est un choix pratique pour modéliser cette fonction dans le cas où $\mathbf{x}$ est une image. En général, on choisit une erreur quadratique moyenne pour modéliser le terme de reconstruction, de sorte que

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{x} \mid \mathbf{z})\Big] \simeq -\frac{1}{L} \sum_{i=1}^{L} \|\mathbf{x} - \hat{\mathbf{x}}^{(i)}\|_2^2 \tag{1.27}$$

Je note que la fondation théorique des auto-encodeurs variationnels repose sur le principe plus général du goulot d'information (Tishby et al., 1999) ; un sujet qui n'est pas abordé dans ce travail, mais qui motive l'utilisation de la version β-VAE du modèle esquissé dans cette section. Sans rentrer dans les détails, on note qu'il est possible de dériver l'objectif de notre auto-encodeur via la théorie de l'information de Shannon (1948) en interprétant l'auto-encodeur comme un système de transmission d'information par compression, avec perte. Une approche naïve pour modéliser ce système serait de maximiser le taux d'information transmise par le système, ç.-à-d. que le nombre de bit moyen encodé dans une variable latente aléatoire Z, mesuré par l'information mutuelle entre le message X et le code Z utilisé pour représenter le message $I(X; Z)$, devrait se rapprocher d'un maximum qu'on nomme la capacité du système $C = \max_{P(X)} I(X;\, Z)$. Toutefois, cet objectif ne mentionne rien sur la qualité ou la pertinence de cette information. Pour obtenir un message pertinent, on veut contraindre la complexité de Kolmogorov du message, ce qui peut être accomplit en contraingnant le code Z à utiliser le moins de bit possible pour encoder le message. C'est le principe de base de la théorie du taux de distortion (Cover and Thomas, 2006). Tishby et al. (1999) observe que la mesure du taux de distortion suivante

$$\mathcal{L}\left[p(\hat{\mathbf{x}} \mid \mathbf{x})\right] = I(\hat{X}; X) - \beta I(\hat{X}; Z) \tag{1.28}$$

1. i.i.d : identiquement et indépendamment distribué.

Le paramètre β est un multiplicateur de Lagrange qui contrôle le niveau de compression désiré Le lecteur est invité à se référer à la revue sur le sujet par Goldfeld and Polyanskiy (2020).

1.4 Machines à inférence récurrentielles

1.4.1 Formalisme bayésien des problèmes inverses

Les machines à inférence récurentielles (RIM) ont été introduites par Putzky and Welling (2017) pour résoudre des problèmes inverses pour lesquels le terme de régularisation est nécessaire mais inconnue a priori et/ou difficile à construire, voir même calculer. Dans cette section, j'introduis le formalisme bayésien des problèmes inverses sur lequel ce modèle repose, puis j'introduis l'algorithme d'inférence et les concepts d'apprentissage machine qui motivent l'utilisation d'une RIM pour des problèmes inverses mal-posés et sous-déterminés.

Les problèmes inverses en astrophysique prennent généralement la forme

$$\mathbf{y} = F(\mathbf{x}) + \boldsymbol{\eta}\,, \tag{1.29}$$

où $\mathbf{y} \in \mathcal{Y}$ est un vecteur d'observables (comme l'image capturé par les détecteurs CCD dans un téléscope), $\mathbf{x} \in \mathcal{X}$ est un vecteur de paramètres qui gouverne le phénomène physique qui nous intéresse, modélisé par le modèle physique$F : \mathcal{X} \to \mathcal{Y}$. Le vecteur $\boldsymbol{\eta}$ est une réalisation d'un bruit additif. On suupose qu'on connait la distribution de ce bruit, de sortes qu'on peut modéliser la fonction de vraisemblance de l'observable

$$\mathbf{y} - F(\mathbf{x}) \sim p(\boldsymbol{\eta}) = p(\mathbf{y} \mid \mathbf{x})\,. \tag{1.30}$$

Le problème d'inférence est celui de déterminer les paramètres $\mathbf{x}$ qui reproduisent l'observation $\mathbf{y}$, ç.-à-d. l'estimé des paramètres $\hat{\mathbf{x}}_{\text{MLE}}$ qui maximisent la fonction de vraisemblence (MLE de l'anglais *maximum likelihood estimate*), ou de façon équivalente ceux qui maximisent le log de la vraisemblence

$$\hat{\mathbf{x}}_{\text{MLE}} = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}} \log p(\mathbf{y} \mid \mathbf{x})\,. \tag{1.31}$$

Dans le cas général, ce problème est mal posé et n'a pas de solutions. En effet, tel que l'observe Hadamard (1902), un problème aux dérivée partielles comme (1.31) ne possède une solution que si le problème est déterminé, ç.-à-d. que, dans le language de Hadamard (1902), le problème doit correspondre en entier à une situation physique. Cette connection remarquable s'exprime en trois conditions qui déterminent si un problème inverse est bien posé

(H_1) Une solution existe ;

(H_2) Cette solution est unique ;

(H_3) La fonction $G_\varphi : \mathcal{Y} \to \mathcal{X}$ qui infère les paramètres $\mathbf{x}$ satisfait la condition de Lipshitz.

Le troisième critère (H_3) requière que la fonction d'inférence soit stable, ç.-à-d. qu'un petit changement dans le vecteur d'observations devrait correspondre à un petit changement de la solution, mesuré par la constante de Lipshitz $L \geq 0$

$$\|G_\varphi(\mathbf{y}_1) - G_\varphi(\mathbf{y}_2)\|_\mathcal{X} \leq L\|\mathbf{y}_1 - \mathbf{y}_2\|_\mathcal{Y}\,, \tag{1.32}$$

où $\|\cdot\|_\mathcal{V}$ est une métrique de distance définit pour l'espace vectoriel $\mathcal{V}$.

Pour un problème mal-posé, ce qui est le cas pour le problème d'inférence des paramètres d'une lentille gravitationnelles de type galaxie-galaxie ou la reconstruction d'image dans le contexte de l'interférométrie par masque non-réguliers, on assume a priori que la première condition de Hadamard (H_1) est respectée. C'est-à-dire qu'on assume que les quantités observées ou mesurées sont causées par un phénomène unique (solution physique). Toutefois, comme les problèmes qui nous intéressent sont sous-determinées, ç.-à-d. que $\dim_\mathbb{R}(\mathcal{X}) > \dim_\mathbb{R}(\mathcal{Y})$, la seconde condition de Hadamard (H_2) n'est pas respectée ; la fonction de vraisemblence ne peut pas distingué la solution physique du nombre infini de solutions non-physiques au problème (1.31).

La condition d'unicité de la solution est résolue par la construction d'une mesure de probabilité a priori sur l'espace des paramètres d'intérêts $p_\theta : \mathcal{X} \to \mathbb{R},$ t.q. $\int_\mathcal{X} p_\theta(\mathbf{x})d\mathbf{x} = 1$, tel que les solutions non-physiques sont exclues de la région de haute densité de cette distribution. On peut alors modifier le problème (1.31) en introduisant cette distribution a priori comme un terme de régularisation de la vraisemblance

$$\hat{\mathbf{x}}_\text{MAP} = \underset{\mathbf{x}\in\mathcal{X}}{\text{argmax}}\ \log p(\mathbf{y} \mid \mathbf{x}) + \log p_\theta(\mathbf{x})\,. \tag{1.33}$$

La solution $\hat{\mathbf{x}}_\text{MAP}$ maximise la distribution a posteriori $p_\theta(\mathbf{x} \mid \mathbf{y})$, tel que définit par le théorème de Bayes

$$p_\theta(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x})p_\theta(\mathbf{x})}{\int_\mathcal{X} p(\mathbf{y} \mid \mathbf{x})p_\theta(\mathbf{x})d\mathbf{x}}\,. \tag{1.34}$$

Le dénominateur est une constante qu'on nomme l'évidence bayesienne. Pour les applications qui nous intéresse, cette constante n'est pas calculée car elle n'est pas nécessaire (et souvent impossible à calculer) pour la recherche d'un maximum de la distribution a posteriori ou la comparaison de solutions par le ratio de la fonction de vraisemblence (ou de la distribution a posteriori).

On note que la stratégie la plus commune pour résoudre les problèmes inverses qui nous intéresse est plutôt de choisir judicieusement l'espace de solution $\mathcal{X}$ tel que $\dim_\mathbb{R}(\mathcal{X}) \leq \dim_\mathbb{R}(\mathcal{Y})$. Dans ce cas, le problème inverse est balancé ou sur-déterminé. Par exemple, pour modéliser la masse d'une lentille gravitationnelle, il est commun de choisir un modèle singulier isotherme ou une loi de puissance elliptique (e.g. Koopmans et al., 2006; Barnabè et al., 2009; Auger et al., 2010), soit une fonction de type $f_\mathbf{x} : \mathbb{R}^2 \to \mathbb{R}_+$ modélisée par quelques paramètres seulement $\dim_\mathbb{R}(\mathcal{X}) \sim 10$, tandis que l'observation $\mathbf{y}$ est une image avec $\dim_\mathbb{R}(\mathcal{Y}) \gtrsim 10^4 \gg \dim_\mathbb{R}(\mathcal{X})$. Cette approche est considérablement plus stable, particulièrement pour les observations de basses qualités. Toutefois, les modèles analytiques deviennent rapidement complexes et difficiles à construire, voir justifier, lorsque

l'observation des systèmes qui nous intéresse sont de haute qualité, ce qui révèle la complexité cachée de ces systèmes (e.g. Schuldt et al., 2019). De plus, ce cadre nous limite à seulement considérer les hypothèses construites par des humains ou par régression symbolique (e.g. Lemos et al., 2022), et non l'ensemble des hypothèses possibles. C'est cette observation qui nous motive à utiliser l'approche esquissée plus haut, où l'espace $\mathcal{X}$ est construit de manière presque agnostique à la solution physique recherchée (e.g. une grille de pixels pour modéliser une distribution de masse), de manière à contenir toutes, ou au moins la plupart, des solutions physiques. Ce genre d'approche a le potentiel de produire des résultats surprenant ou intéressant, puisque l'exploration de l'espace des solutions physiques peut être ajustée via la distribution a priori, $p_\theta(\mathbf{x})$, selon la complexité de l'observation.

1.4.2 La relation de récurrence

Pour résoudre l'équation différentielle ordinaire sous-entendue par le problème (1.33), on considère la méthode de discrétisation d'Euler

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \alpha \boldsymbol{\nabla}_{\hat{\mathbf{x}}^{(t)}} p_\theta(\hat{\mathbf{x}}^{(t)} \mid \mathbf{y}) \,, \tag{1.35}$$

où α est le taux d'apprentissage dans la littérature sur l'apprentissage machine. On est garantie d'obtenir une solution au problème à valeur initiale $\hat{\mathbf{x}}^{(0)} = \mathbf{x}_0$ si l'algorithme, après T itérations, satisfait la condition de Lipschitz. Pour la relation de récurrence (1.35), ceci revient à assumer que l'erreur locale de chaque itération est proportionnelle à α^2, ce qui est satisfait si le gradient $\boldsymbol{\nabla}_{\mathbf{x}} \log p_\theta(\mathbf{x} \mid \mathbf{y})$ satisfait la condition de Lipschitz dans la région de $\mathcal{X}$ explorée par l'algorithme (Atkinson, 1989; Butcher, 2016), en encore si la norme de la dérivée seconde de $\log p_\theta(\mathbf{x} \mid \mathbf{y})$ est bornée dans cette région.

Putzky and Welling (2017) observent qu'on peut réécrire (1.35) de la façon suivante

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \alpha\big(\boldsymbol{\nabla}_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} \mid \hat{\mathbf{x}}^{(t)}) + \boldsymbol{\nabla}_{\hat{\mathbf{x}}^{(t)}} \log p_\theta(\hat{\mathbf{x}}^{(t)})\big); \tag{1.36}$$

$$\implies \hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}} + g_{\varphi^{(t)}}\big(\hat{\mathbf{x}}^{(t)},\ \boldsymbol{\nabla}_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} \mid \hat{\mathbf{x}}^{(t)})\big) \tag{1.37}$$

où $g_{\varphi^{(t)}} : \mathcal{X}^2 \to \mathcal{X}$ est le modèle du gradient de la distribution a posteriori. On remarque que la relation de récurrence (1.35) est un cas spécial de la relation (1.37), soit le cas où on a un modèle explicite pour la distribution a priori (ou son gradient) $\boldsymbol{\nabla}_{\mathbf{x}} \log p_\theta(\mathbf{x})$ et la taux d'apprentissage α. Dans la relation (1.37), les paramètres α et θ sont absorbés dans les paramètres d'inférence $\varphi^{(t)}$, ce qui nous donne une plus grande liberté pour modéliser la distribution a priori en utilisant le théorème d'approximation universelle (Cybenko, 1989; Hornik, 1991). Selon ce nouveau point de vue, le problème de modéliser la distribution a priori, ou plus directement le gradient de la distribution a priori, est équivalent à construire un modèle pour le gradient de la distribution a posteriori dans une relation de récurrence.

Pour le problème de reconstruction d'image, les modèles neuronaux convolutif avec une archi-

tecture de sablier ou encore une architecture en forme de U (Ronneberger et al., 2015) sont des choix naturels pour modéliser $g_{\varphi^{(t)}}$. Toutefois, la troisième condition d'Hadamard (H_3) est respectée seulement si $g_{\varphi^{(t)}}$ satisfait la condition de Lipschitz, ce qui n'est pas trivialement respecté pour un réseau de neurones. Dans ce travail, cette condition n'est pas explicitement imposée au modèle. On note toutefois que l'analyse de la condition de Lipschitz pour les réseaux neuronnaux est un sujet de recherche actif (e.g.), particulièrement dans l'étude des attaques antagonistes de réseaux de neuronnes (e.g.). Nous reportons l'étude de la troisième condition d'Hadamard pour des travaux futurs.

Finalement, on note un aspect important du modèle $g_{\varphi^{(t)}}$, soit la possible dépendance envers t. Cet aspect est directement inspiré des succès récents d'algorithmes d'optimisations comme la méthode d'accélération de Nesterov (1983), RPROP (Riedmiller and Braun, 1993), AdaGrad (Duchi et al., 2011), RMSProp [2] (Hinton, 2012) et ADAM (Kingma and Ba, 2014), qui utilisent explicitement l'information des gradients d'itérations antérieurs à t pour calculer la mise à jour dans la relation de récurrence (1.37). Cette propriété permet à ces algorithmes de collecter de l'information par rapport à la seconde dérivée de la fonction objective, sans la calculer directement. Ainsi, il est important de considérer une classe de modèles avec une mémoire des itérations précédentes. Pour ce faire, on utilise des unités récurrentielles à porte (de l'anglais *gated recurrent units* : Cho et al., 2014) pour modéliser une fonction g_φ augmentée d'un ensemble d'états latents $\{\mathbf{h}_i^{(t)}\}_{i=1}^H$ qui agissent comme une mémoire des activations précédentes du réseau de neuronnes. Les détails de cette couche neuronale sont données dans l'annexe D.

Comme ADAM est considéré comme l'algorithme le plus performant parmis ceux énuméré précédemment, une machine à inférence récurentielle bénéficie énormément de son utilisation pour prétraiter le gradient de la vraisemblence $\boldsymbol{\nabla}_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$ avant de le passer en entrée au réseau de neuronnes g_φ. Cette idée à fait une première apparition dans les travaux de Modi et al. (2021), puis dans notre travail présenté au chapitre 2.

1.4.3 Méta-apprentissage

Le méta-apprentissage est un sujet de recherche qui a une longue histoire dans le champ de recherche sur l'apprentissage machine, qu'on peut tracer jusqu'aux travaux de Marvin Minsky, puis Schmidhuber 1991 (LSTM and thesis and meta algorithm) et Bengio 1990 (). Le lecteur peut se référer à la revue de Hospedales pour une vue moderne sur le sujet (). L'approche qui nous intéresse est classée dans la catégorie de méta-apprentissage par optimisation.

La première apparition concrète de cette méthode est Younger 2001 et Hochreiter 2001, où le théorème de l'approximation universelle est utilisée pour justifier l'utilisation de cellules à mémoire longues et courtes (LSTM, Schidhuber) pour découvrir un algorithme d'optimisation pour un classe de fonctions (e.g. un modèle neuronnal). L'observation qui est faite est précisément que l'algorithme

2. L'algorithme apparaît en premier dans le cours CSC321 à l'Université de Toronto, donné par Geoffrey Hinton en 2011.

d'Euler est un cas particulier d'une classe plus générale de relations de récurrences qui permettent de résoudre des problèmes de type (**??**). Ainsi, un réseau de neuronnes récurrent est une classe de fonctions qui peuvent représenter, en principe, une large portion de cette classe de fonctions. Ce genre d'approche est motivé par le *no free lunch theorem* pour l'optimisation, qui stipule qu'il n'existe aucun algorithme général d'optimisation en mesure de résoudre toutes les classes de problèmes. Dans ca cas, la solution à ce problème est d'introduire des biais inductifs ou des connaissances a priori pour contraindre l'espace des solutions recherchées à un espace où au moins une solution existe. Le problème de méta-apprentissage est donc précisément d'apprendre ou encoder ces biais inductifs dans un modèle d'apprentissage, de sortes que les problèmes d'optimisations subséquents, sur des taches d'essai, sont garanties d'avoir une solution.

Le travail de Andrychowicz et al. (2016) utilise ces idées pour construire un algorithme d'optimisation, aussi basé sur les cellules LSTM, qui performe beaucoup mieux que les algorithmes d'optimisations traditionnelles (e.g. ADAM) pour entraîner un second réseau de neuronnes pour les tâches spécifiques sur lesquelles l'algorithme de méta apprentissage est entrainé (style transfer etc.). Le travail de Putzky et Welling est une généralisation de cette approche aux problèmes inverses en général.

Pour un problème de méta-apprentissage, l'ensemble de donné d'entrainement est légèrement différent d'une tâche d'interpolation ou de classification, où $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ est construit à partir d'exemples dans le domaine $\mathcal{X}$ et l'image $\mathcal{Y}$ connecté par la fonction qu'on essai d'approximer. Pour le méta-apprentissage, l'ensemble d'entraînement est constitué de tâche à performer. Dans notre cas, la tâche à performer est l'optimisation d'une fonction de vraisemblence. On a donc

$$\mathcal{D} = \{\mathbf{x}_i, \log p_i(\mathbf{y} \mid \mathbf{x})\}_{i=1}^N \tag{1.38}$$

où $\mathbf{x}_i$ est la solution qu'on cherche et $\log p_i(\mathbf{y} \mid \mathbf{x})$ est la fonction de vraisemblence que l'algorithme doit optimiser pour obtenir la solution. Les paramètres d'inférence φ sont optimisés sur toute l'ensemble de la trajectoire construire par la relation de récurrence par une erreur quadratique moyenne

$$\mathcal{L}_\varphi(\mathbf{x}, \log p(\mathbf{y} \mid \mathbf{x})) = \sum_{t=1}^{T} w^{(t)} \|\mathbf{x} - \hat{\mathbf{x}}^{(t)}\|_{\mathcal{X}}^2 \tag{1.39}$$

où $w^{(t)}$ est un poids qu'on associe à l'itération t de la relation de récurrence. Dans la plupart des travaux, $w^{(t)} = \frac{1}{T}$. Cet objectif est optimisé par la rétropropagation temporelle des gradients (BPTT, de l'anglais *backpropagation through time*). Le problème de méta-apprentissage est donc un problème de minimisation du risque empirique observé de la fonction objective

$$\varphi_{\mathcal{D}}^\star = \operatorname*{argmin}_{\varphi} \mathbb{E}_{\mathcal{D}}\big[\mathcal{L}_\varphi\big] \tag{1.40}$$

Il est important de discuter de la notion de généralisation dans le contexte de méta-apprentissage.

Dans le contexte où on cherche à construire une interpolation, la notion de généralisation réfère généralement au test où un point $\mathbf{x}$ en dehors du support implicite définit par l'ensemble d'entraînement $\mathcal{D}$ est donné en entré à la fonction f_φ qu'on a construit. Un modèle est en mesure de généraliser si l'erreur quadratique moyenne sur la prédiction est similaire au risque empirique observé sur l'ensemble d'entraînement.

Dans notre contexte, la généralisation réfère plutôt au concept de transfert d'apprentissage, c'est à dire transferrer les connaissances apprises dans un certain contexte en transferrant la structure du problème pour vers des tâches d'essais. Ainsi, on comprend que la généralisation, dans notre contexte, est équivalente à la notion de transfert de connaissance. Les paramètres d'inférences φ, plutôt que d'encoder les détails d'une fonction, encode des biais inductifs, ou autrement des connaissances a priori sur la structure du problème qui sont transferrable d'une problème à un autre. Cette réalisation est particuliè

Finalement, on note que la notion d'initialisation dans la relation de récurrence $\mathbf{x}^{(0)} = \mathbf{x}_0$ est particulièrement importante dans notre traitement. En effet, on assume que la fonction g_φ se comporte bien dans une région de $\mathcal{X}$ qui connecte $\mathbf{x}^{(0)}$ à $\mathbf{x}^{(T)}$. Or, un mauvais choix d'initialisation fait en sortes que la troisième condition d'Hadamard est difficilement respectée. La notion d'apprendre une initialisation qui accélère l'apprentissage de la descente de gradient est un sujet actif du champ de recherche de méta-apprentissage. MAML et Reptile approchent se problème via une boucle double d'optimisation. Or, une approche beaucoup plus simple peut être mise en place si on fait utilisation de l'observation dans notre problème d'inférence. Ici, on peut prendre le point de vue qu'une fonction approximative inverse du modèle physique $\hat{F}_{\varphi'}^{-1}$ est un bon point de départ pour $\mathbf{x}_0$, en particulier si l'image de cette fonction se situe dans la région de haute densité de distribution a prior empirique déterminée par $\mathcal{D}$.

Chapitre 2

Pixelated Reconstruction of Gravitational Lenses using Recurrent Inference Machine

2.1 Introduction

Strong gravitational lensing is a natural phenomenon through which multiple distorted images of luminous background objects, i.e. early-type star-forming galaxies, are formed by massive foreground objects along the line of sight (e.g., Vieira et al., 2013; Marrone et al., 2018; Rizzo et al., 2020; Sun et al., 2021). These distortions are tracers of the distribution of mass in foreground objects, independent of the electromagnetic behaviour of these overdensities. As such, this phenomenon offers a powerful probe of the distribution of dark matter and its properties outside of the Milky Way (e.g., Dalal and Kochanek, 2002; Treu and Koopmans, 2004; Hezaveh et al., 2016; Gilman et al., 2020, 2021).

Lens modeling is the process of inferring the parameters describing both the mass distribution in the foreground lens and undistorted image of the background source. This has traditionally been a time- and resource-consuming procedure. A common practice to model the mass of lensing galaxies is to assume that their density profiles follow simple parametric forms, e.g., a power law $\rho \propto r^{-\gamma'}$. These profiles generally provide a good fit to low-resolution data and are easy to work with due to their small number of parameters (e.g., Koopmans et al., 2006; Barnabè et al., 2009; Auger et al., 2010). However, as high-resolution and high signal-to-noise ratio (SNR) images become available, lens analysis with simple models requires the introduction of additional parameters representing the true complexity of the mass distribution in lensing galaxies and their immediate environments (e.g., Sluse et al., 2017; Wong et al., 2017; Birrer et al., 2019; Rusu et al., 2020, 2017; Li et al., 2021). This approach becomes intractable as the quality of images increases. For example, no simple parametric model of the Hubble Space Telescope (HST) Wide Field Camera 3 (WFC3) images of

the Cosmic Horseshoe (J1148+1930) — initially discovered by Belokurov et al. (2007) — has been able to model the fine features of the extended arc (e.g., Bellagamba et al., 2016; James et al., 2018; Cheng et al., 2019; Schuldt et al., 2019).

Free-form methods — also misleadingly called nonparametric methods — attempt to relax the assumptions about the smoothness and symmetries of these parametric profiles by changing their parametric support to more expressive families like regular (or adaptive) grid representations and meshfree representations. But, this added flexibility comes at a price, which is (often) a high-dimensional inference problem that is under-constrained, meaning that imposing a prior on the reconstructed parameters becomes essential to penalize unphysical solutions and avoid overfitting the data. Free-form methods have a rich history in cluster-scale lensing (Bartelmann et al., 1996; Seitz et al., 1998; Abdelsalam et al., 1998a,b; Bradač et al., 2005; Diego et al., 2005; Cacciato et al., 2006; Diego et al., 2007; Liesenborgs et al., 2006, 2007; Jee et al., 2007; Coe et al., 2008; Merten et al., 2009; Deb et al., 2012; Merten, 2016; Ghosh et al., 2020; Torres-Ballesteros and Castañeda, 2022) and weak lensing (Kaiser and Squires, 1993; Marshall, 2001; Massey et al., 2007; Deb et al., 2008; Simon et al., 2012; Leonard et al., 2012; Lanusse et al., 2016; Jeffrey et al., 2020; Starck et al., 2021; Remy et al., 2022). They strive to make better use of the information contained in lensing features like resolved arc details, multiple image position, flux ratios, image deformation (i.e. weak lensing constraints) or even the null space of an image in order to place better constraints on the morphology of the mass density of the lens. On the other hand, comparatively less work has been done in the context of galaxy-galaxy lensing system to tackle free-form mass modelling directly (Saha and Williams, 1997, 2004; Birrer et al., 2015; Coles et al., 2014). The main reason for this state of affairs is the difficulty of specifying an appropriate prior which regularizes the problem over its non-linear parameters while maintaining computational tractability and sufficient flexibility to model a large variety of systems.

In view of this, the focus of the field has instead been on using free-form methods for the background source reconstruction only. A number of well-established procedures exists for linear inversion of pixellated-source models in the context of traditional maximum likelihood modeling. These methods, originally developed by Warren and Dye (2003); Suyu et al. (2006), are based mainly on imposing a quadratic-log prior to the source pixels to regularize the optimisation. Subsequently, there have been multiple attempts at building a bridge toward free-form methods in order to correct simplistic assumptions on the mass density model by using linear corrections of the lensing potential (Koopmans, 2005; Suyu and Blandford, 2006; Vegetti and Koopmans, 2009; Vegetti et al., 2012), or iteratively specified priors for shapelets corrections (Birrer and Amara, 2018; Nightingale et al., 2018). But these extensions are generally limited by assumptions regarding the accuracy of the initial parametric models, and thus we are left wanting for a more flexible approach.

Over the recent years, deep learning methods have proven extremely successful at modeling quickly and accurately strong lensing systems (Hezaveh et al., 2017; Perreault Levasseur et al., 2017; Morningstar et al., 2018; Coogan et al., 2020; Park et al., 2021; Legin et al., 2021; Wagner-Carena

et al., 2021; Schuldt et al., 2022; Wagner-Carena et al., 2022; Karchev et al., 2022; Anau Montel et al., 2022; Mishra-Sharma and Yang, 2022). More specifically, Morningstar et al. (2019) demonstrated that recurrent convolutional neural networks can learn implicitly complex prior distributions from their training data to successfully reconstruct pixelated undistorted images of strongly lensed sources, circumventing the need to specify explicitly a prior distribution over those parameters. Motivated by this success, we propose a method that extends this framework to solve the full lensing problem and simultaneously reconstruct a pixelated lensing mass map and a pixelated undistorted background source.

The method we propose here is based on the Recurrent Inference Machine (thereafter refered to as RIM), originally developed by Putzky and Welling (2017). In this framework, we aim to learn an iterative inference algorithm, moving away from hand-chosen inference algorithms and hand-crafted priors. Instead, the prior is learned implicitly through the dataset used to train the neural network that update the solution parameters at each iteration.

In this paper, we present a new architecture for the neural network based on a U-net architecture (Ronneberger et al., 2015), tailored to our highly non-linear inverse problem. We also introduce a fine-tuning prescription which allows us to exploit directly the prior encoded in the neural network parameters in order to perform statistically significant — noise-level — reconstructions of high-SNR galaxy-galaxy lensing systems simulated using IllustrisTNG (Nelson et al., 2019) projected density maps and background galaxy images collected from the COSMOS survey (Koekemoer et al., 2007; Scoville et al., 2007).

The paper is organised as follows. Section 2.2 details the inference pipeline. In Section 2.3, we present the data creation and preprocessing for training the RIM and the generative models used in this paper. In Section 2.4, we report the training strategy for the models used in this work. In Section 2.5, we report and discuss our results on a held-out test set of gravitational lenses. Section 2.6 concludes and situates our finding within the larger context of studying gravitational lensing.

2.2 Methods

In this section, we present the steps to build a free-form inference pipeline with a RIM, beginning with a general introduction about MAP inference with a Gaussian likelihood in Section 2.2.1. In Section 2.2.2, we motivate the use of a Recurrent Inference Machine to solve this problem and describe the computational graph of the RIM as well as the optimisation problem of learning the gradient model. The architecture of the gradient model is described in Section 2.2.3. We describe the raytracing simulation in Section 2.2.4. Finally, we describe the fine-tuning procedure and transfer learning technique applied to reach noise level reconstructions in Section 2.2.5.

2.2.1 Maximum a posteriori

The task of reconstructing a signal vector $\mathbf{x} \in \mathcal{X}$ given observed data $\mathbf{y} \in \mathcal{Y}$ is formulated as an ill-posed inverse problem with a known forward model F and additive noise distribution. We assume a Gaussian distribution with known covariance matrix C, such that

$$\begin{aligned} \mathbf{y} &= F(\mathbf{x}) + \boldsymbol{\eta}; \\ \boldsymbol{\eta} &\sim \mathcal{N}(0, C). \end{aligned} \tag{2.1}$$

In our case study, F is a many-to-one non-linear mapping between the model space $\mathcal{X}$ and the data space $\mathcal{Y}$. Finding a unique solution for this ill-posed inverse problem requires strong inductive biases to be introduced in the inference procedure in order to favour certain hypotheses over others. In the maximum a posteriori (MAP) optimization framework, the inductive biases are encoded in a prior $p(\mathbf{x})$ that is introduced as a probability distribution over $\mathcal{X}$ which reduces the relevant space to explore during inference. The MAP solution is the hypothesis that maximizes the product of the likelihood $p(\mathbf{y} \mid \mathbf{x})$ and the prior :

$$\hat{\mathbf{x}}_{\text{MAP}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \ \log p(\mathbf{y} \mid \mathbf{x}) + \log p(\mathbf{x}). \tag{2.2}$$

By restricting ourselves to Gaussian noise models, the likelihood can be calculated directly and takes the form

$$\log p(\mathbf{y} \mid \mathbf{x}) \propto -\big(\mathbf{y} - F(\mathbf{x})\big)^T C^{-1} \big(\mathbf{y} - F(\mathbf{x})\big) \tag{2.3}$$

However, the prior distribution is harder to define. It is problem-dependent and requires expert knowledge of the model domain.

2.2.2 Recurrent Inference Machine

Instead of handcrafting such a distribution, we attempt to build an inference machine with an implicit prior built in the training set $\mathcal{D}$ and encoded in a deep neural network architecture (Bengio, 2009). The RIM (Putzky and Welling, 2017) is a form of learned gradient-based inference algorithm intended to solve inverse problems of the form of equation (2.1). This framework has mainly been applied in the context of linear and under-constrained inverse problems — i.e. where the function F can be represented in a matrix form and where a prior on the solution space is required to produce a unique solution — for which the prior on the parameters $\mathbf{x}$, $p(\mathbf{x})$, is either intractable or hard to compute (Morningstar et al., 2018, 2019; Lønning et al., 2019). The use of the RIM to solve non-linear inverse problems was first investigated in (Modi et al., 2021). In our case, the inverse problem is non-linear, as it is given by equation (2.7).

The governing equation for the RIM is a recurrent relation that takes the general form

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + g_\varphi\big(\hat{\mathbf{x}}^{(t)}, \mathbf{y}, \boldsymbol{\nabla}_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} \mid \hat{\mathbf{x}}^{(t)})\big). \tag{2.4}$$

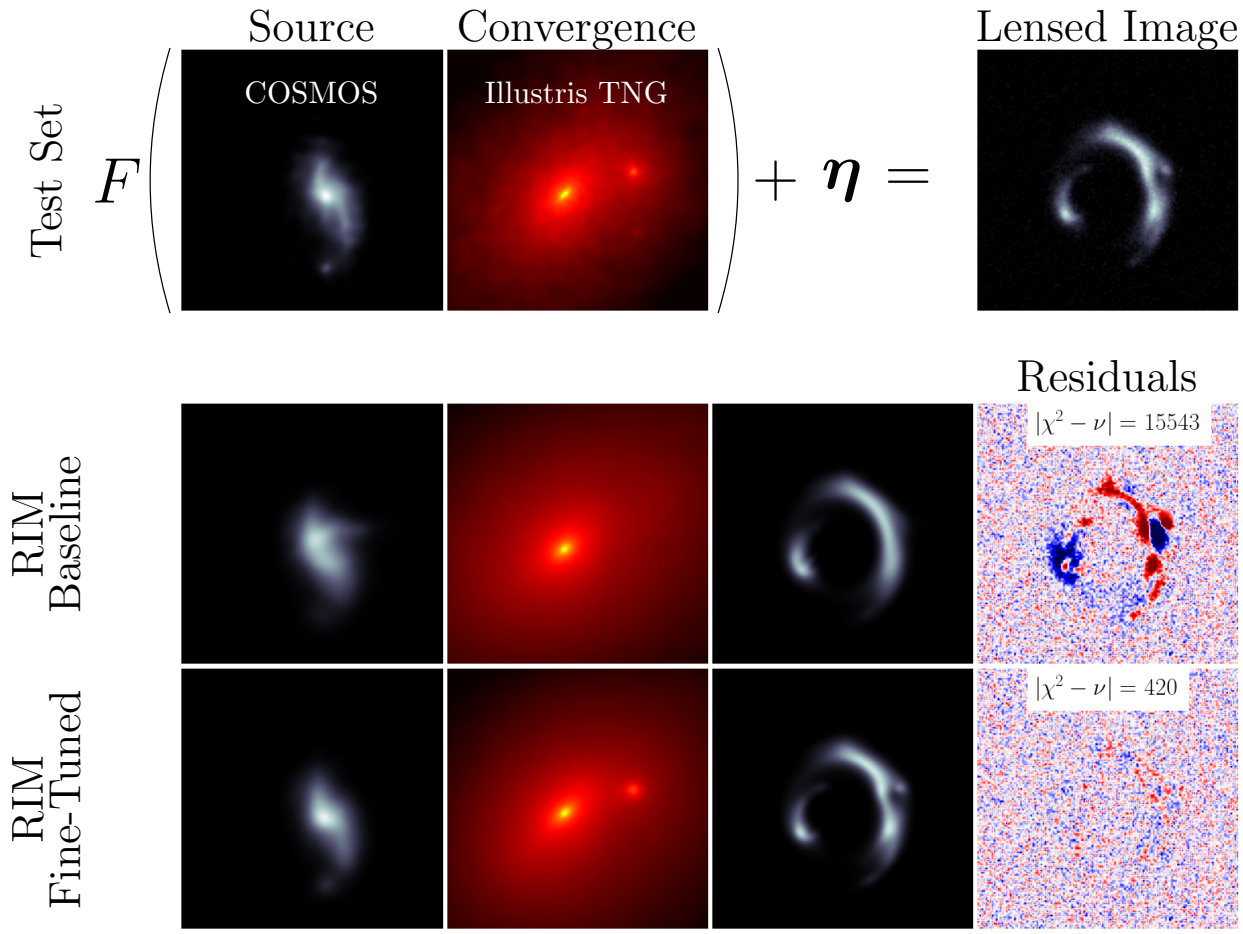

FIGURE 2.1 – Example of a simulated lensed image in the test set that exhibits a large deflection in its eastern arc which indicates the presence of a massive object — in this case a dark matter subhalo. The fine-tuning procedure is able to recover this subhalo because of its strong signal in the lensed image and reduces the residuals to noise level.

In the text, we will often use the shorthand notation $\boldsymbol{\nabla}_{\mathbf{y}|\mathbf{x}} \equiv \boldsymbol{\nabla}_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$ to refer to the gradient of the likelihood. By minimizing a weighted mean squared loss backpropagated through time,

$$\mathcal{L}_\varphi(\mathbf{x}, \mathbf{y}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{M} \mathbf{w}_i (\hat{\mathbf{x}}_i^{(t)} - \mathbf{x}_i)^2, \tag{2.5}$$

the neural network g_φ learns to optimize the parameters $\mathbf{x}$ given a likelihood function. The converged parameters of the neural network given the training set $\mathcal{D}$, $\varphi^\star_\mathcal{D}$, are those that minimize the cost — or empirical risk — which is defined as the expectation of the loss over $\mathcal{D}$

$$\varphi^\star_\mathcal{D} = \underset{\varphi}{\operatorname{argmin}}\ \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\big[\mathcal{L}_\varphi(\mathbf{x}, \mathbf{y})\big]. \tag{2.6}$$

Unlike previous works (Andrychowicz et al., 2016; Putzky and Welling, 2017; Morningstar et al.,

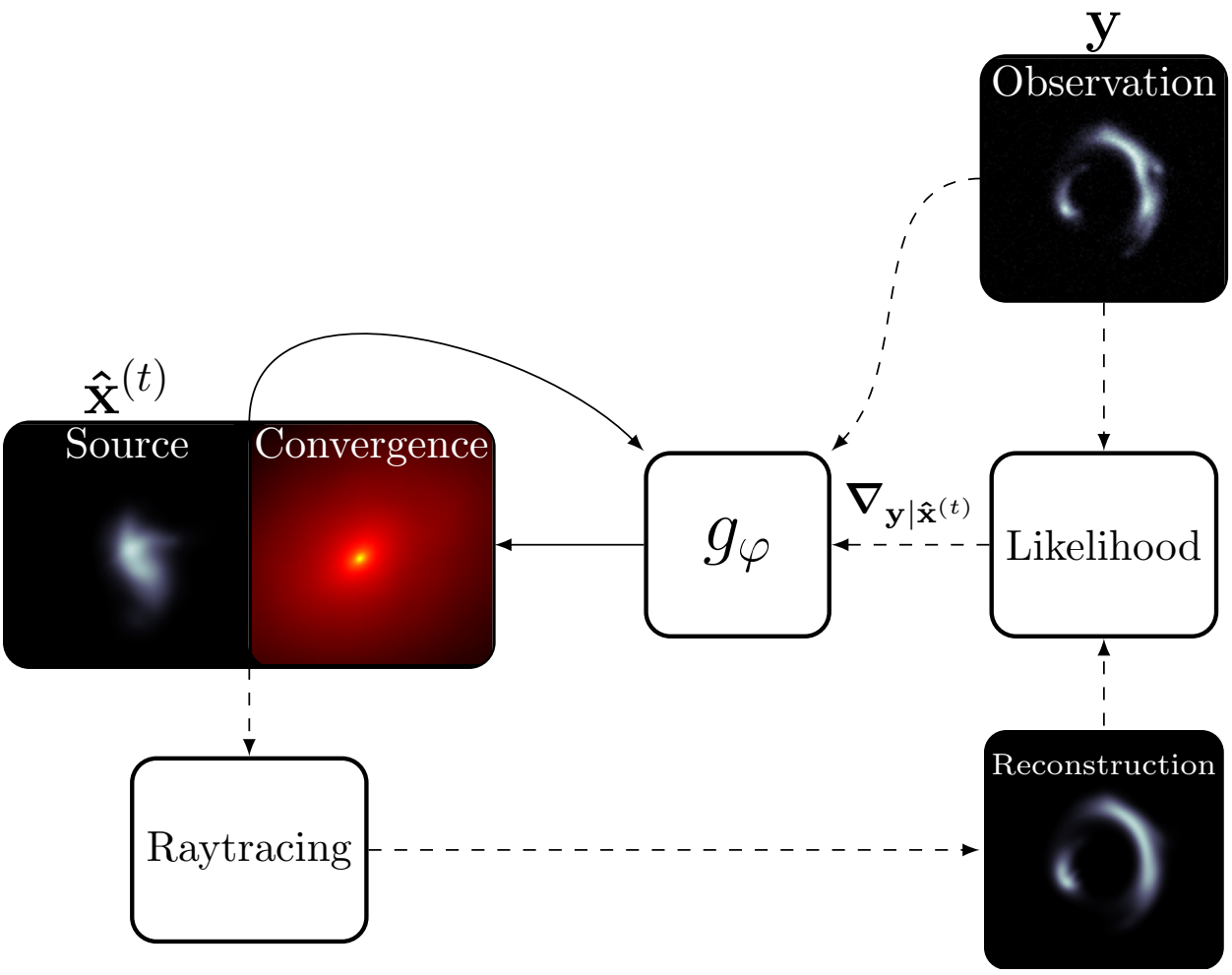

FIGURE 2.2 – Rolled computational graph of the RIM. Dashed arrows represent operations not recorded for BPTT.

2018, 2019; Lønning et al., 2019), the data vector $\mathbf{y}$ — or observation — is fed to the neural network in order to learn the initialization of the parameters ($\mathbf{x}^{(0)} = g_\varphi(0, \mathbf{y}, 0)$) as well as their optimization. We found empirically that this significantly improves the performance of the model for our problem and it avoids situations where the model would get stuck in local minima at test time due to poor initialization.

We follow previous works in setting a uniform weight over the time steps ($\mathbf{w}^{(t)} = \frac{\mathbf{w}}{T}$). The choice of the pixel weights $\mathbf{w}_i$ is informed by our empirical observations when training the network. Details are reported in appendix C.

In Figure 2.2, we show the rolled computational graph of the RIM. During training of the gradient model g_φ, operations along the solid arrows are being recorded for backpropagation through time. The recording is stopped along the dashed arrow since these operations are part of the forward

modelling process. By avoiding the computation of these gradients, training time is reduced and knowledge about the inner workings of a specific likelihood (and forward model) is insulated from the optimization algorithm. This is analogous to a common RNN use-case like text generation, where the process responsible for producing the next element in a time series is a black box to the optimization algorithm.

The gradient of the likelihood is computed using automatic differentiation. Following (Modi et al., 2021), we preprocess the gradients using the Adam algorithm (Kingma and Welling, 2013). For clarity, we only illustrated this step in Figure 2.3.

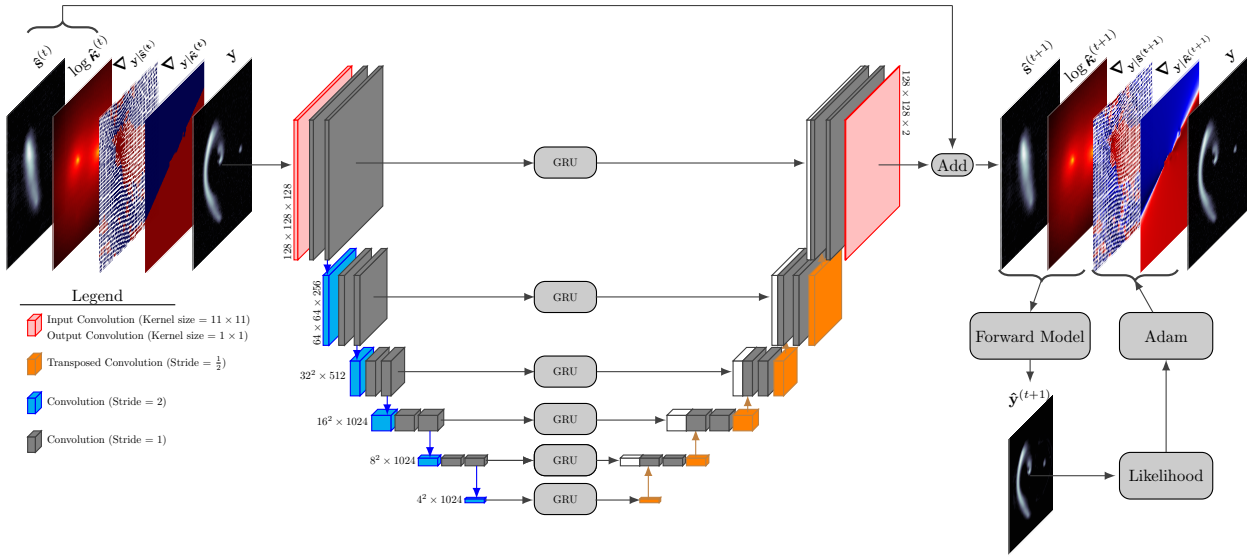

FIGURE 2.3 – A single time step of the unrolled computation graph of the RIM. GRU units are placed in the skip connections to guide the reconstruction of the source and convergence. A schematic of the steps to compute the likelihood gradients is shown in the bottom right of the figure, including the Adam processing step of the likelihood gradient.

2.2.3 The Gradient Model

The neural network architecture is illustrated in Figure 2.3, which shows a single time step of the unrolled computation graph of the RIM. We use a U-net (Ronneberger et al., 2015) architecture with Gated Recurrent Units (GRU : Cho et al., 2014) placed in each skip connections.

Each GRU cell has it's own memory tensor that is updated through time at each iteration of equation 2.4. The shape of a memory tensor is set to match the feature tensor fed into it from the parent layer in the network graph. Instead of learning a compressed representation like in the hourglass architecture (i.e. autoencoder), the U-net architecture naturally separates the spatial frequency components of the signal into its vertical levels. The first level generally encodes high frequency features while the lower level encodes low frequency features (due to downsampling of the feature maps). Adding an independent memory unit at each level preserve this property.

Convolutional layers with a stride of 2 are used for downsampling and stride of $\frac{1}{2}$ for upsampling of the feature maps (identified in blue and orange respectively in figure 2.3). Half-stride convolutions are implemented in practice with the transposed convolution layers from `Tensorflow` (Abadi et al., 2015). Most layers use a kernel size of 3×3, except the first and last layer. The first layer has larger receptive field (11×11) in order to capture more details in the input tensor. The last layer has kernels of size 1×1. A tanh activation function is used for each convolutional layer, including strided convolutions, except for the output layer. The U-net outputs an image tensor with two channels, one dedicated for the update of the source and the other to the update of the convergence (see figure 2.3).

2.2.4 The Forward Model

An observation is simulated by ray tracing the brightness distribution of the background source to the foreground coordinate system. In our case, the coordinate systems have discretized representations. Each pixel of an image is labeled with a subscript index i, which we distinguish from a parenthesized superscript index (i) that refers to the member of a set or list of tensors. For clarity, we omit the superscript index in what follows.

Each pixel is associated with an intensity value and a coordinate vector. The foreground pixel coordinates $\boldsymbol{\theta}_i$ and the source pixel coordinates $\boldsymbol{\beta}_i$ are related by the lens equation

$$\boldsymbol{\beta}_i = \boldsymbol{\theta}_i - \boldsymbol{\alpha}(\boldsymbol{\theta}_i), \tag{2.7}$$

where $\boldsymbol{\alpha}$ is a deflection angle. It is obtained from the projected surface density field κ — also referred to as convergence — by the integral

$$\boldsymbol{\alpha}(\boldsymbol{\theta}_i) = \frac{1}{\pi} \int_{\mathbb{R}^2} \kappa(\boldsymbol{\theta}') \frac{\boldsymbol{\theta}_i - \boldsymbol{\theta}'}{\|\boldsymbol{\theta}_i - \boldsymbol{\theta}'\|^2} d^2\boldsymbol{\theta}'. \tag{2.8}$$

The intensity of a pixel in a simulated observation is obtained by bilinear interpolation of the source brightness distribution at the coordinate $\boldsymbol{\beta}_i$. In this work, the convergence also has a discrete representation. Thus, we approximate this integral by a discrete global convolution. Taking advantage of the convolution theorem, this operation can be computed in near-linear time using the Fast Fourier Transform (FFT).

Assuming the observation has M^2 pixels, the convolution kernel would have $(2M+1)^2$ pixels. Both the convergence tensor and the kernel tensor are zero-padded to a size of $(4M+1)^2$ pixels in order to approximate a linear convolution and significantly reduce aliasing.

A blurring operator — convolution by a point spread function (PSF) — is then applied to the lensed image to replicate the response of an imaging system. This operator is implemented as a GPU-accelerated matrix operation since the blurring kernels used in this paper have a significant proportion of their energy distribution encircled inside a small pixel radius.

2.2.5 Fine-Tuning

Objective function

Once the gradient model is trained, the RIM is a baseline estimator of the parameters $\mathbf{x}$ given a noisy observation $\mathbf{y}$, a PSF and a noise covariance matrix. We now concern ourselves with a strategy to improve this estimator. This is important for observations with high SNR, for which the estimator must be extremely accurate to model all the fine features present in the arcs. The metric for the goodness of fit is the reduced chi squared $\chi^2_\nu = \frac{\chi^2}{\nu}$, where ν is the total number of degrees of freedom which corresponds to the total amount of pixels in $\mathbf{y}$ in this work. Generally, our goal will be to reach $\chi^2_\nu = 1$, or equivalently $|\chi^2 - \nu| = 0$, which indicates that the RIM hypothesis has reconstructed all the signal to be recovered from the observation. We note that such a problem is exceedingly difficult at high SNR. In this regime, our hypothesis has to be both precise and accurate in order to satisfy the chi squared test. This is unlike noisy observations, where strong assumptions about the model parameters are required and often sufficient to reconstruct the remaining information left in the observed vector $\mathbf{y}$.

We observe that we can optimize the network parameters w.r.t to the chi squared directly

$$\hat{\varphi}_{\text{MAP}} = \underset{\varphi}{\text{argmax}} \frac{1}{T} \sum_{t=1}^{T} \log p(\mathbf{y} \mid \hat{\mathbf{x}}^{(t)}) + \log p(\varphi). \tag{2.9}$$

Unlike the loss in equation (2.5), this objective function makes no use of labels ($\mathbf{x}$). This allows us to use equation (2.9) at test time in order to fine-tune the RIM's weights to a specific test example.

Transfer Learning

We now address the issue of transferring knowledge from the training task, problem (2.6), to a test task specific to an observation, problem (2.9). The reader might refer to reviews on transfer learning (Pan and Yang, 2010; Zhuang et al., 2019) for a broad overview of the field. The strategy we outline fall into the category of inductive transfer learning.

Since the data likelihood $p(\mathbf{y} \mid \mathbf{x})$ does not contain *a priori* information about the solution $\hat{\varphi}_{\text{MAP}}$, inductive biases must be introduced to make the problem (2.9) well-posed. Thus, we

($\mathcal{H}_1$) initialize the network parameters with $\varphi^\star_{\mathcal{D}}$;

($\mathcal{H}_2$) apply early stopping when a maximum number of steps is reached or $\chi^2_\nu \leq 1$;

($\mathcal{H}_3$) use a small learning rate.

($\mathcal{H}_2$) and ($\mathcal{H}_3$) encode the assumption that the optimal estimator is to be found *near* the initialization.

As it turns out, ($\mathcal{H}_1$) is not strong enough to preserve the knowledge learned from the training task. This has long been observed in the literature and was coined as the catastrophic interference

phenomenon in connectionist networks (McCloskey and Cohen, 1989; Ratcliff, 1990). In summary, a sequential learning problem exhibits catastrophic forgetting of old knowledge when confronted with new examples (possibly from a different distribution or process), in a manner

(CF_1) proportional to the amount of learning ;

(CF_2) strongly dependant to the disruption of the parameters involved in representing the old knowledge.

While ($\mathcal{H}_2$) and ($\mathcal{H}_3$) can potentially alleviate (CF_1), (CF_2) is not trivially addressed by the inductive biases introduced so far.

We follow the work of Kirkpatrick et al. (2016) to define a prior distribution over φ that address this issue

$$\log p(\varphi) \propto -\frac{\lambda}{2} \sum_j \text{diag}(\mathcal{I}(\varphi^\star_{\mathcal{D}}))_j (\varphi_j - [\varphi^\star_{\mathcal{D}}]_j)^2 \,. \tag{2.10}$$

$\text{diag}(\mathcal{I}(\varphi^\star_{\mathcal{D}}))$ is the diagonal of the Fisher information matrix encoding the amount of information that some set of gravitational lensing systems from the training set, and similar to the observed test task, carries about the baseline RIM weights $\varphi^\star_{\mathcal{D}}$ — the parameters that minimize the empirical risk (equation 2.6). We can also understand this prior using the Cramér-Rao lower bound (Rao, 1945; Cramér, 1946). The prior can thus be framed as a multivariate Gaussian distribution characterised by a diagonal covariance matrix with $\text{diag}(\mathcal{I})$ as its inverse and by $\varphi^\star_{\mathcal{D}}$ as its first moment. Within this view, the Lagrange multiplier is tuning our estimated uncertainty about the neural network weights for the particular task at hand. We've included a derivation of this term in the appendix A.

Examples are drawn from the set of training examples similar to the test task by sampling the latent space of both the source VAE and the convergence VAE near the baseline prediction of the RIM. Figure 2.4 illustrates what we mean by *similar*.

2.3 Data

2.3.1 COSMOS

The background source brightness distributions are taken from the Hubble Space Telescope (HST) Advanced Camera for Surveys Wide Field Channel COSMOS field (Koekemoer et al., 2007; Scoville et al., 2007), a $1.64\,\text{deg}^2$ contiguous survey acquired in the F814W filter. A dataset of mag limited ($\text{F814W} < 23.5$) deblended galaxy postage stamps (Leauthaud et al., 2007) was compiled as part of the GREAT3 challenge (Mandelbaum et al., 2014). The data is publicly available (Mandelbaum et al., 2012), and the preprocessing is done through the open source software `GALSIM` (Rowe et al., 2015).

We applied the `marginal` selection criteria (see the `COSMOSCatalog` class) and imposed a flux per image greater than 50 photons $\text{cm}^{-2}\,\text{s}^{-1}$. This final set has a total of 13 321 individual images. Each image is convolved with its original PSF and drawn into a postage stamps of 158^2 pixels.

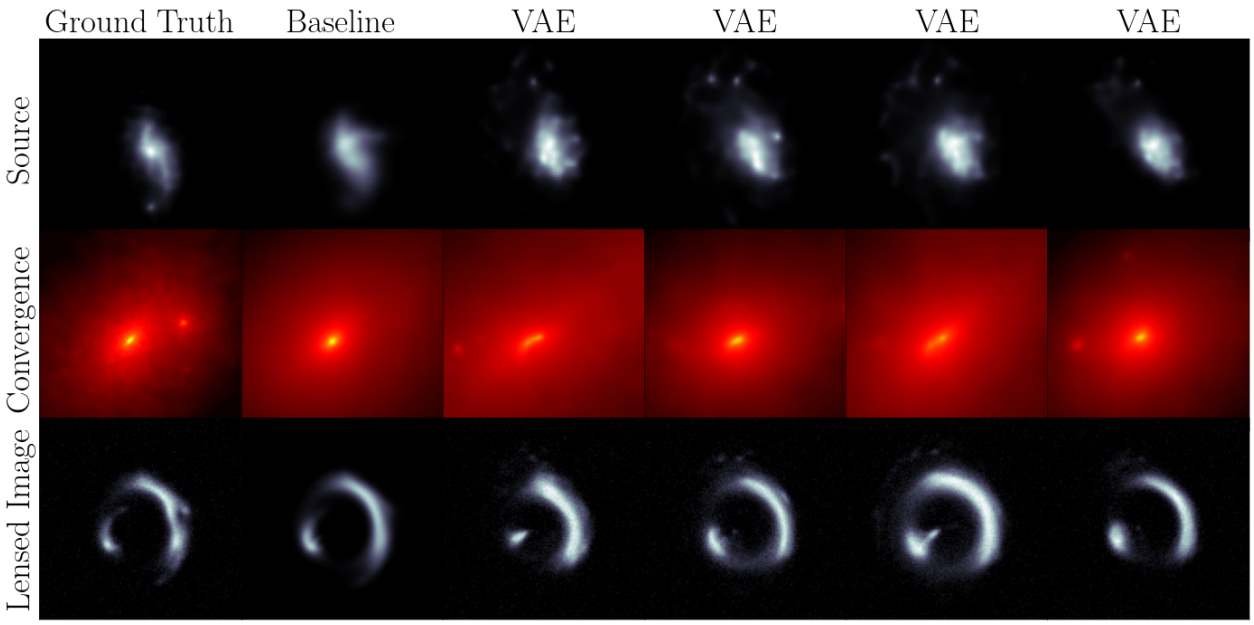

FIGURE 2.4 – Examples similar to the test task, also shown in Figure 2.7. The first column shows the ground truth used to simulate the lensed image. The second column shows the baseline prediction that is then encoded in the latent space of the VAE in order to sample the next 4 columns.

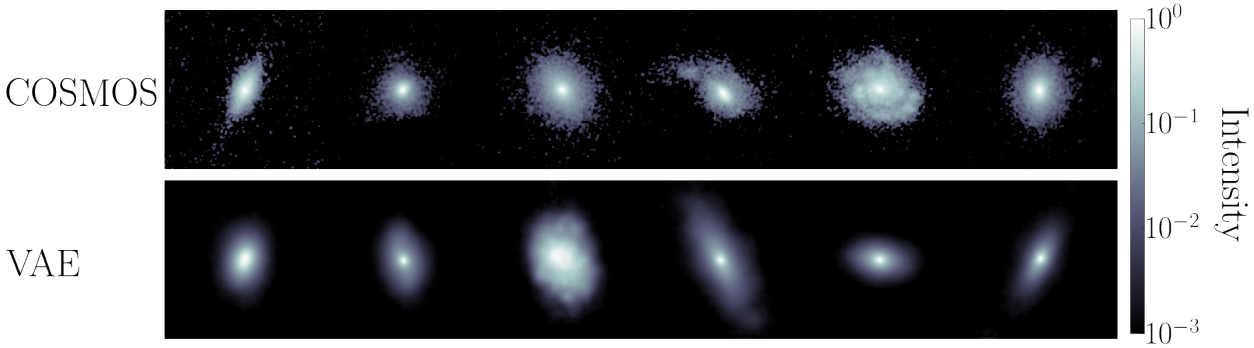

FIGURE 2.5 – Examples of COSMOS galaxy images (top row) and VAE generated samples (bottom row) used as labels in $\mathcal{D}$.

Each image is then background subtracted, randomly shifted, rotated by an angle multiple of 90°, cropped cropped down to 128^2 pixels and normalized to pixel intensities in the range $[0, 1]$.

We split each unique galaxies into a training set (90%) and a test set (10%). The training set is used to train a VAE and produce simulated observations to train the RIM.

2.3.2 IllustrisTNG

Smooth Particle Lensing

To compute a convergence map from an N-body simulation, we follow Aubert et al. (2007) in treating each particle as flow tracers instead of describing their density as Dirac delta functions. Smoothing each particle density on a non-singular kernel reduces the particle noise affecting all important lensing quantities — most importantly the convergence. At the same time, the choice of the kernel size is important to preserve substructures in the lens that we might potentially be interested in. Following Rau et al. (2013), we use Gaussian smoothing with an adaptive kernel size determined by the distance of the 64^{th} nearest neighbours of a given particle $D_{64,i}$.

$$\begin{aligned} \kappa(\mathbf{x}) &= \frac{1}{\Sigma_{\text{crit}}} \sum_{i=1}^{N_{\text{part}}} \frac{m_i}{2\pi\hat{\ell}_i^2} \exp\left(-\frac{1}{2}\frac{(\mathbf{x}-\mathbf{x}_i)^2}{\hat{\ell}_i^2}\right) \\ \hat{\ell}_i &= \sqrt{\frac{103}{1024}} D_{64,i}. \end{aligned} \tag{2.11}$$

The nearest neighbours are found by fitting a k-d tree — implemented in `scikit-learn` (Pedregosa et al., 2011) — to the N_{part} particles in a cylinder centered on the centre of mass of the halo of interest. The critical surface density is defined as

$$\Sigma_{\text{crit}} = \frac{4\pi G}{c^2} \frac{D_\ell D_{\ell s}}{D_s}, \tag{2.12}$$

where D_ℓ, D_s and $D_{\ell s}$ are angular diameter distance to the lens, source and between the lens and the source respectively, G is the gravitational constant and c the speed of light.

Preprocessing

The projected surface density maps (convergence) of lensing galaxies were made using the redshift $z = 0$ snapshot of the IllustrisTNG-100 simulation (Nelson et al., 2019) in order to produce physically realistic realizations of dark matter and baryonic matter halos. We selected 1604 halos with the criteria that they have a total dark matter mass of at least $9 \times 10^{11} M_\odot$. We then collected all dark matter, gas, stars and black holes particles from the data associated to the galaxy cluster within which the halo resides in to create a smoothed projected surface density maps around the centroid of the halo as prescribed in section 2.3.2.

We adopt the ΛCDM cosmology from Planck Collaboration (2020) with $h = 0.68$ to compute angular diameter distances. We also fix the source redshift to $z_s = 1.5$ and the deflector redshift to $z_\ell = 0.5$. We note that changing the redshifts or the cosmology only amount in a rescaling of the κ map by a global scalar, not the morphology of the profiles. Thus, this choice does not change the generality of our method. The smoothed distributions from equation (2.11) are rendered into a

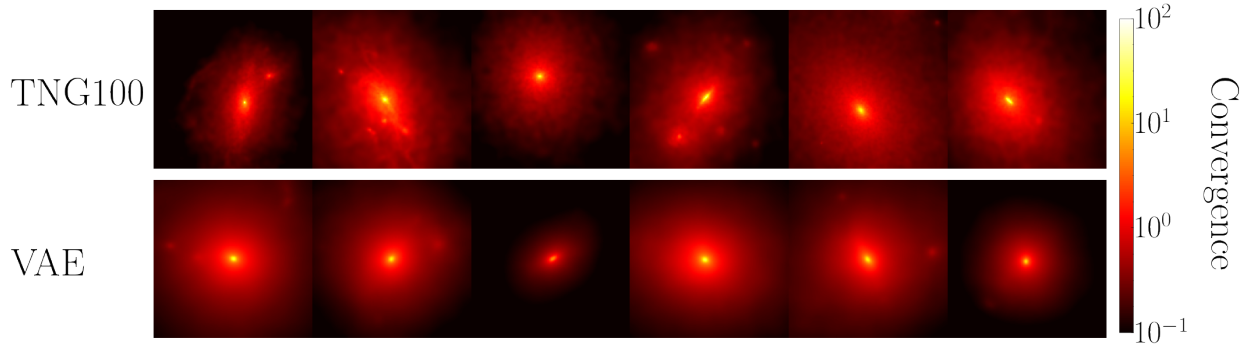

FIGURE 2.6 – Examples of smoothed Illustris TNG100 convergence map (top row) and VAE generated samples (bottom row) used as labels in $\mathcal{D}$.

regular grid of 188^2 pixels with a comoving field of view of 105 kpc/h. To avoid edge effects in the pixelated maps, we include particles outside of the field of view in the sum of equation (2.11).

Before applying augmentation or considering different projections, our dataset of halos is split into a training set (90%) and a test set (10%) in order to make sure that the test set consists only of convergence maps unseen by the RIM during training. We take 3 different projections (xy, xz and yz) of each 3D particle distributions, which amounts to a dataset with a total of 4 812 individual convergence maps. Random rotations by an angle multiple of 90° and random shifts to the pixel coordinates are applied to each image. The κ maps are then rescaled by a random factor to change their estimated Einstein radius to the range [0.5, 2.5] seconds of arc. The Einstein radius is defined as

$$\theta_E = \sqrt{\frac{4GM(\theta_E)}{c^2}\frac{D_{\ell s}}{D_\ell D_s}} \tag{2.13}$$

where $M(\theta_E)$ is the mass enclosed inside the Einstein radius. In practice, we estimate this quantity by summing over the mass of pixels with a value greater than the critical density ($\kappa > 1$). For data augmentation purposes, this procedure gives a good enough estimate of the size of the lensed image that will be produced by some κ map. We test multiple scaling factors for each κ map, then uniformly sample between those that produce an estimated Einstein radius within the desired range. This step is used to remove any bias in the Einstein radius that might come from the mass function of the simulation.

The final maps are cropped down to 128^2 pixels. Placed at a redshift $z_\ell = 0.5$, a κ map will thus span an angular field of view of $7.69''$ with a resolution similar to HST. With these augmentation procedures, a total of 50 000 maps are created from the training split and 5 000 from the test split. The training set is used to train a VAE and produce simulated observations to train the RIM.

2.3.3 Simulated Observations

Having defined a source map and a convergence map, we apply the ray tracing simulation prescribed in section 2.2.4 to produce an observation with observational effects that crudely correspond to HST images.

For each observation, a Gaussian point spread function is created with a full width at half maximum (FWHM) randomly generated from a truncated normal distribution. The support of the distribution is truncated below by the angular size of a single pixel and above by the angular size of 4 pixels. White noise with a standard deviation randomly generated from a truncated normal distribution is then added to the convolved observation to simulate SNR conditions between 10 dB and 30 dB. For simplicity, we define SNR $= \frac{1}{\sigma}$. This definition is equivalent to the peak signal-to-noise ratio.

As a validation criteria for each simulated image, we impose a minimum magnification of 3. Thus, we make sure that most pixel coordinates in the image plane will be mapped inside the source coordinate system through the lens equation (2.7).

TABLE 2.1 – Physical model parameters.

Parameter	Distribution/Value
Lens redshift z_ℓ	0.5
Source redshift z_s	1.5
Field of view (")	7.69
Source field of view (")	3
PSF FWHM (")	$\mathcal{TN}(0.06, 0.3; 0.08, 0.05)$ [1]
Noise amplitude σ	$\mathcal{TN}(0.001, 0.1; 0.01, 0.03)$

400 000 observations are simulated from random pairs of COSMOS sources and IllustrisTNG convergence training splits in order to train the RIM. An additional 200 000 observations are created from pairs of COSMOS source and pixelated SIE convergence map. The parameters for these κ maps are listed in table 2.2. We found this addition to be beneficial to learning since it adds an inductive bias in the learning favoring isothermal profiles. We expect some lensing configurations like large Einstein rings or double images to poorly constrain the inner structure of the mass distribution. Building an inference pipeline with strong constraints on the slope of the profile, other than the lensed image, goes beyond the scope of this work. As such, imposing an implicit prior for the slope through the dataset is sufficient for our goal. It is also motivated by the *bulge-halo conspiracy* — the observation that most lensing configurations observed in the sky can be explained to first order approximation by an average slope consistent with an isothermal profile (Auger et al., 2010; Dutton and Treu, 2014).

1 600 000 simulated observations are generated from the VAE background sources and convergence maps as part of the training set. In principle, we could continuously generate examples from the VAE. However, having a fixed amount let us apply some validation check to each examples

TABLE 2.2 – SIE parameters.

Parameter	Distribution
Radial shift (")	$\mathcal{U}(0, 0.1)$
Azimutal shift	$\mathcal{U}(0, 2\pi)$
Orientation	$\mathcal{U}(0, \pi)$
θ_E (")	$\mathcal{U}(0.5, 2.5)$
Ellipticity	$\mathcal{U}(0, 0.6)$

in order to avoid configurations like a single image of the background source or an Einstein ring cropped by the field of view.

2.4 Training

2.4.1 VAE

As mentionned in Kingma and Welling (2019), direct optimisation of the ELBO can prove difficult because the reconstruction term $\log p_\theta(\mathbf{x} \mid \mathbf{z})$ is relatively weak compared to the Kullback Leibler (KL) divergence term. To alleviate this issue, we follow the work of Bowman et al. (2015) and Kaae Sønderby et al. (2016) in setting a warm-up schedule for the KL term in thstarting from $\beta = 0.1$ up to β_{max}.

Usually, $\beta_{\text{max}} = 1$ is considered optimal since it matches the original ELBO objective derived by Kingma and Welling (2013). But, we are more interested in the sharpness of our samples and accurate inference around small regions of the latent space for fine-tuning. Thus, setting $\beta_{\text{max}} < 1$ allows us to increase the size of the information bottleneck (or latent space) of the VAE and improve the reconstruction cost of the model. This is a variant of the β-VAE (Higgins et al., 2017), where $\beta > 1$ was found to improve disentangling of the latent space (Burgess et al., 2018).

The value for β_{max} and the steepness of the schedule are grid searched alongside the architecture for the VAE. Our criteria for an optimal model is a VAE that achieve the lowest reconstruction error in order to produce sharp images. At the same time, the KL divergence value should be smaller than an empirically defined threshold to respect the latent space prior. This value is found in practice by manually looking at the quality of generated samples for different VAE hyperparameters. A similar method is explored and formalized in the InfoVAE framework (Zhao et al., 2017).

A notable element of the VAE architecture is the use of a fully connected layer to reshape the features of the convolutional layer into the chosen latent space dimension. Following the work of Lanusse et al. (2021), we introduce an ℓ_2 penalty between the input and output of the bottleneck dense layers to encourage an identity mapping. This regularisation term is slowly removed during training.

2.4.2 RIM

The architecture of the gradient model was grid searched on smaller dataset ($\lesssim 10\,000$ examples) in order to quickly identify a small grid of valid hyperparameters. Then, the best hyperparameters were identified using a two-stage training process on the training dataset. In the first stage, we trained 24 different architectures from this small hyperparameter set for approximately 4 days (wall time using a single Nvidia A100 gpu). Different architectures would have a training time much longer than others, and this was factored in the architecture selection process. For example, adding more time steps (T) to the recurrent relation (2.4) would yield better generalisation on the test set, but this would come at great costs to training time until convergence. Following this first stage, 4 architectures were deemed efficient enough to be trained for an additional 6 days. We only report the results for the best architectures out of these 4.

Each reconstruction is performed by fine-tuning the baseline model on a test task composed of an observation vector, a PSF and a noise covariance. In practice, fine-tuning the test of 3 000 examples can be accomplished in parallel so as to be done in at most a few days by spreading the computation on ~ 10 Nvidia A100 GPUs (or 10 hours on ~ 100 GPUs). Each reconstruction uses at most 2000 steps, which turns out to be approximately 20 minutes (wall-time) per reconstruction. Early stopping is applied when the χ^2 reaches noise level. The hyperparameters for this procedure are reported in Table 2.3.

TABLE 2.3 – Hyperparameters for fine-tuning the RIM.

Parameter	Value
Optimizer	RMSProp
Learning rate	10^{-6}
Maximum number of steps	2 000
λ	2×10^5
ℓ_2	0
Number of samples from VAE	200
Latent space distribution	$\mathcal{N}(\mathbf{z}^{(T)}, \sigma = 0.3)^2$

2.5 Results

In this section, we present the performance of our approach on the held out test set. A sample of 3000 reconstruction problems is generated from the held-out HST and IllustrisTNG data with noise conditions and PSFs similar to the training set.

2.5.1 Goodness of Fit

Figure 2.7 is a cherry picked sample of the reconstruction from the test set. The samples are selected to showcase the wide range of lensing configuration that our approach can successfully

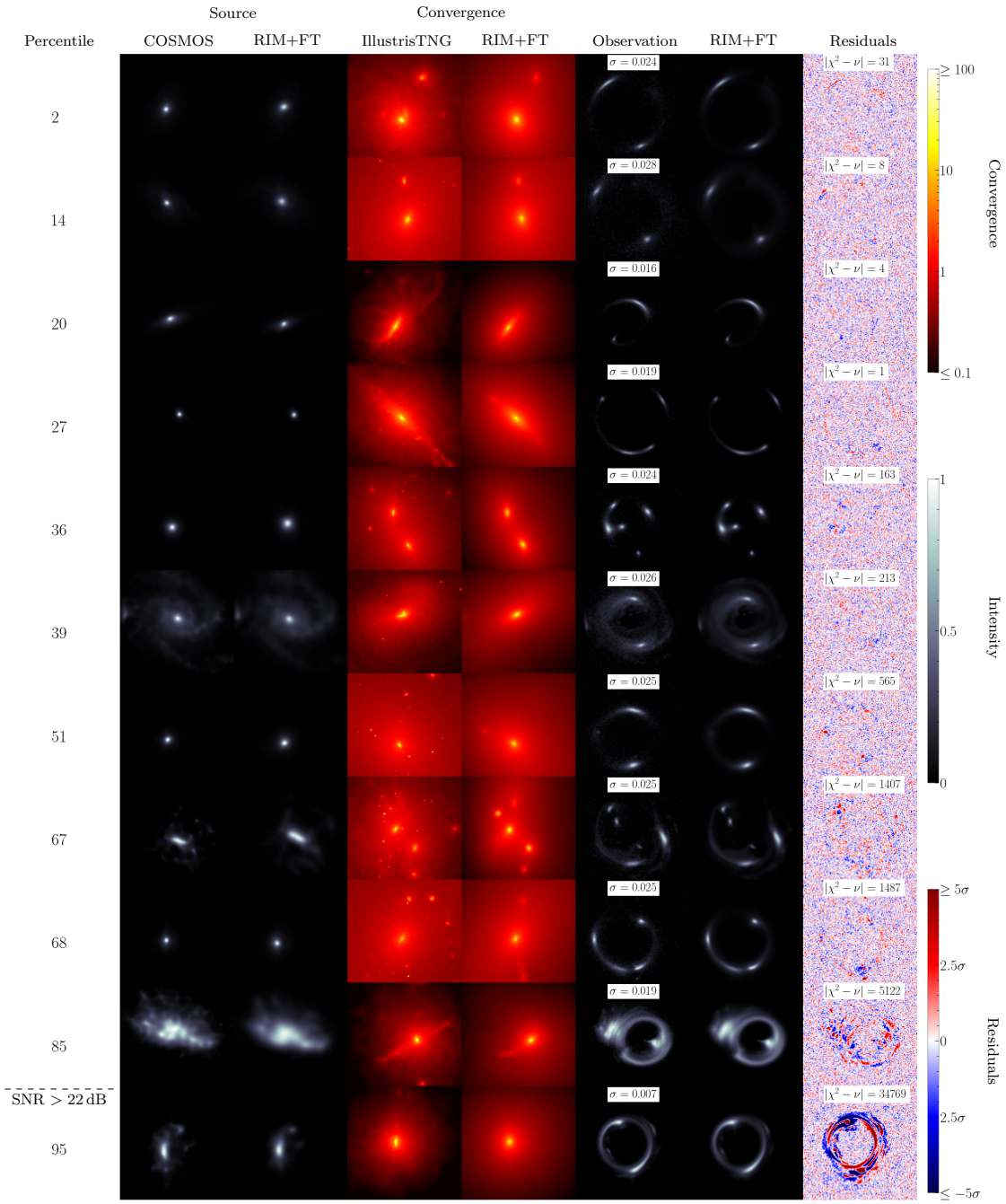

FIGURE 2.7 – Cherry-picked sample of the fine-tuned RIM reconstructions on a test set of 3000 examples. Examples are ordered from the best χ^2 (top) to the worst (bottom). The percentile rank of each example is in the leftmost column. The last example shown has SNR above the threshold defined in Figure 2.9.

solve at high SNR. We made a point to select mostly examples that have a lot of structure in their convergence map to distinguish our approach from existing analytical methods. We did not make an emphasis in selecting complicated sources since we judge that free-form reconstruction of the source is essentially a solved problem. Indeed, many methods can reconstruct free-form sources once the

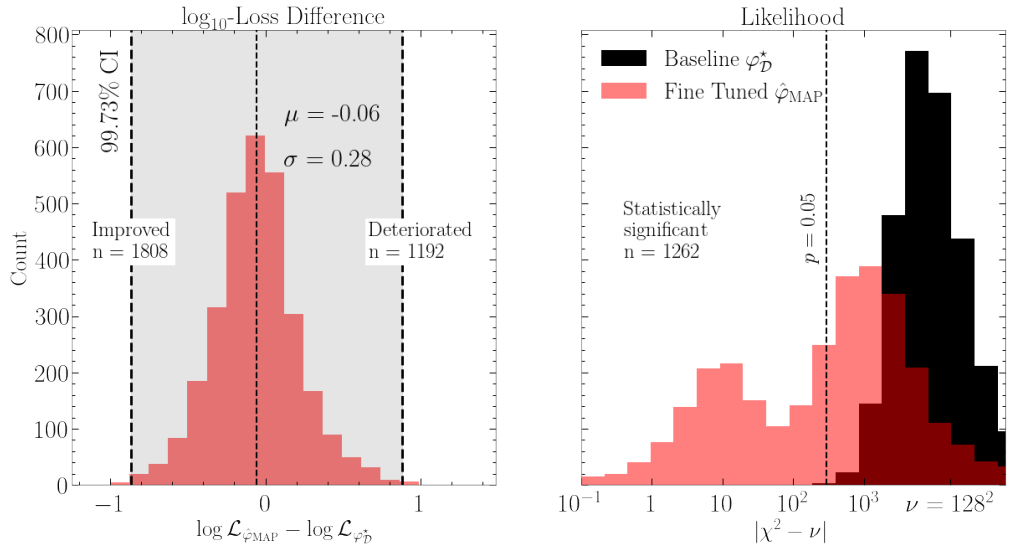

FIGURE 2.8 – Distribution of the goodness of fit for the baseline and fine-tuned network (right panel), as well as log-loss difference between the two network for a given example in the test set (left panel).

convergence map is known or well constrained (Warren and Dye, 2003; Suyu et al., 2006; Vegetti and Koopmans, 2009; Birrer and Amara, 2018; Morningstar et al., 2019; Galan et al., 2021; Karchev et al., 2022; Mishra-Sharma and Yang, 2022).

To offset our selection bias, we selected samples in different percentile from the test set rank ordered by the χ^2 metric. We also show a randomly selected sample from the test set in Figure C.2.

Figure 2.8 shows a comparison between the goodness of fit of the baseline model and the fine-tuned prediction. Since we empirically observe that the distribution of the loss on the test set (and the training set) follows a log-normal distribution, we find much more informative to look at the log-loss distribution to extract information about the fine-tuning procedure. The left panel of Figure 2.8 shows the distribution of the log-loss difference between the fine-tuned prediction and the baseline model. This distribution shows that the fine-tuning procedure loss is constrained within ~ 1 order of magnitude of the original loss with a 99.73 % probability. We find that the log-loss difference has a scatter of $\sigma = 0.28$, which is smaller than the scatter of the baseline log-loss over the entire test set $\sigma(\log \mathcal{L}(\varphi^\star_{\mathcal{D}})) = 0.36$ reported in Table 2.4. We note that this metric is not optimized during fine-tuning, and is only computed as an oracle metric — still, the fine-tuning procedure does not significantly deteriorate or improve the loss of the baseline prediction in average. We report the first 2 moments of the loss log-normal distribution for the baseline and the fine-tuned reconstructions in Table 2.4 in order to compare explicitly the loss distributions. As can be seen in this table, there is no significant difference between the two distribution. This statement can be proven for the measured mean values — $\mu(\log \mathcal{L}(\hat{\varphi}_{\mathrm{MAP}})) = \mu(\log \mathcal{L}(\varphi^\star_{\mathcal{D}}))$ — using the two-sided normal p-value test (Casella and Berger, 2001), which we find satisfy the null hypothesis with $p = 0.87288$ ($Z = -0.16$). All those observations support our claim that EWC regularisation preserves the prior learned during

TABLE 2.4 – $\log_{10}$-normal moments of the loss on the test set

Model	$\mu(\log \mathcal{L})$	$\sigma(\log \mathcal{L})$
Baseline ($\varphi^{\star}_{\mathcal{D}}$)	-1.96	0.36
Fine-tuned ($\hat{\varphi}_{\text{MAP}}$)	-2.02	0.37

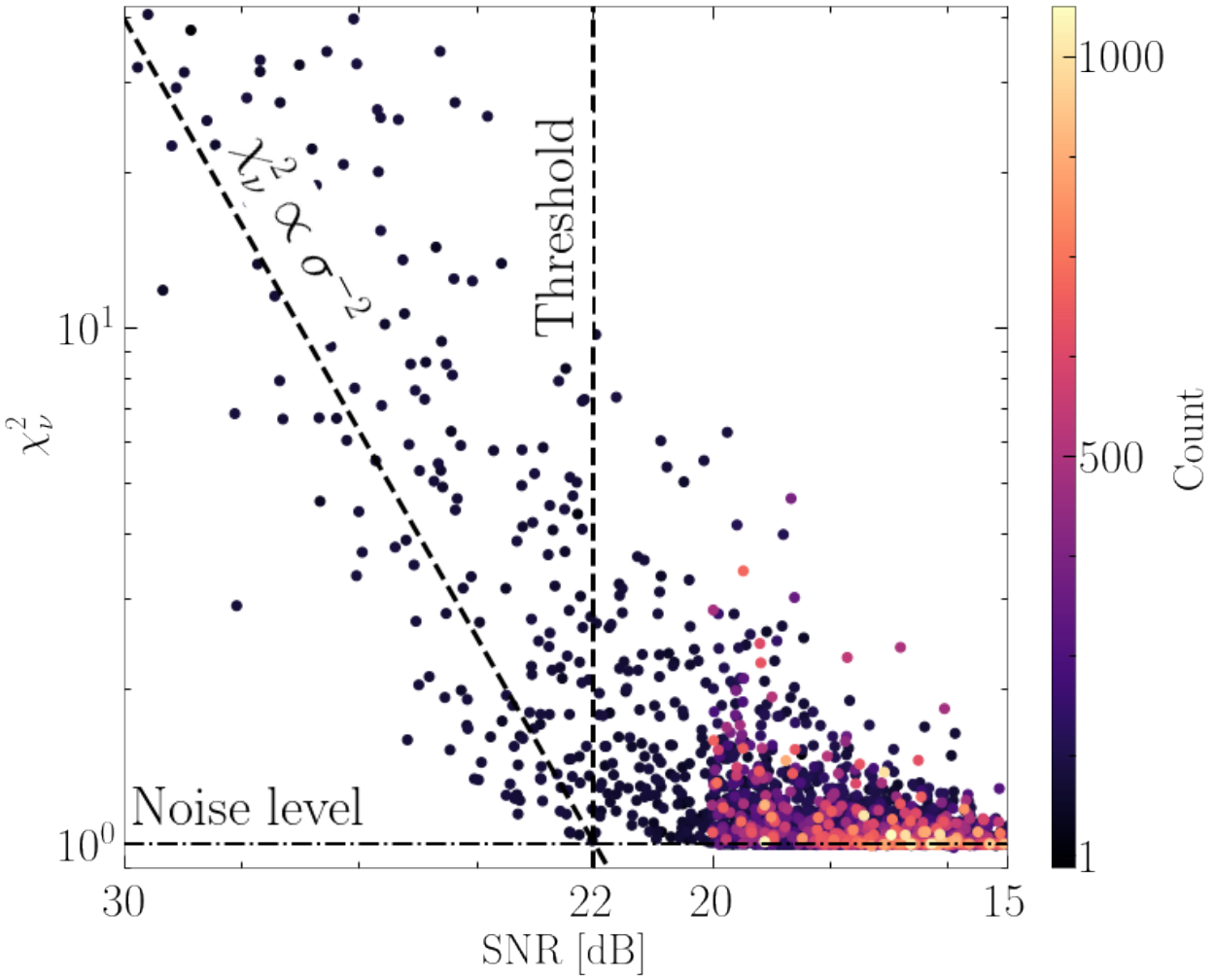

FIGURE 2.9 – Goodness of fit as a function of SNR shows a threshold behavior where our method reaches its limit.

pretraining, or at least that it preserves the surrogate measures of the prior we reported.

The right panel of Figure 2.8 shows that the χ^2 of the reconstruction is improved substantially compared to the baseline model. The χ^2 is directly optimized by the fine-tuning procedure (equation (2.9)). Thus, this improvement is to be expected. Out of the 3000 reconstructions, 1262 reach $|\chi^2 - \nu| \leq 296$, which is the criteria that satisfy the null hypothesis for the number of degrees of freedom $\nu = 128^2$. For those reconstructions that do not satisfy this criteria, we still observe a significant improvement in the distribution of goodness of fit compared to the baseline (black histogram in Figure 2.8).

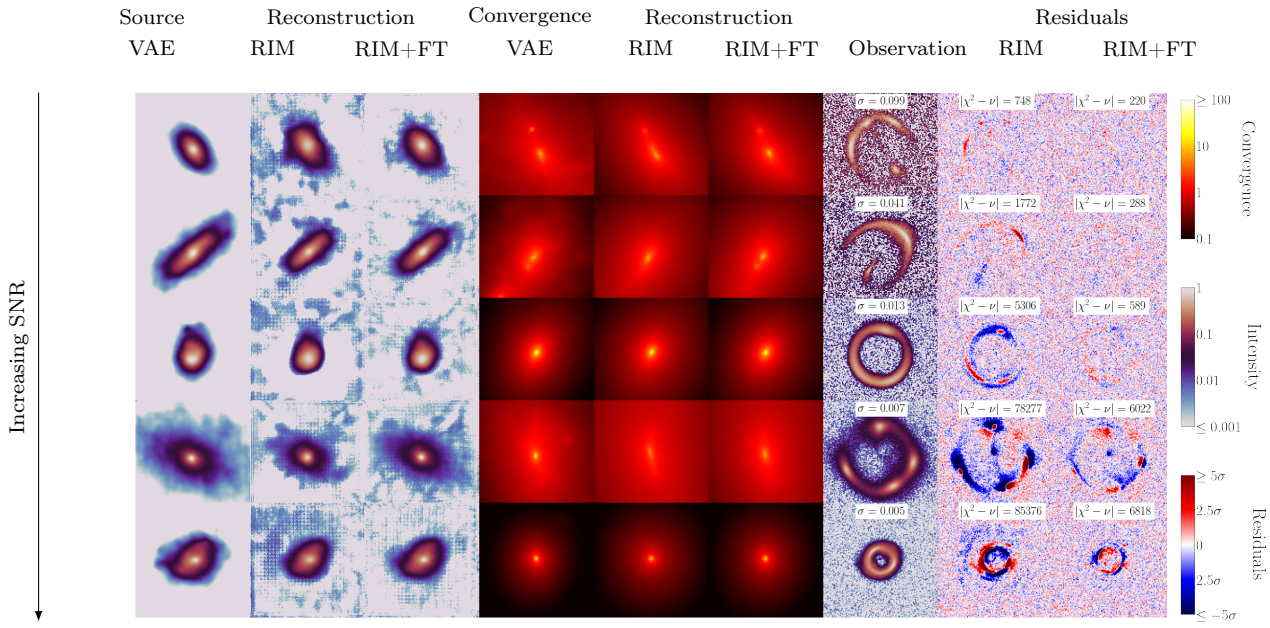

FIGURE 2.10 – Comparison between baseline (RIM) and fine-tuned (RIM+FT) reconstructions for VAE generated gravitational lensing systems. From top to bottom, we increase SNR. The first 2 rows have noise level reconstruction, while the last 3 row show significant improvement over the baseline. The intensity color scale is chosen to show the reconstruction down to the third decimal place, where the baseline prediction breaks down.

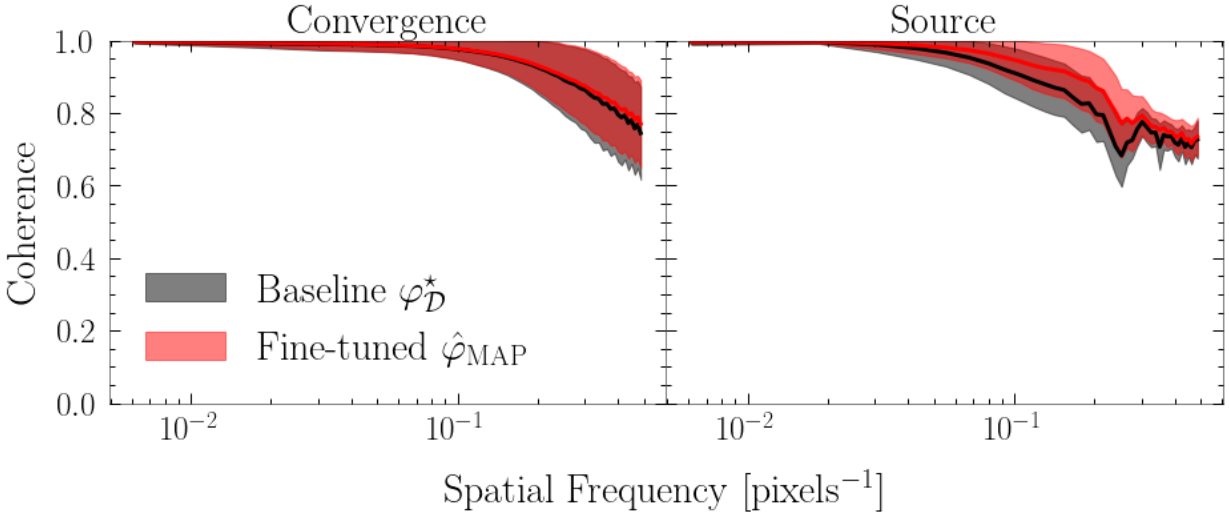

FIGURE 2.11 – Statistics of the coherence spectrum on the test set. The solid line is the average coherence. The transparent region is the 68% confidence interval. The fine-tuning procedure yields a noticeable improvement on the coherence of the source at all frequencies.

To understand why some of the reconstructions do not reach noise level, we report the joint distribution of the reduced chi squared and the SNR in Figure 2.9. Two behaviors can be identified. For SNR below a certain threshold, the goodness of fit of the fine-tuned model is essentially

flat, with a certain scatter, around the noise level. This scatter increases as a function of SNR, which reflects the fact that high-SNR reconstructions are much more difficult to perform than low-SNR observations (for any model or algorithm, not just ours) because they require very precise assumptions about the solution. Near the threshold, this scatter is at its peak. This is SNR regime ($18\,\mathrm{dB} \lesssim \mathrm{SNR} \leq 22\,\mathrm{dB}$) is where most of the fine-tuned reconstructions which do not reach noise level are found ($|\chi^2 - \nu| > 296$). For SNR above the threshold, the goodness of fit follows the trend $\chi^2 \propto \sigma^{-2}$, which means the reconstructions have stopped improving on par with the noise level. We define an optimistic threshold at the data point with largest SNR value which still have statistically significant residuals ($|\chi^2 - \nu| < 296$). We find this threshold value to be 22 dB based on the test set.

To understand this threshold behavior, we inspected some source reconstructions with intensity plotted in log scale in Figure 2.10. As can seen in this figure, the source pixel intensities of the RIM prediction in the regime $\mathrm{SNR} \gtrsim 22\,\mathrm{dB}$, meaning the third decimal place of the reconstructed intensity values shown in pale blue, do not have a coherent structure and correspond mostly to numerical errors, even for the fine-tuned network. Reducing the noise level in the observed lensed image does not change this behavior, which indicates that this is a limitation of the neural network itself. Thus, we argue that the threshold can be explained in part because the inductive biases learned by the neural network during training prohibits the fine-tuning procedure to reconstruct the signal past the threshold value via the EWC prior term. Reducing the strength of the prior will not improve the reconstruction since fine-tuning with $\lambda \ll 2 \times 10^5$ will quickly produce noise leakage in the source reconstruction.

Since the threshold value found is large enough to apply our method to most known gravitational lens images (Bolton et al., 2008; Shu et al., 2017), we leave further investigations to future works. We note that we purposefully ignored the low SNR regime since analytical methods perform well in that regime — at low SNR, simple Sersic (Sérsic, 1963) and SIE (Keeton, 2001) models will often give good fit to the data since most of the information about the source and the convergence morphology is lost.

2.5.2 Quality of the Reconstructions

In addition to a visual inspection of the reconstructed sources and convergences, we compute the coherence spectrum to quantitatively assess the quality of the reconstructions

$$\gamma(k) = \frac{P_{12}(k)}{\sqrt{P_{11}(k) P_{22}(k)}}\,. \tag{2.14}$$

$P_{ij}(k)$ is the cross power spectrum of images i and j at the wavenumber k. Figure 2.11 shows the mean value and the 68% inclusion interval of $\gamma(k)$ for the convergence and source maps in a test set of 3000 examples. The fine-tuning procedure, shown in red, is able to improve significantly the coherence of the baseline background source, shown in black, at all scales. The coherence spectrum of the convergence remains unchanged by the fine-tuning procedure. Still, we note that many examples

in the dataset showcase significant improvement which we illustrate in Figure 2.1.

2.6 Conclusion

The results obtained here demonstrate the effectiveness of machine learning methods for inferring pixelated maps of the distribution of mass in lensing galaxies. Since this is a heavily underconstrained problem, stringent priors are needed to avoid overfitting the data, a task that has traditionally been difficult to accomplish (e.g., Saha and Williams, 1997). The model proposed here can implicitly learn these priors from a set of training data.

The flexible and expressive form of the reconstructions means that, in principle, any lensing system (e.g., a single simple galaxy, or a group of complex galaxies) could be analyzed by this model, without any need for pre-determining the model parameterization. This is of high value given the diversity of observed lensing systems, and their relevance for constraining astrophysical and cosmological parameters.

Another significant advantage of the proposed method is that the true physical model (a ray-tracing simulation code here) is used at every iteration to calculate the likelihood of the RIM predictions given the observations, which significantly helps with the interpretability of the model and the obtained results.

Finally, perhaps the most important limitation of the method is the fact that, in its current form, the model only provides point estimates of the parameters of interest. Quantifying the posteriors of such high-dimensional data may require a generative process. In future work, we will explore the possibility of sampling from latent variables within the RIM to obtain samples from the posteriors.

Software and data

The source code, as well as the various scripts and parameters used to produce the model and results is available as open-source software under the package `Censai` [3]. The model parameters, as well as convergence maps used to train these models and the test set examples and reconstructions results are also available as open-source datasets hosted by Zenodo [4]. This research made use of `Tensorflow` (Abadi et al., 2015), `Tensorflow-Probability` (Dillon et al., 2017), `Numpy` (Harris et al., 2020), `Scipy` (Virtanen et al., 2020), `Matplotlib` (Hunter, 2007), `Scikit-image` (Van der Walt et al., 2014), `IPython` (Pérez and Granger, 2007), `Pandas` (Wes McKinney, 2010), `Scikit-learn` (Pedregosa et al., 2011), `Astropy` (Astropy Collaboration et al., 2013, 2018) and `GalSim` (Rowe et al., 2015).

3. https://github.com/AlexandreAdam/Censai
4. https://doi.org/10.5281/zenodo.6555463

Acknowledgements

We thank Ronan Legin for fruitful discussion and insights about training the neural network and comments about the manuscript. This research was supported by the Schmidt Futures Foundation. The work was also enabled in part by computational resources provided by Calcul Quebec, Compute Canada and the Digital Research Alliance of Canada. Y.H. and L.P. acknowledge support from the National Sciences and Engineering Council of Canada grant RGPIN-2020-05102, the Fonds de recherche du Québec grant 2022-NC-301305, and the Canada Research Chairs Program. A.A. was supported by an IVADO scholarship.

Bibliographie

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow : Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

H. M. Abdelsalam, P. Saha, and L. L. R. Williams. Nonparametric Reconstruction of Abell 2218 from Combined Weak and Strong Lensing. AJ, 116(4) :1541–1552, Oct. 1998a. doi : 10.1086/300546.

H. M. Abdelsalam, P. Saha, and L. L. R. Williams. Non-parametric reconstruction of cluster mass distribution from strong lensing : modelling Abell 370. MNRAS, 294 :734–746, Mar. 1998b. doi : 10.1046/j.1365-8711.1998.01356.x.

N. Anau Montel, A. Coogan, C. Correa, K. Karchev, and C. Weniger. Estimating the warm dark matter mass from strong lensing images with truncated marginal neural ratio estimation. *arXiv e-prints*, art. arXiv :2205.09126, May 2022.

M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. de Freitas. Learning to learn by gradient descent by gradient descent. *arXiv e-prints*, art. arXiv :1606.04474, June 2016.

Astropy Collaboration, T. P. Robitaille, E. J. Tollerud, P. Greenfield, M. Droettboom, E. Bray, T. Aldcroft, M. Davis, A. Ginsburg, A. M. Price-Whelan, W. E. Kerzendorf, A. Conley, N. Crighton, K. Barbary, D. Muna, H. Ferguson, F. Grollier, M. M. Parikh, P. H. Nair, H. M. Unther, C. Deil, J. Woillez, S. Conseil, R. Kramer, J. E. H. Turner, L. Singer, R. Fox, B. A. Weaver, V. Zabalza, Z. I. Edwards, K. Azalee Bostroem, D. J. Burke, A. R. Casey, S. M. Crawford, N. Dencheva, J. Ely, T. Jenness, K. Labrie, P. L. Lim, F. Pierfederici, A. Pontzen, A. Ptak, B. Refsdal, M. Servillat, and O. Streicher. Astropy : A community Python package for astronomy. A&A, 558 :A33, Oct. 2013. doi : 10.1051/0004-6361/201322068.

Astropy Collaboration, A. M. Price-Whelan, B. M. Sipőcz, H. M. Günther, P. L. Lim, S. M. Crawford, S. Conseil, D. L. Shupe, M. W. Craig, N. Dencheva, A. Ginsburg, J. T. Vand erPlas, L. D. Bradley, D. Pérez-Suárez, M. de Val-Borro, T. L. Aldcroft, K. L. Cruz, T. P. Robitaille, E. J. Tollerud, C. Ardelean, T. Babej, Y. P. Bach, M. Bachetti, A. V. Bakanov, S. P. Bamford, G. Barentsen, P. Barmby, A. Baumbach, K. L. Berry, F. Biscani, M. Boquien, K. A. Bostroem, L. G. Bouma, G. B. Brammer, E. M. Bray, H. Breytenbach, H. Buddelmeijer, D. J. Burke, G. Calderone, J. L. Cano Rodríguez, M. Cara, J. V. M. Cardoso, S. Cheedella, Y. Copin, L. Corrales, D. Crichton, D. D'Avella, C. Deil, É. Depagne, J. P. Dietrich, A. Donath, M. Droettboom, N. Earl, T. Erben, S. Fabbro, L. A. Ferreira, T. Finethy, R. T. Fox, L. H. Garrison, S. L. J. Gibbons, D. A. Goldstein, R. Gommers, J. P. Greco, P. Greenfield, A. M. Groener, F. Grollier, A. Hagen, P. Hirst, D. Homeier, A. J. Horton, G. Hosseinzadeh, L. Hu, J. S. Hunkeler, Ž. Ivezić, A. Jain, T. Jenness, G. Kanarek, S. Kendrew, N. S. Kern, W. E. Kerzendorf, A. Khvalko, J. King, D. Kirkby, A. M. Kulkarni, A. Kumar, A. Lee, D. Lenz, S. P. Littlefair, Z. Ma, D. M. Macleod, M. Mastropietro, C. McCully, S. Montagnac, B. M. Morris, M. Mueller, S. J. Mumford, D. Muna, N. A. Murphy, S. Nelson, G. H.

Nguyen, J. P. Ninan, M. Nöthe, S. Ogaz, S. Oh, J. K. Parejko, N. Parley, S. Pascual, R. Patil, A. A. Patil, A. L. Plunkett, J. X. Prochaska, T. Rastogi, V. Reddy Janga, J. Sabater, P. Sakurikar, M. Seifert, L. E. Sherbert, H. Sherwood-Taylor, A. Y. Shih, J. Sick, M. T. Silbiger, S. Singanamalla, L. P. Singer, P. H. Sladen, K. A. Sooley, S. Sornarajah, O. Streicher, P. Teuben, S. W. Thomas, G. R. Tremblay, J. E. H. Turner, V. Terrón, M. H. van Kerkwijk, A. de la Vega, L. L. Watkins, B. A. Weaver, J. B. Whitmore, J. Woillez, V. Zabalza, and Astropy Contributors. The Astropy Project : Building an Open-science Project and Status of the v2.0 Core Package. AJ, 156(3) :123, Sept. 2018. doi : 10.3847/1538-3881/aabc4f.

K. E. Atkinson. *An Introduction to Numerical Analysis*, chapter 6, pages 341–357. John Wiley & Sons, New York, second edition, 1989. ISBN 0471500232. URL `http://www.worldcat.org/isbn/0471500232`.

D. Aubert, A. Amara, and R. B. Metcalf. Smooth Particle Lensing. MNRAS, 376(1) :113–124, Mar. 2007. doi : 10.1111/j.1365-2966.2006.11296.x.

M. W. Auger, T. Treu, A. S. Bolton, R. Gavazzi, L. V. E. Koopmans, P. J. Marshall, L. A. Moustakas, and S. Burles. The Sloan Lens ACS Survey. X. Stellar, Dynamical, and Total Mass Correlations of Massive Early-type Galaxies. ApJ, 724(1) :511–525, Nov. 2010. doi : 10.1088/0004-637X/724/1/511.

M. Barnabè, O. Czoske, L. V. E. Koopmans, T. Treu, A. S. Bolton, and R. Gavazzi. Two-dimensional kinematics of SLACS lenses - II. Combined lensing and dynamics analysis of early-type galaxies at z = 0.08-0.33. MNRAS, 399 (1) :21–36, Oct. 2009. doi : 10.1111/j.1365-2966.2009.14941.x.

M. Bartelmann, R. Narayan, S. Seitz, and P. Schneider. Maximum-likelihood Cluster Reconstruction. ApJ, 464 : L115, June 1996. doi : 10.1086/310114.

F. Bellagamba, N. Tessore, and R. B. Metcalf. Zooming into the Cosmic Horseshoe : new insights on the lens profile and the source shape. *Monthly Notices of the Royal Astronomical Society*, 464(4) :4823–4834, 10 2016. ISSN 0035-8711. doi : 10.1093/mnras/stw2726. URL `https://doi.org/10.1093/mnras/stw2726`.

V. Belokurov, N. W. Evans, A. Moiseev, L. J. King, P. C. Hewett, M. Pettini, L. Wyrzykowski, R. G. McMahon, M. C. Smith, G. Gilmore, S. F. Sanchez, A. Udalski, S. Koposov, D. B. Zucker, and C. J. Walcher. The Cosmic Horseshoe : Discovery of an Einstein Ring around a Giant Luminous Red Galaxy. ApJ, 671(1) :L9–L12, Dec. 2007. doi : 10.1086/524948.

Y. Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1) :1–127, jan 2009. ISSN 1935-8237. doi : 10.1561/2200000006. URL `https://doi.org/10.1561/2200000006`.

S. Birrer and A. Amara. lenstronomy : Multi-purpose gravitational lens modelling software package. *Physics of the Dark Universe*, 22 :189–201, Dec. 2018. doi : 10.1016/j.dark.2018.11.002.

S. Birrer, A. Amara, and A. Refregier. Gravitational Lens Modeling with Basis Sets. ApJ, 813(2) :102, Nov. 2015. doi : 10.1088/0004-637X/813/2/102.

S. Birrer, T. Treu, C. E. Rusu, V. Bonvin, C. D. Fassnacht, J. H. H. Chan, A. Agnello, A. J. Shajib, G. C. F. Chen, M. Auger, F. Courbin, S. Hilbert, D. Sluse, S. H. Suyu, K. C. Wong, P. Marshall, B. C. Lemaux, and G. Meylan. H0LiCOW - IX. Cosmographic analysis of the doubly imaged quasar SDSS 1206+4332 and a new measurement of the Hubble constant. MNRAS, 484(4) :4726–4753, Apr. 2019. doi : 10.1093/mnras/stz200.

A. S. Bolton, S. Burles, L. V. E. Koopmans, T. Treu, R. Gavazzi, L. A. Moustakas, R. Wayth, and D. J. Schlegel. The Sloan Lens ACS Survey. V. The Full ACS Strong-Lens Sample. ApJ, 682(2) :964–984, Aug. 2008. doi : 10.1086/589327.

S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating Sentences from a Continuous Space. *arXiv e-prints*, art. arXiv :1511.06349, Nov. 2015.

M. Bradač, P. Schneider, M. Lombardi, and T. Erben. Strong and weak lensing united. I. The combined strong and weak lensing cluster mass reconstruction method. A&A, 437(1) :39–48, July 2005. doi : 10.1051/0004-6361:20042233.

C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in β-VAE. *arXiv e-prints*, art. arXiv :1804.03599, Apr. 2018.

J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*, chapter 2, pages 21–26. John Wiley & Sons, Hoboken, New Jersey, third edition, 2016. ISBN 9781119121503. doi : 10.1002/9781119121534. URL `https://onlinelibrary.wiley.com/doi/book/10.1002/9781119121534`.

M. Cacciato, M. Bartelmann, M. Meneghetti, and L. Moscardini. Combining weak and strong lensing in cluster potential reconstruction. A&A, 458(2) :349–356, Nov. 2006. doi : 10.1051/0004-6361:20054582.

S. Carroll. *Spacetime and Geometry : An Introduction to General Relativity*. Benjamin Cummings, 2003. ISBN 0805387323.

G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, June 2001. ISBN 0534243126.

J. Cheng, M. P. Wiesner, E.-H. Peng, W. Cui, J. R. Peterson, and G. Li. Adaptive Grid Lens Modeling of the Cosmic Horseshoe Using Hubble Space Telescope Imaging. ApJ, 872(2) :185, Feb. 2019. doi : 10.3847/1538-4357/ab0029.

K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv e-prints*, art. arXiv :1406.1078, June 2014.

D. Coe, E. Fuselier, N. Benítez, T. Broadhurst, B. Frye, and H. Ford. LensPerfect : Gravitational Lens Mass Map Reconstructions Yielding Exact Reproduction of All Multiple Images. ApJ, 681(2) :814–830, July 2008. doi : 10.1086/588250.

J. P. Coles, J. I. Read, and P. Saha. Gravitational lens recovery with GLASS : measuring the mass profile and shape of a lens. MNRAS, 445(3) :2181–2197, Dec. 2014. doi : 10.1093/mnras/stu1781.

A. Coogan, K. Karchev, and C. Weniger. Targeted Likelihood-Free Inference of Dark Matter Substructure in Strongly-Lensed Galaxies. *arXiv e-prints*, art. arXiv :2010.07032, Oct. 2020.

T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

H. Cramér. *Mathematical methods of statistics*, volume 9. Princeton University Press, Princeton, NJ, 1946.

G. V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2 :303–314, 1989.

N. Dalal and C. S. Kochanek. Direct Detection of Cold Dark Matter Substructure. ApJ, 572(1) :25–33, June 2002. doi : 10.1086/340303.

S. Deb, D. M. Goldberg, and V. J. Ramdass. Reconstruction of Cluster Masses Using Particle Based Lensing. I. Application to Weak Lensing. ApJ, 687(1) :39–49, Nov. 2008. doi : 10.1086/590544.

S. Deb, A. Morandi, K. Pedersen, S. Riemer-Sorensen, D. M. Goldberg, and H. Dahle. Mass Reconstruction using Particle Based Lensing II : Quantifying substructure with Strong+Weak lensing and X-rays. *arXiv e-prints*, art. arXiv :1201.3636, Jan. 2012.

J. M. Diego, P. Protopapas, H. B. Sandvik, and M. Tegmark. Non-parametric inversion of strong lensing systems. MNRAS, 360(2) :477–491, June 2005. doi : 10.1111/j.1365-2966.2005.09021.x.

J. M. Diego, M. Tegmark, P. Protopapas, and H. B. Sandvik. Combined reconstruction of weak and strong lensing data with WSLAP. MNRAS, 375(3) :958–970, Mar. 2007. doi : 10.1111/j.1365-2966.2007.11380.x.

J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, S. Vasudevan, D. Moore, B. Patton, A. Alemi, M. Hoffman, and R. A. Saurous. TensorFlow Distributions. *arXiv e-prints*, art. arXiv :1711.10604, Nov. 2017.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12(null) :2121–2159, jul 2011. ISSN 1532-4435.

A. A. Dutton and T. Treu. The bulge–halo conspiracy in massive elliptical galaxies : implications for the stellar initial mass function and halo response to baryonic processes. *Monthly Notices of the Royal Astronomical Society*, 438 (4) :3594–3602, 01 2014. ISSN 0035-8711. doi : 10.1093/mnras/stt2489. URL `https://doi.org/10.1093/mnras/stt2489`.

A. Einstein. Lens-Like Action of a Star by the Deviation of Light in the Gravitational Field. *Science*, 84(2188) : 506–507, 1936. doi : 10.1126/science.84.2188.506. URL `https://www.science.org/doi/abs/10.1126/science.84.2188.506`.

A. Galan, A. Peel, R. Joseph, F. Courbin, and J. L. Starck. SLITRONOMY : Towards a fully wavelet-based strong lensing inversion technique. A&A, 647 :A176, Mar. 2021. doi : 10.1051/0004-6361/202039363.

A. Ghosh, L. L. R. Williams, and J. Liesenborgs. Free-form grale lens inversion of galaxy clusters with up to 1000 multiple images. MNRAS, 494(3) :3998–4014, May 2020. doi : 10.1093/mnras/staa962.

D. Gilman, S. Birrer, A. Nierenberg, T. Treu, X. Du, and A. Benson. Warm dark matter chills out : constraints on the halo mass function and the free-streaming length of dark matter with eight quadruple-image strong gravitational lenses. MNRAS, 491(4) :6077–6101, Feb. 2020. doi : 10.1093/mnras/stz3480.

D. Gilman, J. Bovy, T. Treu, A. Nierenberg, S. Birrer, A. Benson, and O. Sameie. Strong lensing signatures of self-interacting dark matter in low-mass haloes. MNRAS, 507(2) :2432–2447, Oct. 2021. doi : 10.1093/mnras/stab2335.

Z. Goldfeld and Y. Polyanskiy. The Information Bottleneck Problem and Its Applications in Machine Learning. *arXiv e-prints*, art. arXiv :2004.14941, Apr. 2020.

J. Hadamard. Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13 :49–52, 1902.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825) :357–362, Sept. 2020. doi : 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

Y. D. Hezaveh, N. Dalal, D. P. Marrone, Y.-Y. Mao, W. Morningstar, D. Wen, R. D. Blandford, J. E. Carlstrom, C. D. Fassnacht, G. P. Holder, A. Kemball, P. J. Marshall, N. Murray, L. Perreault Levasseur, J. D. Vieira, and R. H. Wechsler. Detection of Lensing Substructure Using ALMA Observations of the Dusty Galaxy SDP.81. ApJ, 823(1) :37, May 2016. doi : 10.3847/0004-637X/823/1/37.

Y. D. Hezaveh, L. Perreault Levasseur, and P. J. Marshall. Fast automated analysis of strong gravitational lenses with convolutional neural networks. Nature, 548(7669) :555–557, Aug. 2017. doi : 10.1038/nature23463.

I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae : Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

G. Hinton. Neural networks for machine learning. `http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf`, 2012. Accèdé le 2022-07-10.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2) :251–257, 1991. ISSN 0893-6080. doi : https://doi.org/10.1016/0893-6080(91)90009-T. URL `https://www.sciencedirect.com/science/article/pii/089360809190009T`.

J. D. Hunter. Matplotlib : A 2d graphics environment. *Computing in Science & Engineering*, 9(3) :90–95, 2007. doi : 10.1109/MCSE.2007.55.

B. L. James, M. Auger, M. Pettini, D. P. Stark, V. Belokurov, and S. Carniani. Mapping UV properties throughout the Cosmic Horseshoe : lessons from VLT-MUSE. MNRAS, 476(2) :1726–1740, May 2018. doi : 10.1093/mnras/sty315.

M. J. Jee, H. C. Ford, G. D. Illingworth, R. L. White, T. J. Broadhurst, D. A. Coe, G. R. Meurer, A. van der Wel, N. Benítez, J. P. Blakeslee, R. J. Bouwens, L. D. Bradley, R. Demarco, N. L. Homeier, A. R. Martel, and S. Mei. Discovery of a Ringlike Dark Matter Structure in the Core of the Galaxy Cluster Cl 0024+17. ApJ, 661(2) : 728–749, June 2007. doi : 10.1086/517498.

N. Jeffrey, F. Lanusse, O. Lahav, and J.-L. Starck. Deep learning dark matter map reconstructions from DES SV weak lensing data. MNRAS, 492(4) :5023–5029, Mar. 2020. doi : 10.1093/mnras/staa127.

C. Kaae Sønderby, T. Raiko, L. Maaløe, S. Kaae Sønderby, and O. Winther. Ladder Variational Autoencoders. *arXiv e-prints*, art. arXiv :1602.02282, Feb. 2016.

N. Kaiser and G. Squires. Mapping the Dark Matter with Weak Gravitational Lensing. ApJ, 404 :441, Feb. 1993. doi : 10.1086/172297.

K. Karchev, A. Coogan, and C. Weniger. Strong-lensing source reconstruction with variationally optimized Gaussian processes. MNRAS, 512(1) :661–685, May 2022. doi : 10.1093/mnras/stac311.

C. R. Keeton. A Catalog of Mass Models for Gravitational Lensing. *arXiv e-prints*, art. astro-ph/0102341, Feb. 2001.

D. P. Kingma and J. Ba. Adam : A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv :1412.6980, Dec. 2014.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *arXiv e-prints*, art. arXiv :1312.6114, Dec. 2013.

D. P. Kingma and M. Welling. An Introduction to Variational Autoencoders. *arXiv e-prints*, art. arXiv :1906.02691, June 2019.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *arXiv e-prints*, art. arXiv :1612.00796, Dec. 2016.

A. M. Koekemoer, H. Aussel, D. Calzetti, P. Capak, M. Giavalisco, J.-P. Kneib, A. Leauthaud, O. Le Fevre, H. J. McCracken, R. Massey, B. Mobasher, J. Rhodes, N. Scoville, and P. L. Shopbell. The COSMOS Survey : Hubble Space Telescope Advanced Camera for Surveys Observations and Data Processing. *The Astrophysical Journal Supplement Series*, 172(1) :196–202, sep 2007. ISSN 0067-0049. doi : 10.1086/520086.

L. V. E. Koopmans. Gravitational imaging of cold dark matter substructures. MNRAS, 363(4) :1136–1144, Nov. 2005. doi : 10.1111/j.1365-2966.2005.09523.x.

L. V. E. Koopmans, T. Treu, A. S. Bolton, S. Burles, and L. A. Moustakas. The Sloan Lens ACS Survey. III. The Structure and Formation of Early-Type Galaxies and Their Evolution since z ~1. ApJ, 649(2) :599–615, Oct. 2006. doi : 10.1086/505696.

F. Lanusse, J. L. Starck, A. Leonard, and S. Pires. High resolution weak lensing mass mapping combining shear and flexion. A&A, 591 :A2, June 2016. doi : 10.1051/0004-6361/201628278.

F. Lanusse, R. Mandelbaum, S. Ravanbakhsh, C.-L. Li, P. Freeman, and B. Póczos. Deep generative models for galaxy image simulations. MNRAS, 504(4) :5543–5555, July 2021. doi : 10.1093/mnras/stab1214.

A. Leauthaud, R. Massey, J.-P. Kneib, J. Rhodes, D. E. Johnston, P. Capak, C. Heymans, R. S. Ellis, A. M. Koekemoer, O. L. Fèvre, Y. Mellier, A. Réfrégier, A. C. Robin, N. Scoville, L. Tasca, J. E. Taylor, and L. V. Waerbeke. Weak Gravitational Lensing with COSMOS : Galaxy Selection and Shape Measurements. *The Astrophysical Journal Supplement Series*, 172(1) :219, sep 2007. ISSN 0067-0049. doi : 10.1086/516598.

R. Legin, Y. Hezaveh, L. Perreault Levasseur, and B. Wandelt. Simulation-Based Inference of Strong Gravitational Lensing Parameters. *arXiv e-prints*, art. arXiv :2112.05278, Dec. 2021.

P. Lemos, N. Jeffrey, M. Cranmer, S. Ho, and P. Battaglia. Rediscovering orbital mechanics with machine learning. *arXiv e-prints*, art. arXiv :2202.02306, Feb. 2022.

A. Leonard, F. X. Dupé, and J. L. Starck. A compressed sensing approach to 3D weak lensing. A&A, 539 :A85, Mar. 2012. doi : 10.1051/0004-6361/201117642.

N. Li, C. Becker, and S. Dye. The impact of line-of-sight structures on measuring H_0 with strong lensing time delays. MNRAS, 504(2) :2224–2234, June 2021. doi : 10.1093/mnras/stab984.

J. Liesenborgs, S. De Rijcke, and H. Dejonghe. A genetic algorithm for the non-parametric inversion of strong lensing systems. MNRAS, 367(3) :1209–1216, Apr. 2006. doi : 10.1111/j.1365-2966.2006.10040.x.

J. Liesenborgs, S. de Rijcke, H. Dejonghe, and P. Bekaert. Non-parametric inversion of gravitational lensing systems with few images using a multi-objective genetic algorithm. MNRAS, 380(4) :1729–1736, Oct. 2007. doi : 10.1111/j.1365-2966.2007.12236.x.

K. Lønning, P. Putzky, J. J. Sonke, L. Reneman, M. W. Caan, and M. Welling. Recurrent inference machines for reconstructing heterogeneous MRI data. *Medical Image Analysis*, 53 :64–78, apr 2019. ISSN 13618423. doi : 10.1016/j.media.2019.01.005.

R. Mandelbaum, C. Lackner, A. Leauthaud, and B. Rowe. COSMOS real galaxy dataset. *Zenodo*, jan 2012. URL `https://zenodo.org/record/3242143`.

R. Mandelbaum, B. Rowe, J. Bosch, C. Chang, F. Courbin, M. Gill, M. Jarvis, A. Kannawadi, T. Kacprzak, C. Lackner, A. Leauthaud, H. Miyatake, R. Nakajima, J. Rhodes, M. Simet, J. Zuntz, B. Armstrong, S. Bridle, J. Coupon, J. P. Dietrich, M. Gentile, C. Heymans, A. S. Jurling, S. M. Kent, D. Kirkby, D. Margala, R. Massey, P. Melchior, J. Peterson, A. Roodman, and T. Schrabback. The Third Gravitational Lensing Accuracy Testing (GREAT3) Challenge Handbook. *The Astrophysical Journal Supplement Series*, 212(1) :5, apr 2014. ISSN 0067-0049. doi : 10.1088/0067-0049/212/1/5. URL `https://iopscience.iop.org/article/10.1088/0067-0049/212/1/5https://iopscience.iop.org/article/10.1088/0067-0049/212/1/5/meta`.

D. P. Marrone, J. S. Spilker, C. C. Hayward, J. D. Vieira, M. Aravena, M. L. N. Ashby, M. B. Bayliss, M. Béthermin, M. Brodwin, M. S. Bothwell, J. E. Carlstrom, S. C. Chapman, C.-C. Chen, T. M. Crawford, D. J. M. Cunningham, C. De Breuck, C. D. Fassnacht, A. H. Gonzalez, T. R. Greve, Y. D. Hezaveh, K. Lacaille, K. C. Litke, S. Lower, J. Ma, M. Malkan, T. B. Miller, W. R. Morningstar, E. J. Murphy, D. Narayanan, K. A. Phadke, K. M. Rotermund,

J. Sreevani, B. Stalder, A. A. Stark, M. L. Strandet, M. Tang, and A. Weiß. Galaxy growth in a massive halo in the first billion years of cosmic history. Nature, 553(7686) :51–54, Jan. 2018. doi : 10.1038/nature24629.

P. J. Marshall. Maximum-entropy reconstruction of the distribution of mass in cluster MS1054-03 from weak lensing data. In D. M. Neumann and J. T. V. Tran, editors, *Clusters of Galaxies and the High Redshift Universe Observed in X-rays*, page 47, Jan. 2001.

R. Massey, J. Rhodes, R. Ellis, N. Scoville, A. Leauthaud, A. Finoguenov, P. Capak, D. Bacon, H. Aussel, J.-P. Kneib, A. Koekemoer, H. McCracken, B. Mobasher, S. Pires, A. Refregier, S. Sasaki, J.-L. Starck, Y. Taniguchi, A. Taylor, and J. Taylor. Dark matter maps reveal cosmic scaffolding. Nature, 445(7125) :286–290, Jan. 2007. doi : 10.1038/nature05497.

M. McCloskey and N. J. Cohen. Catastrophic Interference in Connectionist Networks : The Sequential Learning Problem. In G. H. Bower, editor, *Psychology of Learning and Motivation - Advances in Research and Theory*, volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi : 10.1016/S0079-7421(08)60536-8. URL `https://www.sciencedirect.com/science/article/pii/S0079742108605368`.

M. Meneghetti. *Introduction to Gravitational Lensing*. Springer Cham, 2013. doi : 10.1007/978-3-030-73582-1.

J. Merten. Mesh-free free-form lensing - I. Methodology and application to mass reconstruction. MNRAS, 461(3) : 2328–2345, Sept. 2016. doi : 10.1093/mnras/stw1413.

J. Merten, M. Cacciato, M. Meneghetti, C. Mignone, and M. Bartelmann. Combining weak and strong cluster lensing : applications to simulations and MS 2137. A&A, 500(2) :681–691, June 2009. doi : 10.1051/0004-6361/200810372.

S. Mishra-Sharma and G. Yang. Strong lensing source reconstruction using continuous neural fields, 2022. URL `https://arxiv.org/abs/2206.14820`.

C. Modi, F. Lanusse, U. Seljak, D. N. Spergel, and L. Perreault-Levasseur. CosmicRIM : Reconstructing Early Universe by Combining Differentiable Simulations with Recurrent Inference Machines. *arXiv e-prints*, art. arXiv :2104.12864, Apr. 2021.

W. R. Morningstar, Y. D. Hezaveh, L. P. Levasseur, R. D. Blandford, P. J. Marshall, P. Putzky, and R. H. Wechsler. Analyzing Interferometric Observations of Strong Gravitational Lenses with Recurrent and Convolutional Neural Networks. *arXiv e-prints*, 2018.

W. R. Morningstar, L. P. Levasseur, Y. D. Hezaveh, R. Blandford, P. Marshall, P. Putzky, T. D. Rueter, R. Wechsler, and M. Welling. Data-driven Reconstruction of Gravitationally Lensed Galaxies Using Recurrent Inference Machines. *The Astrophysical Journal*, 883(1) :14, 2019. ISSN 1538-4357. doi : 10.3847/1538-4357/ab35d7.

D. Nelson, V. Springel, A. Pillepich, V. Rodriguez-Gomez, P. Torrey, S. Genel, M. Vogelsberger, R. Pakmor, F. Marinacci, R. Weinberger, L. Kelley, M. Lovell, B. Diemer, and L. Hernquist. The IllustrisTNG simulations : public data release. MNRAS, 6(1), 2019. ISSN 2197-7909. doi : 10.1186/s40668-019-0028-x. URL `www.tng-project.org/data`.

Y. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269 :543–547, 1983.

J. W. Nightingale, S. Dye, and R. J. Massey. AutoLens : automated modeling of a strong lens's light, mass, and source. MNRAS, 478(4) :4738–4784, Aug. 2018. doi : 10.1093/mnras/sty1264.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22 : 1345–1359, 2010.

J. W. Park, S. Wagner-Carena, S. Birrer, P. J. Marshall, J. Y.-Y. Lin, A. Roodman, and LSST Dark Energy Science Collaboration. Large-scale Gravitational Lens Modeling with Bayesian Neural Networks for Accurate and Precise Inference of the Hubble Constant. ApJ, 910(1) :39, Mar. 2021. doi : 10.3847/1538-4357/abdfc4.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

F. Pérez and B. E. Granger. IPython : a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3) :21–29, May 2007. ISSN 1521-9615. doi : 10.1109/MCSE.2007.53. URL `https://ipython.org`.

L. Perreault Levasseur, Y. D. Hezaveh, and R. H. Wechsler. Uncertainties in Parameters Estimated with Neural Networks : Application to Strong Gravitational Lensing. ApJ, 850(1) :L7, Nov. 2017. doi : 10.3847/2041-8213/aa9704.

Planck Collaboration. Planck 2018 results. VI. Cosmological parameters. A&A, 641 :A6, Sept. 2020. doi : 10.1051/0004-6361/201833910.

P. Putzky and M. Welling. Recurrent Inference Machines for Solving Inverse Problems. *arXiv e-prints*, 2017.

C. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin fo the Calcutta Mathematical Society*, 1945.

R. Ratcliff. Connectionist models of recognition memory : Constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2) :285–308, 1990. doi : 10.1037/0033-295X.97.2.285.

S. Rau, S. Vegetti, and S. D. White. The effect of particle noise in N-body simulations of gravitational lensing. *Monthly Notices of the Royal Astronomical Society*, 430(3) :2232–2248, apr 2013. ISSN 13652966. doi : 10.1093/mnras/stt043.

B. Remy, F. Lanusse, N. Jeffrey, J. Liu, J.-L. Starck, K. Osato, and T. Schrabback. Probabilistic Mass Mapping with Neural Score Estimation. *arXiv e-prints*, art. arXiv :2201.05561, Jan. 2022.

M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning : the rprop algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591 vol.1, 1993. doi : 10.1109/ICNN.1993.298623.

F. Rizzo, S. Vegetti, D. Powell, F. Fraternali, J. P. McKean, H. R. Stacey, and S. D. M. White. A dynamically cold disk galaxy in the early Universe. Nature, 584(7820) :201–204, Aug. 2020. doi : 10.1038/s41586-020-2572-6.

O. Ronneberger, P. Fischer, and T. Brox. U-Net : Convolutional Networks for Biomedical Image Segmentation. *arXiv e-prints*, art. arXiv :1505.04597, May 2015.

B. T. Rowe, M. Jarvis, R. Mandelbaum, G. M. Bernstein, J. Bosch, M. Simet, J. E. Meyers, T. Kacprzak, R. Nakajima, J. Zuntz, H. Miyatake, J. P. Dietrich, R. Armstrong, P. Melchior, and M. S. Gill. GalSim : The modular galaxy image simulation toolkit. *Astronomy and Computing*, 10 :121–150, apr 2015. ISSN 22131337. doi : 10.1016/j.ascom.2015.02.002.

B. T. P. Rowe, M. Jarvis, R. Mandelbaum, G. M. Bernstein, J. Bosch, M. Simet, J. E. Meyers, T. Kacprzak, R. Nakajima, J. Zuntz, H. Miyatake, J. P. Dietrich, R. Armstrong, P. Melchior, and M. S. S. Gill. GALSIM : The modular galaxy image simulation toolkit. *Astronomy and Computing*, 10 :121–150, Apr. 2015. doi : 10.1016/j.ascom.2015.02.002.

C. E. Rusu, C. D. Fassnacht, D. Sluse, S. Hilbert, K. C. Wong, K.-H. Huang, S. H. Suyu, T. E. Collett, P. J. Marshall, T. Treu, and L. V. E. Koopmans. H0LiCOW - III. Quantifying the effect of mass along the line of sight to the gravitational lens HE 0435-1223 through weighted galaxy counts. MNRAS, 467(4) :4220–4242, June 2017. doi : 10.1093/mnras/stx285.

C. E. Rusu, K. C. Wong, V. Bonvin, D. Sluse, S. H. Suyu, C. D. Fassnacht, J. H. H. Chan, S. Hilbert, M. W. Auger, A. Sonnenfeld, S. Birrer, F. Courbin, T. Treu, G. C. F. Chen, A. Halkola, L. V. E. Koopmans, P. J. Marshall, and A. J. Shajib. H0LiCOW XII. Lens mass model of WFI2033-4723 and blind measurement of its time-delay distance and H_0. MNRAS, 498(1) :1440–1468, Oct. 2020. doi : 10.1093/mnras/stz3451.

P. Saha and L. L. R. Williams. Non-parametric reconstruction of the galaxy lens in PG 1115+080. MNRAS, 292(1) : 148–156, Nov. 1997. doi : 10.1093/mnras/292.1.148.

P. Saha and L. L. R. Williams. A Portable Modeler of Lensed Quasars. AJ, 127(5) :2604–2616, May 2004. doi : 10.1086/383544.

S. Schuldt, G. Chirivì, S. H. Suyu, A. Yıldırım, A. Sonnenfeld, A. Halkola, and G. F. Lewis. Inner dark matter distribution of the Cosmic Horseshoe (J1148+1930) with gravitational lensing and dynamics. A&A, 631 :A40, Nov. 2019. doi : 10.1051/0004-6361/201935042.

S. Schuldt, R. Cañameras, Y. Shu, S. H. Suyu, S. Taubenberger, T. Meinhardt, and L. Leal-Taixé. HOLISMOKES – IX. Neural network inference of strong-lens parameters and uncertainties from ground-based images. *arXiv e-prints*, art. arXiv :2206.11279, June 2022.

N. Scoville, H. Aussel, M. Brusa, P. Capak, C. M. Carollo, M. Elvis, M. Giavalisco, L. Guzzo, G. Hasinger, C. Impey, J.-P. Kneib, O. LeFevre, S. J. Lilly, B. Mobasher, A. Renzini, R. M. Rich, D. B. Sanders, E. Schinnerer, D. Schminovich, P. Shopbell, Y. Taniguchi, and N. D. Tyson. The Cosmic Evolution Survey (COSMOS) : Overview. *The Astrophysical Journal Supplement Series*, 172(1) :1–8, sep 2007. ISSN 0067-0049. doi : 10.1086/516585.

S. Seitz, P. Schneider, and M. Bartelmann. Entropy-regularized maximum-likelihood cluster mass reconstruction. A&A, 337 :325–337, Sept. 1998.

J. L. Sérsic. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina*, 6 :41–43, Feb. 1963.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 :379–423, 1948. URL `http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf`.

Y. Shu, J. R. Brownstein, A. S. Bolton, L. V. E. Koopmans, T. Treu, A. D. Montero-Dorta, M. W. Auger, O. Czoske, R. Gavazzi, P. J. Marshall, and L. A. Moustakas. The Sloan Lens ACS Survey. XIII. Discovery of 40 New Galaxy-scale Strong Lenses. ApJ, 851(1) :48, Dec. 2017. doi : 10.3847/1538-4357/aa9794.

P. Simon, C. Heymans, T. Schrabback, A. N. Taylor, M. E. Gray, L. van Waerbeke, C. Wolf, D. Bacon, M. Barden, A. Böhm, B. Häußler, K. Jahnke, S. Jogee, E. van Kampen, K. Meisenheimer, and C. Y. Peng. Spatial matter density mapping of the STAGES Abell A901/2 supercluster field with 3D lensing. MNRAS, 419(2) :998–1016, Jan. 2012. doi : 10.1111/j.1365-2966.2011.19760.x.

D. Sluse, A. Sonnenfeld, N. Rumbaugh, C. E. Rusu, C. D. Fassnacht, T. Treu, S. H. Suyu, K. C. Wong, M. W. Auger, V. Bonvin, T. Collett, F. Courbin, S. Hilbert, L. V. E. Koopmans, P. J. Marshall, G. Meylan, C. Spiniello, and M. Tewes. H0LiCOW - II. Spectroscopic survey and galaxy-group identification of the strong gravitational lens system HE 0435-1223. MNRAS, 470(4) :4838–4857, Oct. 2017. doi : 10.1093/mnras/stx1484.

J. L. Starck, K. E. Themelis, N. Jeffrey, A. Peel, and F. Lanusse. Weak-lensing mass reconstruction using sparsity and a Gaussian random field. A&A, 649 :A99, May 2021. doi : 10.1051/0004-6361/202039451.

A. Stockton. The lens galaxy of the twin QSO 0957+561. ApJ, 242 :L141–L144, Dec. 1980. doi : 10.1086/183419.

F. Sun, E. Egami, P. G. Pérez-González, I. Smail, K. I. Caputi, F. E. Bauer, T. D. Rawle, S. Fujimoto, K. Kohno, U. Dudzevičiūtė, H. Atek, M. Bianconi, S. C. Chapman, F. Combes, M. Jauzac, J.-B. Jolly, A. M. Koekemoer, G. E. Magdis, G. Rodighiero, W. Rujopakarn, D. Schaerer, C. L. Steinhardt, P. Van der Werf, G. L. Walth, and J. R. Weaver. Extensive Lensing Survey of Optical and Near-infrared Dark Objects (El Sonido) : HST H-faint Galaxies behind 101 Lensing Clusters. ApJ, 922(2) :114, Dec. 2021. doi : 10.3847/1538-4357/ac2578. URL `https://ui.adsabs.harvard.edu/abs/2021ApJ...922..114S`.

S. H. Suyu and R. D. Blandford. The anatomy of a quadruply imaged gravitational lens system. MNRAS, 366(1) : 39–48, Feb. 2006. doi : 10.1111/j.1365-2966.2005.09854.x.

S. H. Suyu, P. J. Marshall, M. P. Hobson, and R. D. Blandford. A Bayesian analysis of regularized source inversions in gravitational lensing. MNRAS, 371(2) :983–998, Sept. 2006. doi : 10.1111/j.1365-2966.2006.10733.x.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999. URL `https://arxiv.org/abs/physics/0004057`.

D. A. Torres-Ballesteros and L. Castañeda. relensing : Reconstructing the mass profile of galaxy clusters from gravitational lensing. *arXiv e-prints*, art. arXiv :2201.10076, Jan. 2022.

T. Treu and L. V. E. Koopmans. Massive Dark Matter Halos and Evolution of Early-Type Galaxies to z ~1. ApJ, 611(2) :739–760, Aug. 2004. doi : 10.1086/422245.

S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image : image processing in python. *PeerJ*, 2 :e453, 2014.

S. Vegetti and L. V. E. Koopmans. Bayesian strong gravitational-lens modelling on adaptive grids : objective detection of mass substructure in Galaxies. MNRAS, 392(3) :945–963, Jan. 2009. doi : 10.1111/j.1365-2966.2008.14005.x.

S. Vegetti, D. J. Lagattuta, J. P. McKean, M. W. Auger, C. D. Fassnacht, and L. V. E. Koopmans. Gravitational detection of a low-mass dark satellite galaxy at cosmological distance. Nature, 481(7381) :341–343, Jan. 2012. doi : 10.1038/nature10669.

J. D. Vieira, D. P. Marrone, S. C. Chapman, C. De Breuck, Y. D. Hezaveh, A. Weiβ, J. E. Aguirre, K. A. Aird, M. Aravena, M. L. N. Ashby, M. Bayliss, B. A. Benson, A. D. Biggs, L. E. Bleem, J. J. Bock, M. Bothwell, C. M. Bradford, M. Brodwin, J. E. Carlstrom, C. L. Chang, T. M. Crawford, A. T. Crites, T. de Haan, M. A. Dobbs, E. B. Fomalont, C. D. Fassnacht, E. M. George, M. D. Gladders, A. H. Gonzalez, T. R. Greve, B. Gullberg, N. W. Halverson, F. W. High, G. P. Holder, W. L. Holzapfel, S. Hoover, J. D. Hrubes, T. R. Hunter, R. Keisler, A. T. Lee, E. M. Leitch, M. Lueker, D. Luong-van, M. Malkan, V. McIntyre, J. J. McMahon, J. Mehl, K. M. Menten, S. S. Meyer, L. M. Mocanu, E. J. Murphy, T. Natoli, S. Padin, T. Plagge, C. L. Reichardt, A. Rest, J. Ruel, J. E. Ruhl, K. Sharon, K. K. Schaffer, L. Shaw, E. Shirokoff, J. S. Spilker, B. Stalder, Z. Staniszewski, A. A. Stark, K. Story, K. Vanderlinde, N. Welikala, and R. Williamson. Dusty starburst galaxies in the early Universe as revealed by gravitational lensing. Nature, 495(7441) :344–347, Mar. 2013. doi : 10.1038/nature12001.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0 : Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17 :261–272, 2020. doi : 10.1038/s41592-019-0686-2.

S. Wagner-Carena, J. W. Park, S. Birrer, P. J. Marshall, A. Roodman, R. H. Wechsler, and LSST Dark Energy Science Collaboration. Hierarchical Inference with Bayesian Neural Networks : An Application to Strong Gravitational Lensing. ApJ, 909(2) :187, Mar. 2021. doi : 10.3847/1538-4357/abdf59.

S. Wagner-Carena, J. Aalbers, S. Birrer, E. O. Nadler, E. Darragh-Ford, P. J. Marshall, and R. H. Wechsler. From Images to Dark Matter : End-To-End Inference of Substructure From Hundreds of Strong Gravitational Lenses. *arXiv e-prints*, art. arXiv :2203.00690, Mar. 2022.

D. Walsh, R. F. Carswell, and R. J. Weymann. 0957+561 A, B : twin quasistellar objects or gravitational lens ? Nature, 279 :381–384, May 1979. doi : 10.1038/279381a0.

S. J. Warren and S. Dye. Semilinear Gravitational Lens Inversion. ApJ, 590(2) :673–682, June 2003. doi : 10.1086/375132.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi : 10.25080/Majora-92bf1922-00a.

K. C. Wong, S. H. Suyu, M. W. Auger, V. Bonvin, F. Courbin, C. D. Fassnacht, A. Halkola, C. E. Rusu, D. Sluse, A. Sonnenfeld, T. Treu, T. E. Collett, S. Hilbert, L. V. E. Koopmans, P. J. Marshall, and N. Rumbaugh. H0LiCOW - IV. Lens mass model of HE 0435-1223 and blind measurement of its time-delay distance for cosmology. MNRAS, 465(4) :4895–4913, Mar. 2017. doi : 10.1093/mnras/stw3077.

P. Young, J. E. Gunn, J. Kristian, J. B. Oke, and J. A. Westphal. The double quasar Q0957+561 A, B : a gravitational lens image formed by a galaxy at z=0.39. ApJ, 241 :507–520, Oct. 1980. doi : 10.1086/158365.

P. Young, J. E. Gunn, J. Kristian, J. B. Oke, and J. A. Westphall. Q0957+561 : detailed models of the gravitational lens effect. ApJ, 244 :736–755, Mar. 1981. doi : 10.1086/158751.

S. Zhao, J. Song, and S. Ermon. InfoVAE : Information Maximizing Variational Autoencoders. *arXiv e-prints*, art. arXiv :1706.02262, June 2017.

F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A Comprehensive Survey on Transfer Learning. *arXiv e-prints*, art. arXiv :1911.02685, Nov. 2019.

F. Zwicky. Nebulae as gravitational lenses. *Phys. Rev.*, 51 :290–290, Feb 1937. doi : 10.1103/PhysRev.51.290. URL `https://link.aps.org/doi/10.1103/PhysRev.51.290`.

Annexe A

Elastic Weight Consolidation

Suppose we are given a training set $\mathcal{D}$ and a test task $\mathcal{T}$. The posterior of the RIM parameters φ can be rewritten using the Bayes rule as

$$p(\varphi \mid \mathcal{D}, \mathcal{T}) = \frac{p(\mathcal{T} \mid \mathcal{D}, \varphi) p(\varphi \mid \mathcal{D})}{p(\mathcal{T} \mid \mathcal{D})}. \tag{A.1}$$

We suppose that φ encode information about $\mathcal{D}$, while $\mathcal{T}$ was unseen by φ. It follows that $\mathcal{T}$ and $\mathcal{D}$ are conditionally independent when given φ. We do not make the stronger assumption that $\mathcal{D}$ and $\mathcal{T}$ are completely independent. In fact, such an assumption would contradict the premiss of our work that building a dataset $\mathcal{D}$ can inform a machine (RIM) about task $\mathcal{T}$ — or that, more broadly, $\mathcal{D}$ contains information about $\mathcal{T}$.

We rewrite the marginal $p(\mathcal{T} \mid \mathcal{D})$ using the Bayes rule in order to extract $p(\mathcal{D} \mid \mathcal{T})$, the sampling distribution used to compute the Fisher diagonal elements

$$p(\varphi \mid \mathcal{D}, \mathcal{T}) = \frac{p(\mathcal{T} \mid \varphi) p(\varphi \mid \mathcal{D})}{p(\mathcal{D} \mid \mathcal{T})} \frac{p(\mathcal{D})}{p(\mathcal{T})}. \tag{A.2}$$

The log-likelihood $\log p(\mathcal{T} \mid \varphi)$ is equivalent to the negative of the loss function for the particular task at hand. In this work, we assign a uniform probability density to $p(\mathcal{T})$ and $p(\mathcal{D})$ in order to ignore them.

We now turn to the prior $p(\varphi \mid \mathcal{D})$, which appears as a conditional relative to the training dataset. We use the Laplace approximation around the maxima $\varphi^\star_{\mathcal{D}}$ to evaluate the prior, where $\varphi^\star_{\mathcal{D}}$ are the trained parameters of the RIM that minimize the empirical risk (equation (2.6)). The Taylor expansion of the prior around this maxima yields

$$\log p(\varphi \mid \mathcal{D}) \approx \log p(\varphi^\star_{\mathcal{D}} \mid \mathcal{D}) + \frac{1}{2}(\varphi - \varphi^\star_{\mathcal{D}})^T \underbrace{\left(\frac{\partial^2 \log p(\varphi \mid \mathcal{D})}{\partial^2 \varphi} \bigg|_{\varphi^\star_{\mathcal{D}}} \right)}_{\mathbf{H}(\varphi^\star_{\mathcal{D}})} (\varphi - \varphi^\star_{\mathcal{D}}). \tag{A.3}$$

Since $\varphi^{\star}_{\mathcal{D}}$ is an extrema of the prior, the linear term vanishes. The empirical estimate of the negative hessian matrix is the observed Fisher information matrix which can be written as

$$\mathcal{I}(\varphi^{\star}_{\mathcal{D}}) = -\,\mathbb{E}_{\mathcal{D}|\mathcal{T}}[\mathbf{H}(\varphi^{\star}_{\mathcal{D}})] = \mathbb{E}_{\mathcal{D}|\mathcal{T}}\left[\left(\left(\frac{\partial \log p(\varphi \mid \mathcal{D})}{\partial \varphi}\right)\left(\frac{\partial \log p(\varphi \mid \mathcal{D})}{\partial \varphi}\right)^{T}\right)\Bigg|_{\varphi^{\star}_{\mathcal{D}}}\right]. \tag{A.4}$$

The expectation is taken over the sample space $p(\mathcal{D} \mid \mathcal{T})$ since the network parameters are held fixed during sampling. In order to compute the Fisher score, we apply the Bayes rule to the prior to extract a loss function, which we take to be proportional to the training loss (equation (2.5)) and the χ^2 :

$$\log p\big(\varphi \mid (\mathbf{x}, \mathbf{y}) = \mathcal{D}\big) \propto -\mathcal{L}_{\varphi}(\mathbf{x}, \mathbf{y}) + \frac{1}{T}\sum_{t=1}^{T} \log p(\mathbf{y} \mid \hat{\mathbf{x}}^{(t)}) - \frac{\ell_2}{2}\|\varphi\|_2^2 \tag{A.5}$$

We find in practice the the ℓ_2 term has little effect on the Fisher diagonal and our results. Thus, we set $\ell_2 = 0$.

Since the full Fisher matrix is intractable for a neural network, we approximate the quadratic term of the prior with the diagonal of the Fisher matrix following Kirkpatrick et al. (2016). For an optimisation problem, the first term of (A.3) is constant. Thus, the posterior becomes proportional to

$$\log p(\varphi \mid \mathcal{D}, \mathcal{T}) \propto \log p(\mathcal{T} \mid \varphi) - \frac{\lambda}{2}\sum_{j} \operatorname{diag}(\mathcal{I}(\varphi^{\star}_{\mathcal{D}}))_j (\varphi_j - [\varphi^{\star}_{\mathcal{D}}]_j)^2. \tag{A.6}$$

The Lagrange multiplier λ is introduced to tune our uncertainty about the network parameters during fine-tuning.

Annexe B

VAE Architecture and optimisation

For the following architectures, we employ the notion of *level* to mean layers in the encoder and the decoder with the same resolution. In each level, we place a block of convolutional layers before downsampling (encoder) or after upsampling (decoder). These operations are done with strided convolutions like in the U-net architecture of the RIM.

TABLE B.1 – Hyperparameters for the background source VAE.

Parameter	Value
Input preprocessing	$\mathbb{1}$
Architecture	
Levels (encoder and decoder)	3
Convolutional layer per level	2
Latent space dimension	32
Hidden Activations	Leaky ReLU
Output Activation	Sigmoid
Filters (first level)	16
Filters scaling factor (per level)	2
Number of parameters	3 567 361
Optimization	
Optimizer	Adam
Initial learning rate	10^{-4}
Learning rate schedule	Exponential Decay
Decay rate	0.5
Decay steps	30 000
Number of steps	500 000
β_{max}	0.1
Batch size	20

Table B.2 – Hyperparameters for the convergence VAE.

Parameter	Value
Input preprocessing	$\log_{10}$
Architecture	
Levels (encoder and decoder)	4
Convolutional layer per level	1
Latent space dimension	16
Hidden Activations	Leaky ReLU
Output Activation	$\mathbb{1}$
Filters (first level)	16
Filters scaling factor (per level)	2
Number of parameters	1 980 033
Optimization	
Optimizer	Adam
Initial learning rate	10^{-4}
Learning rate schedule	Exponential Decay
Decay rate	0.7
Decay steps	20 000
Number of steps	155 000
$\beta_{\max}$	0.2
Batch size	32

Annexe C

RIM architecture and optimisation

The notion of link function $\Psi : \Xi \to \mathcal{X}$, introduced by Putzky and Welling (2017), is an invertible transformation between the network prediction space $\boldsymbol{\xi} \in \Xi$ and the forward modelling space $\mathbf{x} \in \mathcal{X}$. This is a different notion from preprocessing, discussed in section 2.3, because this transformation is applied inside the recurrent relation 2.4 as opposed to before training. In the case where the forward model has some restricted support or it is found that some transformation helps the training, then the link function chosen must be implemented as part of the network architecture as shown in the unrolled computational graph in Figure C.1. Also, the loss $\mathcal{L}_\varphi$ must be computed in the Ξ space in order to avoid gradient vanishing problems when Ψ is a non-linear mapping, which happens if the non-linear link function is applied in an operation recorded for backpropagation through time (BPTT).

For the convergence, we use an exponential link function with base 10 : $\hat{\boldsymbol{\kappa}} = \Psi(\boldsymbol{\xi}) = 10^{\boldsymbol{\xi}}$. This Ψ encodes the non-negativity of the convergence. Furthermore, it is a power transformation that leaves the linked pixel values $\boldsymbol{\xi}_i$ normally distributed, thus improving the learning through the non-linearities in the neural network. The pixel weights $\mathbf{w}_i$ in the loss function (2.5) are chosen to encode the fact that the pixel with critical mass density ($\boldsymbol{\kappa}_i > 1$) have a stronger effect on the lensing configuration than other pixels. We find in practice that the weights

$$\mathbf{w}_i = \frac{\sqrt{\boldsymbol{\kappa}_i}}{\sum_i \boldsymbol{\kappa}_i}, \tag{C.1}$$

encode this knowledge in the loss function and improved both the empirical risk and the goodness of fit of the baseline model on early test runs.

For the source, we found that we do not need a link function — its performance is generally better compared to other link function we tried like sigmoid and power transforms — and we found that the pixel weights can be taken to be uniform, i.e. $\mathbf{w}_i = \frac{1}{M}$.

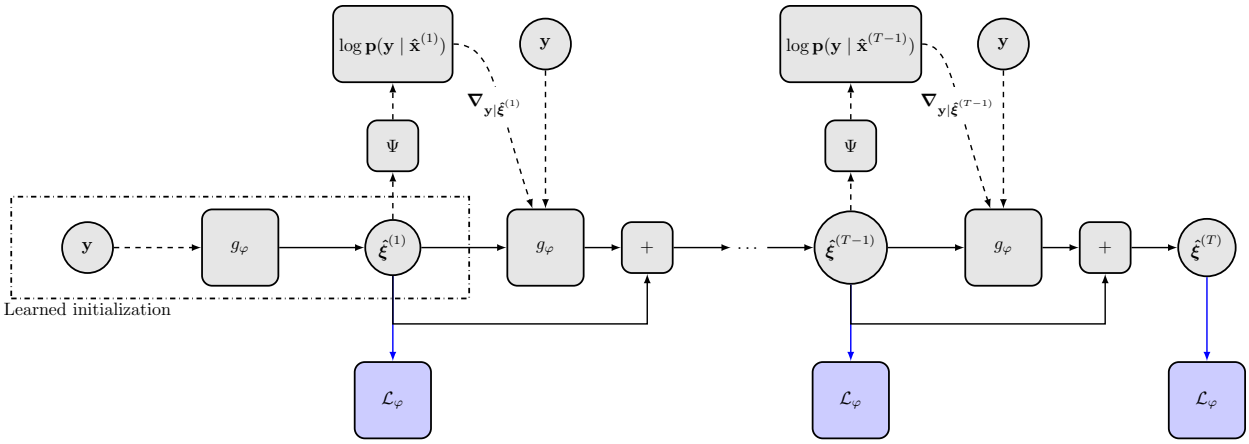

FIGURE C.1 – Unrolled computational graph of the RIM. Operations along solid arrows are being recorded for BPTT, while operations along dashed arrows are not. The blue arrows are only used for optimisation during training. During fine-tuning or testing, the loss is computed only as an oracle metric to validate that our methods can recover the ground truth.

TABLE C.1 – Hyperparameters for the RIM.

Parameter	Value
Source link function	$\mathbb{1}$
κ link function	$10^{\boldsymbol{\xi}}$
Architecture	Figure 2.3
Recurrent steps (T)	8
Number of parameters	348 546 818
First Stage Optimisation	
Optimizer	Adamax
Initial learning rate	10^{-4}
Learning rate schedule	Exponential Decay
Decay rate	0.95
Decay steps	100 000
Number of steps	610 000
Batch size	1
Second Stage Optimisation	
Optimizer	Adamax
Initial learning rate	6×10^{-5}
Learning rate schedule	Exponential Decay
Decay rate	0.9
Decay steps	100 000
Number of steps	870 000
Batch size	1

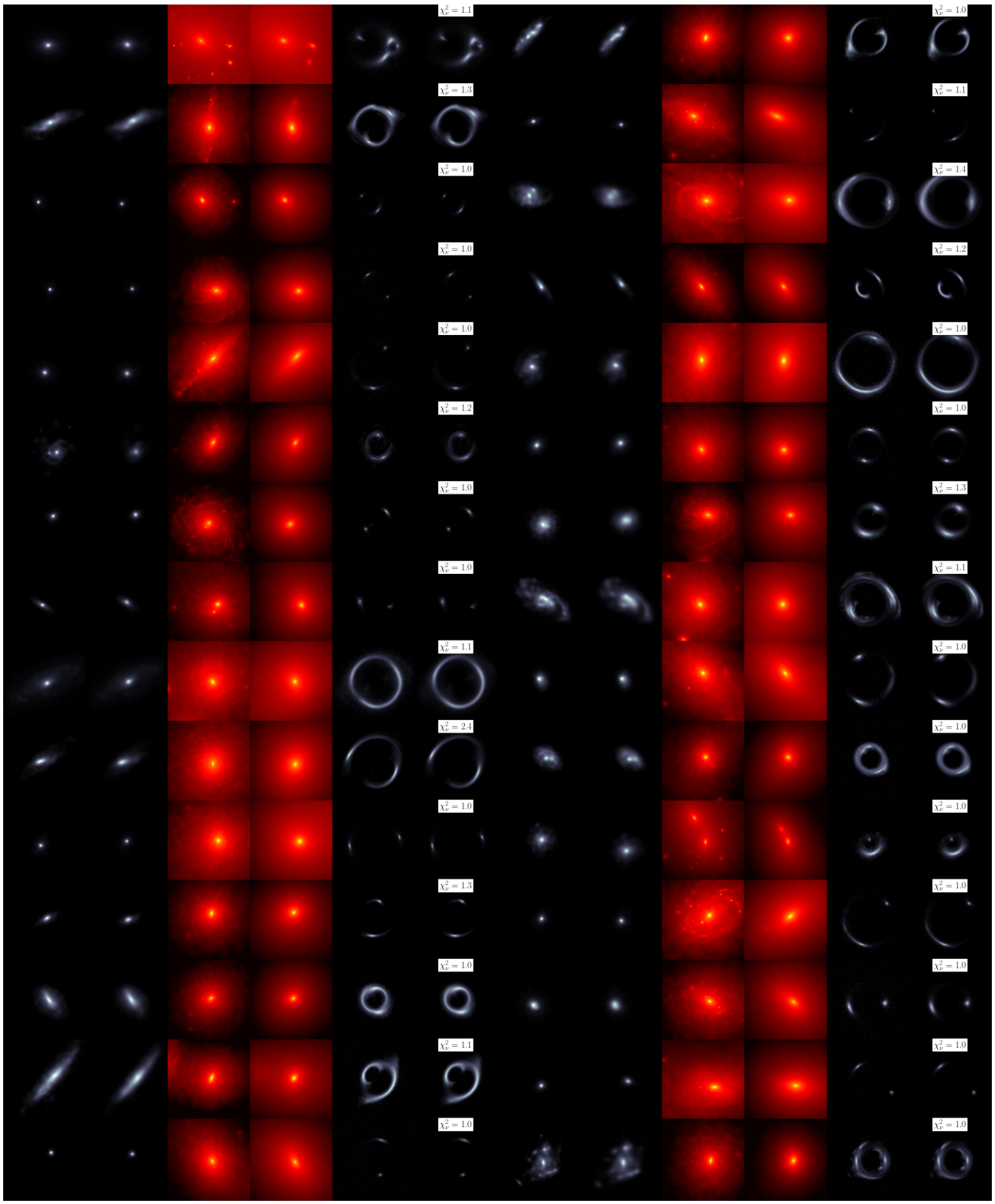

FIGURE C.2 – 30 reconstructions taken at random from the test set of 3000 examples simulated from COSMOS and IllustrisTNG data at high SNR. The colorscale are the same as in Figure 2.7.

Annexe D

GRU

Une unité récurrente à porte convolutionnelles est décrite par les opérations

$$\tilde{\mathbf{x}} = S\Big(\mathbf{w}_o * (\mathbf{h}^{(t-1)} \oplus \mathbf{x}^{(t)}) + \mathbf{b}_o\Big) \qquad \{\text{Porte d'oubli}\} \qquad \text{(D.1)}$$

$$\mathbf{z} = S\Big(\mathbf{w}_z * (\mathbf{h}^{(t-1)} \oplus \mathbf{x}^{(t)}) + \mathbf{b}_z\Big) \qquad \{\text{Porte de mise à jour}\} \qquad \text{(D.2)}$$

$$\tilde{\mathbf{h}} = \tanh\Big(\mathbf{w}_h * \big((\mathbf{h}^{(t-1)} \odot \tilde{\mathbf{x}}) \oplus \mathbf{x}^{(t)}\big)\Big) + \mathbf{b}_h\Big) \qquad \{\text{État candidat}\} \qquad \text{(D.3)}$$

$$\mathbf{h}^{(t)} = \mathbf{h}^{(t-1)} \odot \mathbf{z} + \tilde{\mathbf{h}} \odot (1 - \mathbf{z}) \qquad \{\text{Nouvel état}\} \qquad \text{(D.4)}$$

où $S(x) = \frac{1}{1+e^{-x}}$ est une fonction sigmoïde et $\mathbf{x}^{(t)}$ est un tenseur à l'entrée de l'unité. Les noyaux de convolution $\mathbf{w}$ et les vecteurs de biais $\mathbf{b}$ sont des paramètres libres appris par descente de gradient stochastique. $\oplus$ symbolise l'opération de concatenation. Le tenseur de sortie de cette unité, soit le nouvel état latent $\mathbf{h}^{(t)}$, est une combinaison de l'état latent précédent $\mathbf{h}^{(t-1)}$ et de l'état candidat $\tilde{\mathbf{h}}$, pesée élément par élément par le vecteur à la sortie de la porte de mise à jour $\mathbf{z}$.