Ecole Centrale de Lille

Stage de fin d'étude

Rapport final

Deep Learning for downscaling satellite imagery on the ocean

Author
Jean Le Goff

Supervisors
Patrick Gallinari (MLIA)
& Sylvie Thiria (LOCEAN)

Stage réalisé dans le cadre de la troisième année du cursus ingénieur de l'école Centrale de Lille

en partenariat avec

ISIR - Institut des Systèmes Intelligents et de Robotique
MLIA - Machine Learning & Deep Learning for Information Access
&
LOCEAN - Laboratoire d'Oceanographie et du Climat : Expérimentations et Approches Numériques

Résumé

Ce rapport est le compte-rendu des travaux réalisés lors de mon stage de fin d'étude de l'Ecole Centrale de Lille au sein du laboratoire MLIA de Sorbonne Université et sous la supervision de Patrick Gallinari. Durant ce stage, j'ai été amené à travailler sur la problématique de l'utilisation d'algorithme d'apprentissage statistique pour réaliser la descente d'échelle de données ocanique, en particulier de mesure et simulations de la hauteur de la surface de l'océan (Sea Surface Temperature). Dans ce but, différentes stratégies ont été testé, à la croisée des domaines du traitement de l'image, des réseaux de neurones implicites et de l'apprentissage statistique basé sur la physique.

Remerciements

Je tiens tout d'abord à remercier mon tuteur Patrick Gallinari, pour le suivi tout au long du stage, sa bienveillance et sa pertinence lors de nos échanges qui m'a permis de produire le travail présenté ici, et d'apprendre beaucoup. Merci également à mes co-stagiaire Lise et Raphael, mes collègues doctorants, Marie, Tristan, Yuan, Matthieu, Louis, Thomas, Agnes, Etienne avec qui j'ai eu la chance de partager ces six mois et qui ont tant apporté à mon expérience lors de discussion ou autre partie de baby-foot endiablées. Merci à Christophe pour son aide et sa gentillesse. Merci au chercheurs du LIP6 et de LOCEAN, Sylvie, Dominique, Anastase, Carlos et Théo d'avoir pris le temps de discuter et de me conseiller. Je remercie également Pascal Yim pour l'encadrement de ce stage et ses cours qui, comme ceux de Jean, pendant mon cursus centralien m'ont fait découvrir et apprécier ce domaine du Machine Learning et m'ont donné l'envie de l'approfondir.

Contents

1 Introduction

1.1 General Introduction

Ocean monitoring is crucial and a key to understanding its dynamics. Today, ocean data can come from various sources of data separated in two main types, ocean parameters can be observed or modeled. Observation can be obtained via remote sensing techniques like satellite measurements or via in-situ measures with buoys for examples. Models are based on computer simulations, they rely on differential equations that incorporate physical knowledge on the ocean phenomenon.

Recently, advances in data science and increase of the amounts of data available in the climate and oceanography fields offered interesting opportunities to use data intensive techniques like neural networks as alternatives and complements to classical physic for modeling complex dynamics and physical processes like ocean dynamics. The field born from this idea (AI for science or Physic-based AI) is gaining a lot of attention lately.

The emergence of data science technique in ocean sciences is seen as a potential solution to big challenges of the field. Among them is the challenge of downscaling observations and simulations. Both sources can produce data at various resolution, however computational costs tend to rise a lot to reach high resolutions with models and satellites or in-situ observations fail to reach high resolution measures yet. With already strong achievements in fields like Computer Vision or Natural Language Processing, Deep Learning techniques seem to be able to bring solutions to these challenges. Yet data-driven techniques applied to physical challenges also raise recurrent issues, they require large quantities of labeled data to be trained and they offer no guaranty to demonstrate generalization capacities once trained on a source of data.

These problematic were at the heart of my last year internship in a team whose work focuses on deep learning techniques and their applications to physical and climate sciences. During this experience I had to test the use of state of the art deep learning techniques coming from various fields like Computer Vision or Physic Based Deep Learning and investigate their applicability on challenges coming from the oceanographic field.

1.2 Lab presentation

My internship took place inside of the team MLIA (Machine Learning & deep learning for Information Access) at Sorbonne Université in Paris. The team is part of the lab ISIR (Institut des Systèmes Intelligents et de Robotique) and previously was part of the LIP6 (Laboratoire d'Informatique de Paris 6). The MLIA team is specialized on Statistical Machine Learning and has been one of the first group in France to work on Deep Learning and Neural Networks. Today, MLIA's work focuses on three application domains : Computer Vision, Natural Language Processing and Information Retrieval and Physic-based Deep Learning. My work has been conducted under the supervision of Professor Patrick Gallinari and among the group of researchers and PhD students working on Physic-based Deep Learning topics.
On the climate side, my internship was supervised by another lab from Sorbonne Université : LOCEAN (Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques) which is part of IPSL (Institut Pierre Simon Laplace). My referent on this side was Professor Sylvie Thiria.

1.3 Report Organization

In this report I will present the work conducted during the five months spent inside the MLIA team. In the first part I will introduce the context of my study more precisely as well as a brief presentation of the related works, the data and the objectives of the work. Then in a second part, we will see SSH downscaling task as a super-resolution task and focus on the use of implicit neural representations to learn this task. Then, in a third part, we will investigate how techniques from physic-based Deep Learning can help take the results further and towards more precision. Finally, a fourth part will be dedicated to the attempt to use an architecture combining implicit neural representations and the use of complementary physic information. In each part you will find explanations of the methods that are used and how they were adapted to the SSH downscaling task as well as presentations and analysis of the obtained results.

2 Context and objectives

2.1 Context

One of the big actual challenges of ocean and climate sciences is the downscaling of model outputs and observations. In fact, as claimed in [Hewitt et al., 2022] : the small scales of the ocean may hold the key to surprises. Everywhere in the ocean are sharp fronts and eddies that are not taken into account by most of the climate models. And these small scale phenomenons can, in ocean science particularly, have high impacts on large scale events. Therefore, there is a need for acceleration of efforts towards the development of tools allowing to retrieve high resolution (kilometer-scale) models outputs and observations. Fig. 1 illustrates this problematic by showing the differences between ocean dynamic observations at different resolutions.

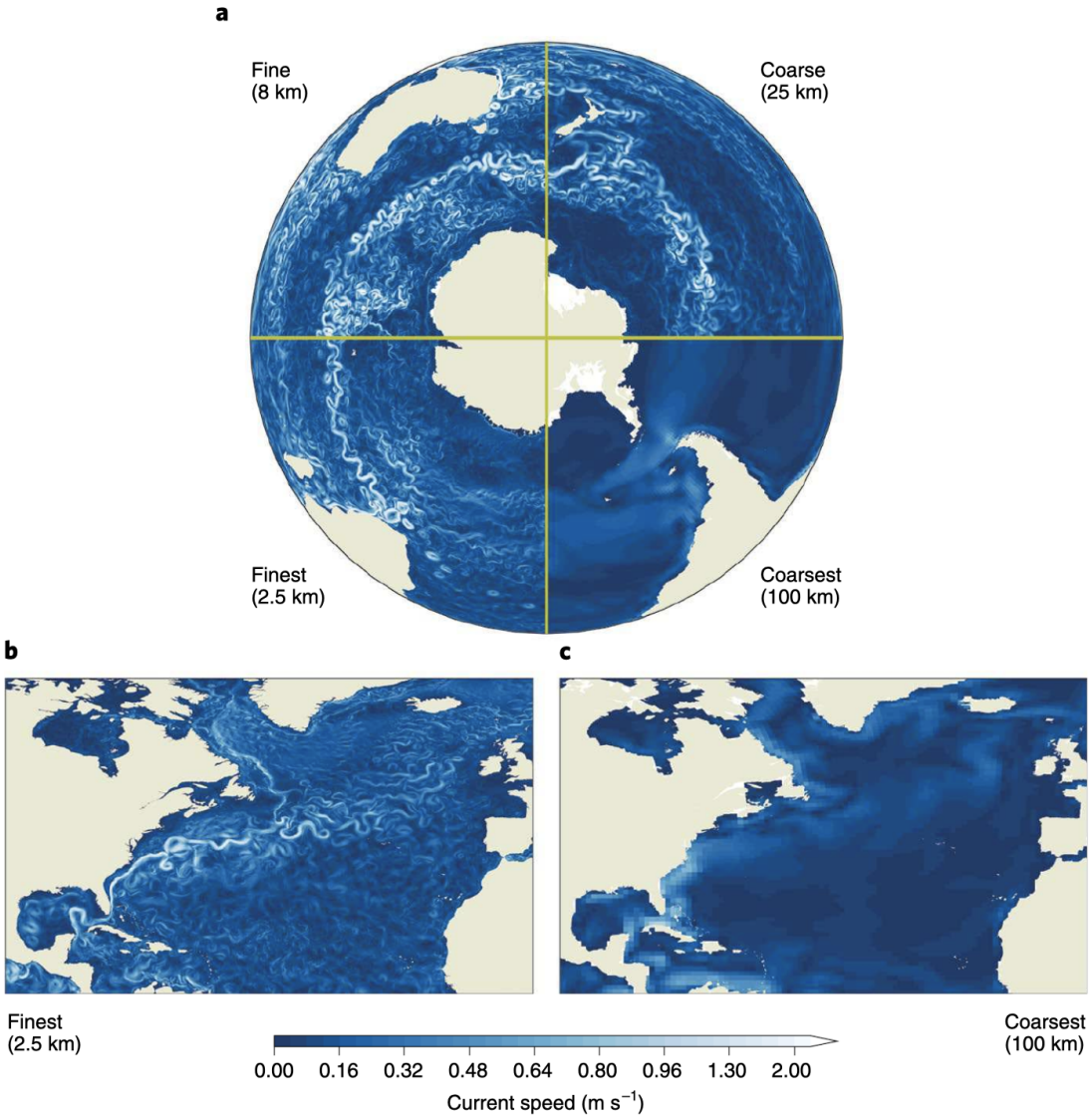
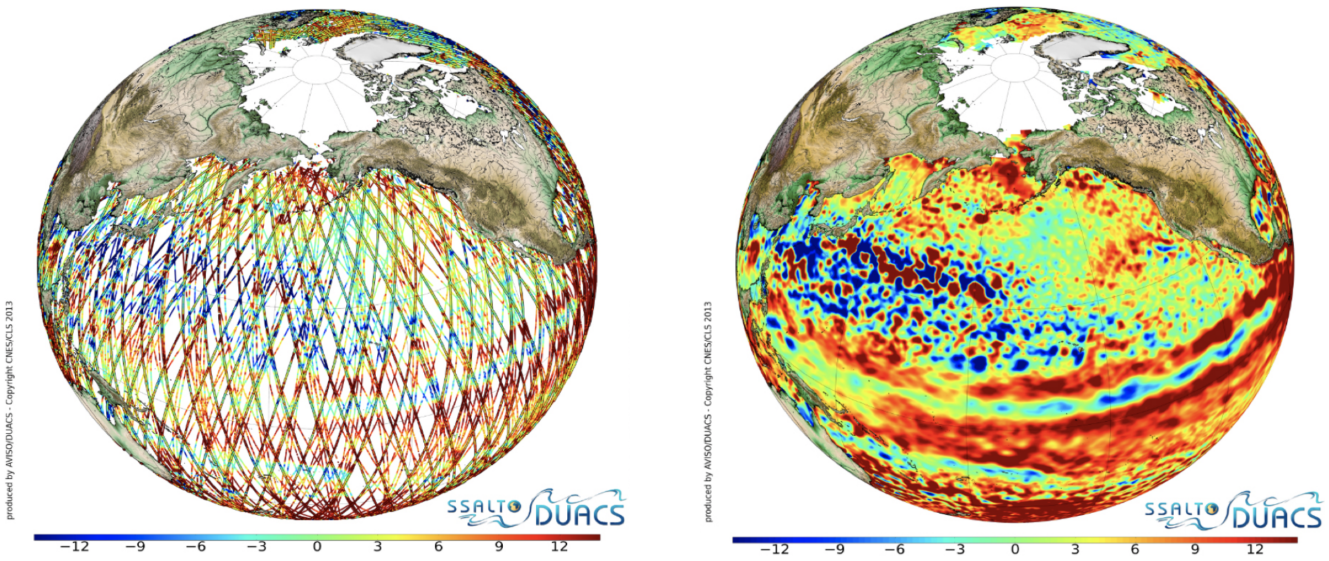


Figure 1: Ocean dynamics observations at different resolutions in the south pole and the northern Atlantic, two highly energetic regions. (Source: [Hewitt et al., 2022]).



(a) Along-track satellite measures.

(b) Interpolated result.

Figure 2: Satellite measures of SSH before and after space-time interpolation on a regular grid.

Among ocean parameters, the Sea Surface Height (SSH) is very important. It is a measure of the sea elevation relatively to a reference mean height of the ocean surface. The variations have amplitudes of about ten to fifty centimeters and rise up to a meter. Its knowledge allows to gain understanding of upper-ocean dynamics and many phenomenons it has influence on : mean currents, heat and salt transport, atmosphere-oceans interactions. Therefore monitoring its variability is a key to many predictions, models and applications. However, it is hard to achieve high resolution for SSH. This parameter is hard to model due to its strong dependency to many external factors, as for example unobserved deep oceans currents. And it is poorly remotely observable. Remote measures are made with satellites equipped with altimeters, those sensors produce along-track SSH measures (Fig. 2a) which are spatially and temporally interpolated to grided SSH data (Fig. 2b). However, the result of the interpolation is still very coarse (around 25 km, 1/4°) and for now, the low density of the measured track and the big return periods of sensors measurements are missing small scale and quick events especially in very energetic regions (e.g. the Gulf Stream in North Atlantic). Therefore, there is a strong need for efficient downscaling algorithm for SSH data.

For this challenge, recent collaborations between ocean sciences and data-driven models can bring solutions. In fact, deep learning has already shown strong performances in Computer Vision for super-resolution task and this task is close to the downscaling task for regularly grided data like the SSH resulting from the interpolation of along-tracks satellite observations. Moreover, Neural Networks have already been applied to climate challenges with great success for example Convolutional Neural Networks. Today, new techniques from this fast-evolving field seem to be able to push new solutions. Implicit neural networks from 3D Computer Vision and models from physic-based deep learning for example are for able to achieve grid-less and physic informed learning. These particularities are strong advantages to deal with the challenges of ocean and climate science. The scope of this work is to study their application to SSH downscaling and to assess their capacities to tackle the complexity inherent to this kind of data as well as their generalization capacities. In fact, as Deep Learning needs labeled

training data for supervised learning, used training data are outputs from high resolution models. Then the results of the models must be adapted to deal with data from new sources (other models or satellite data).

SSH is strongly and physically linked to other parameters of the ocean. One of them is the Sea Surface Temperature (SST). This parameter, contrary to SSH, is well observed via remote sensing and satellite imagery. Indeed, SST satellite observations are available at a scale of 5 km. SST is an important parameter for climate having a big role in ocean-atmosphere interactions. But its high resolution measurements can also be a tool to help SSH super-resolution as both parameters are linked, SST changes over time partly as a result of upper-ocean dynamics that are strongly influenced by SSH. Therefore SST spatial patterns are a big source of information to trace back SSH events like eddies or fronts. Indeed on Fig. 3 we can see that the two parameters have strong spatial similitude.

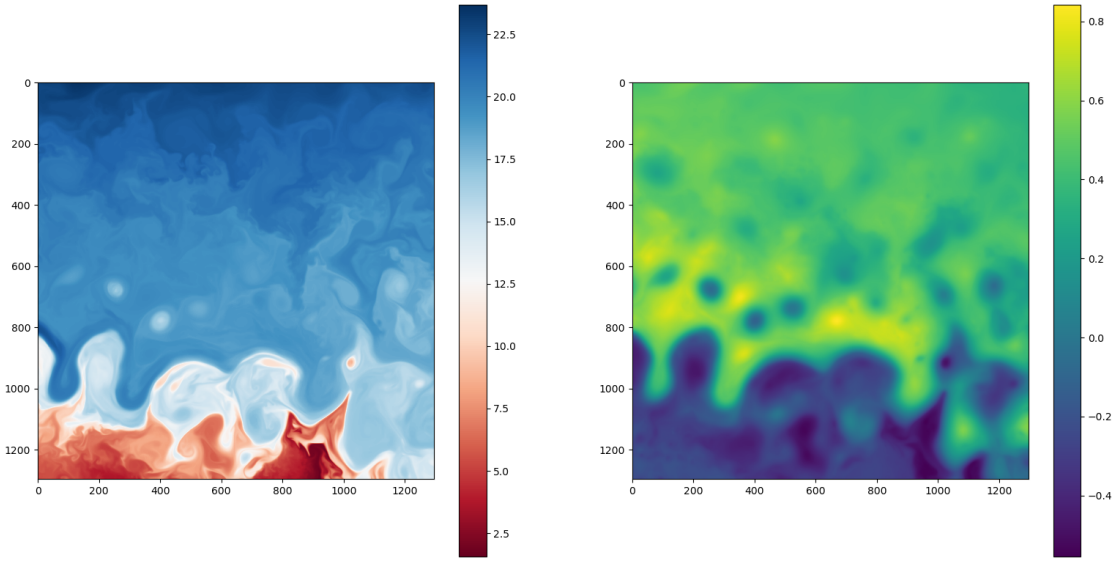


Figure 3: SSH and SST on the study zone of the NATL- SARGAS60 data set.

2.2 Data

The simulated data set that we use for the training of our models on the SSH downscaling task is the SARGAS60 data set [Meija et al., 2021]. This data set is an extraction of the NATL60 model experiments on a domain in the North Atlantic, the Sargasso Sea (26°N up to 66°N in latitude and 65°W to 40°W in longitude) close to America's coast (Fig 4) . This zone has been chosen as it is free of land and the northern part of the area is cut through by the Gulf Stream, a very energetic zone with a lot of surface activity. The data set is composed of one year of data between October 2012 and October 2013 for training, and 4 month in 2008 for testing. The NATL60 model is based on NEMO code, of North Atlantic oceanic circulation. It is the finest simulation (1/60° around 1 km) existing of this area. It required 8 weeks of computation on Occigen super-calculator, which represents 17 millions hours.

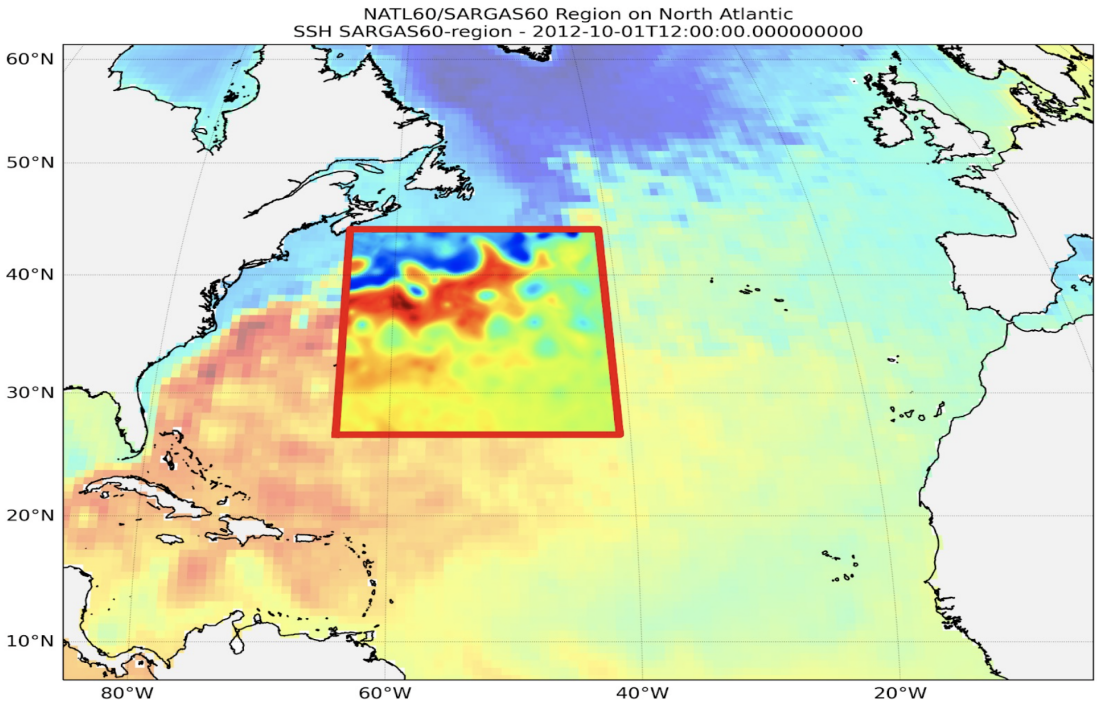


Figure 4: Study zone of the SARGAS60 data set. (Source: [Meija et al., 2021]).

To assess generalization capacities of the models we trained, we had access to another source of data, the Copernicus database. It offers an access to simulated and satellite observations of SSH data. The simulations are at a coarser resolution than the NATL60 data set but are available for a longer period. Furthermore, both models outputs are different since Copernicus outputs are reanalyzed through data assimilation.

Name	Resolution	Temporal coverage	Notes
NATL60	1/60° (1.5 km°)	2012/10/01 - 2013/10/01 and	High-resolution data set of SSH, SST, U, V model outputs in area of northern Atlantic.
Copernicus Satellite SSH	1/4° (25km)	01/01/1993 - 31/12/2020 (28 years)	SSH satellite observations produced by the fusion of several satellites altimeters measures.
Copernicus Satellite SST	1/20° (5km)	01/09/1981 - 30/09/2021 (40 years)	SST satellite observations.
Copernicus MERCATOR GLORYS12V1	1/12° (8km)	01/01/1993 - 31/05/2020 (27 years)	SSH, SST, U, V model outputs reanalyzed with observations.

2.3 Related works

Deep environmental downscaling. Downscaling of environmental data is a task that already often been tackled through the use of deep learning tools. Until now the most used architectures are Convolutional Neural Networks (CNN), especially Residual Networks (Res-Net). [Rodrigues et al., 2018] use a CNN to provide high-resolution weather fields from low-resolution ones on a region of South America, [Vandal et al., 2017] apply a

convolutional single image super resolution model applied to downscale climate change predictions. Same techniques have also been applied to ocean data. [Barthélémy et al., 2022] proposed a workflow of super resolution data assimilation: low-resolution model output is downscale to high-resolution by a CNN in order to improve efficiency of the assimilation of high-resolution observations. Finally, [Archambault et al., 2022] introduced the RESACsub model, a CNN to downscale SSH data using SST information. This work is using the same NATL60 data set on the Sargasso Sea.

Super Resolution. The super-resolution task is a common task of the Computer Vision and Signal Processing field. The NTIRE 2017 Challenge report defines the task as the restoration of rich details (high frequencies) in an image. If this restoration is only based on a set of prior examples with low resolution and corresponding high resolution, this task is known as the Single-Image Super-Resolution (SISR). Lately, this task has been dominated by deep learning models: the Convolutional Neural Networks that revolutionized the Computer Vision field. In 2014, [Dong et al., 2014] first used a supervised CNN with three layers for super-resolution and established new state-of-the-art. Ever since, more sophisticated architecture continued to raised the performances of NN applied to Super Resolution and recently, Residual Networks established stat-of-the-art performances: [Lim et al., 2017] by using Res-Net and multi-scale learning and [Zhang et al., 2018] by using Res-Net and deep local attention. Other architecture like GANs in [Ledig et al., 2017] are being investigated for this task. The super-resolution image can also be performed using not only the low-resolution version as input but also a guiding high-resolution image of the same scene in an different space. For example using high-resolution RGB image of a scene to guide the super-resolution of the depth map of the same scene like in [Tang et al., 2021]. This task is known as fusion-based super-resolution, pan-sharpening or guided super resolution. Although CNN brought revolutionary results to the field, they come with a dependence on the scale at which the supervised learning has been made. Being able to learn a super-resolution at any resolution is the task of the arbitrary-scale super-resolution field. Recently, Implicit Neural Networks (INR) brought strong advances to this field.

Implicit Neural Representations. Implicit neural representation, also known as coordinate-based neural representation, are a new topic of research in deep learning. This novel approach aims at learning continuous representations of signals via a function parameterized by a network. The goal is to map input coordinate to the value of the signal at that point. For example learn to continuously represent an image as a function $f_\theta : \mathbb{R}^2 \to \mathbb{R}^3$ that outputs the RGB pixel values from the pixel (x, y) coordinates. The function f_θ is learned by a Neural Network, usually a MLP (Multi-Layer Perceptron). However, it has been shown that simple MLP struggle to achieve satisfying representations, thus strategies have been introduced to adapt them to successfully learning the signal implicit function. [Sitzmann et al., 2020b] introduced periodic non-linearities as activation functions (SIREN) and well chosen initial parameters in the network and significantly improved the results, they showed promising applications in image and 3D models representation and for learning differential equations as the network is totally differentiable. [Tancik et al., 2020] ameliorate the performances with positional feature encoding of the inputs in Fourier space (Fourier Features) before feeding them to the network. These architecture's huge success in 3D Computer Vision contributed in their fast recent development. For example, [Mildenhall et al., 2020] introduced NeRF, a MLP parameterizing a function mapping 5D inputs (3D

coordinates + 2D viewing direction) to RGB Color and density of the 3D scene. Supervised training can be performed with photos of the 3D scene, thus greatly simplifying the learning of the 3D scene representation and scene reconstruction tasks. NeRF pushed further the use of feature positional encoding. However, naive positional encoding tends to only represent a signal at a single scale. Recent models proposed to learn multi scale representations, [Landgraf et al., 2022] with PINs add incremental frequency feature encoding for better multi scale representation of the network and [Lindell et al., 2021] with BACON also outperforms conventional single-scale implicit networks by choosing a fine analytical Fourier spectrum. However classical INR have the drawback to be scene- or object-specific, thus needing to train a network for each object to represent. Therefore, recent papers tried to generalize MLP-based representation, which means learning implicit function space with a neural network instead of learning only the function representing one signal. It usually means learning a prior over the studied signal in the space of functions. For example, [Sitzmann et al., 2020b] introduce the bases of the generalization of their SIREN networks using a hyper-network taking as input the studied signal and modifying the weight of the MLP implicit network to fit the given signal. For 3D representations, research have been made to be able to generalize to more than one scene, [Yu et al., 2021] condition a NeRF on spatial image feature using an image encoder and a continuous decoder taking into input coordinates and encoded local features, [Jang and Agapito, 2021] use a similar technique learning a code of a picture and reconstructing its associated 3D scene through a NeRF decoder. [Sitzmann et al., 2020a] use a meta-learning technique to perform generalization, claiming to outperform encoder-decoder generalization techniques.

Implicit Image Super-Resolution. For its ability to learn a signal representation without discretization, INR have also been used for 2D Computer Vision offering a interesting alternative to convolutional techniques that are discretization-dependent. Furthermore, learning a mesh-less representation offers the opportunity to reconstruct an image at a finer resolution and thus performing super-resolution. This idea is proposed by [Chen et al., 2021] with the LIIF model. They feed as input to the MLP parameterizing the function of the image not only the pixel coordinates but also a vector of neighbor's features extracted from the image and encoded in a latent space. This method achieved results comparable to convolutional state-of-the-art and brings its arbitrary-scale capacity. Following papers proposed to build amelioration on this new framework like [Xu et al., 2021] with UltraSR which adds spatial encoding to LIIF model and [Liu et al., 2021] which introduce its integral positional encoding module that codes the input coordinates depending on the super-resolution scale to try to encode the new frequency appearing when augmenting the scale. Furthermore, architecture based on LIIF have also been applied to tasks close to SISR like guided super resolution in [Tang et al., 2021] or spatio-temporal super-resolution [Jiang et al., 2020].

2.4 Objectives

The initial objectives were to first, develop a model for super-resolution task on SSH with flexible generalization capacities by using implicit representation architecture, train it on NATL60 data set and then assess model performances on new data (satellite observations) and examine transfer strategies and adaptation methods. It turned out that testing the use of these new models for the SSH downscaling task was an interesting lead to

follow and eventually the second objective of domain transfer was replaced by a deeper investigation of the capacities of implicit method and physic-based models to achieve downscaling of SSH data. In particular, the objective was to leverage physical information to improve the results of a Computer Vision super-resolution model downscaling SSH data.

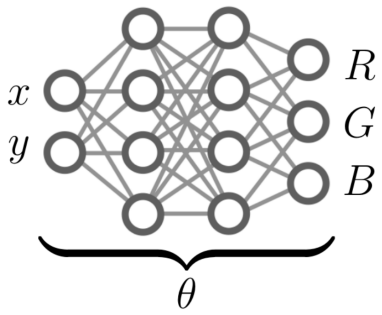


Figure 5: Network architecture of an image implicit representation.

3 Implicit Neural Representations for SSH downscaling

3.1 Method

We first had the objective to investigate the use of Implicit Neural Networks for the SSH downscaling. These architectures learn continuous representations of signals via a function parameterized by a neural network (multi-layer perceptron). The learned function, maps input coordinates to signal value. For example, for an image, it predicts the RGB values from the pixel coordinates (Fig. 5). The learned representation of the signal is the parameterized neural network, often a Multi Layer Perceptron. Moreover this representation is supposed to be independent of any discretization. For the training, it uses a discretized version of the signal, but then the learned function can be evaluated at any point. This capacity can be really useful to perform super-resolution on a signal, an image for example. Yet, this method requires to fit an individual function for each object. Recent developments proposed to generalize Implicit Neural Representations by learning a space of function and share knowledge across objects of the same type for the same task. [Chen et al., 2021] proposed a method for image super-resolution that uses local features to help the generalization of the implicit representation. Their arbitrary-scale method reached results comparable to convolutional state-of-the-art models. They keep the network mapping coordinates to the value of the signal, but also includes a supplementary encoding module to allow generalization to a space of images by conditioning the implicit network to the addition of features of the considered object. Its global architecture is presented in Fig. 6.

The model can be viewed as separated in two parts : a convolutional encoder and an implicit decoder named Local Implicit Image Function (LIIF). The convolutional encoder E_ϕ takes as input the low-resolution image and encodes it into a latent space. The goal is to create a new representations of the low-resolution image with features extracted and present in the D channels at each pixel (D¿3). Then the LIIF module will locally decode the latent representation of the image into a pixel value. LIIF is a MLP who takes as input the coordinates (x, y) of the queried pixel and the encoded features of its nearest neighbor in the low-resolution space. The prediction is thus conditioned to the vector of features allowing the LIIF model to generalize and learn the representation of a space of functions. However, to avoid a too big dependence of the prediction on only one neighbor, the predicted signal is actually the result of the sum of the networks predictions at the four nearest neighbors of the queried pixel (Fig. 7). For each of the four neighbors of the queried pixel at coordinate x_q,

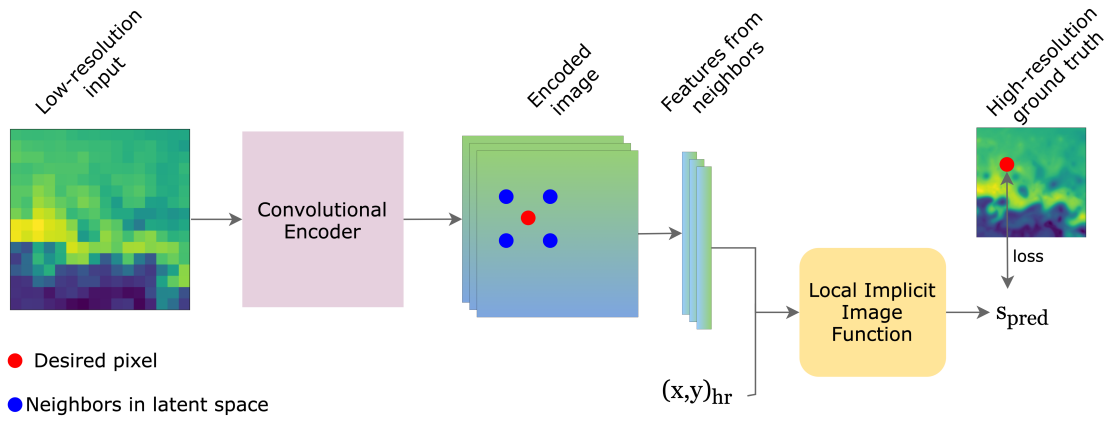


Figure 6: Architecture of the LIIF model used to downscale SSH data.

the function f_θ, parameterized by an MLP will predict the value of the signal :

$$I(x_q) = f_\theta(z_n, x_q - z_n) \tag{1}$$

where I is the continuous studied image, z_n is the vector of feature at the neighbor n in the neighborhood $N = \{00, 01, 10, 11\}$, and x_n the coordinates of the neighbor. Then the final value of the signal is interpolated as a weighted sum of the results for each of the four neighbors :

$$I(x_q) = \sum_{t \in N} \frac{S_t}{S} . f_\theta(z_t, x_q - x_t) \tag{2}$$

Here the weight S_t corresponds to the area of the square between the queried pixel and the neighbor on the side opposite to the one at x_t, as showed on figure Fig. 7. The architecture has other features that allows better results. To extend the neighborhood taken into account by LIIF, a feature unfolding is realized on the encoded low-resolution image, each vector of features is concatenated with the vector of features of its 8 neighbors. Moreover, the network is arbitrary-scale and though doesn't have the knowledge of the targeted scale when doing the prediction, to add this information, LIIF introduces cell decoding : the size of the targeted pixel is added to the input of LIIF thus making a difference between super resolution scales.

The encoder used to translate the input low resolution image in the latent space is a convolutional residual network, we use the architecture of EDSR from [Lim et al., 2017]. The output feature from the encore has a dimension of 64. Then the implicit decoder function is learned by a Multi-Layer Perceptron. This network has 4 layers with a hidden dimension of 256.

3.2 Results

For the training of our SSH-LIIF model, we take conditions similar to the one used in the testing of RESAC from [Archambault et al., 2022] as far as this convolutional method will be our baseline. Therefore, the LIIF model is trained on the NATL dataset. The size of the input low-resolution image is 16x16 and the target super-resolution is 432x432 meaning a super-resolution scale of x27. The training dataset is created by degrading the

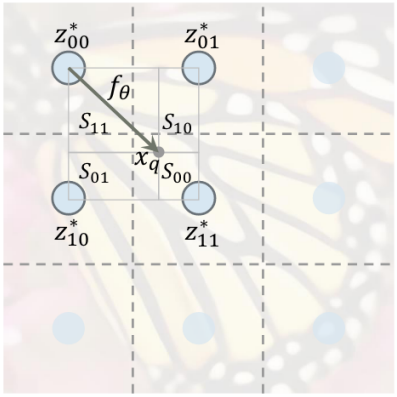


Figure 7: LIIF interpolation between the value predicted by f_θ at the four neighbors. (Source : [Chen et al., 2021])

images from the NATL60 dataset to the two desired resolution : the input and the ground truth. However, we need the same procedure as LIIF which consist in learning the super-resolution task for a random set of scales rather than a fixed one. In practice this is performed by randomly picking a super resolution scale s for each image of the data set and then degrading it, for the ground truth, to the resolution $16xs$. For our experiments we choose scales uniformly distributed between two chosen scales : 25 and 28. This technique aims at having a model which is not specialized for one super-resolution scale but rather generalizable to a broader range of scales.

The metric used for the result is the mean root mean square error on the testing set which is made of four months of simulations in 2008. The results are compared to the ones of the RESAC method but without the use of the SST information. It must be noted that RESAC is, however, scale-dependent. Furthermore, we also choose to compare the results to a classic bi-cubic interpolation. Results can be viewed on Table 1 and visualized on Fig.8 for one sample of the test set.

Model	Resac-SSH	SSH-LIIF	Resac-SSH-SST	Bicubic
RMSE(m)	0.0611	0.0650	0.0418	0.0672

Table 1: Comparison of the results for a scale of 27 and an input size of 16x16 on NATL test set.

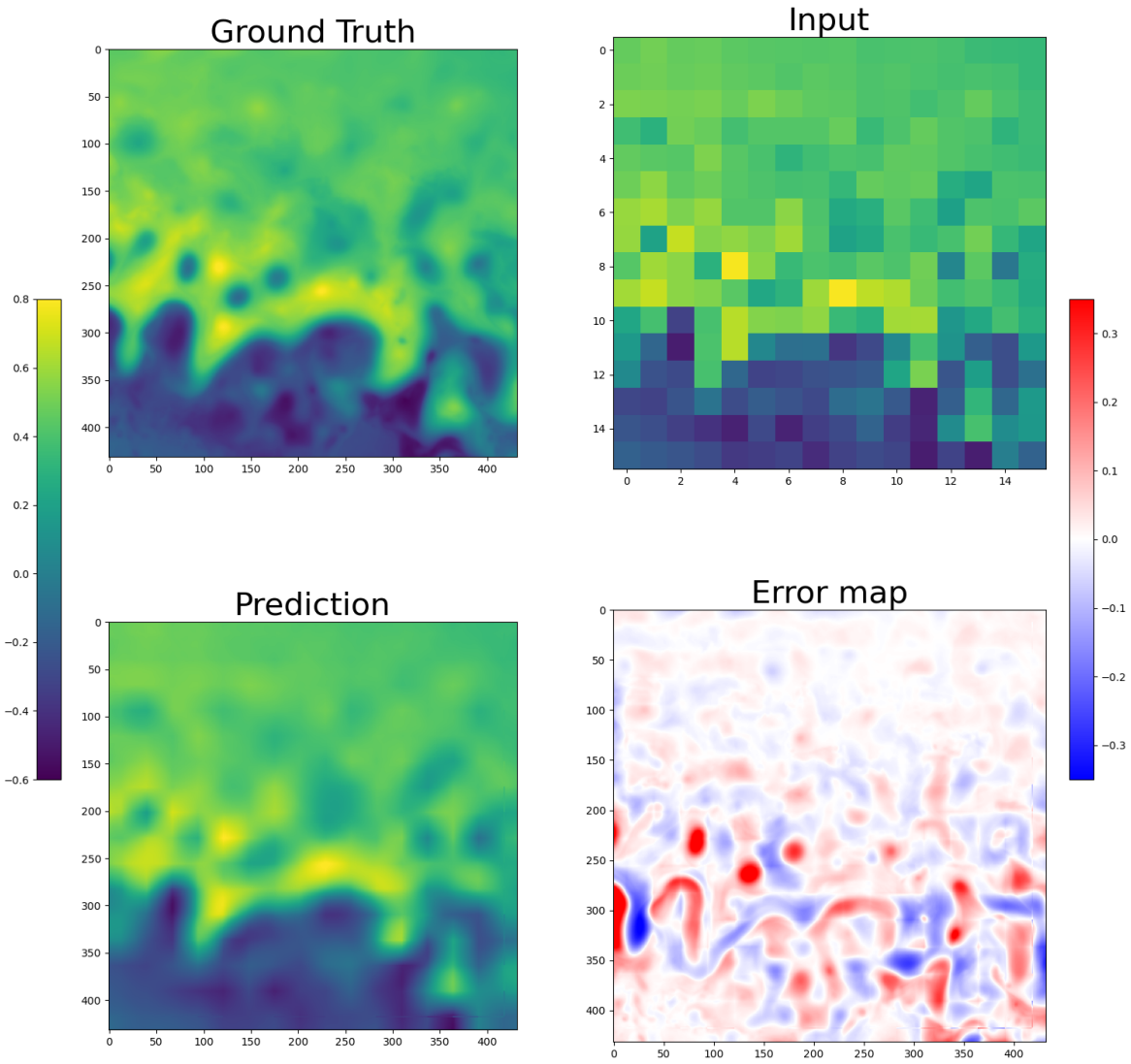


Figure 8: Results of the super-resolution for an input of size 16 and a scale of 27.

We see that our method give results that are comparable to RESAC architecture trained without the SST, furthermore, our model brings the convenience to be scale-arbitrary which is not the case for RESAC. However, both models only weakly outperform a standard Bicubic super-resolution which is not trained and also scale-arbitrary.

3.3 Discussion

We see in the results that the LIIF model has results comparable to the one of the RESAC model without the SST, furthermore, the LIIF model has the additional advantage to be scale-arbitrary and to be able to perform the super-resolution at any scales, even scales unseen at training time. However, the model is outperformed by the RESAC model taking into account the SST. Furthermore, it doesn't perform much better than the bi-cubic interpolation that is not trained but a classic universal interpolation tool. This means that LIIF only learns a smooth interpolation between the points at low-resolution, and struggles to capture and reconstruct high frequencies and small phenomenons that are unobserved at low resolutions and appear at higher resolutions. This can be explained by the fact that LIIF is designed to be a very local function, thus performing an implicit interpolation knowing only about the neighborhood. The model doesn't even have the knowledge of its position in the global image as the coordinate are given relatively to the ones of the neighbor. The model, thus, seems unable, as such, to learn the dynamics of the underlying phenomenons and reconstruct high frequency structure, it is only able to perform a smooth interpolation between known points.

4 Fourier Neural Operator to learn the SST/SSH link

4.1 Method

We saw earlier the limited results of the LIIF architecture on the SSH downscaling task. They point the fact that the LIIF architecture, which comes from Computer Vision, may not be able to handle the complexity of the SSH data, and it's high density of high frequencies in upper-scales. Indeed, the SSH downscaling task is not exactly a Computer Vision task, and the SSH data is not an image, on the contrary we are dealing with physical data. On the one hand, this data is more complex than classical images, but on the other hand we also have more insight and knowledge on these kinds of data. This task, therefore, falls more under the field of physic-based deep learning. For SSH data, we know the link that exist with SST as it has been showed in RESAC ([Archambault et al., 2022]) for example. The idea is to use high resolution information from the SST to help the downscaling of the SSH. This idea comes from the fact that satellite observations of the SST have higher resolution than SSH observations, thus allowing to use SST finer observation in real SSH observation downscaling tasks.

In the physic-based deep learning field, Neural Operators ([Kovachki et al., 2021]) have had recently a big success, notably for their ability to learn the solution to parametric partial differential equations, but also more generally for their capacity to learn a mapping between function spaces while being discretization invariant. The Fourier Neural Operator (FNO) from [Li et al., 2020] is able to learn complex non-local components of the function through by operating multiplication in the spectral domain. Its architecture is presented on Fig. 10, the function is fed as the input (as an image for example), first it is projected to a higher dimensional space P, then it will iteratively pass in Fourier layer. The Fourier Layers are made of two branches, one is a learned linear transformation, and in the other one the function is decomposed in a Fourier base, multiplied and returned to the initial space. Both outputs of the branches are summed and composed to a non-linearity. Finally, after the last Fourier Layer, the last hidden representation is mapped to the output function.

We use FNO to learn how to reconstruct SSH with only the SSH as input. The training is performed at a low resolution where both data are available. The network will learn an operator mapping SST to SSH at any resolution, we can, thus, apply it to high-resolution SST to recover high-resolution SSH. This is based on two strong hypothesis : first that such an operator between SST and SSH is learnable and then that this operator is generalizable to upper resolutions.

4.2 Results

We train a Fourier Neural Operator on NATL data. We train four models at different resolution : 16x16, 48x48, 144x144, 432x432. Then we evaluate the four models on high resolution data from NATL60, meaning data at resolution 1296x1296. For each model, we compute a Mean Root Mean Square Error (MRMSE) on the test set (4 months in 2008). Those results are shown on table Tab. 2. Fig. 11 shows the prediction at high resolution of the four models for a same data point. We see that the higher the training resolution is, the better the reconstruction is done.

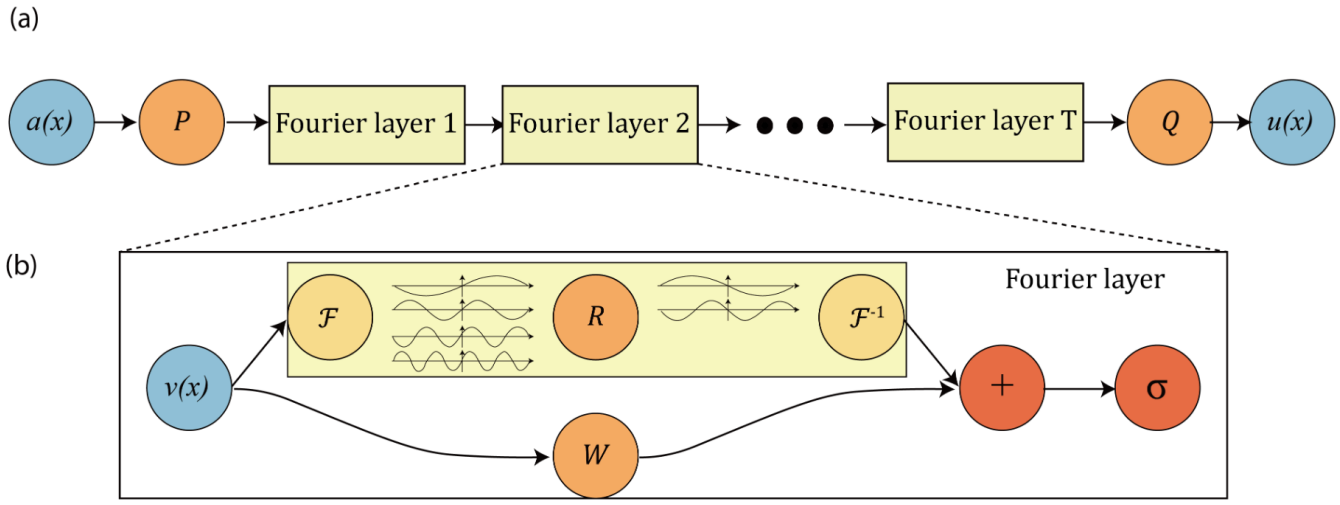


Figure 9: Architecture of a Fourier Neural Operator.

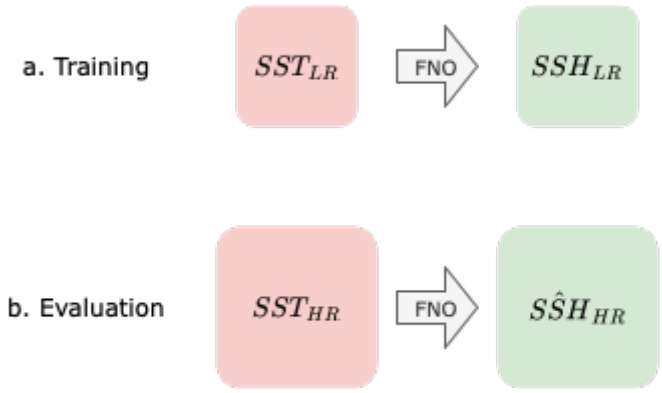


Figure 10: Super-resolution method using FNO.

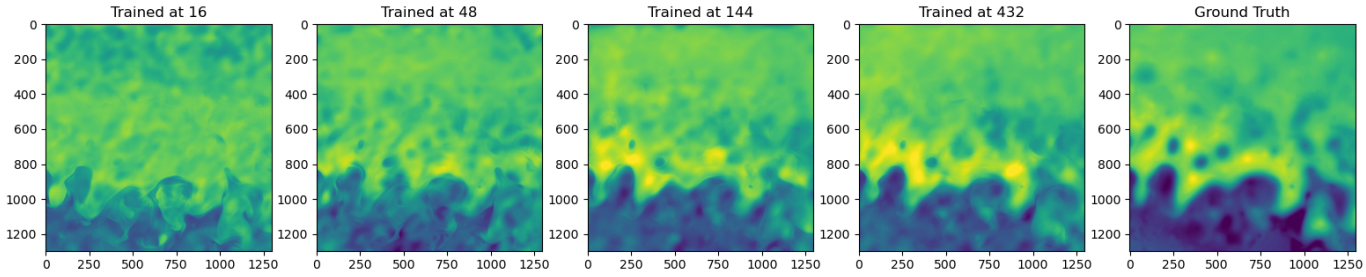


Figure 11: Results of the FNO super-resolution for different train set resolutions.

Training resolutions	16x16	48x48	144x144	432x432
MRMSE(m)	0.2620	0.1987	0.1770	0.1748

Table 2: Comparison of the MRMSE of predictions, at resolution 1296x1296, on the test set, made by FNO trained at different training resolutions.

4.3 Discussion

Trying to learn the link between SST and SSH at a low resolution and use the learned arbitrary-scale function to retrieve SSH from known SST at higher resolution was based on two strong hypothesis. First, we assumed that the link between SST and SSH was learnable, and then that it was generalizable through different resolutions. The results seem to point that FNO is able to reconstruct the SSH with good precision considering that it is only fed with the SST and as no knowledge of the SSH event at low-resolution. Furthermore, the precision of the reconstruction tends to decrease when the difference between the training data set and the evaluation resolution gets bigger. This suggests that the link learned at one resolution is becomes less relevant when the resolution increases. It seems confirmed that the link between SST and SSH is strong and learnable, however this link seems to highly depend on the resolution.

Elsewhere, it can be noted that visually the super-resolution made by FNO is really different from the one made by LIIF. LIIF tends to output a very smoothed super-resolution, missing out some high-frequency components, on the contrary FNO seems to better predict small phenomenons, and high frequency components, despite a less good global value prediction. It seems that the prior complementary information brought by the SST helps the model reconstruct smaller components of the SST. However, FNO is also really different from LIIF as it is a global model thanks to its transformations performed in spectral space, where LIIF is specially a local space.

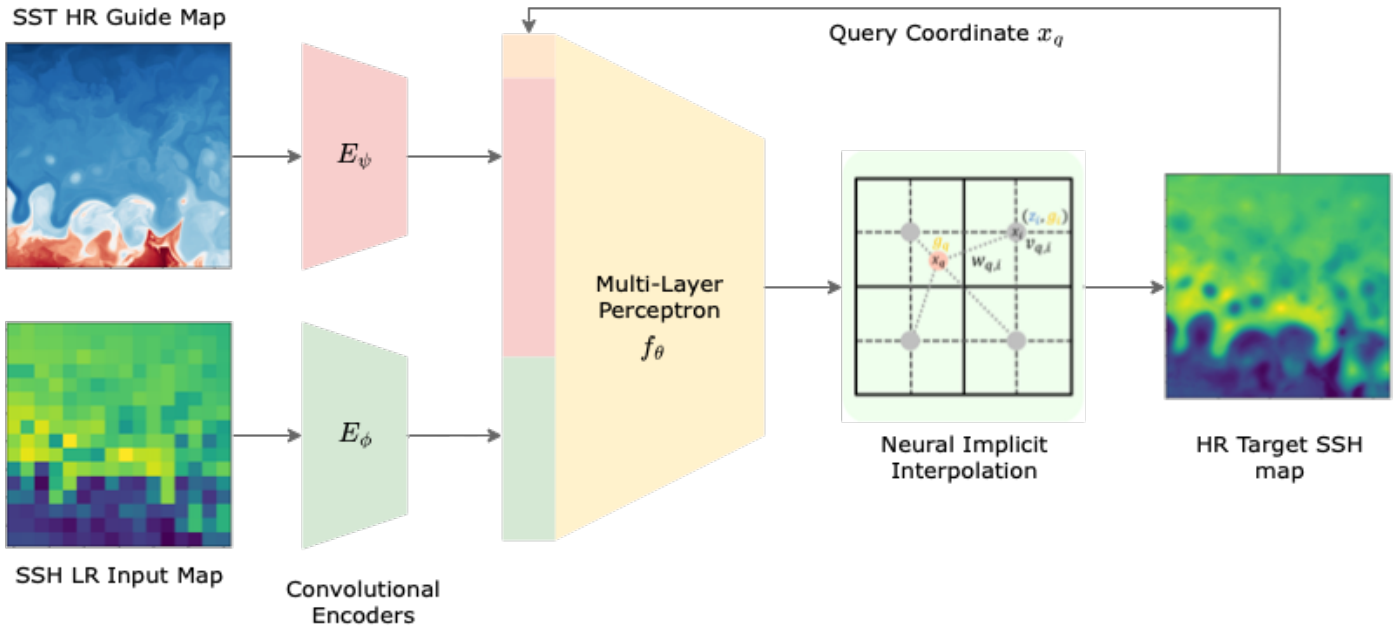


Figure 12: Architecture of the JIIF network for SSH super-resolution using SST.

5 SSH guided implicit super-resolution using SST

5.1 Method

We saw with FNO that adding SST high resolution prior greatly helps the downscaling of the SSH especially for high frequency and small phenomenons and dynamics, having visually better results. However, without the information of the low-resolution SSH, FNO had troubles to be precise enough when reconstruction the values, thus being worse than LIIF on quantitative indicators like MRMSE. Therefore, we would want to have a network leveraging the information of the low-resolution SSH and the one from high-resolution SST. This task is close to the guided super-resolution task.

JIIF architecture, proposed by [Tang et al., 2021], is based on the idea of LIIF but oriented toward the guided super-resolution task. Their motivation is the super-resolution of low-resolution depth maps of a scene using the associated high-resolution RGB images of the same scene, in our case we have quite the same configuration with low-resolution SSH maps and a high-resolution SST maps of the same zone. JIIF adds the super-resolution image to the input of LIIF to guide the super-resolution. To that end, they modified some parts of the LIIF network : the high-resolution input is also encoded to a higher dimension by an other convolutional encoder and the features are added to the input of the Multi Layer Perceptron. The new architecture of JIIF is shown on Fig. 12 More precisely, the inputs of the implicit neural network in JIIF are : the features of the encoded low-resolution image at the considered neighbor z_i, the features of the encoded high-resolution image g_i, the difference between the coordinates of the queried pixel and the one of the neighbor $x_q - x_i$, and finally the difference between the high resolution features at the queried pixel position and the ones at the neighbor position $g_q - g_i$. Furthermore, the JIIF network not only predicts the value of the signal for the considered neighbor $a_{q,i}$ but also the weight of this value in the weighted average $v_{q,i}$ that follows. Thus, the network learns

the following function :

$$a_{q,i}, v_{q,i} = f_\theta(z_i, g_i, g_q - g_i, x_q - x_i), \tag{3}$$

for i a neighbor in the ensemble of neighbors : $N = \{00, 01, 10, 11\}$. The two vectors of features are taken from the images encoded by the two convolutional encoders (E_ϕ and E_ψ), for example : $z_i = E_\phi(SSH_{LR})[x_i]$ and $g_i = E_\psi(SST_{HR})[x_i]$. Then, the final predicted value of the pixel at x_q in the high-resolution image I is :

$$I(x_q) = \sum_{t \in N} w_{q,t}.a_{q,t} \tag{4}$$

with the weights resulting from the following transformation on the prediction $v_{q,i}$ made by the network :

$$w_{q,i} = \frac{exp(v_{q,i})}{\sum_{t \in N} exp(a_{q,t})} \tag{5}$$

To account for the increase of the input dimension, we changed the structure of the Multi-Layer Perceptron to four layers of respective dimensions : 1024, 512, 256, 128.

5.2 Results

The JIIF model is more complex than LIIF, having a supplementary convolutionnal encoder, and bigger input features due to the presence of high-resolution information in the inputs. Thus it turned out that this network was harder to train and took more time to converge. For this purpose, and taking into account the little time left for these experiments, we chose to train the network on a smaller super-resolution than LIIF. The low-resolution inputs are of size 16x16 like LIIF, but the high-resolution target is 144x144 which corresponds to a super-resolution scale of x9. Moreover, we decided first to only train the network to predict the value of the signal and not the weight of the weighted sum to accelerate the training. We evaluated the trained algorithm on the x9 and the x27 task to see its generalization abilities.

Super-resolution scale	x3	x9	x27
MRMSE(m)	0.0671	0.0620	0.0704

Table 3: Comparison of the MRMSE of predictions, at resolution 1296x1296, on the test set, made by FNO trained at different training resolutions.

The results are comparable to the one obtained with the LIIF architecture. The model performs better on the resolution it has been trained on. Qualitatively, the results (Fig. 13) seems to still struggle to reconstruct small structures.

5.3 Discussion

The results seem to show that the new JIIF network has problems, like JIIF to reconstruct high frequency and learn a super-resolution going further than an interpolation. However, JIIF already showed results comparable to LIIF while trained on a smaller scale and without all its features (predicting the weights of the weighted sum for example). In spite of all, JIIF has the same limitation than LIIF regarding the local nature of the function,

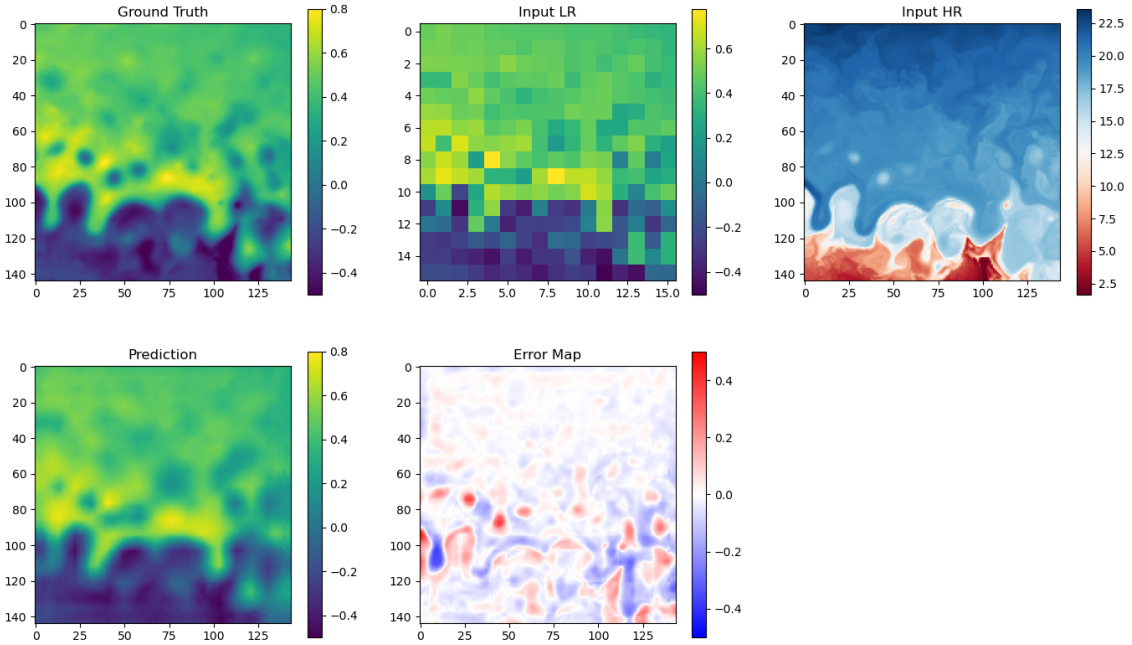


Figure 13: Results of the JIIF model for x9 super-resolution on a data point of NATL60 test set.

we saw with FNO that having a global or spectral point of view was a big help in learning to reconstruct the small phenomenons and to translate them from SST to SSH. Maybe the LIIF/JIIF architecture, as such, is to local to really downscale the physic of the SSH. A lead would be to try incorporate a global or spectral component to LIIF.

6 Conclusion

6.1 Technical Conclusion

First idea was to apply naively a computer vision method to the SSH-downscaling task. But it appeared that this task falls more in the physic-based domain than the computer vision one. In fact, the study objects are not images but observations of complex dynamics, thus while descending the scale, new high frequency components like small eddies appear and the classical computer vision models seems unable to capture and predict those dynamics at high resolution with the prior given by the input low resolution image given. But an the other had, we also have a deeper physical understanding of the SSH data making it possible to add priors to the models. This motivated us to follow the path of [Archambault et al., 2022] and use SST high-resolution to guide SSH downscaling. A first test with neural operators allowed us to confirm the fact that the link was learnable between SST and SSH. Therefore, we then decided to adapt the initial used architecture to take as input high-resolution SST information in order to add complement physical information to the model and refine the super-resolution task. Further developments could be to push the hybridation between computer vision model and physic based techniques even further and add global knowledge in LIIF as done in FNO through spectral manipulations of the signal.

6.2 Experience feedback

During my study at the Ecole Centrale de Lille, I got the chance to dive into the Machine Learning field and chose to deepen my understanding of this field. Later, other experiences led me to discover the oceanography field, its challenges and how Machine Learning and Data Science could play a role in their resolution. Therefore, it was for me a perfect choice to do my end-of-study internship in a research lab on Machine Learning with an application side on oceanography problematic, it allowed me in addition to get a deeper understanding of both studied fields to discover the academic field. This experience has been a perfect addition to my engineer pathway until now. Evolving in a public Machine Learning allowed me to discover a different approach of the Deep Learning models and the domain, there was a constant search for a real understanding of the architecture and the intuition behind them, trying to understand what the model does, the hypothesis it makes. This mindset allows to take a step back, have a wider understanding of the model and thus have a better idea of how appropriate it will to tackle a given problem. I think that those learning are really valuable in the Machine Learning field even in the private sector, and more particularly in Research and Development teams where I see myself working. Moreover, even if a big part of the work and most of the programming part is done on its own, I've had numerous of passionate conversations with colleagues interns, PhD students, professors that were always an awesome source of ideas and new point of views. It has also be the opportunity to spend a consequent time on bibliography and studying very recent works, thus learning how to apprehend and understand a scientific paper and lead a research on a subject and its state of the knowledge. Furthermore, it has also been a challenge to be able to communicate on my research, orally or in writing and to make understandable precise technical topics to people not familiar with them, but I'm convinced that it is a good ability to learn, particularly to evolve in the future in teams composed with people from different background and communicate about its results and its work in a credible way. Finally, it has been a really wonderful experience to evolve in a stimulating

environment, surrounded by passionate people, and it have been a great opportunity to learn so much and even further than my field of research, when meeting with people working on various subject like Computer Vision, NLP, Physic-based Deep Learning, Data Assimilation, Oceanography, Physical Modeling, Complex Dynamical Systems... I had the chance to learn from many people on many different subject in discussion, meetings, seminars, thesis defense and I loved to be a part of this scientific emulation.

References

[Archambault et al., 2022] Archambault, T., Charantonis, A. A., Béréziat, D., Mejia, C., and Thiria, S. (2022). SSH Super-Resolution using high resolution SST with a Subpixel Convolutional Residual Network. In *Climate Informatics*, Asheville, NC, United States.

[Barthélémy et al., 2022] Barthélémy, S., Brajard, J., Bertino, L., and Counillon, F. (2022). Super-resolution data assimilation. *Ocean Dynamics*, pages 1–18.

[Chen et al., 2021] Chen, Y., Liu, S., and Wang, X. (2021). Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8628–8638.

[Dong et al., 2014] Dong, C., Loy, C. C., He, K., and Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer.

[Hewitt et al., 2022] Hewitt, H., Fox-Kemper, B., Pearson, B., Roberts, M., and Klocke, D. (2022). The small scales of the ocean may hold the key to surprises. *Nature Climate Change*, 12(6):496–499.

[Jang and Agapito, 2021] Jang, W. and Agapito, L. (2021). Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958.

[Jiang et al., 2020] Jiang, C. M., Esmaeilzadeh, S., Azizzadenesheli, K., Kashinath, K., Mustafa, M., Tchelepi, H. A., Marcus, P., Prabhat, and Anandkumar, A. (2020). MeshfreeFlowNet: A Physics-Constrained Deep Continuous Space-Time Super-Resolution Framework.

[Kovachki et al., 2021] Kovachki, N. B., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A. M., and Anandkumar, A. (2021). Neural operator: Learning maps between function spaces. *CoRR*, abs/2108.08481.

[Landgraf et al., 2022] Landgraf, Z., Hornung, A., and Cabral, R. (2022). Pins: Progressive implicit networks for multi-scale neural representations.

[Ledig et al., 2017] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690.

[Li et al., 2020] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. (2020). Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.

[Lim et al., 2017] Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. *CoRR*, abs/1707.02921.

[Lindell et al., 2021] Lindell, D. B., Van Veen, D., Park, J. J., and Wetzstein, G. (2021). Bacon: Band-limited coordinate networks for multiscale scene representation. *arXiv preprint arXiv:0000.00000*.

[Liu et al., 2021] Liu, Y.-T., Guo, Y.-C., and Zhang, S.-H. (2021). Enhancing multi-scale implicit learning in image super-resolution with integrated positional encoding. *arXiv preprint arXiv:2112.05756*.

[Meija et al., 2021] Meija, C., Jean-Marc Molines, and Sorror, C. (2021). The resac-sargas60 dataset on the area of the sargasso sea.

[Mildenhall et al., 2020] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

[Rodrigues et al., 2018] Rodrigues, E. R., Oliveira, I., Cunha, R., and Netto, M. (2018). Deepdownscale: a deep learning strategy for high-resolution weather forecast. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 415–422. IEEE.

[Sitzmann et al., 2020a] Sitzmann, V., Chan, E., Tucker, R., Snavely, N., and Wetzstein, G. (2020a). Metasdf: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems*, 33:10136–10147.

[Sitzmann et al., 2020b] Sitzmann, V., Martel, J. N., Bergman, A. W., Lindell, D. B., and Wetzstein, G. (2020b). Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*.

[Tancik et al., 2020] Tancik, M., Srinivasan, P. P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. T., and Ng, R. (2020). Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*.

[Tang et al., 2021] Tang, J., Chen, X., and Zeng, G. (2021). Joint implicit image function for guided depth super-resolution. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4390–4399.

[Vandal et al., 2017] Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., and Ganguly, A. R. (2017). Deepsd: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 1663–1672.

[Xu et al., 2021] Xu, X., Wang, Z., and Shi, H. (2021). Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. *arXiv preprint arXiv:2103.12716*.

[Yu et al., 2021] Yu, A., Ye, V., Tancik, M., and Kanazawa, A. (2021). pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587.

[Zhang et al., 2018] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. (2018). Residual dense network for image super-resolution. *CoRR*, abs/1802.08797.