# Enhancing Detail Recovery in ICF Radiographs: A Transformer-based Approach with ViXReg

**Nga T. T. Nguyen-Fotiadis**
Information Sciences
Computer, Computational, and Statistical Sciences Division
Los Alamos National Laboratory
Los Alamos, NM 87545
nga.nguyen@lanl.gov

**Bradley Wolfe**
Thermonuclear Plasma Physics
Physics Division
Los Alamos National Laboratory
Los Alamos, NM 87545

**Zhehui Wang**
Thermonuclear Plasma Physics
Physics Division
Los Alamos National Laboratory
Los Alamos, NM 87545

## Abstract

We introduce ViXReg, a framework that adapts Vision Transformers (such as Google ViT, Swin, BEiT) to tackle image analysis challenges in Inertial Confinement Fusion (ICF) radiography. ViXReg introduces a novel approach by adapting Vision Transformers, originally designed for pixel-level classification, to handle complex image regression tasks. This transformation, which is not commonly explored in current methods, enables precise reconstruction of asymmetric double-shell structures, essential for diagnosing nuclear fusion dynamics and identifying instabilities in high-energy-density plasmas. Our investigation explores architectural adaptations, including nonlinear and linear mappings, and advanced fine-tuning strategies like multi-scale pre-training and knowledge distillation, enhancing model scalability and generalization across diverse data distributions. Evaluating 60,000 synthetic ICF radiographs and 115 radiographs captured from 6 ICF experimental shots, we further craft domain learning techniques with weakly pseudo-labeled data, enabling ViXReg to transfer robust representations effectively to experimental dataset. The results from the above tasks demonstrate considerable advancements in using transformers as backbone architectures for fusion imagery, effectively capturing the subtle double-shell structures identified in plasma physics. Additionally, fine-tuning the pre-trained ViXReg model accelerates training convergence and enhances the accuracy of double-shell reconstructions, surpassing the performance of traditional convolutional neural networks and generative adversarial models. These findings demonstrate ViXReg's potential as a candidate for a foundation model component in scientific modeling for nuclear fusion research.

## 1 Introduction

Inertial Confinement Fusion (ICF) represents a significant branch of high-energy physics, primarily focusing on initiating fusion reactions. Central to this process are ingeniously designed double-shell capsules[1, 2]. These tiny yet complex structures, composed of an external ablator layer and an internal shell, play a crucial role to compress and heat a mixture of Deuterium and Tritium, two

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

hydrogen isotopes[2], creating the extreme conditions necessary for fusion governed by essential physical parameters, such as neutron scattering and gamma radiation, impacting how we analyze related imagery[1, 3, 4]. However, conventional analysis methods, including correlations[5], often struggle with these datasets. Specifically, they can fail to adequately capture subtle but crucial elements like variations in neutron scattering or the finer details in gamma radiation profiles, critical for plasma diagnostics[6]. This challenge is compounded by the typically limited availability of high-resolution and reliable experimental data, posing a significant hurdle to achieving accurate image analysis.

Machine learning (ML)[7–10] techniques are increasingly applied to these datasets, providing insights into ICF and contributing to the advancement of nuclear fusion as a sustainable energy source. Vision Transformers (VTs)[11–13] have rapidly become a new standard in the field of computer vision, offering a paradigm shift in how we approach image analysis. They serve as core structures in visual foundation models (FMs)[14–16]. [*To avoid confusion, we use the term "Vision Transformers" (VTs) to refer to vision architectures that adapt the self-attention mechanism for images, while "ViT" specifically refers to the Vision Transformer variants initially developed by Google.*] The ability to capture long-range dependencies and complex patterns within images makes VTs particularly well-suited for the dynamic imagery found in ICF experiments. This transformative potential of VTs has motivated us to pursue a streamlined and energy-efficient model architecture centered around VTs as foundational components for scientific models, particularly those that heavily involve image analysis.

VTs process images in patch sequences[12] using self-attention mechanisms[17], focusing on key aspects while reducing dependence on pre-processing like super-resolution. This attribute is valuable in ICF imaging, where fine shell details may indicate new physical insights, supporting our goal to develop an energy-efficient model while maintaining or enhancing the accuracy compared to hybrid approach incorporated super-resolution.

In addition, the strategic shift towards employing vision transformer (VT)-centric models in scientific imaging, particularly in plasma physics, requires a deep understanding of their operational strengths and potential limitations. We explore fine-tuning strategies such as multi-scale pre-training and knowledge distillation through a teacher-student model to improve ViXReg's performance in reconstructing double-shell structures in ICF radiography. These approaches leverage scaling laws and model adaptation techniques to enhance convergence and accuracy, even when beginning with modest baseline models like MobileNet[18]. Additionally, we employ domain learning with weakly labeled data, encouraging the model to learn robust representations that transfer from synthetic datasets to noisy experimental images. Our results demonstrate that these techniques not only accelerate training convergence but also improve the accuracy of double-shell reconstructions. This adaptability will examine the architecture's transferability and generalization.

## 2 Methodology

### 2.1 Imagery Architecture within ViXReg

Key to our work is the crafted adaptation and integration of VT models into our specialized ***regression-based*** framework, ViXReg. This includes the adaptation of Google ViT Large and Base instances [12], BEiT [19], and Swin Base and Swin Tiny variants [20]. We inherit parts of these models from their original image classification roles, adapting them for regression tasks by transforming their outputs into mappings $\mathrm{Y} = f(\mathrm{X})$ at the regression head stage. Here, X represents the feature embeddings generated by the model after processing the input ICF radiograph, and Y corresponds to the precise double-shell inference or target. We implement $\mathrm{f}(\mathrm{X})$ as a linear or nonlinear mapping. One strategy to do that is to preserve the foundational layers of these VT models, incorporating them up to certain later learning stages. This is then followed by the customized addition of nonlinear/linear head consisting of concatenated feature layer and dense tables, culminating in a tailored regression mapping. While the core of the model inherently introduces nonlinearity through self-attention and convolutional operations, the distinction lies in the structure of the final regression mapping.

For Convolutional Neural Networks (CNNs), we apply a similar adaptation. For instance, in Xception [21], we freeze the high-level feature extraction layers and introduce a dense layer with 1024 neurons to effectively interpret ICF double shells. The outputs from the last middle and exit flow layers are concatenated into a single layer, which then connects to a custom 1024-neuron dense

| Custom Models | Xception | InceptionV3 | Swin Tiny | ResNet152 | ViT Base | BEiT Base | Swin Base | ViT Large |
|---|---|---|---|---|---|---|---|---|
| Parameters (M) | 19.4 | 20 | 26 | 58 | 83 | 83 | 86 | 305 |

Table 1: Complexity of the core models used in our ViXReg framework, sorted in ascending order by the number of parameters (in millions, M).

table. Fine-tuning the kernel sizes at this stage further enhances performance. Our control experiments indicate that this customized architecture achieves comparable accuracy to models with denser configurations. Additionally, we extended this adaptation strategy to other CNN models, such as InceptionV3 [22] and ResNet152 [23], for a comprehensive comparison with our VT-based regression architectures.

Our approach is a departure from traditional methods commonly used in scientific imaging tasks, such as denoising [24, 25] or conditional generative adversarial networks (c-GANs)[26–28]. While many existing methods focus on indirect tasks like denoising to enhance signal quality, our direct regression approach provides a more efficient solution for extracting meaningful patterns and structures from scientific imagery.

### 2.2 ICF radiograph dataset

We developed a dataset of 60,000 synthetic radiographs paired with corresponding double-shell inferences, both of image size $256 \times 256 \times 3$, serving as reference standards. These synthetic radiographs $\hat{\mathrm{X}}$ were generated from their inference capsules using the Beer-Lambert Law, computed numerically via the Tomographic Iterative GPU-based Reconstruction Toolbox (TIGRE)[29] incorporating physical features characteristic of ICF shots, ensuring realism in the dataset, $\mathrm{Y}$. In our ViXReg framework, we address an inverse regression problem: raw inputs $\hat{\mathrm{X}}$ are processed through the model's core layers (e.g., self-attention or convolutional layers) to generate feature representations, denoted as X, which are used to predict the target $\mathrm{Y}$.

For clarity, examples of the raw paired images are shown in fig. 3, where the simulated radiograph input, depicted as $\hat{\mathrm{X}}$, is shown in the first column, and the corresponding double-shell inference as $\mathrm{Y}$ in the second column. Additionally, we obtained 115 experimental ICF radiographs from six distinct ICF shots, used as an extra test set to evaluate our model's generalization capability. These experimental images lack corresponding ground truth inferences, presenting a unique challenge for validation. We consulted plasma physicists to help valuate the double-shell reconstructions for this real-world case. In scientific imaging of X-rayed ICF, we tackle the challenge of noise, evident in low Peak Signal-to-Noise Ratio (PSNR), that inhibits image processing. We applied unsupervised noise2noise [30] algorithm to remove salt-and-pepper adversarial noise to our synthetic and experimental ICF datasets.

The synthetic ICF dataset, discussed above, is divided into training, validation, and test sets in a 0.8/0.1/0.1 ratio. We trained ViXReg on 50,000 denoised ICF images and their corresponding inferences as targets, validated on 5,000 separate images and inferences. All subsequent ViXReg's results shown in this paper are on the test set consisting the remaining 5,000 ICF images and inferences and later on 115 experimental ICF images.

### 2.3 Optimizing regression mapping: model and data scaling

We conducted a series of control experiments by testing different key layers within various architectures on general VT models, identifying two optimal mapping types for $\mathrm{f}(\mathrm{X})$ at the later learning stage of regression. The first type, nonlinear mapping, involves a multi-layer configuration (three dense layers with 256, 512, and 1024 neurons, followed by a concatenated layer) applied during the later stages of training. The second type, linear mapping, directly connects input pixels to target pixels through a single dense layer on top of a concatenated layer, integrated with either a self-attention or convolutional learning architecture. The preserved backbone learning layers and the number of hyperparameters that need optimization when training ViXReg are summarized in table 1. As shown in fig. 1, the nonlinear mapping approach results in consistently higher loss values across epochs compared to the linear mapping strategy. This is shown in the top curves (blue for ViT Base and orange for Swin Base) versus the lower loss values observed in the remaining 12 curves representing the 6 linear mapping cases using the same core architectures of ViT and Swin variants. This suggests that the linear mapping method achieves better performance by an order of magnitude.

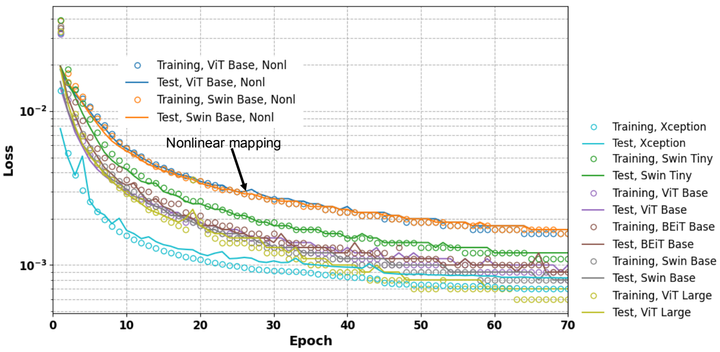

Figure 1: Training and test losses per epoch on a logarithmic scale (base 10). The y-axis represents the loss values to highlight convergence patterns across orders of magnitude. The top four curves/circles (orange and blue) represent losses for nonlinear mapping $\mathrm{f(X)}$, while the remaining twelve curves/circles in six colors (right legend) correspond to linear mapping $\mathrm{f(X)}$.

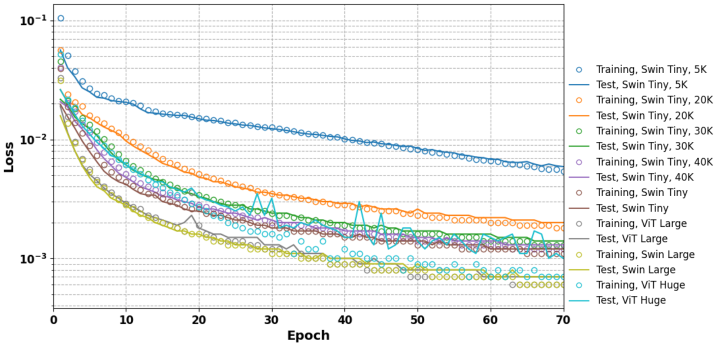

Figure 2: Training and test losses per epoch on a logarithmic scale (base 10) to highlight convergence patterns across multiple orders of magnitude. Losses are shown for the linear mapping protocol $\mathrm{f(X)}$ of Swin and ViT families, evaluating model and data scaling.

| Errors | Xception | InceptionV3 | Swin Tiny | ResNet152 | ViT Base | BEiT Base | Swin Base | ViT Large | Swin Large |
|---|---|---|---|---|---|---|---|---|---|
| $\bar{\epsilon}$ ($10^{-3}$) | 1.10 | 29.97 | 0.97 | 30.50 | 0.84 | 0.86 | 0.82 | 0.63 | **0.58** |
| $\epsilon$ hard ($10^{-3}$) | 27.46 | 59.90 | 20.9 | 62.21 | 25.31 | 22.53 | **17.62** | 19.20 | 18.67 |

Table 2: Errors measured by MSE $\epsilon$ for ViXReg's study models presented in table 1 performed on the 5,000 test radiograph images. The top row shows the average values of $\epsilon$ over the entire test set, denoted as $\bar{\epsilon}$, and the bottom row represents the hard study case shown in fig. 4.

We trained our most effective models using the direct linear mapping over 120 epochs with 50K data pairs (radiographs and double-shells), and the final error metrics are presented in table 2 (top row). For visualization purposes, we plotted the first 70 epochs (out of 120 epochs) to clearly capture the acceleration rate.

The scaling law on the models' size and dataset aspects are shown in figs. 1 and 2. In fig. 1, the performance on learning our 50K pair dataset across different variants of Swin and ViT is explored. Although Xception performs well, its dense learning structure shows limitations in our control runs on different datasets; the compact structure is less effective than VTs when handling similar regression problems with higher resolution data, such as 512×512×3 images. InceptionV3 and ResNet152 exhibit substantial noise, resulting in large fluctuations in both training and test losses (not plotted here due to intensive feature observation). In contrast, VT families, despite having a larger number of hyperparameters, show better efficiency in scaling both model size and input size.

Among the models, Swin Tiny, the smallest VT variant with the least number of parameters on the backbone, has the highest error rate but still outperforms the Swin Base with nonlinear head mapping. The Swin family demonstrates better scalability in model size compared to ViT, as ViT Large shows signs of overfitting (see also fig. 2). When handling data size, the Swin family also performs better than ViT (see fig. 2). We examine data scaling by gradually increasing the training set from 5K to 50K (from magenta to blue curves/circles), with the trained model tested on a fixed set containing 5K pairs of radiograph images and corresponding double-shells. An apparent improvement is observed as expected.

Regarding model size, Swin continues to show scalability up to the largest model considered, Swin Large, which contains roughly 194M parameters in our customized regression model's backbone. Additionally, despite being smaller than ViT Large or ViT Huge ($\approx$ 630M parameters in our customized regression variant), Swin Large shows a smaller error on the test set (see table 2). For our subsequent learning and mechanism testing runs, we will mainly focus on Swin Base due to its size and efficiency in examining our architecture and hybrid learning protocol. We plan to explore ViT Huge and ViT Large to examine their characteristics as suitable FM components within a larger dataset scale.

## 3 ViXReg: Foundational model properties

### 3.1 Model robustness: Understanding ViXReg's effectiveness with low-resolution dataset and non-label dataset

#### 3.1.1 Model accuracy on the ICF synthetic dataset

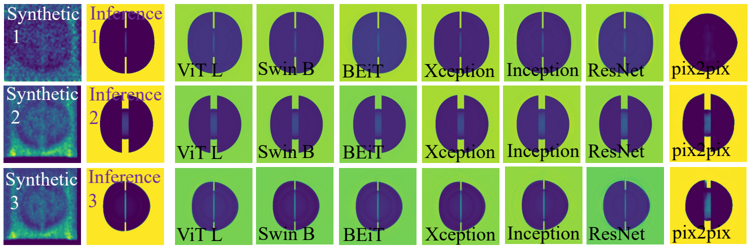

Figure 3: Three synthesized ICF radiographs (far-left column) alongside their corresponding ground truths as inferences (second far-left column). The double shell inference structures are reconstructed with ViXReg, implemented using 6 distinct regression models adapted from ViT Large, Swin Base, BEiT, Xception, InceptionV3, and ResNet152, as well as the cGAN pix2pix mapping, shown respectively, from left to right in the last 7 columns. Images of original size $256 \times 256 \times 3$.

First, we analyzed the generalizability of six models, evenly split between Vision Transformer (VT) and Convolutional Neural Network (CNN) architectures, using their mean squared error (MSE) between reconstructed $\hat{\text{Y}}$ and the original inference Y: $\epsilon = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2$, as shown in table 2. Swin Large appears to be the best backbone for the ViXReg regression architecture, with the smallest MSE error ($\bar{\epsilon} = 5.8 \times 10^{-4}$), followed by ViT Large ($\bar{\epsilon} = 6.3 \times 10^{-4}$), then followed closely by Swin Base ($\bar{\epsilon} = 8.2 \times 10^{-4}$). Despite their larger sizes, detailed in table 1, VTs demonstrated more efficient memory usage with larger inputs in the ViXReg regression framework compared to CNNs. Notably, the ViT family manages memory allocation more effectively than the Swin family. However, due to their significantly longer training times, Swin Large- and ViT Huge-inspired regression models will be excluded from the following numerical experiments (see also fig. 2).

Despite InceptionV3 demonstrating stable convergence, its generalizability, like that of ResNet152, is relatively lower compared to the other models, with errors of $\epsilon = 0.02997$ and $0.0305$, respectively, as shown in table 2. By contrast, our adapted regression models within ViXReg, as detailed in table 1, showed significantly higher precision in reconstructing synthetic double-shell structures on the test ICF images. This performance is illustrated in the double-shell reconstructions in fig. 3, with the corresponding error metrics provided in the top row of table 2.

In fig. 3, we further compared the results obtained using VTs and CNNs with those produced by a conditional Generative Adversarial Network (cGAN), specifically pix2pix [26, 31]. Unlike ViXReg, which optimizes for regression tasks, pix2pix utilizes a cGAN architecture to learn a supervised mapping from radiographs to inference targets. Across all models tested, ViXReg consistently outperformed pix2pix in terms of accuracy. The results from pix2pix, presented in the last column of fig. 3, revealed notable mispredictions, particularly in the double-shell interiors and asymmetries in the two representative cases (top and bottom rows). Additionally, in our control experiments, we evaluated other cGAN architectures, such as CycleGAN [28] and TransGAN [27]. However, pix2pix demonstrated the best performance among these cGAN models for image-to-image translation tasks on our ICF synthetic dataset. This highlights ViXReg's superior adaptability and precision in handling complex regression tasks, especially when compared to traditional cGAN approaches.

### 3.1.2 Model efficiency in low-resolution image analysis

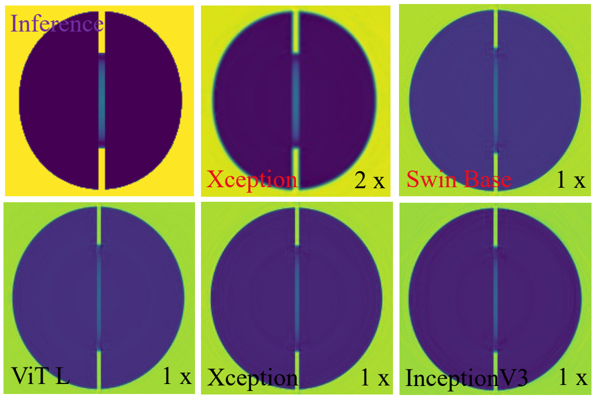

Figure 4: A hard study case of synthesized ICF radiograph with inference ground truth shown in the top left labeld as 'Inference'. The double shell inference is reconstructed using four distinct regression models, Swin Base, ViT Large, Xception, and InceptionV3, shown in the remaining 5 images. 2 x and 1 x refer, respectively, to the enhanced resolution of 2 times and original ICF radiograph. Original image size is $256 \times 256 \times 3$. 2 x resolution is $512 \times 512 \times 3$. To obserfve finer details, it is recommended to use a zoom-in tool.

**Learning the finer details with vision transformers** We previously built a hybrid ML framework that incorporated an unsupervised super-resolution SRCNN [32] to a regression protocol based on, e.g., Xception and received image quality improvement resulting in training speed up by 70%. Super-resolution process is an one-time offline step, and in our previous work, it has been crucial for accurate predictions in synthetic images. To evaluate the advancements of regression framework ViXReg as a *standalone* architecture, we examined a very complex ICF case with inner shell rim "hidden" in the inference, which can be mispredicted as visible inner shell contour. One such study inference is the top left image in fig. 4. This inference appearance serves for both image sizes $256 \times 256 \times 3$ and $512 \times 512 \times 3$. Hereafter, unless specifically referring to the pixel count, we will omit the channel dimension ($\times$ 3) for simplicity when discussing image scale. The performances of our adapted VTs and CNNs on this complex example are characterized in fig. 4 for double-shell reconstructions and table 2 (bottom row) for quantified errors (MSE $\epsilon$). Xception integrated with a super-resolution-enhanced process on $512 \times 512$ ICF dataset in a separate hybrid learning algorithm, Swin Base and ViT Large on the study ICF dataset without super-resolution ($256 \times 256$) within ViXReg, as in top row and bottom left of fig. 4, achieved subtle inner shell details mostly obscured and visible only at certain edges, shown in the top left ground truth in fig. 4. The limited capability of Xception on the original-sized dataset, i.e. within ViXReg, is revealed in the middle image of the bottom row in fig. 4 (MSE $\epsilon = 0.02746$). Both Xception (1x) and InceptionV3 ($\epsilon = 0.0599$), with reconstructions shown in the bottom row of fig. 4 and errors in the bottom row of table 2, exhibit variations in their reconstructions, occasionally missing subtle "details" along the inner shell rim, as discrepancy highlighted against the top left ground truth in fig. 4. This analysis emphasizes the distinct precision levels of each model in complex imaging and examines the energy efficiency of CNNs and VTs as independent regression architectures.

**Model transferability** ViXReg leverages specifically trained regression models on the training synthetic dataset to analyze observable radiographs in ICF dynamics where inferences are not available. In fig. 5, we present the inferred double-shell capsules using all top models under study (ViT Large, Swin Base, ViT Base, BEiT, Swin Tiny, Xception), whose origins and errors are detailed in table 1 and table 2, respectively. Despite broadly similar training convergence patterns across models, their diverse interpretations on two experimental ICF shots highlight significant differences, as seen in fig. 5. Consultation with experts shows that models like Swin Base, Swin Tiny, ViT Base, Xception have distinct generalizability compared to ViT Large and BEiT, each showing relatively unique extrapolation behaviors (double shells, asymmetry) in Experiment 1 (top row

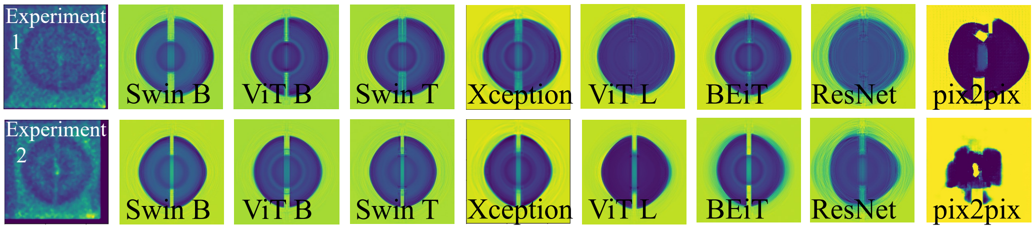

Figure 5: Double shell capsule inference reconstructions on two experimental ICF input radiographs (far left column) using transferred learning weights obtained with 7 ViXReg's regression variants. No ground truths on the inferences are available. We compare our adapted regression instances with cGAN pix2pix. Images of original size 256 × 256 × 3.

of fig. 5). In contrast, these models demonstrate more consistent extrapolation in Experiment 2 (bottom row). Notably, the cGAN pix2pix mapping demonstrates limited effectiveness, evident from disruptions in image reconstructions (far right column in fig. 5). Overall, ViXReg's VT variants achieve reconstructions with significantly less noise, identified as the black and greenish 'fringes' observed outside the active double-shell region in Xception's and ResNet152's predictions (respectively, fifth and eighth column from the left in fig. 5).

The limitations of ViXReg are highlighted by the mismatches observed in double-shell predictions for experimental radiographs, where no ground truths are available, across different VT architectures, indicating areas for improvement. In the remainder of the paper, we discuss several strategies to examine and enhance the performance of ViXReg from a foundation model perspective, including its application to multi-scale tasks and the use of nested learning architectures for better generalization. We explore fine-tuning methods tailored to our regression and imagery analysis to develop efficient building blocks for a foundation model in radiography. These strategies include knowledge distillation, few-shot learning, and pseudo-labeling techniques, all of which are integrated within our ViXReg framework.

### 3.2 Enhancing ViXReg through finetuning and multi-scale training

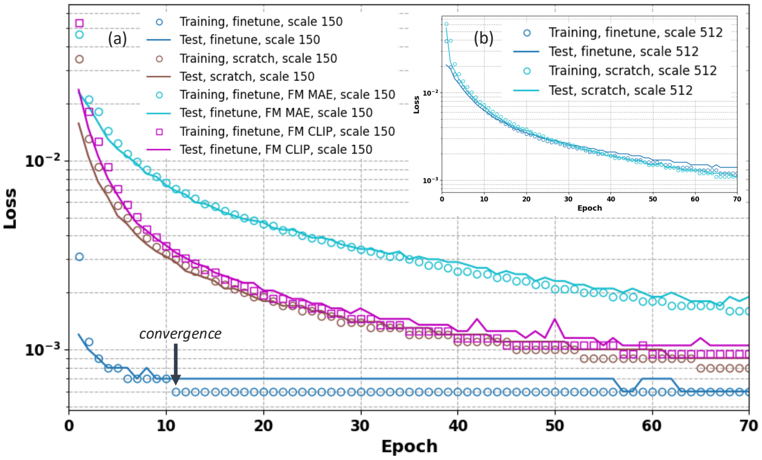

Figure 6: ViXReg's multi-scale variant, built with the Swin Base backbone, is finetuned on the ICF synthetic dataset at two scales: (a) 150 × 150 and (b) 512 × 512. Training and test loss per epoch are displayed on a logarithmic scale of base 10. Label "convergence" indicates the stage at which finetuning begins from the multi-scale pre-trained ViXReg at a smaller scale (150 × 150). Finetuning results from two existing pre-trained foundation models, MAE and CLIP, are also included for scale 150 × 150.

To enhance the robustness and generalization of our ViXReg model, specifically the ViXReg multi-scale variant with the Swin Base backbone, we adopted a multi-scale training strategy designed to enable the model to learn features across different spatial resolutions. This approach allows the model to capture both coarse and fine-grained features in one running, making it more adaptive to varying input sizes. We trained the ViXReg model using images at three different scales: 256 × 256 × 3 (original size), 192 × 192 × 3, and 168 × 168 × 3 pixels. During each epoch, the model was exposed to all three scales, iterating over the resized images and performing gradient updates based on the MSE loss between the predicted outputs and the true targets. The model's weights were updated using the Adam optimizer.

The ViXReg multi-scale model, pre-trained on different scales, was then fine-tuned on the ICF synthetic images at two scales: (a) 150 × 150 and (b) 512 × 512. At the smaller scale (150 × 150), fine-tuning significantly accelerated convergence, achieving faster learning with lower errors compared to: (i) two state-of-the-art pre-trained foundation models, MAE [15] and CLIP [16] (see the blue and cyan curves and circles in fig. 6(a)); and (ii) our ViXReg Swin Base variant trained from scratch. For clarity, "training from scratch" in this context is a relative term, as the ViXReg variant is initialized with partially pre-trained weights from Swin (or ViT, BEiT, see details in section 2) models and then adapted for regression tasks. Although it begins with pre-trained weights, the model undergoes full training from scratch specifically for the regression task within ViXReg. While CLIP outperforms MAE, it still underperforms compared to both the finetuned and from-scratch training of ViXReg, as shown by the magenta, cyan, and brown curves and squares/circles in fig. 6(a). Notably, the finetuned ViXReg achieves converged errors within only 10 epochs ($\bar{\epsilon} = 6.1 \times 10^{-4}$). This overall error $\bar{\epsilon} = 9.1 \times 10^{-4}, 1.5 \times 10^{-3}, 9.8 \times 10^{-4}$, respectively, for training from scratch ViXReg, MAE, CLIP after epoch $\approx 70$ (fig. 6(a)). Further testing at various scales, such as 180 × 180 and 112 × 112, confirms that the finetuned multi-scale ViXReg remains the top-performing model. However, at a larger scale (512 × 512), which is twice the original image size (256 × 256) as shown in fig. 6(b), no significant improvement is observed for the finetuned multi-scale ViXReg variant. The model trained from scratch (cyan curve and circles in fig. 6(b) shows slightly better performance toward the end of the training cycle. Our available 512 × 512 ICF dataset has only 10,000 paired images of radiographs and double-shell inferences in the training set and each of 1,500 paired images in the validation and test set.

In the next subsection, we discuss potential learning strategies to overcome this limitation of finetuning when applied to unseen resolutions significantly higher than the pre-trained scales.

### 3.3 Enhancing generalization with hybrid learning: Knowledge distillation and domain learning

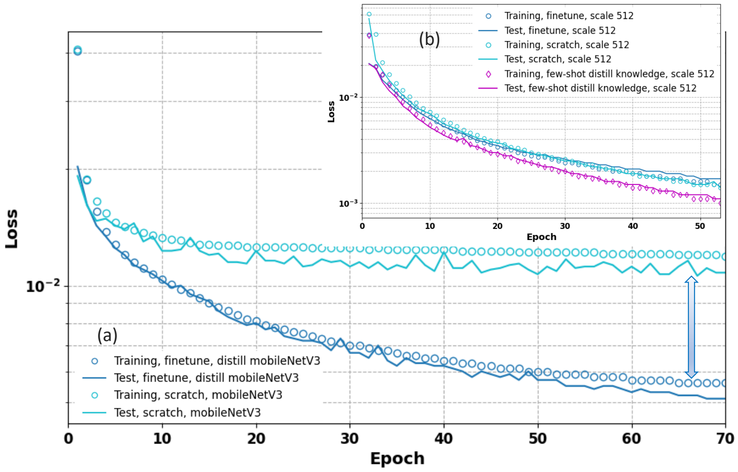

Figure 7: (a) Knowledge distillation (KD) applied to ViXReg Swin Base coupled with MobileNetV3; (b) Application of the KD strategy from (a) to finetuning from the ViXReg multi-scale Swin Base pre-trained model on the 512 × 512 dataset. Training and test loss per epoch are displayed on a logarithmic scale of base 10.

**Distill the knowledge within ViXReg** We first applied a knowledge distillation technique[33, 34] to our original ViXReg model with an image scale of 256 × 256. The motivation behind this approach was to transfer the knowledge from a robust pre-trained model, referred to as the "teacher," to a more compact model, referred to as the "student." The teacher model, in this case, is our customized Swin Base regression within ViXReg and provides "soft labels" to guide the student model. Soft labels are probability distributions over the output space that represent the teacher model's confidence in its predictions. The student model is trained to replicate the teacher's behavior. To examine the pronounced effectiveness of the distillation, we design student model from a relatively modest mode, as a MobileNetV3[18] adapted regression. The student's learning is guided by a training loss as a combination of two components: MSE loss, which captures the difference between the student model's predictions and the actual ground truth targets, and distillation loss (namely Kullback-Leibler divergence)[33]. This loss is written as: $\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{MSE}} + (1-\alpha)\mathcal{L}_{\text{KL}}$ where $\alpha$ is a control hyperparameter that balances the weight of the MSE loss and the distillation loss. This loss function is designed to effectively encourage the student to mimic the teacher's behavior, leveraging both the ground truth labels and the information contained in the soft labels. We demonstrate in fig. 7(a) the effectiveness of this technique applied to ViXReg, with Swin Base as the teacher model and MobileNetV3 as the student model. When trained with our synthetic dataset (50K data points), MobileNetV3 initially exhibits relatively poor performance with high errors (top cyan curve and circles in fig. 7(a) with $\bar{\epsilon} = 10^{-2}$ after 120 epochs. However, when treated as a student model in the hybrid learning framework of knowledge distillation, it shows substantial improvements (bottom blue curve and circles in fig. 7(a)) with $\bar{\epsilon} = 1.8 \times 10^{-3}$ after 120 epochs.

The success of the knowledge distillation approach[33, 35, 36] and our finding above inspired us design of a new learning protocol for ViXReg. In this protocol, the teacher model is a customized Swin Base variant of ViXReg, specifically trained on multiple scales ($256 \times 256$, $192 \times 192$, $168 \times 168$). This teacher model inherits its weights from the pre-trained multi-scale variant, as illustrated in fig. 6. The student model, also a Swin Base variant, is initialized with the pre-trained weights of the teacher model, a good starting point. To further enhance learning, we added a few-shot learning phase on top of the knowledge distillation, a new step compared to our previous architecture fig. 7(a). A small subset (150 samples of paired radiographs and inferences) of the training data was used to fine-tune the student model in this final stage, allowing it to adapt to unseen data. Our results are presented in fig. 7(b). The previous failure to improve fine-tuning for a finer scale not seen during pre-training (see fig. 6(b)) is now resolved using our new hybrid learning framework. The fine-tuning with the distilled ViXReg now outperforms the results from training from scratch (cyan curve and circles in fig. 6(b)). Note the similar trend inherited from the knowledge distill in fig. 7(a) now observed in fig. 7(b) where the loss gap widens with increasing epochs. $\bar{\epsilon}$ attains at $8.4 \times 10^{-4}$ after 120 epochs for the distilled finetuning of ViXReg shown as magenta curve and diamonds in fig. 7(b) compared to $1.2 \times 10^{-3}$ for the two remaining cases shown in fig. 6(b), repost in fig. 7(b).

**Model domain learning: Preliminary results** We augmented the Swin Base-adapted ViXReg regression variant with a semi-supervised learning strategy to evaluate the model's generalizability. The model employs a weakly-labeled domain adaptation approach[37], generating pseudo-labels[38, 37, 39] for 115 *experimental* ICF radiographs that lack ground-truth annotations. Initially, the model learns from synthetic, labeled data and subsequently generates pseudo-labels for the experimental data using its own predictions. These pseudo-labels are then used to refine the model through consistency loss[37]. Various augmentations[40] (such as brightness, contrast, and masking) are applied to the inputs to ensure robust training, allowing the model to generalize effectively despite the absence of labeled experimental data. This approach leverages both supervised learning from labeled synthetic data and unsupervised learning from experimental data with pseudo-labels to improve generalization.

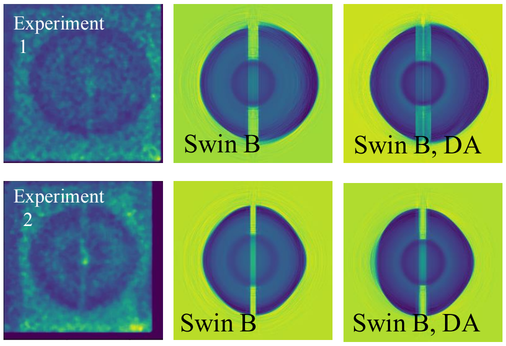

Figure 8: Domain learning results (last column) using weakly labeled data with the ViXReg Swin Base backbone as the regression variant. The first two columns are repost from fig. 5. Images of original size $256 \times 256 \times 3$.

We show in fig. 8 our preliminary results (last column) obtained with this training methodology and repost the experimental inputs (first column) and results obtained with original Swin Base adapted regression ViXReg in fig. 5. We note that we see the improvement in the accuracy on the overall MSE error $\bar{\epsilon} = 7.1 \times 10^{-4}$, compared to Swin Base $\bar{\epsilon} = 8.2 \times 10^{-4}$ (from table 2).

Domain learning with ICF images presents unique challenges, largely due to the complexity of the images themselves, which contain complex, high-dimensional patterns that are difficult to interpret, given the inherent noise in experimental data. The absence of labeled experimental data further complicates the task, requiring the model to generalize from limited synthetic data to real-world experimental conditions. To further evaluate and enhance the model's performance, we plan to use domain adaptation techniques[41–44] and self-supervised learning (SSL) methods [45, 46], to better handle unseen data distributions of noisy inputs. We are also exploring multi-task learning architectures within ViXReg across different multi-modal datasets, including leveraging complementary learning from multiple probe sources within ICF environment.

## 4 Discussion

Our study evaluates VTs and CNNs for ICF double-shell detection, highlighting Swin Base's and ViT's effectiveness in recovering subtle features in low-resolution plasma images and Xception's efficiency despite its simpler design. VT models, particularly those utilizing cutting-edge architectures (e.g. ViT, Swin, BEiT), offer less noise in imaging regression tasks compared to CNNs. However, the varying interpretation obtained from ViXReg's adapted regression models' predictions on the real-world cases of experimental samples highlights areas for improvement.

Leveraging a teacher-student framework, our model integrates knowledge distillation with few-shot learning to efficiently transfer learned representations from a robust multi-scale trained teacher model to a more compact student model, ensuring high accuracy and rapid adaptability even with limited labeled data. Our findings on model scaling and finetuning indicate that ViXReg holds promising potential when integrated with a much larger data scale.

By applying the principles of scientific FMs, we aim to leverage the scalability and adaptability of ViXReg to effectively handle diverse data types encountered in ICF radiography, such as synthetic simulations and noisy experimental images. Our approach aligns with recent efforts to build FM components tailored for scientific tasks by focusing on domain-specific fine-tuning techniques and robust model architectures capable of capturing the complex details of double-shell structures in fusion environments. This work explores certain potential of FMs to accelerate scientific discovery in nuclear fusion research by creating adaptable, efficient tools that can be fine-tuned for specific diagnostic tasks.

## 5 Acknowledgement

This work was supported in part by the Los Alamos National Laboratory (LANL) OES/ICF program under the supervision of Ann J. Satsangi, and in part by the LANL LDRD DI ArtIMIs led by Earl Lawrence. Computational resources were provided by the LANL High Performance System **Darwin**, as part of the Computational Systems and Software Environments (CSSE) subprogram under LANL's ASC program (NNSA/DOE), utilizing both 4- and 8-GPU A-100 Nvidia accelerators. We also thank the anonymous reviewers from the 2024 NeurIPS Workshop Foundation Models for Science for their helpful comments, which have further improved our manuscript.

## References

[1] S Pfalzner. *An Introduction to Inertial Confinement Fusion (1st ed.).* CRC Press, 2006. URL `https://doi.org/10.1201/9781420011845`.

[2] B.M Haines, R.C. Shah, and J.M. et al Smidt. Observation of persistent species temperature separation in inertial confinement fusion mixtures. *Nat Commun*, 11:544, 2020.

[3] D. N. Fittinghoff, N. Birge, and V. Geppert-Kleinrath. Neutron imaging of inertial confinement fusion implosions. *Review of Scientific Instruments*, 94(2), 2 2023. doi: 10.1063/5.0124074.

[4] S Nakai and H Takabe. Principles of inertial confinement fusion - physics of implosion and the concept of inertial fusion energy. *Rep. Prog. Phys.*, 59:1071, 1996.

[5] Harti R., Strobl M., and Betz B et al. Sub-pixel correlation length neutron imaging: Spatially resolved scattering information of microstructures on a macroscopic scale. *Sci Rep*, 7:44588, 2017.

[6] Yong Ho Kim and Hans W. Herrmann. Gamma ray measurements for inertial confinement fusion applications. *Review of Scientific Instruments*, 94(4), 4 2023. doi: 10.1063/5.0126969.

[7] Baolian Cheng and Paul Andrew Bradley. What machine learning can and cannot do for inertial confinement fusion. *Plasma*, 6(2), 6 2023. doi: 10.3390/plasma6020023.

[8] Reabal Najjar. Redefining radiology: A review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17), 2023. ISSN 2075-4418. doi: 10.3390/diagnostics13172760. URL `https://www.mdpi.com/2075-4418/13/17/2760`.

[9] Nga T.T. Nguyen-Fotiadis, Bradley Wolfe, and Zhehui Wang. Deep regression outperforms conditional gan mapping on reconstructing double shells of icf images. In *Optica Imaging Congress (3D, COSI, DH, FLatOptics, IS, pcAOP), Technical Digest Series*, page HM3D.1. Optica Publishing Group, 2023.

[10] Bradley T. Wolfe, Michael J. Falato, Xinhua Zhang, Nga T. T. Nguyen-Fotiadis, J. P. Sauppe, P. M. Kozlowski, P. A. Keiter, R. E. Reinovsky, S. A. Batha, and Zhehui Wang. Machine learning for detection of 3d features using sparse x-ray tomographic reconstruction. *Review of Scientific Instruments*, 94(2), 2023. doi: 10.1063/5.0101681.

[11] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *CoRR*, abs/2103.13413, 2021. URL `https://arxiv.org/abs/2103.13413`.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL `https://arxiv.org/abs/2010.11929`.

[13] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on visual transformer. *CoRR*, abs/2012.12556, 2020. URL `https://arxiv.org/abs/2012.12556`.

[14] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sidney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

[18] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019. URL `https://arxiv.org/abs/1905.02244`.

[19] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. URL `https://arxiv.org/abs/2106.08254`.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. URL `https://arxiv.org/abs/2103.14030`.

[21] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016. URL `http://arxiv.org/abs/1610.02357`.

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL `http://arxiv.org/abs/1512.00567`.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

[24] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

[25] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Dong Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. URL `http://arxiv.org/abs/1411.1784`.

[27] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.

[29] Ander Biguri, Manjit Dosanjh, Steven Hancock, and Manuchehr Soleimani1. Tigre: a matlab-gpu toolbox for cbct image reconstruction. *Biomed. Phys. Eng. Express*, 2:055010, 2016.

[30] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data, 2018.

[31] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. URL `http://arxiv.org/abs/1611.07004`.

[32] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *CoRR*, abs/1501.00092, 2015. URL `http://arxiv.org/abs/1501.00092`.

[33] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[34] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[36] He Yang, Shaoyu Zhang, Xinchao Wang, and Yaowei Li. Focal and global knowledge distillation for detectors. *arXiv preprint arXiv:2203.11890*, 2022.

[37] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pages 596–608, 2020.

[38] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

[39] Bowen Zhang, Yidong Wang, Wenxin Hou, Mingsheng Wu, Tong Liu, Jindong Wang, and Dacheng Tao. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, 2021. URL `https://arxiv.org/abs/2110.08263`.

[40] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. URL `https://arxiv.org/abs/1909.13719`.

[41] Yaroslav Ganin and Victor Lempitsky. Domain-adversarial training of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2096–2104, 2015.

[42] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, pages 97–105, 2015.

[43] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176, 2017.

[44] Abhinav Valada et al. Hierarchical domain adaptation for adaptation of image classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 400–410, 2022.

[45] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21271–21284, 2020.

[46] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.