DDU-Net

A Domain Decomposition-Based CNN for High-Resolution Image Segmentation on Multiple GPUs

Verburg, Corne; Heinlein, Alexander; Cyr, Eric C.

DOI
10.1109/ACCESS.2025.3561033

Publication date
2025

Document Version
Final published version

Published in
IEEE Access

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 March 2025, accepted 30 March 2025, date of publication 15 April 2025, date of current version 23 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3561033

RESEARCH ARTICLE

DDU-Net: A Domain Decomposition-Based CNN for High-Resolution Image Segmentation on Multiple GPUs

CORNÉ VERBURG[1], ALEXANDER HEINLEIN[1], AND ERIC C. CYR[2]
[1]Delft Institute of Applied Mathematics, Delft University of Technology (TU Delft), 2628 CD Delft, The Netherlands
[2]Computational Mathematics Department, Sandia National Laboratories, Albuquerque, NM 87123, USA

Corresponding authors: Alexander Heinlein (a.heinlein@tudelft.nl) and Corné Verburg (c.verburg@tudelft.nl)

This work was supported in part by Sandia National Laboratories funded by U.S. Department of Energy, the Office of Science, and the Office of Advanced Scientific Computing Research, through the SEA-CROGS Project in the MMICCs Program. Sandia National Laboratories is a Multimission Laboratory managed and operated by the National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for U.S. Department of Energy's National Nuclear Security Administration under Contract DE-NA0003525.

ABSTRACT The segmentation of ultra-high resolution images poses challenges such as loss of spatial information or computational inefficiency. In this work, a novel approach that combines encoder-decoder architectures with domain decomposition strategies to address these challenges is proposed. Specifically, a domain decomposition-based U-Net (DDU-Net) architecture is introduced, which partitions input images into non-overlapping patches that can be processed independently on separate devices. A communication network is added to facilitate inter-patch information exchange to enhance the understanding of spatial context. Experimental validation is performed on a synthetic dataset that is designed to measure the effectiveness of the communication network. Then, the performance is tested on the DeepGlobe land cover classification dataset as a real-world benchmark data set. The results demonstrate that the approach, which includes inter-patch communication for images divided into 16×16 non-overlapping subimages, achieves a $2 - 3\,\%$ higher intersection over union (IoU) score compared to the same network without inter-patch communication. The performance of the network which includes communication is equivalent to that of a baseline U-Net trained on the full image, showing that our model provides an effective solution for segmenting ultra-high-resolution images while preserving spatial context. The code is available at `https://github.com/corne00/DDU-Net`

INDEX TERMS Convolutional neural networks, deep learning, ultra-high-resolution images, image processing, parallel processing, semantic segmentation, U-Net, spatial context.

I. INTRODUCTION

The vast majority of deep learning models in computer vision focuses on low-resolution 2D and 3D images, typically 256×256 pixels or smaller. However, the increased utilization of high-resolution image datasets introduces new challenges due to the memory constraints of a single GPU, especially for memory-intensive tasks such as semantic segmentation of images [1]. Semantic segmentation is the computer vision task of classifying the pixels in the input into distinct, non-overlapping semantic categories. Ultra-high-resolution image segmentation holds significance in diverse fields such as object segmentation in satellite images [2], [3], metallic surface defect detection [4], [5], and computer-aided medical diagnosis [6], [7]. Whereas deep convolutional neural networks (CNNs) have achieved remarkable success in image segmentation [8], most of these models are unsuitable for model training and inference for high-resolution images due to their high memory requirements.

To illustrate this, consider computed tomography (CT) scans with sub-millimeter resolution, resulting in voxel image data with a typical resolution of $512 \times 512 \times 512$ voxels.

[0]The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

Even with half-precision floating-point numbers and a modest batch size of 8, processing such images with a 1-layer convolutional neural network with 64 filters demands over 137GB of GPU/TPU memory, as highlighted in [9]. Dealing with such high-resolution inputs using conventional strategies, like down-sampling or patch cropping, generally leads to the loss of detailed information or spatial context, resulting in a lower segmentation accuracy [1], [9], [10], [11].

In this paper, we propose a novel network architecture based on the well-known U-Net [12] developed for biomedical image segmentation tasks. The U-Net is a convolutional neural network (CNN) with a bottleneck structure and skip connections between the encoder and decoder paths. This foundational architecture and variations of it [13], [14], [15], [16] have demonstrated remarkable, state-of-the-art accuracy in semantic segmentation and other image-to-image tasks [17]. However, its substantial memory requirements are prohibitive for application to high-resolution images on computational devices with limited memory [18]. The high memory demands for training stem from the required storage of intermediate feature maps obtained during forward propagation to be used during the subsequent backward propagation pass. This is especially acute when dealing with high-resolution inputs, as it leads to high-dimensional feature maps in the first and last few layers of the network.

Our proposed model integrates the established U-Net architecture with a divide-and-conquer strategy inspired by domain decomposition methods (DDMs) [19] to deal with computational device memory limitations. Previous parallelization strategies employ a decomposition of the image into subimages (subdomains) [9], [20], [21], with communicated margins before each convolution or redundant computations providing global contextual information between subimages. Our approach similarly decomposes the image. However, communication is explicitly limited to only the bottleneck within the U-Net architecture. This minimizes the communication overhead while preserving essential contextual information.

We summarize our main contributions as follows:

- We propose a novel approach combining U-Net architectures with domain decomposition strategies to segment ultra-high-resolution images efficiently while preserving spatial context.
- We show that the communication network, which is an important component of our approach, can be employed to exchange information between different subimages. It enhances understanding spatial context without significant computational overhead and with minimal extra communication and memory cost.
- By evaluating our architecture on a synthetic and a realistic image dataset, we demonstrate competitive segmentation performance compared to baseline U-Net models. Our approach remains scalable, even when training is confined to the largest image portion that the available devices can handle.

The remainder of this paper is organized as follows: In Section II, we discuss the U-Net architecture in more detail as well as existing memory parallelization strategies. Furthermore, we highlight approaches using ideas from the field of domain decomposition to speed up and parallelize CNNs. Section III introduces our novel parallel architecture and training methodology as well as the considered test datasets. In Section IV, we discuss the influence of this new architecture on memory requirements and the receptive field. Then, we evaluate our proposed network model and the related training strategy using these datasets in Section V. We provide experimental results for the different datasets in terms of the segmentation accuracy of our approach compared to global semantic segmentation U-Net models. Finally, in Section VI, we conclude and present possible future research directions.

II. RELATED WORK

In this section, we discuss the U-Net, which serves as the basis for our proposed approach, and the receptive field, a key concept for understanding CNNs. Additionally, we provide an overview of past efforts to tackle the memory challenges associated with the U-Net and other CNN-based semantic segmentation models; we discuss both strengths and drawbacks of these methods. Finally, we introduce the idea of domain decomposition as a natural way to handle memory constraints, highlighting some previous works that adopt this approach for both classification and segmentation.

A. THE U-NET

The new segmentation approach introduced in the this paper is based on the U-Net architecture [12]. Since its publication, many variants and extensions of the U-Net architecture have been developed, for instance, 3D U-Net [13], UNet++ [14], Attention U-Net [15], and ResUNet-a [16]. We assume that the reader is familiar with the fundamental concepts underlying CNNs [22] and the building blocks of such models – convolutional and pooling layers. For a detailed explanation of these concepts, we refer to the rich literature on this topic, for instance, [23, Chapt. 9] or [24, Chapt. 10].

The U-Net architecture [12], depicted in Fig. 1, consists of two pathways: the *contraction path* or encoder path, and the *expansive path* or decoder path. The contraction path involves repeated blocks, each consisting of two successive 3×3 convolutions, followed by a ReLU activation and a max-pooling layer. Conversely, the expansion path employs blocks that up-sample the feature map using 2×2 transposed convolutions. Subsequently, the feature maps from the corresponding layer in the contracting path are cropped (if the up-sampling path discards the boundary pixels) and concatenated to the up-sampled feature maps. This concatenation is followed by two successive 3×3 convolutions and a ReLU activation.

In the final stage, the number of feature maps is reduced to the desired number of classes by a 1×1 convolution, producing the segmented image. The *skip connections*

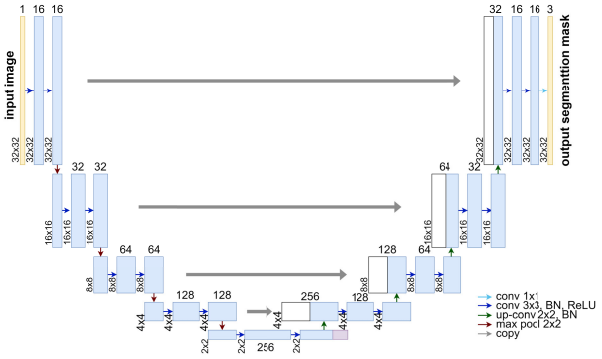


FIGURE 1. U-Net architecture for 32 × 32 pixel input images and corresponding masks. Each blue block represents a multi-channel feature map, with resolutions indicated at the lower left edge of each box. White boxes show copied feature maps from the skip connections (gray arrows). The colored arrows denote different operations. This figure is based on the architecture described in [12]. Image is best viewed online.

between the contraction path, which captures (global) context and features, and the expansive path, which enables precise, fine-grained localization, distinguishes the U-Net from other related network architectures, such as the Fully Convolutional Network (FCN) presented in [25]. Note that the originally proposed U-Net returns a cropped segmentation mask (due to the loss of border pixels in every convolution). However, more recent implementations of the U-Net often use zero padding at the borders to obtain output dimensions equal to the input dimensions, such that the input dimensions are preserved and cropping is not required; cf. [26], [27], [28].

B. THE RECEPTIVE FIELD

An important concept for understanding CNNs is their *receptive field*, or *field of view*. The receptive field of a CNN is the area of the input image that influences the output of a pixel in the CNN output. Unlike fully-connected neural networks, where each input neuron is connected with each output neuron, CNNs have outputs that depend on specific regions of the input layer. In each layer of the CNN, the receptive field expands as we move deeper into the network, allowing higher-level features to be influenced by larger portions of the input image.

The theoretical size of the receptive field can be analyzed both experimentally and theoretically, as has been discussed extensively in [29]. The size of the receptive field significantly impacts the predictive performance of the CNN. If the receptive field is too small, the network may not capture all relevant information, leading to suboptimal predictions. For tasks like segmentation, the receptive field should be large enough to include all pixels in the input image that are relevant for predicting the correct pixel class. Note that the optimal size of the receptive field depends on both the network architecture as well as the data on which the network is trained.

In practice, the receptive field size of a CNN is influenced by several factors, including the number of convolutional layers, the size of the convolutional kernels, and the presence of pooling layers; cf. [30]. Increasing the depth of the network and the size of the kernels can expand the receptive field, allowing the network to capture more context from the input image. However, this also increases computational complexity and the risk of overfitting.

C. OVERCOMING U-NET MEMORY LIMITS

One issue shared by most U-Net variants is the large memory requirements due to the storage of intermediate feature maps during training, making the model unsuitable for high-resolution applications with limited-memory computing devices [18].

Two naive methods to limit memory usage during U-Net training are image patching and image downsizing. To show the downsides of these approaches, we trained a ResNet18-UNet [16] on the DeepGlobe land cover classification dataset. For more details on the training procedure and dataset used, we refer to Section III. Image patching involves training the U-Net model on randomly extracted patches from the full-resolution image. The patch size can be chosen based on the available memory and the optimal batch size. However, this approach does not allow the network to see the patch in a broader spatial context, which may be important. This is illustrated in Table 1, where a ResNet18-U-Net was trained on patches of varying sizes. As shown, larger patch sizes (which include more context) lead to significantly better predictions, both on the training and test datasets.

TABLE 1. Random patching results on the DeepGlobe dataset. All IoU scores are obtained on the full scale images (2 048 × 2 048 pixels). Training happened on different patch sizes. Training and methodology details are outlined in more detail in Section III.

Train patch size	mean IoU (train)	mean IoU (test)
16	0.3866	0.3099
32	0.4651	0.3943
64	0.5371	0.4882
128	0.5943	0.5691
256	0.6741	0.6342
512	0.7383	0.6795
1024	**0.7601**	**0.6847**

A second naive approach to address memory issues is downsampling. By reducing the resolution of the images, memory usage during training can be decreased. However, this results in the loss of fine-grained details, as these are discarded in the downsampling process. While this method yields better results than using random patches of small sizes (see Table 2), we still observe that prediction accuracy is highest when using the largest training images, thus preserving as much fine-grained context as possible.

A natural approach to overcome these constraints is to partition the memory load of the feature maps over different computational devices by using data and/or model parallelism. Several parallelized forms of the U-Net employ parallelism to reduce the memory per device. Both in [9] and in [31], the authors implement a spatial partitioning technique

TABLE 2. Downscaling results on the DeepGlobe dataset: All IoU scores were obtained on the full-scale images (2048 × 2048 pixels), while training was conducted on patches of varying sizes. Training and methodology details are outlined in more detail in Section III.

Downsampled size	mean IoU (train)	mean IoU (test)
16	0.5001	0.4654
32	0.5875	0.5058
64	0.6688	0.5601
128	0.7194	0.6582
256	0.7578	0.6896
512	**0.7740**	0.6923
1024	0.7619	**0.6926**

that decomposes the input and output of the convolutional layers into smaller, non-overlapping subimages. Before each convolution operation, devices exchange patch margins of the feature maps of half the convolution kernel size with each other. While this approach introduces a communication overhead, it effectively distributes memory across devices without altering the fundamental architecture of the U-Net. Our approach, on the other hand, involves architectural modifications to the U-Net to further optimize performance and limit the communication to the bottleneck of the architecture.

The authors of [20] partition the image into overlapping subimages, with the overlap size determined by the receptive field size. Due to redundant computations ($\mathcal{O}(q\frac{4\epsilon}{N})$ for an $N \times N$ pixel image partitioned into $q \times q$ subimages and a U-Net with receptive field size $\epsilon \times \epsilon$), this approach allows for a fully parallelized execution of both forward and backward passes. [11] adopts a similar strategy, but with an application to image classification rather than image segmentation using the ResNet architecture [21].

These approaches have been successful in partitioning the U-Net in such a way that the model can be trained and evaluated in parallel, even for ultra-high-resolution image datasets. However, they either involve communicating margins before each convolutional operation (this is the case for [9], [31]), which leads to communication overhead through the many point-to-point messages, or they entail many redundant computations (this is the case for [11], [21]).

D. MEMORY OPTIMIZATION FOR CNNs USING DOMAIN DECOMPOSITION APPROACHES

For improving memory efficiency, DDMs are inspiring due to their inherent parallelization and scalability properties, resulting from localization of the computations. DDMs are effective iterative solvers for (discretized) partial differential equations (PDEs), exhibiting scalability through a divide-and-conquer strategy that partitions the computational domain into overlapping or non-overlapping subdomains; see, for instance, [19], [32], [33], [34]). The PDE problem is partitioned into subproblems defined on these subdomains. This enables parallel execution of computations within the subdomains. Whereas global convergence is ensured by well-balanced neighbor communication at potentially overlapping subdomain boundaries, and a limited amount of global communication. Highly scalable state-of-the-art DDMs are, for instance, variants of overlapping Schwarz methods [35], [36] or balancing domain decomposition by constraints (BDDC) [37], [38] and finite element tearing and interconnecting - dual primal (FETI-DP) [39], [40].

Recently, there has been an increasing interest in integrating DDMs with machine learning (ML) algorithms, combining the strengths of both fields [41], [42]. Up to now, the majority of works focuses on combining DDMs and ML for solving PDEs. The present paper, however, pursues the combination of CNN with DDM techniques for semantic image segmentation. Notably, to the best of the authors' knowledge, no prior studies have tackled image segmentation tasks using ML explicitly based on domain decomposition strategies, although some methods have been proposed that have similarities with domain decomposition strategies. Existing research mainly focuses on image classification, which is closely related to semantic segmentation; however, due to the different model output, network architectures are somewhat different. Here, we provide a concise overview of methods combining DDM strategies with ML for image classification and semantic image segmentation tasks.

In [43], the authors propose a image partitioning approach for image segmentation, reducing network complexity and enhancing parallelizability. They partition input images into non-overlapping subimages and subsequently train smaller local CNNs on these subimages instead of a large global CNN. This enhances model specialization to specific image regions and reduces overall complexity without a significant impact on test errors compared to a global CNN. Importantly, the authors process the subimages completely independently, with no coupling between them. In [44], DDM strategies are applied to train a CNN-DNN (convolutional neural network - deep neural network) architecture for image classification. The authors partition the input image into subimages, each used to train an individual CNN for predicting the class locally. A DNN then aggregates these local predictions into one global prediction. The authors interpret the DNN as a coarse problem solver that combines the finer-grained information from the local networks. In [45], the authors decompose a global CNN for image classification into a finite number of smaller, independent local subnetworks along its width. The weights obtained from training these local subnetworks serve as an initialization for the subsequent training of the global network, employing a transfer learning strategy. In [46], the authors propose a physically-motivated neural network topology for estimating extensive parameters such as energy or entropy. They decompose the domain into subdomains with *focus* and *context* regions, overlapping based on the *locality* of the extensive parameter estimated. The same subnetwork weights are used on the different subdomains (we employ this strategy as well in our approach), leading to a reduction of the number of parameters, enhanced parallelizability, and faster inference.

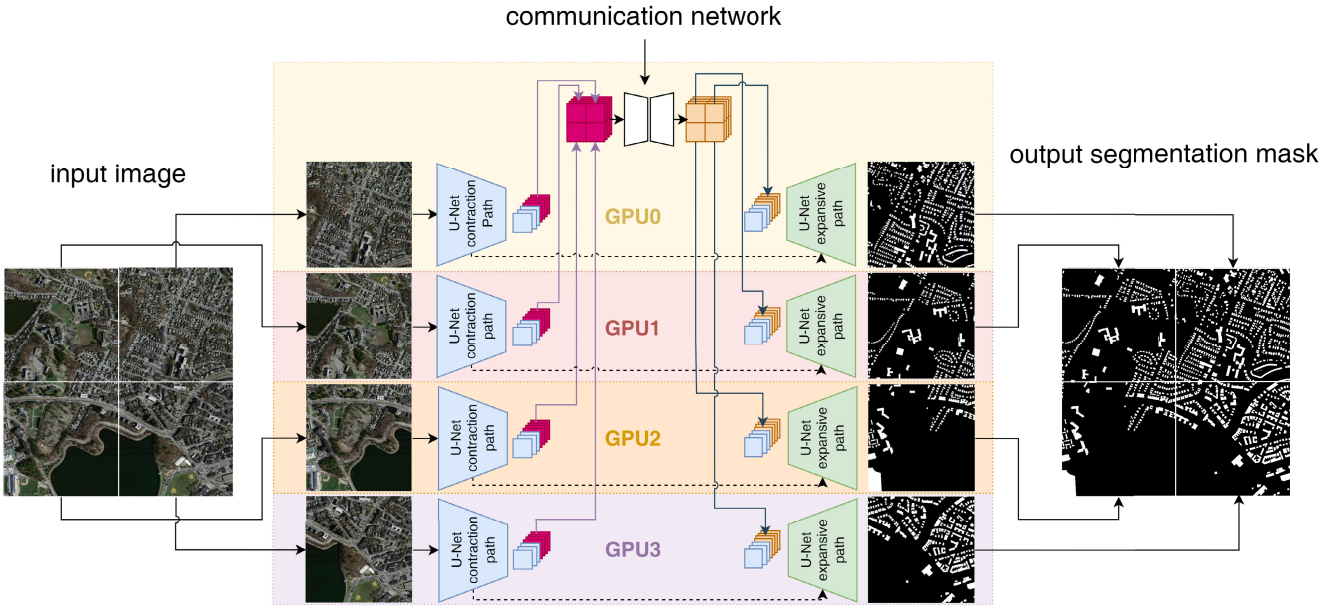


FIGURE 2. Schematic of the proposed network architecture. Input images are partitioned into subimages that are processed independently in the encoder paths. After encoding, a number of encoded feature maps is communicated to the device containing the communication network and then processed via the communication network. The output of this network replaces the input feature maps. The decoding is also done in parallel without communication between the computational devices. Dashed arrows indicate skip connections. Detailed architectures of the encoder-decoder network and communication network are shown in Fig. 3 and 4, respectively.

Finally, in [47], a DDM-inspired segmentation algorithm is proposed that divides the input image into several overlapping subimages and trains one segmentation network on these subimages, allowing for parallel inference. The authors conclude that this approach leads to better segmentation accuracy of small objects.

III. METHODOLOGY

In this section, we introduce our domain decomposition-based U-Net (DDU-Net) architecture, along with our training approach. We also describe the datasets used for testing the model and outline the model training procedure.

A. NETWORK ARCHITECTURE

1) DECOMPOSITION OF THE IMAGES AND MASKS

Fig. 2 shows a schematic representation of our proposed network architecture. Starting with a dataset of (high-resolution) images and corresponding segmentation masks, the model processes a 2D pixel image with dimensions $H \times W$ (left) and outputs a probability distribution for $K \in \mathbb{N}$ classes for each pixel of the image (right).

Following DDM strategies, we decompose the input data into $N \times M$ non-overlapping subimages, with $N, M \in \mathbb{N}$. In particular, each image is divided into $N \times M$ subimages with heights H_i and widths W_j, $i = 1, \ldots, N$ and $j = 1, \ldots, M$, such that $\sum_{i=1}^{N} H_i = H$ and $\sum_{j=1}^{M} W_i = W$. The corresponding segmentation masks are partitioned similarly. The subimages and sub-masks are distributed across the computational devices (e.g., GPUs/TPUs) to balance the workload evenly.

After partitioning, the subimages are processed independently, in parallel, by the encoders on the separate computational devices. For more details on the architectures of encoders and decoders, we refer to Section III-A2. A chosen number of feature maps from the last layer of the encoder is sent to the communication network (see Section III-A3), allowing for the exchange of relevant context between subimages. The communication network concatenates the encoded feature maps from all subimages corresponding to the arrangement of the subimages in the original full resolution image and processes them. Each output feature map of the communication network is again partitioned into $N \times M$ sub-feature maps, that are then sent back to the devices. The decoders then produce sub-predictions for each subimage, which are concatenated to form a global predicted mask for the entire high-resolution image.

2) ENCODER-DECODER NETWORKS

Fig. 3 shows the encoder-decoder architecture used in this work, which closely aligns with the classical U-Net architecture [12]; cf. Fig. 1. The primary difference is a modification in the deepest layer of the encoder path to facilitate communication between the local encoder-decoder clones, which will be discussed below.

As discussed in Section III-A1, each computational device contains a separate encoder-decoder network, allowing for parallel processing of the subimages. However, these networks share their weights to ensure consistent segmentation across subimages, making them local clones of a global encoder-decoder network. This implies that during training,

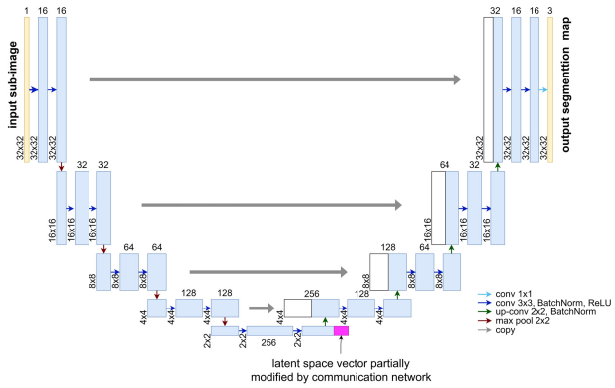


FIGURE 3. The proposed encoder-decoder architecture. The architecture of the encoder-decoder is nearly identical to the architecture of U-Net [12]. The only difference is located in the latent space vector, where a number of the feature maps are modified by the communication network. For optimal detail resolution, view this figure on a digital device.

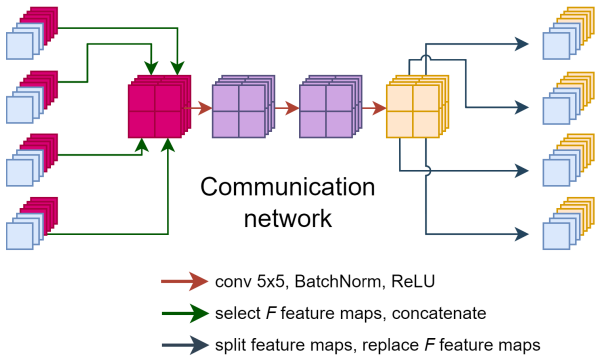


FIGURE 4. The proposed communication network for four subimages.

the weights and gradients need to be synchronized after each backward propagation step.

Each local encoder network processes its corresponding input subimage with several blocks of 3×3 convolutions, each followed by a batch normalization layer and a ReLU activation. During training, a dropout layer is also applied to prevent over-fitting. After two convolution blocks, max-pooling is performed to reduce the spatial dimension of the feature maps. The deepest encoder layer produces 256 spatially coarse feature maps (this number can vary with changes in the network architecture), some of which are employed for communication between the encoder-decoder networks. The output of each encoder network are 256 feature maps, with small spatial dimensions. The last F of these 256 feature maps are directed to the communication network, which processes them and returns modified feature maps with the same dimensions that replace the original input feature maps; see Section III-A3. After communicating, the feature maps then progress through the expansive path, resulting in a final segmented sub-mask for each subimage. These sub-masks are concatenated to form a full mask covering the entire full-resolution image.

3) COMMUNICATION NETWORK

The communication network, depicted in Fig. 4, transfers contextual features, captured by the encoders operating on subimages, between the encoded subimage feature maps.

For this paper, the communication network consists of three layers of 5×5 convolutions. The DDU-Net architecture is not restricted to this specific network configuration. Adjustments such as altering the number of layers, dilation, and kernel size may enhance the model's receptive field size for specific applications. We choose our communication network to be fully convolutional to ensure adaptability to arbitrary input sizes as well as scaling the number of subimages.

The communication network receives as input F feature maps from the deepest encoder layers for each of the $N \times M$ subimages; cf. Section III-A2. These feature maps are concatenated along the height and width dimensions, aligning with the subimages' positions in the full-resolution input image; cf. Fig. 2. The communication network processes this concatenated input and produces output feature maps with the same spatial dimensions as the inputs. These feature maps are partitioned back along height and width into $N \times M$ sub-feature maps, which are then sent back to the corresponding encoder-decoder network for further processing; they replace the original feature maps generated by the encoder and used as inputs to the communication network.

B. NOTATION

In the remainder of this paper, we compare various model configurations using a naming convention for our domain decomposition-based U-Net that encapsulates key parameters. Each model is denoted as DDU-Net(D, F, C), where:

- D ($\in \mathbb{N}_{>0}$) indicates the number of up- and down-sampling blocks in the encoder-decoder network.
- F ($\in \mathbb{N}_{\geq 0}$) specifies the number of feature maps processed by the communication network.
- C (Y or N) denotes whether communication between communication feature maps is enabled ($C = Y$) or disabled ($C = N$)).

Note that, if communication is disabled (C = N) but the number F of feature maps processed by the communication network is non-zero, this means that F feature maps are sent through the communication network independently, without concatenation. This effectively results in a baseline U-Net architecture with extra convolutions in the bottleneck layer.

For instance, DDU-Net(4, 64, Y) refers to a model where the encoder-decoder network has a depth of 4, 64 feature maps are sent to the communication network, and communication is enabled. The case DDU-Net(D, 0, N) operating on 1×1 subimages is equivalent to the baseline U-Net architecture. DDU-Net(D, F, N), with $F > 0$ refers to a baseline U-Net with extra convolutions in the bottleneck layer. Additionally, for the case where $F = 0$, it does not

matter whether communication is enabled (Y) or not (N), as the communication network uses 0 feature maps.

C. MODEL TRAINING

1) LOSS FUNCTION

For the network training, we employ the dice loss function, as this function addresses the issue of class imbalance for semantic segmentation of images; cf. [48], [49]. The dice loss is defined as

$$DL = 1 - 2\frac{\sum_{p=1}^{P}\sum_{k=1}^{K} y_{k,p}\hat{y}_{k,p} + \epsilon}{\sum_{p=1}^{P}\sum_{k=1}^{K} y_k + \sum_{p=1}^{P}\sum_{k=1}^{K} \hat{y}_k + \epsilon}. \quad (1)$$

Here, K represents the number of classes, P is the total number of pixels in a batch, and $\hat{y}_{k,p}$ is the predicted probability of pixel p for class k, obtained by applying a softmax function to the model's output so that all the output logits are in the range $[0, 1]$. Furthermore, $y_{k,p}$ denotes the true probability of pixel p for class k, with values restricted to $\{0, 1\}$ as the true mask is known, and ϵ serves as a small numerical stability constant (to avoid division by zero), set to $\epsilon = 10^{-7}$ in this paper.

2) PARAMETER OPTIMIZATION

The weights of the network are initialized using the method proposed by He et al. [50]. We employ the Adam optimizer [51], with momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a plateau learning rate decay strategy, where the learning rate decays with a factor of 2 when the validation loss does not decrease, maintaining a patience of a chosen number of epochs. The initial learning rate, batch size, and early stopping criterion were chosen depending on the task and available memory and will be provided in the results section.

3) TRAINING PROCEDURE

For training the DDU-Net architecture, we first initialize the encoder-decoder network with shared weights and distribute clones of this network across all $M \times N$ devices. The coarse network for communication is initialized on a selected device and, due to the small size of the communicated feature maps, may optionally reside on the same device as one of the encoder-decoder networks. During forward propagation, each device independently computes the encoded feature maps. After processing a selected number F of those feature maps using the communication network, the decoding of the feature maps again takes place fully in parallel. Backward propagation through the encoders and decoders can, again, be done in parallel without dependencies between the local encoder-decoder networks. However, after backward propagation through the decoders, communication is necessary for backward propagating through the communication network. After backward propagation, gradients are accumulated centrally on a main device and used to update weights, ensuring uniform updates across all devices. The updated weights are broadcasted to the other devices for synchronization.

4) IMPLEMENTATION

The implementation was done using PyTorch [52] (version `1.12.0`), an open-source machine learning library. We carried out the training and testing on the DelftBlue supercomputer [53] at the Delft University of Technology, and we employed NVIDIA Tesla V100S GPUs, with a memory of 32 GB, for the training.

D. DATASETS

For testing our model, we use two different image datasets: 1) a synthetically generated dataset designed to test the capabilities of the communication network and 2) a realistic image semantic segmentation dataset for multi-class land cover segmentation to assess the effectiveness of the proposed model both in terms of segmentation quality and memory efficiency.

1) SYNTHETIC DATASET

In our approach, only deep feature maps, with low spatial resolution, are exchanged between encoder-decoder networks. This differs significantly from previous U-Net parallelization methods that involve the exchange of a number of feature maps at each U-Net layer. To assess the level of spatial context that these low-resolution feature maps can capture, we designed a synthetic dataset. This dataset comprises one-channel gray-scale images with dimensions $(k \cdot 32) \times 32$ pixels, where $k \in \{2, 3, 4, 6, 8, 16\}$. For each k, we generate 4 000 training, 1 000 testing and 1 000 validation images, resulting in a total synthetic dataset size of 36 000 images. This design allows the decomposition of images into k subimages of 32×32 pixels; see Fig. 5 for examples with $k = 4$ subimages.

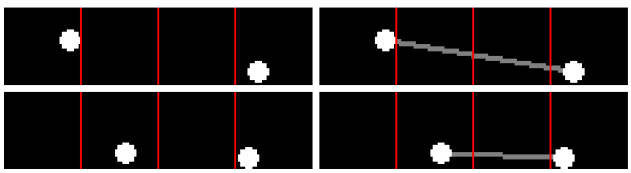


FIGURE 5. Two example images (left) and masks (right) from the synthetic dataset. The subimage boundaries used for these images are shown by the red vertical lines.

Each image in this gray-scale dataset displays a black background with two randomly placed white circles, each with a 4-pixel radius, placed completely within a subimage such that the subimage boundaries do not intersect the circles. The corresponding mask resembles the image, except for the addition of a third class of pixels: a line segment drawn between the centers of the two circles. Two examples of images and corresponding masks are shown in Fig. 5.

When processing an image in this dataset with the DDU-Net, the segmentation of the line segment connecting the two circles relies entirely on the communication network. As the two subimages are processed independently and only linked by the communication network, the effectiveness of

the segmentation is a direct reflection of the communication network's capability to transfer global information accurately.

2) DeepGlobe LAND COVER CLASSIFICATION DATASET

The DeepGlobe land cover classification dataset [54] is a semantic segmentation dataset for land cover types. The dataset contains 803 high-resolution (2 448 × 2 448 pixels) annotated satellite images with 7 classes: urban, agriculture, rangeland, forest, water, barren, and unknown. The images have 50 cm/pixel resolution and span a total area of 1 716.9 km^2. In addition to the high resolution of the images, segmenting this dataset is challenging due to the large class imbalance; see Table 3. In Fig. 6, we show two example images and masks from the dataset.

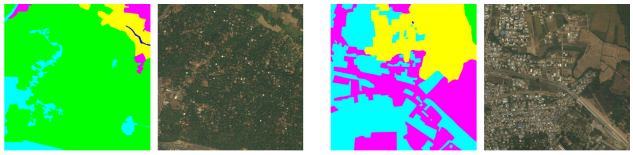


FIGURE 6. Two example images (right) and their corresponding masks (left) from the DeepGlobe land cover classification dataset [54].

Other challenges include the limited number of images, inexact ground truth, and the presence of multi-scale relevant contexts; cf. [55], [56]. To illustrate this, note that trees can exist in various land types such as urban, rangeland, or forest areas. As a result, the network needs to assimilate contextual information from a wider region around the tree to predict the correct class for the tree pixels accurately.

TABLE 3. Class distributions in the DeepGlobe land cover classification dataset. Table retrieved from [54].

class	pixel count	proportion
urban	642.4 M	9.35 %
agriculture	3898.0 M	56.76 %
rangeland	701.1 M	10.21 %
forest	944.4 M	13.75 %
water	256.9 M	3.74 %
barren	421.8 M	6.14 %
unknown	3.0 M	0.04 %

E. EVALUATION METRICS

After training the model, we analyze the results. Because of the large class imbalance, we used the (mean) class-wise intersection over union (IoU) score as a metric, as was suggested in the paper introducing the DeepGlobe dataset [54]. Given a dataset with n images, the IoU score for class j is defined as:

$$IoU_j = \frac{\sum_{i=1}^{n} TP_{ij}}{\sum_{i=1}^{n} TP_{ij} + \sum_{i=1}^{n} FP_{ij} + \sum_{i=1}^{n} FN_{ij}}, \quad (2)$$

where:

- TP_{ij} is the number of pixels in the i-th image correctly predicted as class j,
- FP_{ij} is the number of pixels in the i-th image incorrectly predicted as class j, and
- FN_{ij} is the number of pixels in the i-th image that belong to class j but were predicted as another class.

The mean IoU score ($mIoU$) provides a single metric by averaging the IoU scores across all classes, and is then defined as:

$$mIoU = \frac{1}{K} \sum_{j=1}^{K} IoU_j, \quad (3)$$

where K is the number of classes.

IV. ARCHITECTURE DISCUSSION

In this section, we analyze some architectural properties of the DDU-Net. First, we investigate the memory requirements of our approach both experimentally and theoretically and compare the results against a standard U-Net model. Then, we analyze the size of the receptive field of DDU-Net models with different architectures, that is, with varying depth of the subnetworks and the communication network.

A. MEMORY REQUIREMENTS

We compare the memory requirements for the baseline U-Net depicted in Fig. 1 and the proposed DDU-Net model architecture, both with a depth of 4 up-sampling and down-sampling blocks, operating on 2 subimages; cf. Section III, Fig 2 to 4. We specifically present results for a configuration with $F = 64$ communicated feature maps. The channel distribution for the encoder-decoder networks in both models follows the same scheme depicted in Fig. 1.

Table 5 presents a detailed analysis of the memory requirements for the encoder and decoder of the U-Net architecture, which are identical to the encoder and decoder used in the DDU-Net, during training on a 1 024 × 1 024 image. The table also includes the memory requirements for the proposed communication network with $F = 64$. The values in this table were derived theoretically (see, for instance, [57]) and validated experimentally using the `torch` library [52].

From the analysis, it is evident that storing the feature maps demands significantly more memory than the storing of model weights. The shallow layers, including the input block, the first decoder block, and the last decoder block, collectively contribute to nearly half of the total memory allocation for feature maps. Conversely, the number of weights increases for the deeper layers of the U-Net due to the higher number of kernels and larger kernel sizes in these blocks. The coarse feature maps used in the communicated layer result in relatively low memory requirements for the communication network, demonstrating the efficiency of the proposed DDU-Net in terms of memory utilization.

Figure 7 shows the peak memory usage during inference for the baseline U-Net architecture and for a GPU containing both the U-Net encoder-decoder and the proposed communication network, for various image resolutions. For smaller image resolutions, the memory allocation can be largely

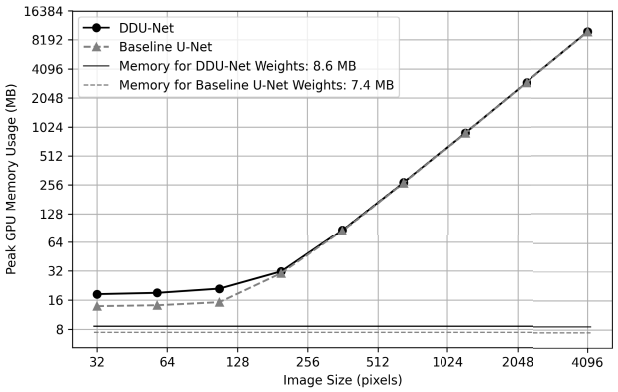


FIGURE 7. Measured peak memory of the proposed DDU-Net(4, 64, Y) evaluated on two sub-images, compared to the baseline U-Net (with depth 4) evaluated on a single subimage, for various subimage resolutions during inference. The lines with markers represent the peak total memory usage for both networks, while the two horizontal lines only indicate the part of memory used for storing model weights. The experimental peak memory was measured using `torch`. For the DDU-Net, the GPU contains both the communication network (processing $F = 64$ feature maps) and an encoder-decoder network. It is important to note that the U-Net peak memory usage is measured on a single GPU, while the DDU-Net peak memory usage is the maximum of the peak memory usage across two GPUs.

attributed to model weights. However, as image size surpasses a moderate size of $2^7 \times 2^7$ (128×128) pixels, the memory required to store the feature maps begins to dominate the total memory cost. Beyond this point, a noticeable increase in memory requirements is observed, as illustrated in Fig. 7. It is evident from Fig. 7 that the memory requirements for the DDU-Net and U-Net become relatively significantly closer at higher resolutions: for a resolution of 32 pixels, the relative difference is 25.9%, whereas for 4 096 pixels, this difference reduces to only 0.65%. Additionally, it is important to note that peak memory scales quadratically with resolution; therefore, doubling the resolution results in a fourfold increase in memory consumption. The results show that the proposed DDU-Net, with its communication network, requires only a small memory overhead compared to the baseline U-Net, despite the added cost of the communication network. This demonstrates the DDU-Net's efficiency in memory utilization: it facilitates communication between sub-images while increasing the memory requirements only slightly.

B. RECEPTIVE FIELD ANALYSIS

The proposed communication network operates on the coarse bottleneck layer of the encoder-decoder networks. Outputs of those layers typically have large receptive fields, as they are the result of numerous convolutional and downsampling operations. Therefore, the communication network is very effective in increasing the receptive field size of the model architecture. In Table 4, we compare the (theoretical) receptive field size of a standard U-Net with that of our proposed model at different depths of the encoder-decoder network, following the theoretical approach as presented in [29]. For the DDU-Net, we calculate the theoretical receptive field size for one infinitely large subdomain, providing an upper boundary for the true receptive field size when the network operates on multiple subimages. The comparison shows that the communication network significantly enlarges the receptive field size of the encoder-decoder network.

TABLE 4. Receptive field size for different encoder-decoder depths for baseline U-Net (cf. Fig. 1) the DDU-Net (cf. Fig. 2 to 4). Theoretical analysis of the receptive field was done following the approach described in [29]. Note that the number of feature maps F sent to the coarse network does not influence the receptive field size as long as $F > 0$.

model / depth D	2	3	4
baseline U-Net	44×44	92×92	188×188
DDU-Net(D,F,Y)	92×92	188×188	380×380

V. EXPERIMENTAL RESULTS

In this section, we compare the DDU-Net approach against the corresponding baseline U-Net model based on the segmentation quality on the datasets introduced in Section III. Moreover, we conduct an ablation study to examine the effectiveness of the communication between encoder-decoder networks processing subimages.

A. SYNTHETIC DATASET RESULTS

In this section, we present the results obtained using our synthetic dataset, which is described in Section III-D1. The dataset comprises gray-scale images of varying dimensions, specifically $(k \cdot 32) \times 32$ pixels, where $k \in \{2, 3, 4, 6, 8, 16\}$. For each k, we trained separate baseline U-Nets and DDU-Nets.

For our experiments, we trained baseline U-Nets using entire $k \cdot 32 \times 32$ images as inputs. In contrast, the DDU-Net architecture received k subimages of size 32×32 pixels, processed by k separate encoder-decoder clones. Both the U-Net and DDU-Net models were designed for 1-channel grayscale input images, generating outputs with three channels corresponding to the logits for the pixel classes: *background*, *line segment*, and *circle*.

We evaluate

- the impact of varying the number of down- and up-sampling blocks (depth D) in the encoder-decoder networks,
- the influence of the depth of the communication network, which is equal to the number of feature maps sent to the communication network (denoted as F), and
- the effect of inter-feature map communication (C, with values Y for enabled and N for disabled).

As mentioned in Section III-B, we use the notation DDU-Net(D, F, C) to represent a DDU-Net with parameters $D \in \mathbb{N}_{\geq 0}$, $F \in \mathbb{N}_{\geq 0}$, and $C \in \{Y, N\}$. For a baseline U-Net with depth D we use the notation U-Net(D). The training hyperparameters used to train the networks are listed in Table 10. The number of weights in the encoder, decoder,

TABLE 5. Theoretical analysis of the memory requirements and number of weights of a 4-blocks deep U-Net encoder and decoder architecture (see Section II-A) training for an RGB image of 1 024 × 1 024 pixels and a 3-class segmentation task. The table displays the size, number of input and output channels, memory usage (in terms of number of values and megabytes), as well as weight count and size (in terms of values and megabytes) for each block of the encoder and decoder part of the CNN architecture as well as the communication network. The communication network memory is based on 4 subimages and communication of 64 feature maps per subimage. Note that the number of weights and their memory cost is independent of the image size, whereas the number of values and memory of the feature maps will increase as image size increases.

name	size	# of input channels	# of output channels	memory feature maps (# of values)	memory feature maps (MB)	memory weights (# of values)	memory weights (MB)
input	1 024	3	3	3.1 M	12,0	-	0.00
input block	1 024	3	16	67 M	256.0	2 800	0.01
encoder block 1	512	16	32	42 M	176.0	13 952	0.05
encoder block 2	256	32	64	21 M	88.0	55 552	0.21
encoder block 3	128	64	128	10 M	44.0	221 696	0.85
encoder block 4	64	128	256	5.2 M	22.0	885 760	3.38
communication network	64	64	64	7.6 M	29.0	307 776	0.04
decoder block 1	64	256	128	13 M	48.0	573 952	2.19
decoder block 2	128	128	64	25 M	96.0	143 616	0.55
decoder block 3	256	64	32	50 M	192.0	35 968	0.14
decoder block 4	512	32	16	101 M	384.0	9 024	0.03
output block	1 024	16	3	3.1 M	12.0	51	0.00
labels	1 024	3	3	1.0 M	8.0	-	-

and communication network for the different experiments can be found in the appendix, specifically in Table 8.

1) QUALITATIVE RESULTS

To show the effectiveness of the coarse network with communication, we compare the qualitative segmentation masks for three architectures: the U-Net(3), the DDU-Net(3, 32, Y) and the DDU-Net(3, 32, N) for the 32 × 64 pixel images ($k = 2$) in Fig. 8. We observe that the DDU-Net(3, 32, N) is unable to predict the position of the line segment correctly, as it is unaware of the circle position in the other subimage due to the lack of communication. In contrast, the DDU-Net(3, 32, Y) predicts the location of the line segment correctly.

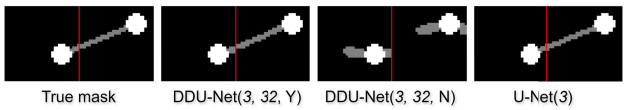


FIGURE 8. Predictions for different models. The DDU-Net operates on two subimages, whereas the U-Net is trained on the full 64 × 32 image. We remind the reader here that the DDU-Net(3, 32, N) processes the subdomains independently without communication (as indicated by the "N"), whereas the DDU-Net(3, 32, Y) includes communication. The red vertical line in the horizontal center of the image indicates the subimage border.

Fig. 8 shows that the proposed encoder-decoder network can capture the relevant features in the $F = 32$ communicated feature maps, and that the communication network is able to transfer this information between the encoder-decoder networks. While we currently communicate all 32 feature maps in the bottleneck of the DDU-Net(3, 32, Y) for this example, we will examine the impact of varying the number of feature maps in Section V-A2. This is particularly noteworthy, as the communication only happens on a very coarse level (with 4 × 4 pixel feature maps) and the line segments are drawn on the finer grained 64 × 32 resolution. Apparently, communication on this coarse level is sufficient for the decoder to produce a good segmentation result on the fine level.

2) VARYING THE NUMBER OF SUBIMAGES AND COMMUNICATED FEATURE MAPS

In Fig. 9, the IoU score (Eq. (2)) for the line segment class is depicted for the baseline U-Net(3) as well as the DDU-Net(3, F, Y), with $F \in \{0, 1, 2, 4, 8, 16, 32\}$ feature maps communicated, for different numbers of subimages k. Recall that the case $F = 0$ corresponds to a DDU-Net where the communication network has a depth of 0, meaning it is not used, also implying that there is no communication between encoder-decoder clones. We observe two phenomena. First,

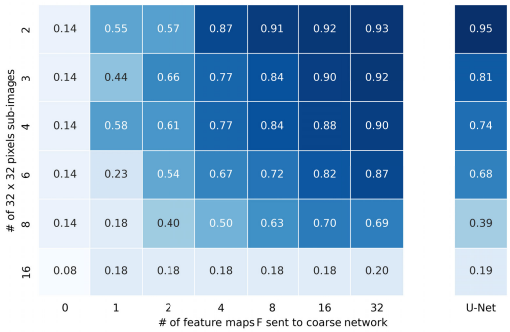


FIGURE 9. IoU score for the line segment class for a U-Net(3) and DDU-Net(3, F, Y) for different numbers of feature maps F in the communication network and different image dimensions. Note the increase in IoU score when the number of communicated feature maps F increases (horizontal).

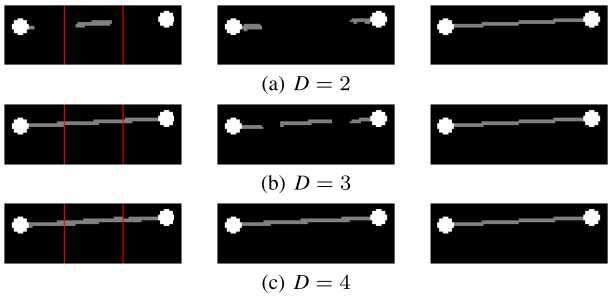
there exists a clear positive correlation between the number of communicated feature maps F and the quality of the results. Secondly, the DDU-Net performs even better than the baseline U-Net for larger images and numbers and subimages, respectively. This is related to the sizes of the receptive fields as reported in Table 4: the DDU-Net has a

larger receptive field (188×188 pixels versus 92×92 pixels), yielding better results when the circles are further apart.

3) IMPACT OF THE DEPTH OF THE ENCODER-DECODER NETWORK

In the DDU-Net, the encoder-decoder network has depth D, that is, the number of up- and down-sampling blocks. The ideal depth is a trade-off: while shallow depths cannot capture all relevant features due to a limited receptive field size (cf. Table 4), overly deep encoder-decoders contain more parameters and may produce too coarse-grained and less informative feature maps for communication.

To see the effect of the depth of the encoder-decoder on the predictions, we compare the segmentation masks generated by a DDU-Net(D, 16, Y) to the segmentation masks generated by a U-Net(D), with $D \in \{2, 3, 4\}$. The resulting predictions of those predictions are shown in Fig. 10. For encoder-decoder networks with $D = 2$, we observe that the line segment is only predicted correctly for the pixels that are in the center region between the two circles. This can also be explained by the receptive field: shallow networks have a limited receptive field, such that pixels far away from one of the two circles are not affected by this circle; this leads to the incorrect segmentation result. It can also be seen that the DDU-Net with communication



(a) $D = 2$

(b) $D = 3$

(c) $D = 4$

FIGURE 10. From left to right: mask predicted by the DDU-Net(*D*, 16, Y), mask predicted with baseline U-Net(*D*), and true mask for different encoder-decoder depths *D*. The subimage borders are indicated by the red vertical lines. Note that the baseline model is trained on the full image, so there are no subimages present for this case. The broken lines for the baseline U-Net are caused by the limited receptive field size of this network.

enabled gives better results than the U-Net for $D = 2$ and $D = 3$, as the communication network increases the receptive field size; cf. Table 4.

4) GENERALIZATION TO DIFFERENT NUMBERS OF SUBIMAGES

The baseline U-Net and the DDU-Net are both size-agnostic, as both are fully convolutional neural networks. This allows them to process images with sizes different from the training images. In case of the DDU-Net, this means that we can vary both the size of subimages and the number of subimages. In order to test the generalization of the DDU-Net with respect to the number of subimages for different amounts of communication, we train for each $k \in \{2, 3, 4, 6, 8, 16\}$ a DDU-Net(3, F, Y) on input images with spatial dimensions of $(k \cdot 32) \times 32$ pixels, as described in section Section III-D1, with $F \in \{0, 1, 2, 4, 8, 16, 32\}$. Each of these models is then evaluated on a dataset with 6 subimages. The results are shown in Fig. 11.

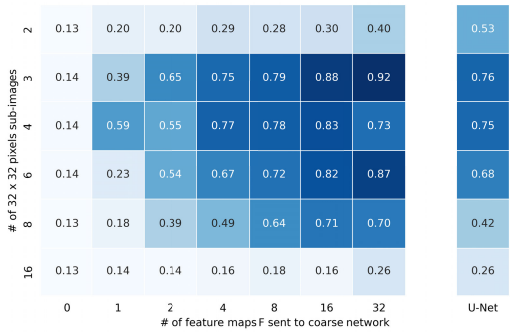


FIGURE 11. IoU score for the line segment class for a 3-deep encoder-decoder network in the DDU-Net for models trained on varying numbers of subimages with varying numbers and numbers of communicated feature maps. All models are evaluated on the same test dataset of $(6 \cdot 32) \times 32$ subimages.

We can make the following observations:

1) Models trained on 3, 4, and 6 subimages perform the best on the 6 subimage test dataset. This is likely because the images in these datasets are small enough to be (almost) fully covered by the 188×188 pixels receptive field of the DDU-Net with $D = 3$; cf. 4.
2) The more inaccurate results for the 8 and 16 subimages can be explained by the limited receptive field size, 188×188 pixels, of the DDU-Net with $D = 3$; cf. Table 4.
3) The IoU score deteriorates when the model is trained on only two subimages and then evaluated on a larger number of subimages. This is likely due to the fact that every subimage in the training data set contains one circle and a part of the line segment; however, there are no samples with subimages that do not contain circles or line segments.

These findings suggest that the DDU-Net can be trained on a different number of subimages than on which it is evaluated.

5) SUMMARY OF RESULTS ON THE SYNTHETIC DATASET

We summarize our findings for the synthetic dataset as follows:

- The communication network is able to transfer contextual information across subimages, which becomes clear both from qualitative comparison in Fig. 8 and quantitative results in Fig. 9.
- The segmentation quality increases when the number of feature maps involved in communication increases.
- The larger receptive field size of the DDU-Net with communication leads to better results compared to the baseline U-Net; cf. Fig. 9 and 10.

- The DDU-Net can be trained successfully on a fixed number of subimages and evaluated on different number of subimages.

B. DEEPGLOBE DATASET

Now, we assess the effectiveness of the DDU-Net for a real-world dataset with high-resolution images: the DeepGlobe dataset; cf. Section III-D2.

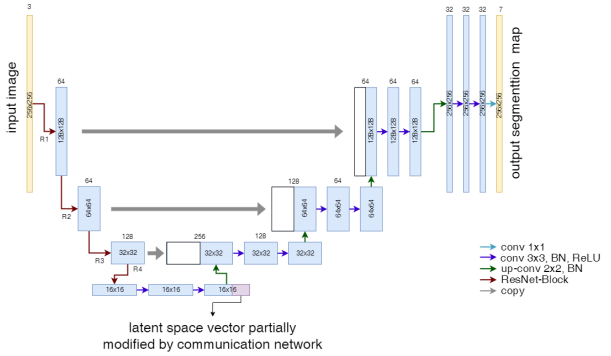


FIGURE 12. Architecture of the encoder-decoder ResNet-UNet achitecture for image segmentation. The pre-trained blocks of the ResNet-18 are shown as $R1$, $R2$, $R3$, $R4$. Note that the inner architecture of these blocks is not shown completely, but is largely simplified here.

Given the small number of 803 images in the DeepGlobe dataset relative to the segmentation task's complexity, we made several adjustments to the encoder-decoder architecture, including batch normalization, random dropout layers, and data augmentation (horizontal and vertical flipping, random rotation of R radians, with $R \in \{0, \pi/2, \pi, 3\pi/4\}$), and we incorporated a pre-trained image encoder model, the ResNet-18 [21]. This model, trained on over 1 000 classes on the ImageNet data set [58], offers a rich feature representation for diverse images. The ResNet-18 model consists of residual blocks with skip connections allowing an effective gradient flow. We used the first four residual blocks of the ResNet-18 to initialize our encoder. This strategy leverages the pre-existing knowledge within the pre-trained model to enhance the network's ability to generalize patterns from the limited dataset. A visualization of the employed model architecture is given in Fig. 12. For a detailed overview of the distribution of model parameters within each component of the network, we refer to Table 9.

During training, we kept the ResNet-18 model's weights fixed, only adjusting the weights of the decoder and the communication network. Additionally, we inserted two extra 3×3 convolutional layers in the bottleneck layer. These layers allow the network to restructure and refine the feature maps produced by the ResNet encoder generate relevant information for the communication network.

To ensure a fair comparison between the segmentation quality of the baseline U-Net and the DDU-Net, we trained both models on equally large images. Due to the baseline U-Net architecture's inherent inability to parallelize across different GPUs and the limited training device memory (32 GB), we cropped $1\,024 \times 1\,024$ non-overlapping patches from the DeepGlobe dataset, resulting in a training dataset consisting of 2 412 images. It is crucial to distinguish between these large "global patches" and the subimages in the DDU-Net architecture. During training of the DDU-Net, the global $1\,024 \times 1\,024$ patches are further partitioned into smaller subimages, which then are distributed across the encoder-decoder networks. Using mixed precision training [59], this approach allowed us to train with a mini-batch size of 12 images on a single GPU. This and other hyperparameters used for training are shown in Table 6. Furthermore, to take into account random initialization of the network parameters, we trained every network repeatedly for three times with the same settings and training dataset. For each configuration, the best performing model (in terms of IoU score on the test dataset) was selected.

TABLE 6. Hyperparameters used for training the model on the DeepGlobe land cover classification Dataset.

hyperparameters	
learning rate	0.001
number of epochs	100
dropout rate	0.1
batch size	12
loss function	dice loss (Eq. (1))

We want to investigate if the DDU-Net can perform equivalently to the baseline U-Net. However, when we include communication in the DDU-Net architecture the additional trainable parameters imply a potential ability to learn more complex patterns. To isolate the effects of (1) the *extra parameters* in the communication network and (2) the *communication itself*, we vary the number of feature maps F for both DDU-Net(4, F, Y) and DDU-Net(4, F, N). The DDU-Net(4, F, N) scenario represents an encoder-decoder network with an extra coarse network at the bottleneck layer (with coarse network chosen the same as the communication network), but *without* information exchange between subdomains. In this case, the coarse network operates solely on the *local* bottleneck feature maps, rather than on the concatenated feature maps of all subimages. Notably, $F = 0$ corresponds to the baseline U-Net(4). The difference between the cases with communication enabled (Y) and disabled (N) shows the impact of communication across subimages.

1) QUANTITATIVE RESULTS

Fig. 13 shows the mean IoU scores for different DDU-Net configurations and image partitions, trained on $1\,024 \times 1\,024$ images and evaluated on the $2\,048 \times 2\,048$ test dataset with fixed subimage size, using the approach discussed in Section V-A4.

From this figure, two main observations emerge. First, the segmentation quality improves with increased number of feature maps F in the coarse network. This is expected since a higher number of feature maps leads to an increase

in the number of parameters and a larger receptive field, enabling the model to capture more complex and distant patterns. Another trend in Fig. 13 is that the quality of the

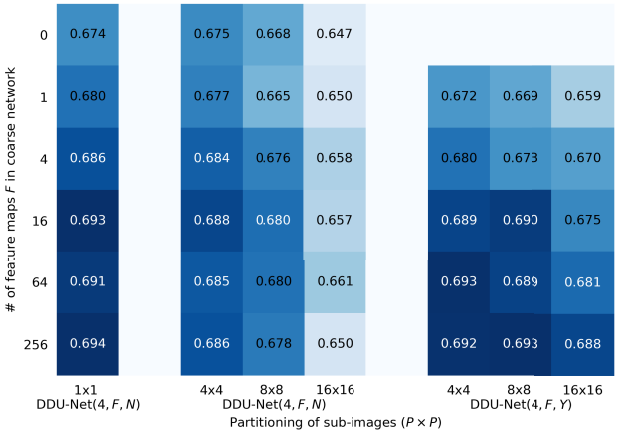


FIGURE 13. Mean IoU scores for various configurations of the DDU-Net architectures on a 2 048 × 2 048 test dataset. The left section presents results for a baseline DDU-Net(4, F, N) model evaluated on entire images ($P = 1$). The middle and right sections show the performance of the DDU-Net(4, F, N) and DDU-Net(4, F, Y) models, respectively, trained on $P \times P$ subimages that form a non-overlapping partition of the full training image, with $P \in \{4, 8, 16\}$. The experiments vary the partitioning $P \times P$ of the subimages and the number of feature maps F in the coarse network. Note that the baseline U-Net is the same model as the DDU-Net(4, 0, N) evaluated on 1 × 1 subimages.

predictions decreases as the number $P \times P$ of subimages increases, or, equivalently, as the subimage size decreases. However, for the DDU-Net(4, F, Y), this decrease in quality is much less pronounced compared to the DDU-Net(4, F, N). This indicates that the communication combined with the coarse network effectively transfers contextual information between subimages, also for this realistic dataset.

2) QUALITATIVE RESULTS

In Fig. 14, example predictions are shown for different training and model configurations. We observe that the DDU-Net trained on 1 × 1 and the DDU-Net trained on 8 × 8 subimages with communication enabled both produce good segmentation results, although there are differences compared to the true mask. Conversely, the DDU-Net trained on 8 × 8 subimages without communication shows poorer predictions.To better illustrate prediction errors, we include Fig. 15 in the Appendix, where black indicates correct predictions and white indicates incorrect predictions. The key distinction lies in the consistency across neighboring subimages: when communication is enabled, predictions for neighboring subimages exhibit smoother boundaries instead of patchy patterns. This highlights the effectiveness of communication between subimages in the DDU-Net architecture.

C. COMPARISON TO OTHER METHODS

We compared the DDU-Net with two high-resolution segmentation methods: GL-Net [10] and the

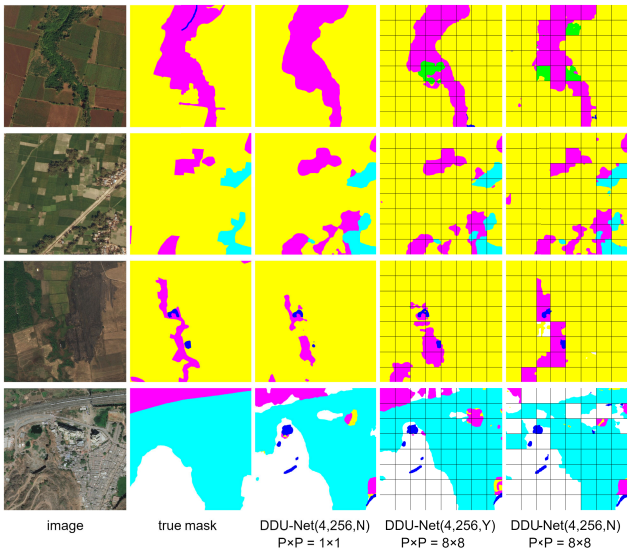


FIGURE 14. From left to right: original image, true mask, predictions by the DDU-Net(4, 256, N) evaluated on the full image, predictions by the DDU-Net(4, 256, Y) evaluated on 8 × 8 subimages, and predictions by the DDU-Net(4, 256, N) also evaluated on 8 × 8 subimages. The borders of the 8 × 8 subimages are indicated by black lines. Prediction errors for these images are visualized in Fig. 15. Note that the DDU-Net(4, 256, N) evaluated on the full image ($P \times P = 1 \times 1$) is equivalent to a 5-deep baseline U-Net.

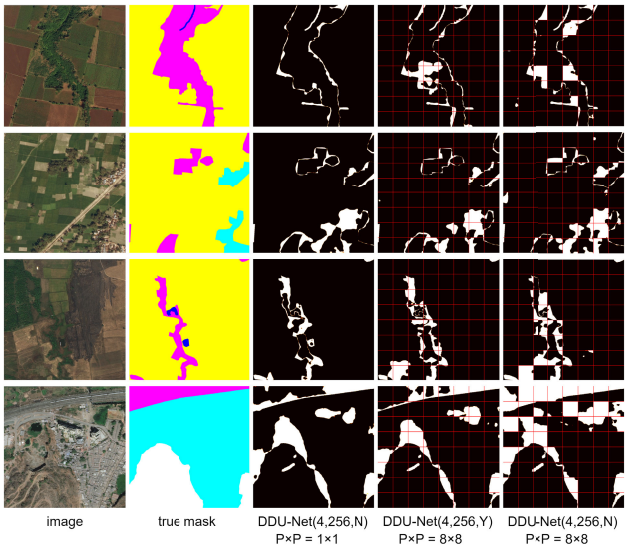


FIGURE 15. Visualization of prediction errors for various network configurations of the DDU-Net. Black indicates correct predictions, white indicates incorrect predictions. From left to right: original image, true mask, error by the DDU-Net(4, 256, N) evaluated on the full image, error by the DDU-Net(4, 256, Y) evaluated on 8 × 8 subimages, and error by the DDU-Net(4, 256, N) also evaluated on 8 × 8 subimages. The borders of the 8 × 8 subimages are indicated by red lines. Note that these predictions correspond to the ones shown in Fig. 14.

From-Contexts-to-Locality (FCtL) network [60]. Unlike these methods, the DDU-Net adopts a fundamentally different approach. Existing methods focus on minimizing computational workload to fit tasks onto a single GPU, often sacrificing critical high-resolution details. In contrast,

TABLE 7. Comparison results with other high-resolution segmenation network architectures FCtL [60] and GL-Net [10]. Note that the DDU-NetResNet on 1×1 subdomain is similar to a baseline ResNet-UNet and the implementation for other network architectures is only suitable for 1 GPU by design, in contrast to the DDU-Net. The depth of the used encoder-decoder networks is $D = 4$ blocks. $F = 256$ feature maps were communicated through the (coarse) communication network.

method (# of GPUs)	mean IoU	number of parameters	peak memory usage for a $2\,048 \times 2\,048$ image (MB)	inference time average (s) $\pm$ std (s)
DDU-NetResNet50 (1×1 subdomains, 1 GPU)	0.704	44M	3 834	0.2637 ± 0.0023
DDU-NetResNet50 (2×2 subdomains, 4 GPUs)	0.703	44M	1 109	0.0983 ± 0.0016
DDU-NetResNet18 (1×1 subdomains, 1 GPU)	0.690	9.7M	2 742	0.0857 ± 0.0004
DDU-NetResNet18 (2×2 subdomains, 4 GPUs)	0.684	9.7M	739	0.0463 ± 0.0011
From Contexts to Locality [60] (1 GPU)	**0.705**	54.1M	3 131	3.3907 ± 0.0150
GL-Net [10] (1 GPU)	0.696	15M	2 170	3.0550 ± 0.0240

TABLE 8. Properties of the encoder-decoder network and coarse network used in the synthetic datset experiments.

	depth D	channel distribution	bottleneck size	# of parameters
subnetworks	2	1-4-8-16-16-8-4-3	8×8	7 487
	3	1-4-8-16-32-32-16-8-4-3	4×4	30 470
	4	1-4-8-16-32-64-64-32-16-8-4-3	2×2	122 031
	# of feature maps communicated (F)	**channel distribution**		**# of parameters**
coarse network	1	1-1-1-1		84
	2	2-2-2-2		318
	4	4-4-4-4		1 236
	8	8-8-8-8		4 872
	16	16-16-16-16		19 344
	32	32-32-32-32		77 088
	64	64-64-64-64		307 776

TABLE 9. Properties of the encoder-decoder network and coarse networks used for the DeepGlobe land cover classification segmentation Dataset.

	module	channel distribution	output feature map size	# of parameters
encoder-decoder network	ResNet18-block1	3-64	128×128	9 536
	ResNet18-block2	64-64-64-64-64	64×64	147 968
	ResNet18-block3	64-128-128-128-128	32×32	525 568
	ResNet18-block4	128-256	16×16	2 099 712
	inter-conv (3×3)	256-256-256	16×16	1 180 672
	up-sampling path	256-128-64-32-7	256×256	845 415
	# of feature maps communicated (F)	**channel distribution**	**kernel size**	**# of parameters**
coarse network	1	1-1-1-1	5×5	0
	4	4-4-4-4	5×5	84
	16	16-16-16-16	5×5	19 344
	64	64-64-64-64	5×5	307 776
	256	256-256-256-256	5×5	4 917 504

TABLE 10. Hyperparameters used for synthetic data experiments.

training hyperparameters	
learning rate	0.008
(max.) number of epochs	40
early stopping patience	8 epochs
learning rate decay patience	3 epochs
minimum stopping epoch	20
dropout rate	0.1
batch size	16
loss function	dice loss (Eq. (1))

the DDU-Net distributes the workload across multiple GPUs, enabling efficient processing of large-scale, high-resolution data without losing detail. Future research could explore integrating the strengths of both approaches, balancing efficient workload distribution with single-GPU compatibility.

We trained a DDU-Net with ResNet50 and ResNet18 encoder backbones, the FCtL network, and GL-Net on the same DeepGlobe dataset of $2\,048 \times 2\,048$ images. DDU-Net training followed the settings in Table 6, while FCtL and GL-Net adhered to their respective papers' configurations.

The results, summarized in Table 7, compare accuracy, peak memory usage, and inference time on the DelftBlue supercomputer [53]. The DDU-Net is uniquely capable of parallelizing tasks across multiple GPUs. For instance, the DDU-NetResNet18 with 2×2 subdomains achieves

a mean IoU of 0.684, requiring just 601 MB of memory and 0.0463 ± 0.0011 seconds per $2\,048 \times 2\,048$ image. In contrast, state-of-the-art methods like GL-Net and FCtL require over 2 GB of memory and inference times exceeding 3 seconds. A significant part of this performance gap stems from fundamental differences in approach. Unlike FCtL and GL-Net, which is designed to process parts of the image sequentially on a single GPU, the DDU-Net distributes the entire image workload across multiple GPUs, which enables more efficient parallelization. While the DDU-Net with a ResNet50 backbone uses more memory than FCtL and GL-Net, its memory usage scales almost linearly with the number of GPUs while maintaining a consistent mean IoU, which is exactly the aim of the DDU-Net.

VI. CONCLUSION

This paper develops a new domain decomposition-based U-Net (DDU-Net) architecture for semantic segmentation tasks. Our results show that by including communication between subimages, the DDU-Net can handle high-resolution image segmentation efficiently without sacrificing accuracy or memory efficiency. Our approach improves segmentation accuracy by leveraging inter-subimage communication. Future research will focus on refining communication strategies, applying DDU-Net to more complex datasets, and further evaluating the benefits of parallelization on computing times and memory usage. Another direction is extending the parallelization strategy developed in this paper to other encoder-decoder architectures.

ACKNOWLEDGMENT

The authors acknowledge the use of computational resources of the DelftBlue supercomputer, provided by Delft High Performance Computing Centre (https://www.tudelft.nl/dhpc) [53]. This article describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the article do not necessarily represent the views of U.S. Department of Energy or U.S. Government.

APPENDIX

The properties of the encoder-decoder and coarse networks used for the synthetic dataset experiments are detailed in Table 8. The properties are shown for different configurations in terms of encoder-decoder depth D. The numbers of weights for the encoder-decoder and coarse network architectures used for the DeepGlobe land cover classification dataset are detailed in Table 9. The encoder-decoder network consists of multiple ResNet18 blocks followed by an inter-convolutional and up-sampling path. We show the number of weights in the coarse network for various choices of its architecture, corresponding to the different experiments as shown, for instance, in Fig. 13. Table 10 summarizes the hyperparameters employed during training for synthetic data experiments. Fig. 15 visualizes the difference between predicted and true masks for different configurations of the DDU-Net.

REFERENCES

[1] A. Bakhtiarnia, Q. Zhang, and A. Iosifidis, "Efficient high-resolution deep learning: A survey," *ACM Comput. Surveys*, vol. 56, no. 7, pp. 1–35, Jul. 2024.

[2] K. A. Korznikov, D. E. Kislov, J. Altman, J. Doležal, A. S. Vozmishcheva, and P. V. Krestov, "Using U-net-like deep convolutional neural networks for precise tree recognition in very high resolution RGB (red, green, blue) satellite images," *Forests*, vol. 12, no. 1, p. 66, 2021.

[3] Y. Li, B. Dang, W. Li, and Y. Zhang, "GLH-water: A large-scale dataset for global surface water detection in large-size very-high-resolution satellite imagery," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 20, pp. 22213–22221.

[4] X. Tao, D. Zhang, W. Ma, X. Liu, and D. Xu, "Automatic metallic surface defect detection and recognition with convolutional neural networks," *Appl. Sci.*, vol. 8, no. 9, p. 1575, Sep. 2018.

[5] G. Lv, Z. Yao, D. Chen, Y. Li, H. Cao, A. Yin, Y. Liu, and S. Guo, "Fast and high-resolution laser-ultrasonic imaging for visualizing subsurface defects in additive manufacturing components," *Mater. Design*, vol. 225, Jan. 2023, Art. no. 111454.

[6] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 137–178, Jan. 2021.

[7] P.-L. Bazin, M. Weiss, J. Dinse, A. Schäfer, R. Trampel, and R. Turner, "A computational framework for ultra-high resolution cortical segmentation at 7 Tesla," *NeuroImage*, vol. 93, pp. 201–209, Jun. 2014.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] L. Hou, Y. Cheng, N. Shazeer, N. Parmar, Y. Li, P. Korfiatis, T. M. Drucker, D. J. Blezek, and X. Song, "High resolution medical image analysis with spatial partitioning," 2019, *arXiv:1909.03108*.

[10] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8916–8925.

[11] A. Tsaris, J. Romero, T. Kurth, J. Hinkle, H.-J. Yoon, F. Wang, S. Dash, and G. Tourassi, "Scaling resolution of gigapixel whole slide images using spatial decomposition on convolutional neural networks," in *Proc. Platform Adv. Scientific Comput. Conf.*, Jun. 2023, pp. 1–11.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany. Cham, Switzerland: Springer, Jan. 2015, pp. 234–241.

[13] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Proc. 19th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Athens, Greece. Cham, Switzerland: Springer, Jan. 2016, pp. 424–432.

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, Granada, Spain. Cham, Switzerland: Springer, Sep. 2018, pp. 3–11.

[15] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[16] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet—A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[17] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[18] R. Azad, E. Khodapanah Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. Paul Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of U-Net," 2022, *arXiv:2211.14830*.

[19] A. Toselli and O. Widlund, *Domain Decomposition Methods—Algorithms and Theory* (Springer Series in Computational Mathematics), vol. 34. Berlin, Germany: Springer, 2005. [Online]. Available: https://mathscinet.ams.org/mathscinet-getitem?mr=2104179

[20] S. K. Seal, S.-H. Lim, D. Wang, J. Hinkle, D. Lunga, and A. Tsaris, "Toward large-scale image segmentation on summit," in *Proc. 49th Int. Conf. Parallel Process.-ICPP*, Aug. 2020, pp. 1–11.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[22] Y. LeCun, "Generalization and network design strategies," in *Connectionism in Perspective*, vol. 19. Zurich, Switzerland: Elsevier, 1989, p. 18. [Online]. Available: https://nyuscholars.nyu.edu/en/publications/generalization-and-network-design-strategies

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[24] C. M. Bishop and H. Bishop, *Deep Learning : Foundations and Concepts*. Cham, Switzerland: Springer, 2024. [Online]. Available: https://link.springer.com/10.1007/978-3-031-45468-4

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[26] G. Cathelain, B. Rivet, S. Achard, J. Bergounioux, and F. Jouen, "U-net neural network for heartbeat detection in ballistocardiography," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 465–468.

[27] H. Zhang, Z. Jiang, G. Zheng, and X. Yao, "Semantic segmentation of high-resolution remote sensing images with improved U-Net based on transfer learning," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, p. 181, Nov. 2023.

[28] S. S. Bangaru, C. Wang, X. Zhou, and M. Hassan, "Scanning electron microscopy (SEM) image segmentation for microstructure analysis of concrete using U-net convolutional neural network," *Autom. Construction*, vol. 144, Dec. 2022, Art. no. 104602.

[29] M. L. Richter, J. Schöning, A. Wiedenroth, and U. Krumnack, "Receptive field analysis for optimizing convolutional neural network architectures without training," in *Deep Learning Applications*, vol. 4. Cham, Switzerland: Springer, 2022, pp. 235–261.

[30] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, Dec. 2016, pp. 4898–4906.

[31] A. Tsaris, J. Hinkle, D. Lunga, and P. A. Dias, "Distributed training for high resolution images: A domain and spatial decomposition approach," in *Proc. IEEE/ACM Redefining Scalability Diversely Heterogeneous Architectures Workshop (RSDHA)*, Nov. 2021, pp. 27–33.

[32] T. F. Chan and T. P. Mathew, "Domain decomposition algorithms," *Acta Numerica*, vol. 3, pp. 61–143, Jan. 1994.

[33] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*. Oxford, U.K.: Academic, Oct. 2023. Accessed: Apr. 15, 2025. [Online]. Available: https://doi.org/10.1093/oso/9780198501787.001.0001

[34] V. Dolean, P. Jolivet, and F. Nataf, *An Introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation*. Philadelphia, PA, USA: SIAM, 2015.

[35] C. R. Dohrmann, A. Klawonn, and O. B. Widlund, "Domain decomposition for less regular subdomains: Overlapping schwarz in two dimensions," *SIAM J. Numer. Anal.*, vol. 46, no. 4, pp. 2153–2168, Jan. 2008.

[36] N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl, "Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps," *Numerische Math.*, vol. 126, no. 4, pp. 741–770, Apr. 2014. [Online]. Available: https://mathscinet.ams.org/mathscinet-getitem?mr=3175183

[37] J.-M. Cros, "A preconditioner for the Schur complement domain decomposition method," in *Domain Decomposition Methods in Science and Engineering*, O. W. I. Herrera, D. Keyes and R. Yates, Eds., Mexico City, MX, USA: National Autonomous University of Mexico, 2003, pp. 373–380.

[38] C. R. Dohrmann, "A preconditioner for substructuring based on constrained energy minimization," *SIAM J. Scientific Comput.*, vol. 25, no. 1, pp. 246–258, Jan. 2003. [Online]. Available: https://mathscinet.ams.org/mathscinet-getitem?mr=2047204

[39] C. Farhat, M. Lesoinne, P. LeTallec, K. Pierson, and D. Rixen, "FETI-DP: A dual–primal unified FETI method—Part I: A faster alternative to the two-level FETI method," *Int. J. Numer. Methods Eng.*, vol. 50, no. 7, pp. 1523–1544, Mar. 2001. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nme.76

[40] C. Farhat, M. Lesoinne, and K. Pierson, "A scalable dual-primal domain decomposition method," *Numer. Linear Algebra Appl.*, vol. 7, nos. 7–8, pp. 687–714, Jan. 2000. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/1099-1506%28200010/12%297%3A7/8%3C687%3A%3AAID-NLA219%3E3.0.CO%3B2-S

[41] A. Heinlein, A. Klawonn, M. Lanser, and J. Weber, "Combining machine learning and domain decomposition methods for the solution of partial differential equations—A review," *GAMM-Mitteilungen*, vol. 44, no. 1, Mar. 2021, Art. no. 202100001.

[42] A. Klawonn, M. Lanser, and J. Weber, "Machine learning and domain decomposition methods—A survey," 2023, *arXiv:2312.14050*.

[43] N. Man, S. Guo, K. F. C. Yiu, and C. K. S. Leung, "Multi-layer segmentation of retina OCT images via advanced U-net architecture," *Neurocomputing*, vol. 515, pp. 185–200, Jan. 2023.

[44] A. Klawonn, M. Lanser, and J. Weber, "A domain decomposition-based CNN-DNN architecture for model parallel training applied to image recognition problems," 2023, *arXiv:2302.06564*.

[45] L. Gu, W. Zhang, J. Liu, and X.-C. Cai, "Decomposition and composition of deep convolutional neural networks and training acceleration via sub-network transfer learning," *ETNA-Electron. Trans. Numer. Anal.*, vol. 56, pp. 157–186, Jun. 2022.

[46] K. Mills, K. Ryczko, I. Luchak, A. Domurad, C. Beeler, and I. Tamblyn, "Extensive deep neural networks for transferring small scale learning to large scale systems," *Chem. Sci.*, vol. 10, no. 15, pp. 4129–4140, 2019.

[47] J. Park, D. Kwon, B. W. Choi, G. Y. Kim, K. Y. Kim, and J. Kwon, "Small object segmentation with fully convolutional network based on overlapping domain decomposition," *Mach. Vis. Appl.*, vol. 30, no. 4, pp. 707–716, Jun. 2019.

[48] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss Odyssey in medical image segmentation," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102035.

[49] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[52] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[53] (2024). *Delft High Performance Computing Centre (DHPC), DelftBlue Supercomputer (Phase 2)*. [Online]. Available: https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2

[54] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–17209.

[55] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.

[56] S. D. Khan, L. Alarabi, and S. Basalamah, "Deep hybrid network for land cover semantic segmentation in high-spatial resolution satellite images," *Information*, vol. 12, no. 6, p. 230, May 2021.

[57] M. Sewak, M. R. Karim, and P. Pujari, *Practical Convolutional Neural Networks: Implement Advanced Deep Learning Models Using Python*. Birmingham, U.K.: Packt, 2018. [Online]. Available: https://books.google.nl/books/about/Practical_Convolutional_Neural_Networks.html?id=bOlODwAAQBAJ&source=kp_book_description&redir_esc=y

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[59] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," 2017, *arXiv:1710.03740*.

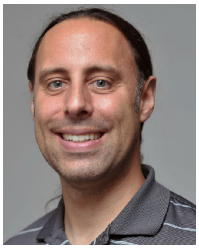
[60] Q. Li, W. Yang, W. Liu, Y. Yu, and S. He, "From contexts to locality: Ultra-high resolution image segmentation via locality-aware contextual correlation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7232–7241.

CORNÉ VERBURG received the first B.S. degree in applied physics and the second B.S. and M.S. degrees in applied mathematics from Delft University of Technology (TU Delft), The Netherlands, in 2022 and 2024, respectively, where he is currently pursuing the Ph.D. degree with a focus on biology-informed data-supported dynamic modeling and system identification for complex biological systems. For his master's thesis, he worked on high-resolution image segmentation using convolutional neural networks (CNNs), integrating domain decomposition strategies, which resulted in the current paper.

His research interests include scientific machine learning (SciML) techniques and the dynamical modeling of complex systems, particularly those that are partially unknown, using a combination of statistical and machine learning-driven methods.

ALEXANDER HEINLEIN received the Diploma degree in mathematics from the University of Duisburg–Essen, in 2011, and the Ph.D. degree in mathematics from the University of Cologne, in 2016.

After his Ph.D. degree, he was a Postdoctoral Researcher at the University of Cologne, until 2021. From 2018 to 2021, he was the Managing Coordinator of the Center for Data and Simulation Science (CDS), University of Cologne. Before becoming an Assistant Professor in numerical analysis with Delft University of Technology (TU Delft), in 2021, he was an acting Full Professor for numerical mathematics for high-performance computing at the University of Stuttgart. His research interests include numerical methods for partial differential equations, scientific and high-performance computing, and scientific machine learning. He focuses on domain decomposition and multiscale methods, model order reduction techniques, and their combination with machine learning. He also works on high-performance implementations on current computer architectures (CPUs and GPUs) and the application to real-world problems. He is a member of American Mathematical Society (AMS), Dutch-Flemish Scientific Computing Society (SCS), European Mathematical Society (EMS), German Association for Computational Mechanics (GACM), the Association of Applied Mathematics and Mechanics (GAMM), and the Society for Industrial and Applied Mathematics (SIAM).



ERIC C. CYR received the B.S. degree in computer science from Clemson University, in 2002, and the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign, in 2008.

He was a Postdoctoral Researcher at Sandia National Laboratories, where he has been continued as a Staff Member, since May 2010. His research interests include scientific machine learning, parallel training algorithms, preconditioning large-scale systems, multigrid methods, and discretization using computational plasma physics. He is a member of the Society of Industrial and Applied Mathematics (SIAM), where he serves as an Associate Editor for the *Journal on Scientific Computing*. In addition, he has been a member of the program committee for the Copper Mountain Conference on Multigrid Methods, since 2023. In 2018, he received U.S. Department of Energy Early Career Award to develop layer-parallel methods for training deep neural networks.