

# Exploration of image and 3D data segmentation methods: an exhaustive survey

Hasnae Briouya, Asmae Briouya, Ali Choukri

Laboratory of Computer Sciences, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

---

## Article Info

### Article history:

Received Sep 27, 2023

Revised Jan 05, 2024

Accepted Jan 12, 2024

---

### Keywords:

2D data

3D data

Convolutional neural network

Image

Segmentation semantic

---

## ABSTRACT

The field of image and 3-dimensional (3D) data segmentation is growing fast and has many uses, like in medicine, and robotics. In this article, we explain how computers understand and divide images and 3D data. We compare different ways of doing this in 2D and 3D, and look at the computer methods used. We also discuss recent work and what they discovered. This article gives a broad overview of what's happening in this area of computer science. It explains the goals of the research, how they do it, and what they've found out. It's a useful guide for researchers to understand what's happening now and what challenges they might face in the future.

*This is an open access article under the [CC BY-SA](#) license.*



---

## Corresponding Author:

Hasnae Briouya

Laboratory of Computer Sciences, Faculty of Sciences, Ibn Tofail University

Kenitra, Morocco

Email: hasnae.briouya@uit.ac.ma

---

## 1. INTRODUCTION

Semantic segmentation stands as a pivotal facet within the realm of computer vision and machine learning, attaining substantial significance across an array of applications. This intricate process involves the meticulous labeling of individual pixels or voxels within images or volumetric data, thereby facilitating the discernment of distinct object classes within a specified visual context [1]. The methodological precision inherent in semantic segmentation empowers computational systems to not only recognize and categorize the contents of an image but also to assign semantic meaning to each pixel or voxel, thereby enhancing the overall understanding of the visual information at hand.

The demand for high-performance semantic segmentation has surged significantly, driven by applications such as autonomous driving [2], indoor navigation [2], environmental monitoring [3], mapping [4], virtual, and augmented reality systems [5]. The accurate delineation of objects within images or volumetric data is crucial for enhancing the immersive experience and functionality of these technologies. However, the field of semantic segmentation faces several challenges that need to be addressed to ensure its effectiveness. These challenges include achieving accurate and real-time segmentation [6], handling diverse datasets with varying complexities, and adapting to both 2-dimensional (2D) and 3D contexts. Overcoming these challenges is essential for the successful deployment of cutting-edge applications [7] relying on semantic segmentation. This article aims to address these challenges by providing a comprehensive overview of deep learning techniques applied to semantic segmentation. It explores the distinctions between 2D and 3D segmentation [8], discusses primary datasets and their complexities, and surveys prevalent neural network architectures dedicated to semantic segmentation.

The motivation behind this research stems from the vital role that accurate semantic segmentation plays in modern technology. From autonomous vehicles [9] to precise navigation systems and immersive digital experiences, the successful implementation of these applications heavily relies on advancements in semantic segmentation. By exploring the latest developments in deep learning, this article aims to empower researchers and practitioners to address current challenges and propel advancements in computer vision and machine learning. This comprehensive examination of semantic segmentation techniques not only assesses their current advantages and limitations but also sets the stage for future investigations. By shedding light on emerging trends and potential areas for improvement, this article contributes to the continuous evolution of semantic segmentation, shaping the technological landscape in the years ahead.

## 2. THE IMPORTANCE OF SEMANTIC SEGMENTATION IN COMPUTER VISION

Semantic segmentation is crucial in computer vision for both 2D and 3D data as it enables the assignment of semantic labels to each pixel [10] or voxel [11] in an image or 3D scene. This step is essential for various applications, including object recognition, autonomous navigation [12], 3D mapping, augmented reality, medical imaging [13], and many more. In the case of 2D data, semantic segmentation is used to extract precise information about objects in an image. It allows for the detection of object boundaries, accurate identification, and precise localization [14]. This can be applied in applications such as traffic surveillance, pedestrian detection in surveillance videos, object detection in medical images [14], and more. For 3D data, semantic segmentation is crucial in extracting information about the structure and semantic meaning of the 3D scene. It is used for object detection, scene understanding, autonomous navigation, 3D mapping, and more. For example, in autonomous navigation [15], semantic segmentation is used to detect obstacles and plan a safe trajectory for the vehicle. In 3D mapping, it helps in reconstructing precise 3D models of the scene with detailed information about the different objects present.

## 3. KEY CHALLENGES OF SEMANTIC SEGMENTATION

Semantic segmentation encounters challenges in handling diverse object appearances due to variations in lighting, scale, and pose. The reliance on manually annotated data introduces potential inaccuracies, hindering precise object labeling. Additionally, the variability in object sizes and the need for adaptability to novel situations pose further complexities in achieving accurate and robust semantic segmentation.

- Object appearance variability [16]: objects can exhibit significant variations in appearance due to lighting conditions, viewing angles, scale, and pose. This makes semantic segmentation challenging as it is difficult to find robust visual features to identify all types of objects.
- Accuracy of annotations: semantic segmentation models are typically trained using manual annotations [17], which can introduce errors and inaccuracies. Specifically, objects may be mislabeled or have ambiguous labels, making semantic segmentation challenging.
- Variation in object sizes: objects can vary greatly in size, making it challenging to determine an appropriate scale for semantic segmentation. Additionally, some objects may be very small and difficult to detect [18], while others may be very large and cover a significant portion of the scene.
- Handling large amounts of data: computer vision data is often large and complex, making real-time processing challenging. For semantic segmentation, this can pose issues with processing time and memory.
- Adaptability to novel situations: semantic segmentation models need to be able to adapt to new situations, such as unknown objects, novel scenes [19] or different environmental conditions. This can be challenging as models are typically trained on specific datasets that may not fully capture the real-world data variability.

## 4. DIFFERENCES BETWEEN 2D AND 3D SEGMENTATION

2D segmentation is a process of image processing where a 2D image is segmented into different regions or objects. This can be achieved using various techniques such as edge detection, pixel classification, and contour detection. 2D segmentation is often used in fields such as pattern recognition, medical imaging, and computer vision [20]. On the other hand, 3D segmentation involves segmenting images in three dimensions, often obtained from medical imaging techniques such as computed tomography (CT) or magnetic resonance

imaging (MRI). 3D segmentation is commonly used in fields such as medical treatment planning, virtual, augmented reality, and 3D modeling [21]. 3D segmentation is more complex than 2D segmentation due to the volumetric nature of 3D images, which contain additional information about depth and object structure. Techniques for 3D segmentation may include threshold-based segmentation, shape-based segmentation, region-based segmentation, and deep neural network-based segmentation.

## 5. DIFFERENCES BETWEEN 2D AND 3D DATASET

This part of the paper is for two types of readers: people who are just starting to learn about the topic and people who already know a lot and want to know what's new. Newcomers need to understand which good datasets to use and some tips for getting the data ready, while experienced researchers might use this section to review the basics or find new information. Now, when it comes to the data itself, we talk about two kinds: 2D data and 3D data. 2D data is like regular pictures-it's flat, with two sides, like a sheet of paper. 3D data, on the other hand, adds depth, like a small box. This depth helps represent things with more detail and complexity. Think of it as the difference between a flat picture and a small object you can turn around and look at from different angles.

### 5.1. 2D

In the field of image analysis and object recognition, the primary focus has traditionally been on two-dimensional images. As a result, datasets containing two-dimensional representations, which include grayscale images and the commonly encountered red green blue (RGB) images, have become the most abundant, and widely used resources. In this context, '2D datasets' refer to collections of these flat images. Understanding the pivotal role of these 2D datasets in computer vision research is crucial, as they serve as the cornerstone of numerous studies. Therefore, this section aims to explore the significance of 2D datasets within the broader context of semantic segmentation.

#### 5.1.1. CamVid

The CamVid database is a pivotal resource for researchers in road/driving scene understanding [22]. It includes four HD video sequences, totaling 22 minutes, and capturing various urban scenarios. What's unique is its controlled camera recording, ensuring consistent settings. With 32 annotated semantic classes and 701 frames, CamVid enables comprehensive road environment analysis, essential for advancing autonomous driving systems and intelligent transportation solutions.

Researchers in computer vision can utilize CamVid's rich annotations, diverse conditions, and partitioning scheme (367 training, 100 validation, and 233 testing) for algorithm development. This dataset fosters innovation in road scene analysis, crucial for advancing autonomous driving systems and intelligent transportation solutions. CamVid's meticulous recording and camera pose tracking offer unique advantages, making it a valuable asset for advancing research in this field.

#### 5.1.2. Cityscapes

The cityscapes dataset is a valuable resource designed to advance the field of computer vision, with a primary focus on semantic understanding of urban environments [23]. This dataset provides semantic, instance-wise, and dense pixel annotations across 30 classes, enabling a wide range of applications in urban scene analysis. It encompasses 5,000 high-quality annotated images and an additional 20,000 images with coarse annotations, covering scenes from 50 different cities. The dataset features rich metadata, including preceding and trailing video frames, stereo views, GPS coordinates, and vehicle odometry data enhancing its utility for tasks like optical flow, tracking, and structure-from-motion. Notably, it offers diverse images captured over several months, varying in weather conditions, and scene complexity. Furthermore, it includes extensions by other researchers, such as bounding box annotations for people, and images augmented with fog and rain.

The cityscapes dataset holds a crucial position in the field of computer vision, providing not only a vast collection of urban scene images but also serving as a comprehensive benchmark suite and evaluation server. Its multifaceted usefulness extends to supporting research in various areas, such as pixel-level semantic labeling, instance-level semantic labeling, and panoptic semantic labeling. Researchers rely on this dataset to advance their work and evaluate the performance of algorithms and models across different dimensions of urban scene understanding.

### 5.1.3. Pascal visual object classes

The pascal pascal visual object classes (VOC) challenge 2006 (VOC2006) was a significant milestone in computer vision [24]. Its primary goal was to recognize objects within realistic scenes across twenty diverse object classes, including vehicles, animals, and humans. The challenge encompassed various tasks: classification, detection, segmentation, and person layout prediction. In the classification task, participants predicted object presence, and provided confidence scores. The detection task required predicting bounding boxes for objects and associated confidences. Additionally, participants could tackle pixel-level object segmentation and person layout, including body parts. The dataset featured 9,963 annotated images across 20 classes, with a strong emphasis on the “person” category. VOC2006 laid the foundation for advancements in object recognition, shaping the landscape of computer vision research and competitions.

### 5.1.4. Sift flow

The sift flow dataset a subset of the LabelMe database, comprises 2,688 fully annotated images, each measuring 256×256 pixels [25]. These images encapsulate the diversity of 8 distinct outdoor scenes, encompassing streets, mountains, fields, beaches, and buildings. Within this dataset, each pixel is meticulously labeled with one of 33 semantic classes, such as “building”, “grass”, “tree”, and more. Additionally, three geometric categories, namely “horizontal”, “vertical”, and “sky”, are also incorporated into the labeling scheme. Of the total images, 2,488 are designated for training, while the remaining 200 are set aside for testing purposes. Pixels that are either unlabeled or incorrectly labeled as a different semantic class are considered as unlabeled, contributing to the dataset’s robustness for scene parsing and segmentation tasks.

## 5.2. 3D

As we delve into the realm of 3D datasets, we open the door to a world of spatial richness and complexity. These datasets, unlike their two-dimensional counterparts, offer a multidimensional perspective, capturing depth, and volume that is essential for understanding the intricate structures of the physical world. In the following sections, we will explore and define various types of 3D datasets, each tailored to specific applications and domains. From medical imaging’s quest for detailed anatomical insights to robotics’ need for precise spatial awareness, these 3D datasets serve as the building blocks for cutting-edge research and technological innovation.

### 5.2.1. ShapeNet

ShapeNet is a high-quality dataset of 3D models that includes over 55 object categories, such as cars, airplanes, furniture, animals, and more [26]. It was created by researchers from Princeton University and Stanford University and contains over 51,000 3D models. The models in ShapeNet are represented as meshes, which are networks of points, edges, and faces that define the shape of the object. Each model is annotated with semantic information, such as the object category and individual parts of the object. ShapeNet has been used for various computer vision tasks, including object classification, object detection, semantic segmentation, and shape generation.

### 5.2.2. ModelNet

ModelNet is a fundamental dataset in the fields of computer vision, machine learning, and computer graphics [27]. It is widely used as a large-scale dataset for 3D computer-aided design (CAD) models. Created and maintained by researchers from Princeton University, this extensive dataset has become a cornerstone for advancing research in diverse domains. Its availability and comprehensiveness have made it a crucial resource for developing and evaluating algorithms and models in the aforementioned fields. The models in ModelNet are represented as meshes, which consist of a set of vertices, edges, and faces that define the shape of the object. Each model is aligned in a canonical pose, facilitating comparison and analysis of different models. ModelNet has been utilized for various tasks, such as object recognition, shape retrieval, and 3D shape synthesis. It has also served as a benchmark dataset for evaluating the performance of different techniques and algorithms in 3D computer vision and machine learning

### 5.2.3. Karlsruhe Institute of Technology and Toyota Technological Institute

The Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) vision benchmark suite is a popular computer vision dataset for vehicle detection, segmentation, and trajectory prediction in urban environments [28]. It includes diverse sensor data like color images, depth images, point clouds, and lidar data, collected from sensor-equipped cars in Karlsruhe, Germany. The dataset is annotated with object

categories, vehicle trajectories, and road signal detections. KITTI is extensively used for training and evaluating machine learning models in various tasks like object detection, semantic segmentation, and vehicle trajectory prediction in urban environments. It is also utilized for evaluating algorithms related to visual odometry, vehicle localization, and road signal detection

#### 5.2.4. ScanNet

ScanNet is a dataset of 3D indoor building scans used for semantic segmentation, 3D reconstruction, and object recognition [28]. It consists of over 1500 scans of indoor environments like apartments, offices, schools, and hospitals covering various architectural styles and sizes. Each scan includes depth maps, RGB images, and annotations for objects such as walls, doors, windows, floors, and ceilings.

The ScanNet dataset has become a fundamental resource in the field of machine learning, offering a comprehensive and diverse collection of data for training and evaluating models in various essential tasks. These tasks include semantic segmentation, 3D reconstruction, and object recognition specifically within indoor building environments. The dataset's versatility extends beyond the academic realm and has found practical applications in industries such as robotics, video game development, and architectural design. By leveraging the ScanNet dataset, machine learning models can be enhanced and their capabilities expanded to tackle real-world challenges in these industries [28].

## 6. NEURAL NETWORK ARCHITECTURES

Let's embark on a journey through the world of neural network architectures. These are like the brains of machines, helping them learn, and solve problems. Just as there are many different tools for different jobs, there are various types of neural networks, each with its own special abilities. In the sections ahead, we'll unravel the secrets of these different neural network architectures. From how convolutional networks make computers understand images to the fascinating world of recurrent networks that handle sequences and time, you'll soon grasp the exciting diversity of neural networks.

### 6.1. Convolutional neural network

A Convolutional neural network (CNN) is a type of neural network architecture specifically designed for image and video processing tasks [29]. CNNs typically comprise convolutional layers, pooling layers, and fully connected layers. CNNs have demonstrated remarkable success in various domains, including facial recognition, object recognition, and semantic image segmentation. They excel at learning and recognizing intricate patterns and structures in images, making them particularly well-suited for visual tasks. Examples of popular CNNs:

#### 6.1.1. AlexNet

AlexNet was a groundbreaking milestone in the field of computer vision [30]. Introduced by Alex Krizhevsky and his team in 2012, this pioneering CNN made a significant impact on the research landscape. Notably, AlexNet achieved a resounding victory in the 2012 ImageNet large scale visual recognition challenge (ILSVRC), surpassing traditional methods with an impressive 84.6% TOP-5 accuracy. In contrast, the closest competitor using conventional approaches achieved only 73.8% accuracy. This achievement showcased the power of deep learning, specifically CNNs, in revolutionizing image recognition tasks and solidified AlexNet's status as a transformative model in the field.

The architecture of AlexNet, depicted in Figure 1, represented a paradigm shift in deep learning. It consisted of five convolutional layers, max-pooling operations, rectified linear unit (ReLU) non-linearities, three fully-connected layers, and dropout regularization. This combination of architectural elements played a crucial role in its success. The use of the ReLU activation function helped alleviate the vanishing gradient problem and enabled faster convergence during training. Additionally, the incorporation of dropout regularization reduced overfitting and improved the model's generalization performance. AlexNet's groundbreaking design and its utilization of these innovative techniques have had a profound impact on the field of computer vision. It has served as a source of inspiration for subsequent CNN architectures, influencing the development of numerous models.

#### 6.1.2. The fully connected layers (also known as dense layers) of neural networks

Fully convolutional networks (FCN) stands for fully convolutional networks, which are neural network architectures designed for semantic image segmentation [31]. First introduced in 2014, they have become

one of the most popular neural networks for this task. The distinctive feature of FCN is the use of a fully convolutional architecture, which consists solely of convolutional layers without any fully connected layers. This architecture allows for the preservation of the spatial geometry of the input image throughout the network, which is crucial for segmentation. FCN also employs transposed convolutional layers to upsample the network output, which is essential for semantic segmentation. Finally, the network activations are converted into a segmented output using a score layer or a sigmoid layer.

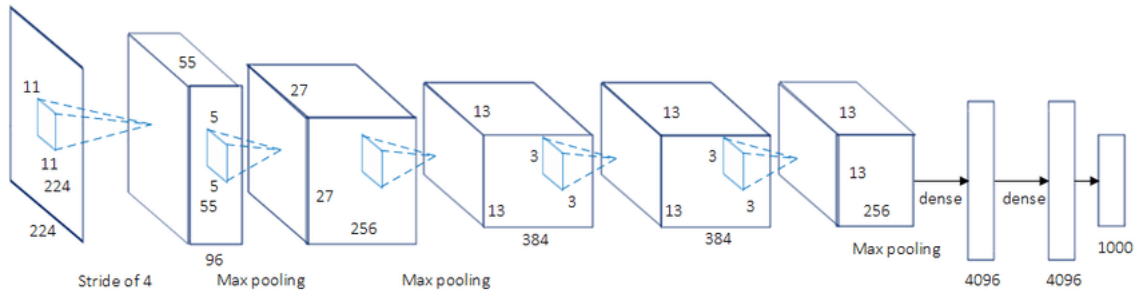


Figure 1. The architecture of AlexNet [24]

### 6.1.3. MobileNet

MobileNet is a CNN architecture designed for mobile devices such as smartphones and tablets [32]. The goal of MobileNet is to provide an efficient CNN architecture in terms of computation and memory that can be implemented on mobile devices with limited resources. MobileNet uses a technique called “depthwise separable convolution”, which separates the convolution operations into two distinct parts: the first part is depthwise convolution as shown in Figure 2, which processes each channel separately, while the second part is pointwise convolution, which combines the results of the depthwise convolution. This technique significantly reduces the number of parameters required compared to traditional CNN architectures while maintaining high performance.

Originally developed by Google, MobileNet has gained significant traction in the realm of mobile applications. It has been seamlessly integrated into well-known platforms such as Google Photos and Google Translate. Leveraging its capabilities, MobileNet has demonstrated exceptional performance in diverse areas, including object recognition, face detection, and semantic image segmentation applications.

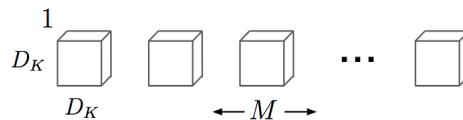


Figure 2. Depthwise convolutional filters [33]

### 6.1.4. VGGNet

VGGNet is a prominent CNN architecture that was introduced in 2015 by a team of researchers from the University of Oxford [34]. It has made a significant impact in the field of deep learning, particularly in the area of image recognition. Known for its distinctive characteristics, VGGNet has left a lasting impression on the landscape of deep learning models. Its architectural design and innovative concepts have contributed to advancements in image recognition tasks, paving the way for further developments in the field (Figure 3).

VGGNet is also known for its modular implementation, where the convolution and pooling layers are organized into blocks [34]. This modularity makes VGGNet easy to adapt to other image processing tasks, it has been widely used in the computer vision community for tasks such as object recognition, semantic image segmentation, and image generation. Although VGGNet is slower to train compared to other CNN architectures such as ResNet and InceptionNet due to its large number of layers, it is still considered a benchmark in the field of image recognition due to its high performance and modularity.

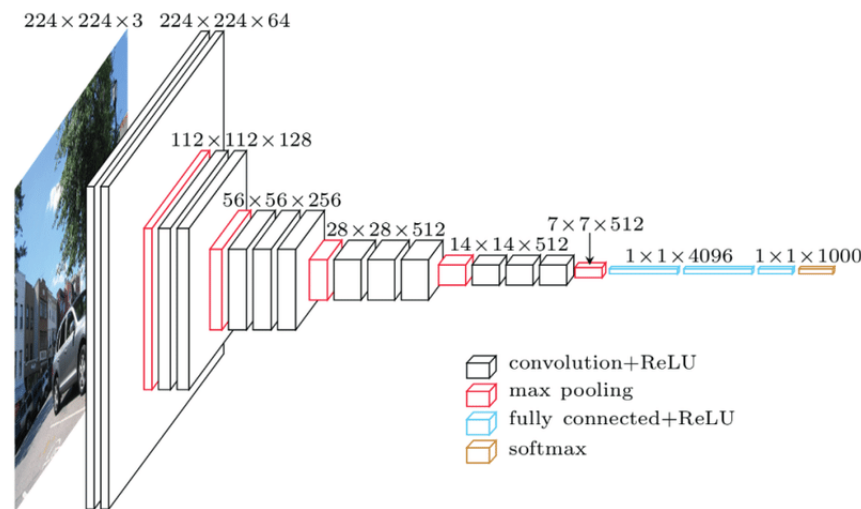


Figure 3. Architecture of VGGNet

### 6.1.5. ResNet

ResNet or residual network, is a CNN architecture introduced in 2015 by a group of researchers from Microsoft Research [35]. ResNet employs a deep-layer architecture with residual blocks to facilitate the training of very deep networks shown in Figure 4. Residual blocks are designed to address the issue of vanishing gradients that occurs when training very deep networks [36]. Residual blocks utilize skip connections to add the activations from the previous layer to the output of the next layer. This enables gradients to propagate more easily through very deep networks, making the training of such networks easier.

ResNet won the 2015 ILSVRC competition using a very deep CNN architecture with 152 layers. Since then, ResNet has been widely used in the computer vision community for tasks such as object recognition, semantic image segmentation, and object detection. ResNet has also inspired many other CNN architectures that use residual blocks to facilitate the training of very deep networks [37].

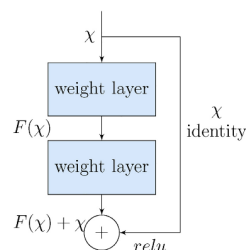


Figure 4. Residual block from the ResNet architecture [35]

### 6.1.6. LeNet-5

LeNet-5 is a CNN architecture introduced in 1998 by Lecun *et al.* [38] for handwritten digit recognition. It was one of the earliest CNNs to have a significant impact on the field of character recognition. LeNet-5 utilizes a deep-layer architecture with multiple layers of convolution and pooling, followed by fully connected layers for the final classification as shown in Figure 5. LeNet-5 also uses non-linear activation functions such as the sigmoid function and the hyperbolic tangent function, which were popular at the time.

LeNet-5, initially developed for recognizing handwritten digits, has had a profound influence on CNNs and beyond. While it was designed with a specific task in mind, its impact has extended far beyond that. LeNet-5's unique architecture and groundbreaking concepts have sparked a wave of innovation, leading to the creation of various CNN architectures specifically designed for a wide range of image-related tasks. As a result, LeNet-5 has become a fundamental building block in the advancement of image processing techniques.

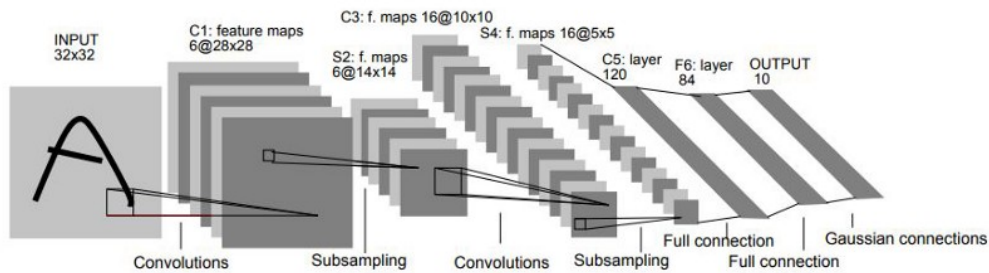


Figure 5. LeNet-5 architecture [38]

### 6.1.7. Clockwork ConvNets

Clockwork ConvNets are a CNN architecture introduced in 2014 by Shelhamer *et al.* [39]. This architecture utilizes a hierarchical approach to solving image and video processing tasks by using clocked computation modules. In clockwork ConvNets, each layer of the network is divided into multiple computation modules, each associated with a specific clock. Each clock is responsible for updating a specific part of the output from the previous layer. Computation modules associated with faster clocks are updated more frequently than modules associated with slower clocks. This approach significantly reduces the computational time required to train very deep neural networks as shown in Figure 6.

Clockwork ConvNets have been evaluated on various image and video processing tasks, including handwritten digit recognition and action recognition in videos. The results have shown that clockwork ConvNets can improve recognition performance by using different clocks for different computation modules. However, this approach is more complex than traditional CNN architectures and requires additional expertise for network design and training.

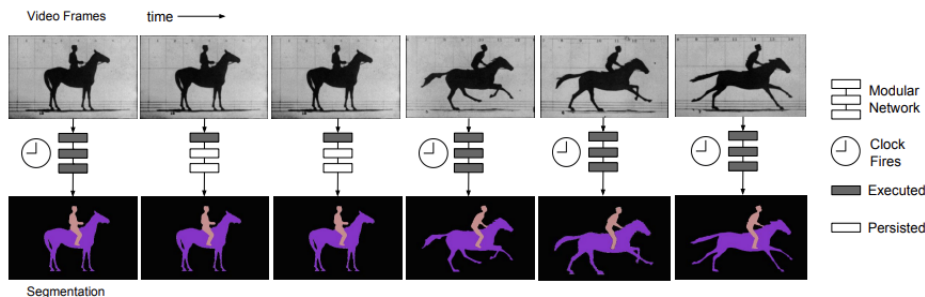


Figure 6. Clockwork method [39]

### 6.1.8. EfficientNet

EfficientNet (EFFNet) is a CNN architecture introduced in 2019 by Tan and Le [40]. EFFNet employs a size optimization approach to achieve CNN architectures that are efficient in terms of computation and memory. EFFNet utilizes a technique called "compound scaling", which involves simultaneously increasing the depth, width, and resolution of neural networks. This technique allows for finding an optimal balance between network complexity and classification accuracy. EFFNet is capable of achieving superior performance compared to other CNN architectures with significantly fewer parameters. EFFNet has been evaluated on various image processing tasks, including object recognition, semantic image segmentation, and object detection, and has achieved state-of-the-art performance on several benchmarks. EFFNet has become highly popular in the computer vision community and is widely used for image processing tasks that require an efficient CNN architecture in terms of computation and memory.

### 6.1.9. Eff-UNet

Eff-UNet is a CNN architecture used for semantic image segmentation. This architecture is based on EFFNet [40]. Eff-UNet utilizes a UNet-like architecture, which consists of a series of convolutional and pooling



layers to extract features from the image, followed by a series of deconvolutional layers to reconstruct the segmented image. However, Eff-UNet uses EfficientNet-like convolution blocks instead of standard convolution blocks to improve network efficiency. Eff-UNet has been evaluated on various semantic image segmentation tasks, including medical image segmentation, road segmentation in satellite images, and cell segmentation in microscopic images. The results have shown that Eff-UNet is capable of achieving state-of-the-art performance with significantly fewer parameters compared to other semantic segmentation architectures. Eff-UNet is, therefore, a promising CNN architecture for semantic image segmentation tasks that require efficiency in terms of computation and memory.

### 6.1.10. U-Net

U-Net is a CNN architecture introduced by Ronneberger *et al.* [14] for semantic image segmentation. The U-Net architecture comprises convolutional and pooling layers for feature extraction and deconvolutional layers for image reconstruction. Notably, U-Net incorporates residual connections to aid information flow as demonstrated in Figure 7. U-Net addresses the challenge of limited annotated data in semantic segmentation by leveraging data augmentation techniques and cropping methods, enabling training with diverse data and fewer labeled samples.

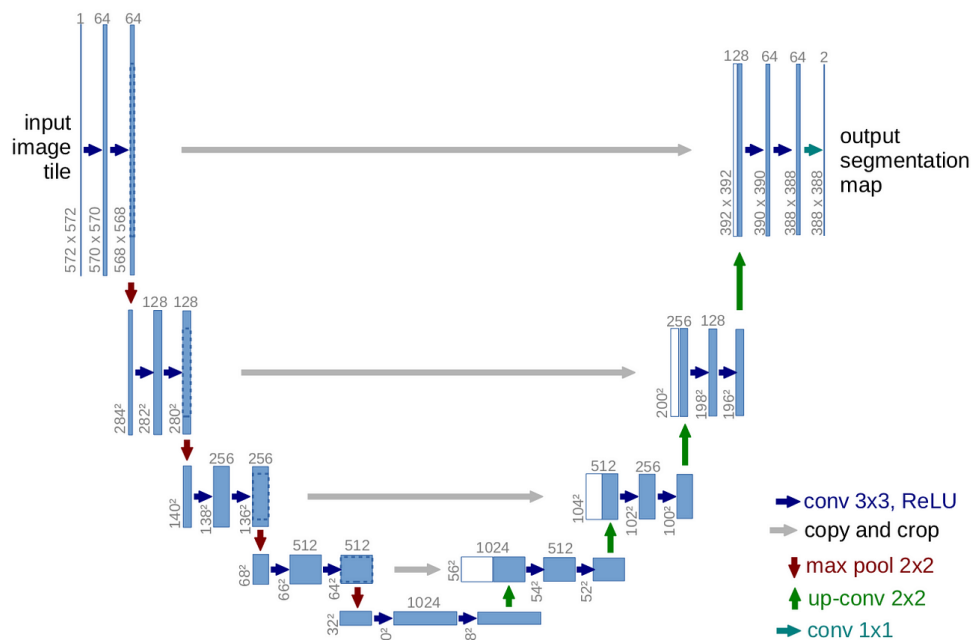


Figure 7. U-Net method

### 6.1.11. SegNet

SegNet introduced by Badrinarayanan *et al.* [41], is a CNN architecture used for semantic image segmentation. Similar to U-Net, SegNet employs convolutional and pooling layers for feature extraction and deconvolutional layers for image reconstruction. However, SegNet adopts an encoder-decoder architecture, where the encoder compresses image information into a latent space, and the decoder reconstructs the segmented image.

A distinguishing feature of SegNet is the use of “max-pooling with unpooling”, which preserves indices of maximum pixels during max-pooling to reconstruct the segmentation map in the subsequent deconvolution phase [42]. SegNet’s performance has been evaluated across diverse semantic segmentation tasks, including medical image, satellite road, and microscopic cell segmentation. The results demonstrate its ability to achieve state-of-the-art performance even with limited annotated data samples.

## 6.2. Recurrent neural network

Recurrent neural network (RNN) is a neural network architecture designed for sequential data processing [42], [43]. It utilizes feedback loops to maintain a representation of the sequence history, making it suitable for tasks like text generation, machine translation, and speech recognition. One challenge with RNNs is the vanishing gradient problem, where gradients for earlier steps become small and hinder weight updates during training. To overcome this, variants like long short-term memory (LSTM) and gated recurrent unit (GRU) have been introduced. LSTM and GRU incorporate advanced memory mechanisms to better handle sequential data. In summary, RNNs and their variants are powerful tools for processing sequential data, offering applications in various domains.

### 6.2.1. Long short-term memory

LSTM is a powerful type of RNN architecture for processing sequential data [44]. It overcomes the vanishing or exploding gradient problem by incorporating an internal memory mechanism with input, output, and forget gates [33]. LSTMs capture relationships between time steps, making them ideal for tasks like text generation, machine translation, and speech recognition. They find applications in various fields, including finance, biology, and social sciences.

### 6.2.2. Gated recurrent unit

GRU is a type of RNN architecture that was proposed to address certain issues faced by LSTM in processing sequential data [45]. GRU is similar to LSTM in that it also has an internal memory that allows it to store long-term information. However, GRU has simpler control gates compared to LSTM, which reduces the number of parameters in the model and makes it faster to train.

GRU consists of two gates: the update gate and the reset gate. The update gate determines which information should be stored in the internal memory based on the input and the previous state. The reset gate decides which information should be forgotten from the internal memory based on the input and the previous state.

## 6.3. RELATED WORK

### 6.3.1. GuessWhat?! game

The proposed method uses a question-answering game between a player and a questioner to guess an object. The player asks questions using images and receives “yes” or “no” responses. A deep CNN extracts image features, and a recurrent neural network models the dialogue. Evaluation was done on the GuessWhat?! dataset, with 155,280 dialogues and 821,889 question/answer pairs. The dataset includes 66,537 unique images and 134,073 unique objects. Dialogues have an average of 5.2 questions, and there are 2.3 dialogues per image. The dataset contains 3,986,192 word tokens and 11,465 different words. Success rates vary, with 84.6% successful, 8.4% unsuccessful, and 7.0% unfinished dialogues. Different subsets are available depending on inclusion criteria.

### 6.3.2. Scene labeling with LSTM RNN

In this excerpt, it is explained that the neural networks used for semantic scene segmentation are composed of three main layers: the input layer, the hidden layer, and the output layer. The hidden layer consists of a 2D LSTM layer and a feed-forward layer and is stacked as a deep network. The input layer receives fixed-size image patches and extracts features using a CNN. These features are then fed into the hidden layer, which consists of a 2D LSTM layer that models the spatial relationships between patches and a feed-forward layer that processes the LSTM layer’s output features. The output layer assigns a class label to each pixel in the image using a softmax activation function.

By stacking multiple hidden layers, the authors created deep neural networks capable of modeling more complex spatial relationships between image patches, which improved the performance of the proposed method. For testing, they used two fully labeled datasets of outdoor scenes: the Stanford background dataset and the sift flow dataset. The LSTM networks achieved comparable results to state-of-the-art methods on the Stanford background and SIFT Flow datasets, with a pixel accuracy of 78.56% for single-scale LSTM networks. The precision differences between LSTM networks and RCNNs with two or three instances were below 1% per class on all datasets.

### 6.3.3. Segmentation-based urban traffic scene understanding

In their influential paper, the authors presented an innovative two-stage methodology specifically tailored for urban street scene classification. The initial stage of their approach focused on patch-based scene classification, which involved systematically categorizing the intricate urban landscape into 13 distinct urban texture classes. This detailed classification not only served as a basis for further analysis but also played a crucial role in building a comprehensive intermediate feature set.

The effectiveness of the approach was evaluated on two challenging sequences. The results demonstrated that while a state-of-the-art scene classifier could accurately identify global classes like road types, a manually designed feature set based on segmentation outperformed it in terms of object classes. The authors emphasized the potential for further enhancements in the system. For instance, the texture classifier could benefit from incorporating additional features such as those based on 3D points or optical flow. The authors concluded by highlighting the promising possibilities this system offers for future research and improvements.

### 6.3.4. Multi-modal medical image retrieval

The main contribution of the article is the generative model-based approach for medical image retrieval using both visual and textual information, which relies on a novel unsupervised learning method using an extended probabilistic latent semantic analysis (pLSA) model. The authors plan to adapt more sophisticated visual analysis techniques to improve the system's performance, model the spatial layout of local features, and explore methods to integrate a medical ontology into the proposed multimodal medical image retrieval system. The system consists of two main components: a learning component to build the model and generate the latent subject representation for each image, and a retrieval component to retrieve images based on queries. The system was successfully tested on the ImageCLEF 2009 medical image retrieval challenge dataset, achieving an average precision of 0.29, which outperformed compared algorithms that only utilized visual or textual features.

### 6.3.5. 3D-R2N2

The 3D-R2N2 recurrent neural network based on LSTMs, learns a correspondence between images and 3D shapes without the need for image annotations or object class labels [46]. The network can reconstruct objects in situations where traditional 3D methods fail, thanks to its ability to handle self-occlusions of objects when multiple views are provided to the network. Two different 2D CNN encoders were introduced for the network: a standard convolutional layer and a deep residual network. The article mentions that the addition of residual connections improved and expedited the optimization process for very deep neural networks. The 3D-R2N2 network then utilizes 3D convolutional LSTMs based on GRU and operates on a spatial structure of distributed units in a 3D grid, enabling it to selectively update its prediction of previously occluded parts of the object. The network was trained on a combination of the ShapeNet and pascal 3D+ datasets to fine-tune hyperparameters and evaluate performance. The proposed approach outperformed the previous method in all categories in terms of intersection over union (IoU) of voxels on the pascal VOC dataset. The limitations of the network were also noted, including its inability to reconstruct as many details as MVS methods and its less precise reconstructions of high-texture objects.

## 7. RESULTS AND DISCUSSION

We present a comprehensive evaluation of CNN models in a unified manner, combining information from a single table that summarizes our investigation. Table 1 contains important details such as the model's inception, developer, top-5 error rate, number of training parameters, and corresponding training time. Additionally, the table includes crucial performance analysis parameters like the number of convolutional layers, strides, fully connected layers, and total multiply-accumulate (MAC) operations [47], [48]. By consolidating the results, we provide a cohesive overview of different CNN models, revealing trends and insights into their relative performance. We further discuss the implications of these findings, exploring factors that contribute to the effectiveness or limitations of the models. This comprehensive analysis offers a nuanced perspective on the landscape of CNN models, fostering a deeper understanding of their strengths and weaknesses in various applications.

Table 1. An analysis comparing different CNN models [47]

Model	Year of inception	Developed by	Top-5 error rate (%)	Time taken to train the model	The popular model	Convolutional layers	Fully connected layers	Total MACs	Total Weights
LeNet	1998	LeNet Yann LeCun <i>et al.</i>	–	–	LeNet-5	2	2	2.3 M	431 k
AlexNet	2012	Alex Krizhevsky <i>et al.</i>	15.3	Five to six days (Two GTX 580 GP)	AlexNet	5	3	724 M	61 M
VGGNet	2014	Simonyan <i>et al.</i>	7.3	Either Two to three weeks (4 Nvidia Titan Black GPUs)	VGG-16	16	3	15.5 G	138 M
ResNet	2015	Kaiming He	3.6	Either Two to three weeks (8 GPU machines)	ResNet-50	50	1	3.9 G	25.5 M

## 8. CONCLUSION

In conclusion, this article provided a comprehensive analysis of 3D data semantic segmentation methods, highlighting the advantages, and limitations of each approach. Various neural network architectures were examined, along with their performances and applications in various fields such as robotics, augmented reality, 3D printing, and medicine. However, applying 3D data semantic segmentation on mobile phones poses significant technical challenges due to limited resources in terms of computational power and memory. Despite this, with advancing technology and improved performance of mobile processors, it is possible that more efficient and lightweight 3D data semantic segmentation architectures will be developed in the future for mobile phone usage. In summary, 3D data semantic segmentation is a promising research area with numerous potential applications in various domains. Future prospects include ongoing improvements in the accuracy and efficiency of neural network architectures, as well as their adaptation to the limited resources of mobile phones for more practical and accessible use.

## REFERENCES




- [1] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene Labeling With LSTM Recurrent Neural Networks," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3547-3555, doi: 10.1109/CVPR.2015.7298977.
- [2] K. Ravishankar, P. Devaraj, and S. Kumar, "Floor Segmentation Approach Using FCM and CNN," *Acadlore Transactions on AI and Machine Learning*, vol. 2, no. 1, pp. 33-45, 2023, doi: 10.56578/ataiml020104.
- [3] J. Tang, Y. Li, M. Ding, H. Liu, D. Yang, and X. Wu, "An Ionospheric TEC Forecasting Model Based on a CNN-LSTM-Attention Mechanism Neural Network," *Remote Sensing*, vol. 14, no. 10, p. 2433, 2022, doi: 10.3390/rs14102433.
- [4] T. Kikuchi, K. Sakita, S. Nishiyama, and K. Takahashi, "Landslide susceptibility mapping using automatically constructed CNN architectures with pre-slide topographic DEM of deep-seated catastrophic landslides caused by Typhoon Talas," *Natural Hazards*, vol. 117, pp. 339-364, 2023, doi: 10.1007/s11069-023-05862-w.
- [5] J. J. Cao, K. Y. Lam, L. H. Lee, X. Liu, P. Hui, and X. Su, "Mobile Augmented Reality: User Interfaces, Frameworks, and Intelligence," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1-36, 2022, doi: 10.1145/3557999.
- [6] A. Lou, S. Guan, and M. H. Loew, "CFPNet-M: A light-weight encoder-decoder based network for multimodal biomedical image real-time segmentation," *Computers in Biology and Medicine*, vol. 154, p. 106579, 2023, doi:10.1016/j.compbiomed.2023.106579.
- [7] A. Sreedevi and C. Manike, "Tomato Leaf Disease Detection Using Cutting-Edge Deep Learning ArchiTectures," *Annals Of Forest Research*, vol. 66, no. 1, pp. 4320-4332, 2023.
- [8] C. Zhao *et al.*, "Context-aware network fusing transformer and V-Net for semi-supervised segmentation of 3D left atrium," *Expert Systems with Applications*, vol. 214, p. 119105, 2023, doi: 10.1016/j.eswa.2022.119105.
- [9] J. Cahill *et al.*, "Exploring the Viability of Bypassing the Image Signal Processor for CNN-Based Object Detection in Autonomous Vehicles," in *IEEE Access*, vol. 11, pp. 42302-42313, 2023, doi: 10.1109/ACCESS.2023.3270710.
- [10] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1473-1481, 2020, doi: 10.1109/CVPRW50498.2020.00187.
- [11] J. Huang and S. You, "Point cloud labeling using 3D Convolutional Neural Network," *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2670-2675, 2016, doi: 10.1109/ICPR.2016.7900038.
- [12] A. Ess, T. Müller, H. Grabner, and L. V. Gool, "Segmentation-Based Urban Traffic Scene Understanding," *In BMVC*, vol. 1, p. 2, 2009.
- [13] Y. Cao *et al.*, "Medical image retrieval: a multimodal approach," *Cancer informatics*, vol. 13, 2014, doi: 10.4137/CIN.S14053.
- [14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *In Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9,*

- 2015, *Proceedings, Part III 18*, vol 9351, pp. 234-241, 2015, doi: 10.1007/978-3-319-24574-4\_28.
- [15] M. Fayyaz, M. Saffar, M. Sabokrou, M. Fathy, R. Klette, and F. Huang, "STFCN: Spatio-Temporal FCN for Semantic Video Segmentation," *CoRR* 2016, doi: 10.48550/arXiv.1608.05971.
- [16] L. Jonathan, S. Evan, and D. Trevor, "Fully convolutional networks for semantic segmentation," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431-3440.
- [17] H. Liu, B. Li, X. Lv, and Y. Huang, "Image Retrieval Using Fused Deep Convolutional Features," *Procedia Computer Science*, vol. 107, pp. 749-754, 2017, doi: 10.1016/j.procs.2017.03.159.
- [18] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, "GuessWhat?! Visual Object Discovery Through Multi-Modal Dialogue," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5503-5512, 2017.
- [19] P. Y. Huang, W. T. Hsu, C. Y. Chiu, T. F. Wu, and M. Sun, "Efficient Uncertainty Estimation for Semantic Segmentation in Videos," *In Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 520-535, 2018.
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [21] N. Alalwan, A. Abozeid, A. A. ElHabshy, and A. Alzahrani, "Efficient 3D Deep Learning Model for Medical Image Semantic Segmentation," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 1231-1239, 2021, doi: 10.1016/j.aej.2020.10.046.
- [22] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88-97, 2009, doi: 10.1016/j.patrec.2008.04.005.
- [23] M. Cordts et al., "The Cityscapes Dataset," *In CVPR Workshop on the Future of Datasets in Vision*, vol. 2, 2015.
- [24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, 2009, doi: 10.1007/s11263-009-0275-4.
- [25] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1972-1979, 2009, doi: 10.1109/CVPR.2009.5206536.
- [26] Z. Wu et al., "3D ShapeNets: A Deep Representation for Volumetric Shapes," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912-1920, 2015.
- [27] A. X. Chang et al., "ShapeNet: An Information-Rich 3D Model Repository," *arXiv:1512.03012v1*, 2015, doi: 10.48550/arXiv.1512.03012.
- [28] A. Geiger, P. Lenz, C. Stillner, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231-1237, 2013, doi: 10.1177/0278364913491297.
- [29] A. W. Salehi et al., "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," *Sustainability* vol. 15, no. 7, p. 5930, 2023, doi: 10.3390/su15075930.
- [30] W. Tang, J. Sun, S. Wang, and Y. Zhang, "Review of AlexNet for Medical Image Classification," *arXiv preprint arXiv:2311.08655*, 2023, doi: 10.48550/arXiv.2311.08655.
- [31] X. Zhou, X. Xia, and G. Qu, "Research on the comparison of FCN and U-Net in remote sensing image change detection," *In Fourteenth International Conference on Graphics and Image Processing (ICGIP 2022)*, vol. 12705, pp. 7-16, doi: 10.1117/12.2679995.
- [32] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv:1704.04861*, 2017, doi: 10.48550/arXiv.1704.04861.
- [33] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [34] E. Prasetyo, N. Suciati, and C. Fatchah, "Multi-level residual network VGGNet for fish species classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no.8, pp. 5286-5295, 2022, doi: 10.1016/j.jksuci.2021.05.015.
- [35] "Evolution of CNN Architectures: LeNet, AlexNet, ZFNet, GoogleNet, VGG and ResNet," *OpenGenus IQ: Learn Computer Science*, 2019, <https://iq.opengenus.org/evolution-of-cnn-architectures/>.
- [36] G. K. Pandey and S. Srivastava, "ResNet-18 comparative analysis of various activation functions for image classification," *2023 International Conference on Inventive Computation Technologies (ICICT)*, pp. 595-601, 2023, doi: 10.1109/ICICT57646.2023.10134464.
- [37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.
- [38] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998, doi: 10.1109/5.726791.
- [39] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, "Clockwork Convnets for Video Semantic Segmentation," *In Computer Vision-ECCV*, vol. 9915, pp. 852-868, 2016, doi: 10.1007/978-3-319-49409-8\_69.
- [40] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *In International conference on machine learning*, pp. 6105-6114, 2019.
- [41] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481-2495, 2017, doi: 10.1109/TPAMI.2016.2644615.
- [42] A. G. Garcia, S. O. Escolano, S. Oprea, V. V. Martinez, P. M. Gonzalez, and J. G. Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41-65, Sep. 2018, doi: 10.1016/j.asoc.2018.05.018.
- [43] M. S. Alam, F. B. Mohamed, A. Selamat, and A. B. Hossain, "A Review of Recurrent Neural Network Based Camera Localization for Indoor Environments," *in IEEE Access*, vol. 11, pp. 43985-44009, 2023, doi: 10.1109/ACCESS.2023.3272479.
- [44] Y. D. Prabowo, H. L. H. S. Warnars, W. Budiharto, A. I. Kistijantoro, Y. Heryadi, and Lukas, "Lstm And Simple Rnn Comparison In The Problem Of Sequence To Sequence On Conversation Data Using Bahasa Indonesia," *2018 Indonesian Association for Pattern Recognition International Conference (INAPR), Jakarta, Indonesia*, pp. 51-56, 2018, doi: 10.1109/INAPR.2018.8627029.
- [45] R. Dey and F. M. Salem, "Gate-variants of Gated Recurrent Unit (GRU) neural networks," *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1597-1600, 2017, doi: 10.1109/MWSCAS.2017.8053243.
- [46] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction," *In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision-ECCV 2016. ECCV 2016. Lecture Notes in*




- Computer Science*, vol 9912, pp. 628–644, 2016, doi: 10.1007/978-3-319-46484-8\_38.
- [47] S. Patel, “A comprehensive analysis of Convolutional Neural Network models,” *International Journal of Advanced Science and Technology*, vol. 29, no.4, pp.771-777, 2020.
- [48] A. M. Zahangir et al., “The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches,” *arXiv preprint arXiv:1803.01164*, 2018, doi: 10.48550/arXiv.1803.01164.

## BIOGRAPHIES OF AUTHORS






**Hasnae Briouya**    is a seasoned software engineer specializing in web and mobile application development. She attained her bachelor’s degree in computer science in 2021, marking the initiation of her academic and professional journey. Concurrently, she is pursuing a Ph.D. program at Ibn Tofail University. Her Ph.D. research is centered on advancing the field of computer science, specifically focusing on the sophisticated exploration of semantic segmentation of 3D data, and natural language processing (NLP). This involves in-depth analysis and comprehension of the structure and meaning of 3D data to enhance the segmentation process. For inquiries or potential collaborations with Hasnae Briouya. She can be contacted at email: [hasnae.briouya@uit.ac.ma](mailto:hasnae.briouya@uit.ac.ma).



**Asmae Briouya**    is a committed software engineer specializing in the development of web and mobile applications. She successfully earned her bachelor’s degree in computer science in 2021, marking the commencement of her academic and professional journey. In parallel with her bachelor’s degree, Asmae pursued a Ph.D program at Ibn Tofail University. Her Ph.D research revolves around two captivating areas: Natural Language Processing (NLP) and 3D data segmentation. In the realm of NLP, Asmae is dedicated to developing algorithms and models that empower computers to understand and process human language. Concurrently, she delves into the intricate field of 3D data segmentation, aiming to partition three-dimensional data into meaningful, and distinct parts. For any inquiries or potential collaborations. She can be contacted at email: [asmae.briouya@uit.ac.ma](mailto:asmae.briouya@uit.ac.ma).



**Ali Choukri**    He is an assistant professor at the National Academy of Applied Sciences, completed his master’s in computer science and telecommunications at the University of Ibn Tofail, Kenitra, Morocco, back in 2008. His academic journey includes a Ph.D from the School of Computer Science and Systems Analysis (ENSIAS) and a degree from ENSET (Higher Normal School of Technical Teaching). Within the SIME laboratory’s MIS team, Choukri focuses his research on mobile intelligent ad hoc communication systems and wireless sensor networks. His diverse interests span across ubiquitous computing, the internet of things (IoT), QoS routing, mathematical modeling, game theory, optimization, and more. To get in touch or for any inquiries. He can be contacted at email: [ali.choukri@uit.ac.ma](mailto:ali.choukri@uit.ac.ma).