Received: 16 June 2023 | Revised: 19 July 2023 | Accepted: 25 July 2023 | Published online: 27 July 2023

REVIEW

Archives of Advanced Engineering Science
2023, Vol. 1(1) 11–26
DOI: 10.47852/bonviewAAES32021220

BON VIEW PUBLISHING

Leveraging Deep Learning Techniques to Obtain Efficacious Segmentation Results

Joy Purohit[1] **and Rushit Dave**[2,]*

[1]*Department of Computer Engineering and Information Technology, Veermata Jijabai Technological Institute, India*

[2]*Department of Computer Information Science, Minnesota State University, USA*

Abstract: Image segmentation is a critical task in the field of computer vision. In the past, traditional segmentation algorithms were frequently used to tackle this problem but had various shortcomings. However, the advent of deep learning has revolutionized this field, leading to the development of novel image segmentation algorithms. This paper presents a comprehensive overview of deep learning-based models applied to medical imaging, iris recognition, pedestrian detection, and autonomous driving. The study encompasses various techniques, such as convolutional neural networks, fully convolutional neural networks, encoder–decoder architectures, multi-scale approaches, attention mechanisms, and image transformers. Moreover, this paper evaluates the performance of these models on relevant datasets, providing insightful recommendations for researchers to integrate promising techniques into their work for specific applications. The discussion also explores the challenges, constraints, and potential research directions in these domains.

Keywords: deep learning, image segmentation, medical imaging, iris recognition, pedestrian detection, autonomous driving

1. Introduction

Image segmentation is a computer vision task in which a digital image is divided into various subgroups called image segments that aid in image processing and classification tasks. Many visual understanding systems depend on the results of image segmentation as a key component. The task of image segmentation has several uses, such as in the fields of medical imaging (e.g., boundary estimation of tumors, segmentation of anomalies in organs), object detection (e.g., pedestrian detection, face detection), recognition systems (e.g., face recognition, iris recognition), satellite image analysis, and video surveillance. The image segmentation problem can be characterized as the task of classifying each individual pixel in an image according to a predefined set of labels. Furthermore, recent image segmentation methods can be mainly classified into two categories: semantic segmentation and instance segmentation. Semantic segmentation is a pixel-level classification problem where each pixel of an image is classified to a class label from a predefined set of labels where several objects of the same class are handled as a single entity. However, in instance-level segmentation, multiple objects of the same class are treated as individual instances. These individual instances are obtained through a sequence of object detection and semantic segmentation operations, which are carried out inside each individual bounding box produced. A bounding box is a hypothetical rectangle that acts as a point of reference for object recognition and creates a collision box for that object.

In the past, various techniques were proposed to represent the input image as a feature vector based on characteristics such as texture, shape, and color. Color moments (Huang et al., 2010), histogram of oriented gradients (Chen et al., 2015; Cherabit et al., 2012; Pan et al., 2015), scale-invariant feature transform (Lowe, 2004), Gabor (Idrissa & Acheroy, 2002), and wavelet transforms (Mallat, 1989) are among the most popular hand-crafted features that extract features such as color, shape, and texture. Additionally, traditional segmentation algorithms based on thresholding (Otsu, 1979), region-based (Lalaoui & Mohamadi, 2013), and edge-based (Al-Amri et al., 2010; Al-Fahoum, 2003) techniques were popular. However, the process of hand-crafted feature extraction is application-specific and sensitive to noise, illumination, scale, translation, and rotation (Al-Fahoum & Reza, 2004). Furthermore, there is limited performance as it is not able to represent all the characteristics of an image. In addition, edge-based techniques fail on smooth edges and images having numerous edges, region-based techniques demand accurate seed-point determination, and threshold-based techniques fall short on images having complex intensity distribution (Al-Fahoum et al., 2014) (e.g., medical images).

In recent times, the use of deep learning methods has gained popularity and shown commendable advancement in the field of segmentation. These techniques are delivering outstanding results and frequently attaining the best accuracy rates on common benchmarks, leading to a paradigm shift in the industry. Commonly, while training the deep learning model, images are vectorized and passed through a neural network where the image features are extracted and a classifier activation function is used at the last layer of the network to generate a segmentation map. This generated segmentation map and the ground truth map are fed to the loss function, which conducts backpropagation on the model to update the weights. Initially, convolutional neural networks (CNNs) were introduced, which showed promising results in extracting the features of an image by learning and labeling. Several researchers have proposed various CNN architectures (He et al., 2016;

*__Corresponding author:__ Rushit Dave, Department of Computer Information Science, Minnesota State University, USA. Email: rushit.dave@mnsu.edu

© The Author(s) 2023. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

Krizhevsky et al., 2017; Simonyan & Zisserman, 2014), which vary in training parameters and the depth of convolution layers. However, a fully connected layer used in CNN to generate an output segmentation map increases the training time and leads to an immense number of parameters in the model. This adds to the computational complexity of the model, which demands more resource consumption and a longer training time.

Unlike CNN, which uses a fully connected layer after the last convolution layer to produce an output segmentation map, a fully convolutional neural network (FC-NN) (Long et al., 2015) was then proposed, which gets rid of the fully connected layer and transforms an intermediate feature map to the size of the input image using transposed convolution. This development has shown improved performance with less computational complexity as there are fewer parameters involved when compared to CNN. In recent times, researchers have proposed numerous techniques such as encoder–decoder-based architectures, attention mechanisms, multi-scale information fusion, and the application of distinct transformers in FC-NN that aid in addressing the segmentation challenge. Additionally, some methods also contain an ensemble of these techniques, enabling the model to gather various kinds of information and leading to a boost in performance. This work provides a comprehensive review of this recent literature, covering a range of ground-breaking initiatives in semantic and instance segmentation, consisting of encoder–decoder architectures, attention models, multi-scale approaches, and image transformers in the applications of medical imaging, iris recognition, pedestrian detection, and autonomous driving.

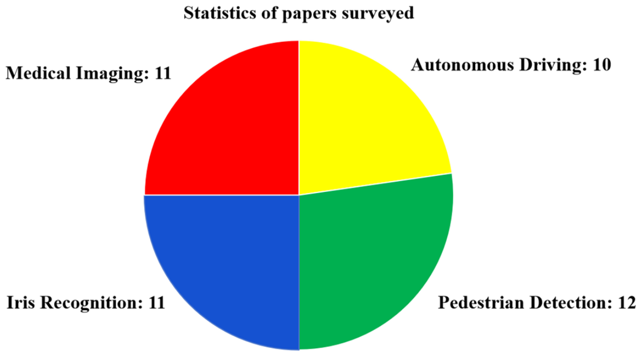
The primary contributions of the paper can be stated as follows:

- Discussion and analysis of distinct models belonging to each application.
- Inference was drawn on the basis of performance of the models on a common dataset to suggest the most promising technique out of all the models reviewed for each application.

2. Literature Review

In Figure 1, the surveyed papers highlight advancements in each domain, addressing research gaps and presenting novel techniques for improved outcomes. The insights gathered from this survey contribute to the progress and future directions in these crucial fields.

Figure 1
Statistics on paper surveyed for each application

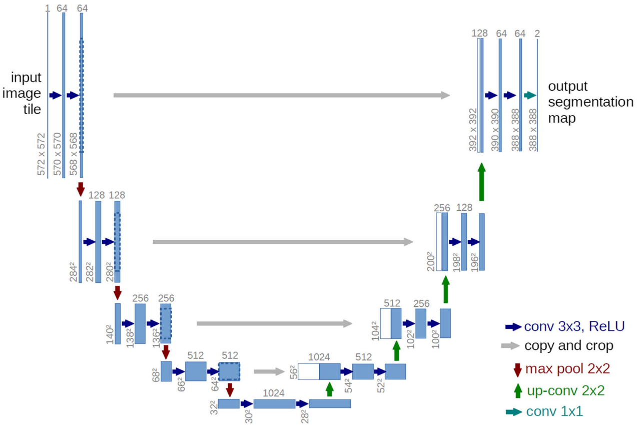


2.1. Medical imaging

2.1.1. Encoder–decoder networks

Figure 2 illustrates the U-Net network proposed by Ronneberger et al. (2015), which is an encoder–decoder type of architecture that consists of a contracting path on the encoder side and an expanding path on the decoder side. An image is given as an input to the encoder, and an output segmentation map is generated through the last layer of the decoder. Each step in the contracting path extracts high-level resolution feature maps through repeated convolutions and downsampling implemented using max pooling. Each step in the expanding path consists of convolution layers and an upsampling of the feature map to extract spatial information.

Figure 2
U-Net. From Arora et al. (2021)



Further, concatenation is performed of the corresponding cropped feature map from the encoder subnetwork (a fusion of high-resolution feature maps from the encoder and semantically rich feature maps from the decoder) with the decoder subnetwork, which is known as skip connections. These skip connections enable the fusion of high-resolution feature maps from the encoder and semantically rich feature maps from the decoder, ensuring the extraction of deep-level features and precise localization of the object in the image. The final layer in the decoder maps the feature vector to the desired number of segmentation classes.

By stacking two U-Net architectures on top of one another to efficiently extract more semantic information, Jha et al. (2020) propose a novel approach. It uses two U-Net architectures in sequence, each having two encoders and decoders, where the first encoder is the pre-trained VGG-19 (Simonyan & Zisserman, 2014), which is trained on ImageNet (Deng et al., 2009). ImageNet is a large dataset of annotated images created with the intention of serving as a tool to support research and the creation of more effective computer vision techniques. The distinguishing factors between Network 1 of Jha et al. (2020) and Ronneberger et al. (2015) are the application of VGG-19 (Simonyan & Zisserman, 2014) in the first encoder, the usage of atrous spatial pyramid pooling (ASPP) (Chen et al., 2017) (between the encoder and the decoder, which helps to extract high-resolution feature maps), and the squeeze-and-excitation (SE) block (Hu et al., 2018), which enhances the quality of the feature maps by passing more appropriate information. In Network 2 of Jha et al. (2020), the difference lies in the use of ASPP and SE block in the decoder. The skip connections to the first decoder are just from the first encoder, but for the second decoder, the skip connections are received from both encoders, with which the feature map's spatial resolution is maintained. The final output segmentation map is computed by concatenating the resulting maps from Network 1 and Network 2 and passing them through a sigmoid function.

By leveraging the strengths of the deep residual model (He et al., 2016), the recurrent convolutional neural network (RCNN) (Liang & Hu, 2015), and U-Net (Ronneberger et al., 2015), the RU-Net and R2U-Net models are proposed (Alom et al., 2018). For implementing RU-Net and R2U-Net, recurrent convolutional layers (RCLs) and RCLs with residual connectivity are used, respectively, replacing the standard forward convolutional layers used in Ronneberger et al. (2015). The residual connection allows for a deeper model that is more effective.

The operations of the RCL are performed with respect to the discrete-time steps implemented according to the RCNN, and for gaining residual connectivity, the output of the RCL is passed through a residual unit (recurrent residual convolutional neural network (RRCNN)). The concatenation of the input to the RCNN with the output of the RCNN computes the output through the RRCNN. Different from Ronneberger et al. (2015), efficient feature accumulation is included in both models, which are done at different time steps, which leads to a stronger feature representation. This helps in extracting low-level features efficiently and improving segmentation performance. The crop and copy method employed in Ronneberger et al. (2015) is replaced by concatenation operations, resulting in a more sophisticated deep learning model.

The Unet++ architecture proposed by Zhou et al. (2018) is constructed by re-designing the skip pathways in Unet (Ronneberger et al., 2015) and further adding a deep supervision module to the network. Unlike Unet, the feature maps from the encoder undergo a dense convolution block employing convolution layers and dense skip connections. These dense convolution blocks bridge the semantic gap between the encoder and decoder feature maps, which helps the optimizer solve an easier optimization problem as compared to Unet. Rather than straightforward connections as in Unet, dense connections are employed in Unet++, which improve the gradient flow in the network. Deep supervision warrants model pruning and enables the model to operate in two modes. Furthermore, a combination of binary cross-entropy and the dice coefficient loss function is attributed to the full-resolution feature maps at different semantic levels. By replacement of the dense skip connections of the Unet++ architecture (Zhou et al., 2018) with full-scale skip connections and the deep supervision module with full-scale deep supervision and further the addition of a novel classification-guided module (CGM), Unet3+ was designed (Huang et al., 2020).

To explore sufficient information at full scale, full-scale skip connections are implemented, in which at each decoder layer it incorporates both same-scale and lower-scale feature maps from the encoder (inter-connection) and high-scale feature maps from the decoder (intra-connection), which combine fine-grained details and coarse-grained semantics at full scale. Further, a feature aggregation mechanism is applied to each decoder layer for the five-scale feature maps obtained. Different from the deep supervision proposed in Unet++, at each decoder stage in Unet3+, a side output is computed that is supervised by the ground truth. Furthermore, to enhance the boundary of the organs, a multi-scale structural similarity index (MS-SSIM) (Wang et al., 2003) loss function is proposed that assigns higher weights to fuzzy boundaries. This allows the network to monitor fuzzy boundaries better, as the greater the regional distribution difference, the higher the MS-SSIM value. A CGM is proposed to detect if the image consists of an organ or not. This is done by feeding the feature map extracted from the deepest encoder layer through a series of dropout, 1 × 1 convolution, max-pooling, and sigmoid activation functions, which produce a 2-dimensional vector indicating probabilities of with or without organs. Further, this classification output is multiplied by each side-segmentation output, and the loss function (binary cross-entropy loss) is guided to prevent segmentation of a non-organ image.

Figure 3 elaborates the Trans-UNet architecture proposed by Chen et al. (2021), which adopts the U-Net shape (Ronneberger et al., 2015) and integrates a hybrid CNN-transformer encoder that generates high-resolution feature maps capturing global context and a cascaded decoder that upsamples the feature maps for precise localization. Skip connections are maintained to output feature maps with high resolution, global context, and precise localization. Transformer blocks are precisely depicted in Figure 4 and are used in the encoder for computing encoded feature maps.

Each block consists of image sequentialization, which is the extraction of a sequence of 2D patches from the input image, followed by patch embedding, which outputs the encoded feature map. In patch embedding, each extracted patch in the sequence is mapped to a D-dimensional embedding space (tokenized patches) using a trainable linear projection that is concatenated with learnable position embeddings. Further, the tokenized patches are fed to a multi-head self-awareness (MSA) module followed by a multi-layer perceptron block. The encoded feature map is then reshaped and fed to the cascaded upsampler, which outputs the final segmentation map. The skip connections are established between each upsampling layer in the upsampler and the convolution layers in CNN, which enable feature aggregation at different resolutions.

Different from the hybrid CNN-transformer encoder and cascaded upsampler in Trans-Unet described above, Cao et al. (2022) propose Swin-Unet which is a hierarchical Swin transformers with shifted windows and a patch merging layer (downsampling of the input) as the encoder, and a symmetric Swin transformer-based decoder with a patch-expanding layer (upsampling of the input) as the decoder. The architecture of Swin-Unet is similar to Unet (Ronneberger et al., 2015), where a bottleneck is introduced additionally. The tokenized patches are calculated in a similar way to Trans-Unet. These tokenized patches are fed to the encoder, which consists of Swin transformer blocks (Liu et al., 2021) (responsible for feature representation learning) and patch merging layers.

Different from the conventional MSA module used in Trans-Unet, the Swin transformer block is constructed based on the shifted windows method (Liu et al., 2021). The window-based multi-head self-attention module and the shifted-window-based multi-head self-attention module are used as self-attention mechanisms in two successive Swin transformer blocks.

Each encoder consists of three sets of two consecutive SWI transformer blocks followed by a patch merging layer for downsampling the feature maps. As the transformers are too deep to converge (Touvron et al., 2021), only two single-phase transformer blocks are used in the bottleneck. Corresponding to the encoder, the decoder is symmetrically built with S-shaped transformer blocks. Patch's expanding layer upsamples the feature maps for precise localization. Finally, the output from the last patch-expanding layer is fed to a linear projection to compute the final pixel-level segmentation map.

Figure 3
Trans-UNet. From Zhao et al. (2022)

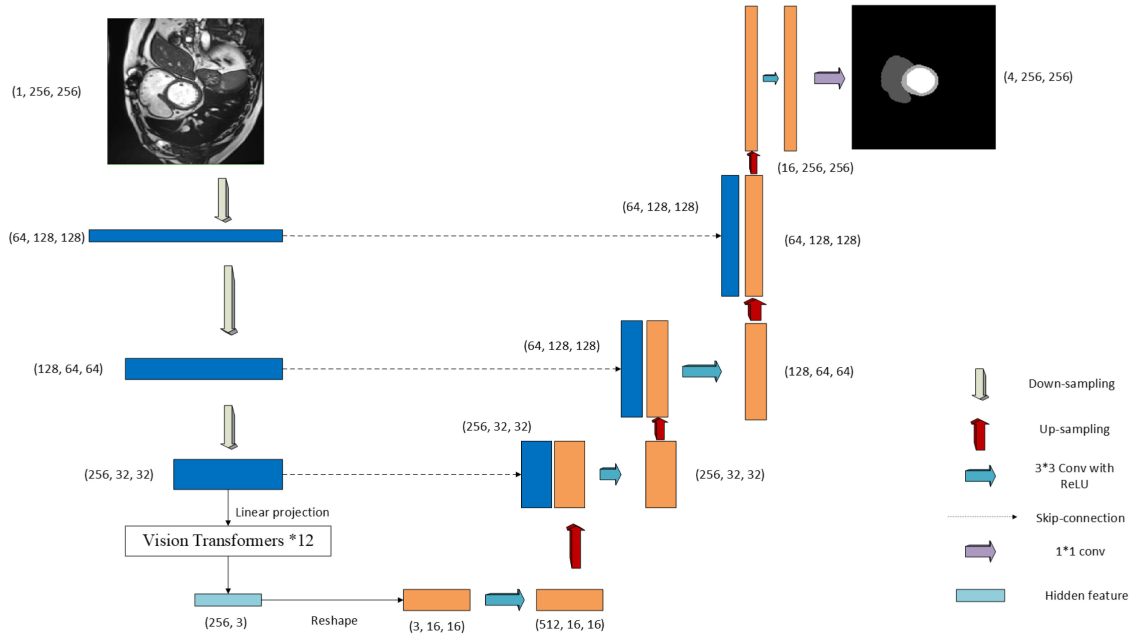
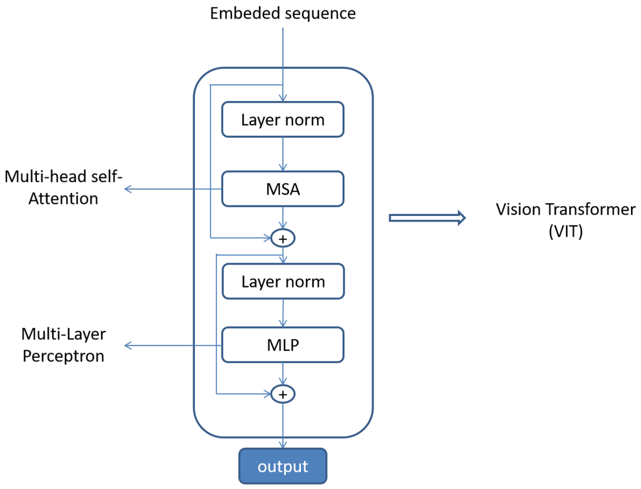


Figure 4
Transformer blocks. From Zhao et al. (2022)



The context-encoder network proposed by Gu et al. (2019) aims to capture more high-level information and maintain spatial information for the segmentation task. The three main modules are the feature encoder module, the context extractor module, and the feature decoder module. In the feature encoder module, the encoder blocks in U-Net (Ronneberger et al., 2015) are replaced by the ResNet-34 (He et al., 2016) (residual connectivity prevents gradient-vanishing problems and accelerates the convergence of the network), retaining the first four blocks.

The context extractor module consists of the dense atrous convolution (DAC) module and the residual multi-kernel pooling (RMP) module, which extract context semantic information and generate high-level feature maps. DAC has four cascade branches with gradual increments in a number of atrial convolutions. Thus, it employs different receptive fields (motivated by inception (Szegedy et al., 2017)), which are able to extract features of various object sizes. The RMP module resolves the issue of the large variation in object size by encoding global context information at four different receptive fields. Hence, the four-level output contains feature maps of various sizes. Similar to Apostolopoulos et al. (2017), a decoder block is adopted consisting of a 1 × 1 convolution, a 3 × 3 transposed convolution, and a 1 × 1 convolution consecutively.

Kaul et al. (2019) propose a general architecture that incorporates a separate self-attention mechanism into a combination of Res-Net (He et al., 2016) and SE-Net (Hu et al., 2018) hybrid architectures. The proposed FocusNet architecture consists of two parallel information flow branches that are in the form of encoder–decoder networks, where one branch is fully devoted to attention. Skip connections are used throughout the architecture to promote better gradient flow, which helps with the easier training of a deep network.

The encoder in the first branch consists of 3 × 3 convolution layers with residual connectivity. The second branch consists of the SE blocks (used extensively to recalibrate the weighing of the output feature maps at intermediate steps), where per-layer decoded output is passed through a sigmoid-gated function that is multiplied by the output of the first SE block. Downsampling is done through stride convolution throughout the architecture, and no bottleneck full pre-activation residual blocks (He et al., 2016) are implemented between the deepest layer of the encoder and the decoder in each branch.

2.1.2. Other developments

Ragab et al. (2019) modified a general computer aided detection system in the image segmentation module, where the suggested two approaches are: (1) determining the region of interest (ROI) manually and (2) using threshold- and region-based techniques. Further, the proposal employs the architecture of the deep convolutional neural network, replacing the last fully connected layer with support vector machine.

A novel segmentation method integrating FC-NN and conditional random fields (CRFs) is proposed by Zhao et al. (2018). Local image and context information on a larger scale is extracted through the FC-NN (Long et al., 2015). Basically, the FC-NN computes the probability of assigning predefined labels to each pixel, and the CRF takes these probabilities as input and globally optimizes the spatial consistency, considering pixel intensity and position as parameters.

2.2. Iris recognition

2.2.1. Traditional segmentation methods

Different from just using edge-based algorithms for iris segmentation as in the method illustrated above, Li et al. (2019) propose a combination of learning-based and edge-based algorithms. First of all, a faster R-CNN (FRCNN) (Ren et al., 2015) with only six layers is used to locate one bounding box having maximum pixels of eye class and an appropriate aspect ratio in the image, inside which the pupil and limbus boundary estimation takes place.

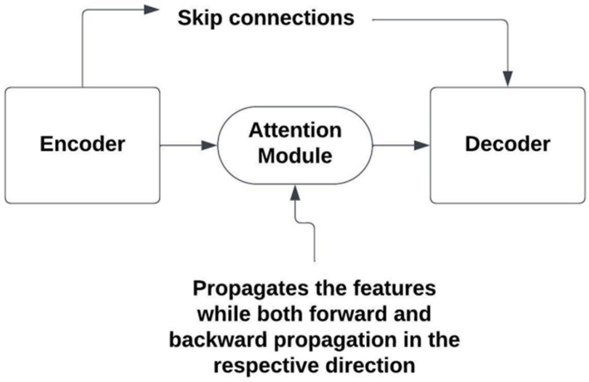
For locating the pupil region, a Gaussian mixture model is applied inside the proposed bounding box, which is followed by image processing to isolate one pupil region. Through pixel scanning, five key boundary points are chosen, including the center of the pupil region, to smooth the circle curve of the pupil and remove the noisy pixels. A sophisticated boundary point selection algorithm is used to select the boundary points of the limbus, and then the limbus circular boundary is approximated using these points.

2.2.2. Fully convolutional neural networks

As shown in Figure 5, the encoder–decoder architecture has been employed in Wang et al. (2020) for iris segmentation purposes, where it additionally incorporates an attention module to help improve the segmentation performance. A central bottleneck part exists between the encoder and the decoder, which comprises an attention module. The attention mechanism allows the model to assign different appropriate weights to different channels in the feature maps and re-estimates the spatial distribution of the feature maps accordingly as mentioned in Woo et al. (2018) and Park et al. (2018) enabling more discriminative features to be learned.

The attention module consists of the ASPP module (Chen et al., 2017), which extracts multi-scale contextual features and then applies the attention mechanism where important feature signals are considered and disturbing noise feature signals are suppressed simultaneously. An attention map is generated by multiplying the original feature map with the importance level of each pixel. Lastly, the original feature map is concatenated with the attention map to restore valuable information in the original feature map. Inner and outer boundary points are computed, and the least-squares circle fitting algorithm (Chernov & Lesort, 2005) is applied to produce circular inner and outer iris boundaries, which are further refined.

Figure 5
IrisParseNet


Skip connections
Encoder
Attention Module
Decoder
Propagates the features while both forward and backward propagation in the respective direction

Based on the same idea of learning more discriminative features by focusing on relevant regions using attention as mentioned in IrisParseNet above, Lian et al. (2018) propose an attention U-Net architecture where it employs the U-Net architecture (Ronneberger et al., 2015) and incorporates a different attention mechanism when compared to IrisParseNet. Attention U-Net regresses a bounding box of the potential iris region and generates an attention mask, which is then fused with feature maps, enabling the model to focus more attention on iris regions.

The regression module is added at the end of the encoder part, which helps generate the attention mask M. After computing the attention mask M, it guides the final segmentation by forcing the model to pay more attention to iris regions, and a soft attention scheme has been applied. The attention mask is multiplied with the feature maps of the second-last layer of the decoder, and now these attention-incorporated features are used to guide the backpropagation of the model and are essential to the attention mechanism.

Jalilian and Uhl (2017) proposes 3 variants of a fully convolutional encoder–decoder network, which is similar to Badrinarayanan et al. (2017). The first variant consists of five stocks in the encoder and the decoder, where each stock comprises blocks whose architectures are formed from convolution layers (body network (BN) and ReLu). Additionally, the blocks in the encoder consist of a max pooling layer for downsampling, and these max pooling indices are stored to pass this information to the respective upsampling layer in the decoder. A softmax classifier is a mathematical function that normalizes the output values, which turns weighted sum values into probabilities that add up to one. Each value in the softmax function's output is regarded as a likelihood that a given class would contain that value. So, this Softmax classifier outputs the probability map corresponding to each pixel.

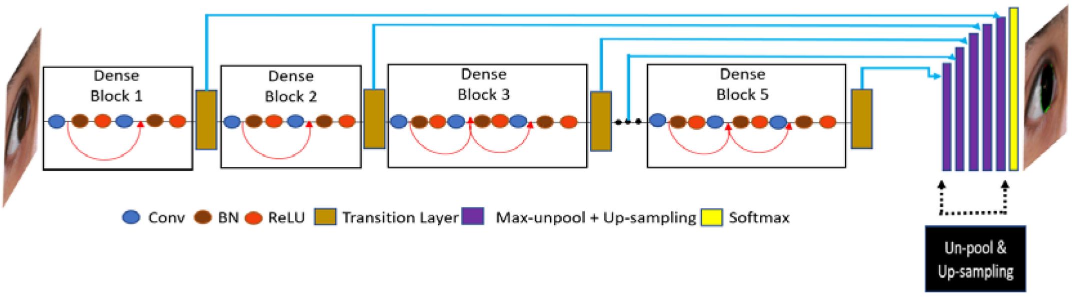
The second variant, the "basic variant," has a lightweight architecture as compared to the original variant and comprises four stocks in the encoder, with the decoder having the same blocks used in the first variant.

The third variant, the Bayesian Basic variant, enables pixel-level segmentation using Monte-Carlo sampling and the drop-out technique (Gal & Ghahramani, 2016; Srivastava et al., 2014). In contrast to the basic variant, two extra drop-out layers are added to the last two blocks of the encoder and the first two blocks of the decoder. At test time, the posterior distribution over the weights is sampled and generates the distribution of softmax class probabilities, where the mean of these samples is for segmentation prediction and the variance denotes the model uncertainty for each class.

To extract more details from an image, Zhang et al. (2019) propose fully dilated convolution combining U-Net (FD-UNet), which in Unet (Ronneberger et al., 2015) replaces all the 3 × 3 convolution layers except the 1 × 1 convolution in the last layer with dilated convolution (Yu & Koltun, 2015) having a dilation rate of 2. This augments the receptive field of the convolution kernel, enabling greater information extraction from the image.

Replacing the series of 3 × 3 convolution layers of the encoder in Unet (Ronneberger et al., 2015) with squeeze-expand modules (Iandola et al., 2016) improves accuracy while reducing computation time and the number of trainable parameters; such an architecture is proposed by Sardar et al. (2020). Further, an interactive learning mode is incorporated. First of all, the

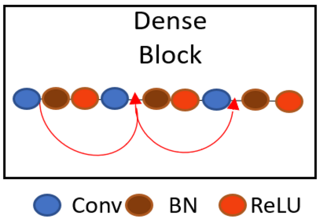
Figure 6
IrisDenseNet. From Arsalan et al. (2018)



incorrectly predicted output images (having misclassified pixels) are refined through user intervention, and these refined images are added to the ground truth dataset. Second, depending on a threshold value set, if the number of incorrectly predicted images exceeds this set threshold value, further fine-tuning takes place and the weights for the model are updated.

Figure 6 illustrates the IrisDenseNet architecture proposed by Arsalan et al (2018), which is distinct from other methods because it leverages the use of dense connectivity (Huang et al., 2017) in the encoder, which reduces the vanishing gradient problem, strengthens features flowing through the model, and also enables feature reuse, resulting in stronger learned features. The essential component of the encoder that enables dense connections is dense blocks. Figure 7 precisely describes the structure and connectivity of a dense block.

Figure 7
Dense connectivity within the dense blocks. From Arsalan et al. (2018)



As can be seen in Figure 6, the encoder consists of five dense blocks comprising convolution and concatenation layers, and each dense block is separated by a transition layer consisting of a 1 × 1 convolution and max pooling for downsampling. Further, for the upsampling purpose, it does not employ direct skip connections such as Ronneberger et al. (2015), but rather pooling indices are fed to the respective decoder block from the transition layer. IrisDenseNet uses the Seg-Net basic (Badrinarayanan et al., 2017) architecture for its decoder, where each decoder block receives pooling indices and dense features from the respective transition layer as input and outputs an upsampled feature map.

Employing the Res-Net 50 (He et al., 2016) contracting path and adding an attention gate (AG) module in the skip connections to identify important feature regions efficiently, Wei et al. (2022) propose a novel architecture. Each residual block in the encoder consists of a convolution block and an identity block, and the residual mapping is performed. Furthermore, the AG module (Schlemper et al., 2019) is implemented in the skip connections, which computes the attention coefficients for each pixel to target iris regions better.

The AG module at each layer takes input as feature maps from the contracting path "x" and a gating feature vector "g" from the respective expansive path (decoder layer), which helps suppress irrelevant noise and generates attention coefficients that are passed as the value of skip connections. Additionally, during backward propagation of the AG module, the background regions are down-weighted, improving the focus of the model on the iris regions.

The bilateral transformer proposed by Wei et al. (2021) is an encoder–decoder architecture with the integration of bilateral self-attention modules and a novel iris segmentation uncertainty learning module. The proposed architecture is a residual encoder–decoder architecture that additionally comprises a bottleneck (two BiTrans blocks). The encoder is stacked with alternating dense downsampling convolutions (for downsampling) and BiTrans blocks, and the decoder is symmetrically stacked with dense upsampling convolutions (for upsampling) and the cross-attention version of BiTrans blocks (which consider multi-scale features from skip connections).

Each BiTrans block consists of bilateral self-attention modules, mostly consisting of spatial and visual branches to extract spatial information and determine discriminative contextual information from visual characteristics, respectively, which are then aggregated. Further, linear projections are replaced by convolutional projections to improve spatial perception.

Furthermore, for two successive transformer blocks, they are incorporated as Swin transformers (Liu et al., 2021). The segmentation head is implemented at the last decoder layer, whereas the auxiliary head is embedded in a hidden layer. Since layers at different depths in the network learn different features, this difference between the two segmentation results determines the uncertainty map. Using this uncertainty map, a weighted scheme is implemented that assigns higher weights to pixels with high uncertainty, enabling the model to pay more attention to high-uncertainty regions. Further, a regularization term is incorporated, which reduces the uncertainty of the segmentation predictions from a global perspective.

2.2.3. Other developments

Bezerra et al. (2018) propose a method which leverages the use of generative adversarial networks (Goodfellow et al., 2020) for iris segmentation, which comprises two networks, namely a generator and discriminator. First of all, the generator receives noise as input and generates samples; these samples and training data are fed to the discriminator, which finds discrepancies between the two inputs. These discrepancies are fed to the loss function, which fine-tunes the generator, enabling the generator to produce more realistic samples and the discriminator to distinguish better with each iteration.

Zhao and Kumar (2017) proposed an architecture that comprises FeatNet (for extracting discriminative iris features) and MaskNet (to provide sufficient information for masking non-iris images) and proposes a novel extended triplet loss function (ETL) to guide the triplet network implemented for learning convolution kernels in FeatNet. This ETL incorporates the bit-shifting of feature maps and non-iris masking, which helps to learn discriminative spatial iris features more efficiently.

2.3. Pedestrian detection

2.3.1. Multi-scale approaches

A multi-scale approach to tackle the task of pedestrian detection can be observed in Figure 8, which leverages the region proposal network (RPN) of the FRCNN (Ren et al., 2015) for pedestrian detection and modifies it to a multi-scale version by having multiple RPN for different layers of the backbone network (VGG16 (Simonyan & Zisserman, 2014)), each concentrating on a different scale. In addition, it employs skip-pooling, where multiple layers are pooled for each ROI simultaneously and are subsequently L2-normalized, concatenated, and scaled to output a fixed-length vector. Furthermore, a bootstrapping sampling strategy is used for each proposal detector in the architecture.

Similar to the approach mentioned above, Cai et al. (2016) aim to perform detection at different layers to have multiple proposals at different scales. Different from the method illustrated above, these detected objects at multiple layers are considered final proposed objects and are not skip-pooled and concatenated. The object detection network comprises ROI pooling, deconvolution, which improves resolution and provides adequate data for ROI pooling, and the fully connected layer. Additionally, pedestrian object features and the context of multiple regions are stacked together post-ROI pooling.

2.3.2. Attention-based approaches

Figure 9 demonstrates an attention-based deep learning model (Lin et al., 2018) named graininess-aware deep feature learning (GDFL)-based detector that uses a pedestrian attention mechanism that generates attention masks for both small and large pedestrians. Input from different layers of the backbone network (Simonyan & Zisserman, 2014) having distinct scales is fed to the attention module, enabling focus on small pedestrians and large pedestrians.

Further, two (small and large) graininess-aware feature maps are computed by multiplying feature maps from the backbone with the respective attention masks. Additionally, to obtain more powerful feature maps having context information as well as local details for small objects, a zoom-in-zoom-out module is proposed. Using the small-pedestrian attention mask generated through layer “x” from the backbone network, the previous and next layers to “x” are encoded with the same attention mask and further concatenated, which explores both small-scale and large-scale information.

Different from GDFL mentioned above, the backbone network in Pang et al. (2019) is leveraged from Ren et al. (2015) by replacing the ROI pooling layer in GDFL with the ROI alignment layer. A novel mask-guided attention network (MGAN) is incorporated, which helps the model pay more attention to visible body features of the pedestrian and suppress the occluded parts in the full body features. The multichannel features from the ROI align layer are fed to the MGAN, which computes the modulated multichannel features through a series of convolutions, ReLu, and sigmoid functions (attention masks), which are further multiplied with the input, and visible-region bounding box information is used to guide the loss function in the MGAN.

Zhang et al. (2018) propose a novel network which has the same backbone network employed by the method mentioned above for detection but consists of a distinct attention mechanism. It is proposed to handle different occlusion patterns by performing channel-wise attention and consider three attention mechanisms: self-attention, visible-box attention, and part attention. In channel-wise attention, channels with occluded regions are assigned lower weights, and channels with visible regions are assigned higher weights, enabling the model to pay more attention to visible regions.

All kinds of attention networks are fed with the top-convolution features. Self-attention exploits the interdependencies between the channels and uses SE blocks (Iandola et al., 2016) to assign higher weights to more informative features. Visible-box attention additionally takes a full-body, visible-body bounding box and explores four kinds of occlusion patterns and weighs the channels accordingly. A fully convolutional part detection network pre-trained on Insafutdinov et al. (2016) gives information on 14 body parts, which is used additionally as an input to the part attention network to guide the attention mask.

2.3.3. Enhancing non-maximum suppression for pedestrian detection

As shown in Figure 10, non-maximum suppression (NMS) is a post-processing algorithm that filters out bounding box proposals based on a threshold value for classification. In Figure 10, (a)

Figure 8
Multi-scale region proposal network (RPN). Three convolution layers are connected to three RPNs. These distinct results are combined and provide regional proposals for the pedestrian detector

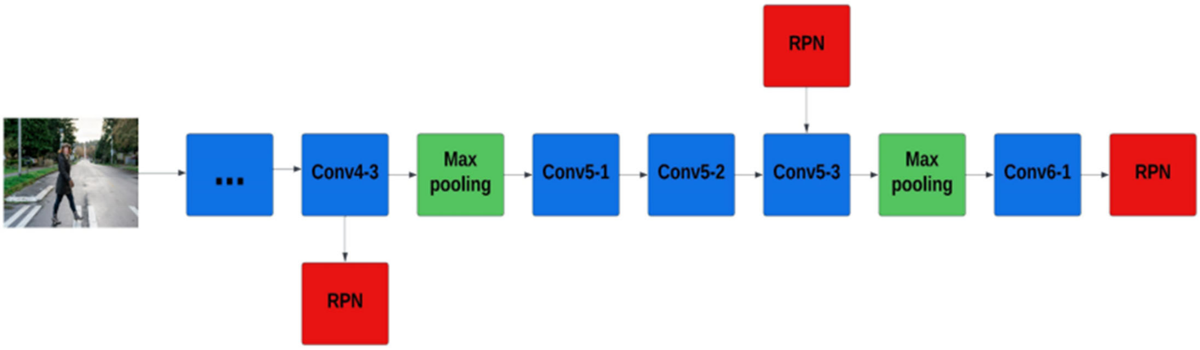


Figure 9
Multi-scale region proposal network (RPN). Three convolution layers are connected to three RPNs. These distinct results are combined and provide regional proposals for the pedestrian detector

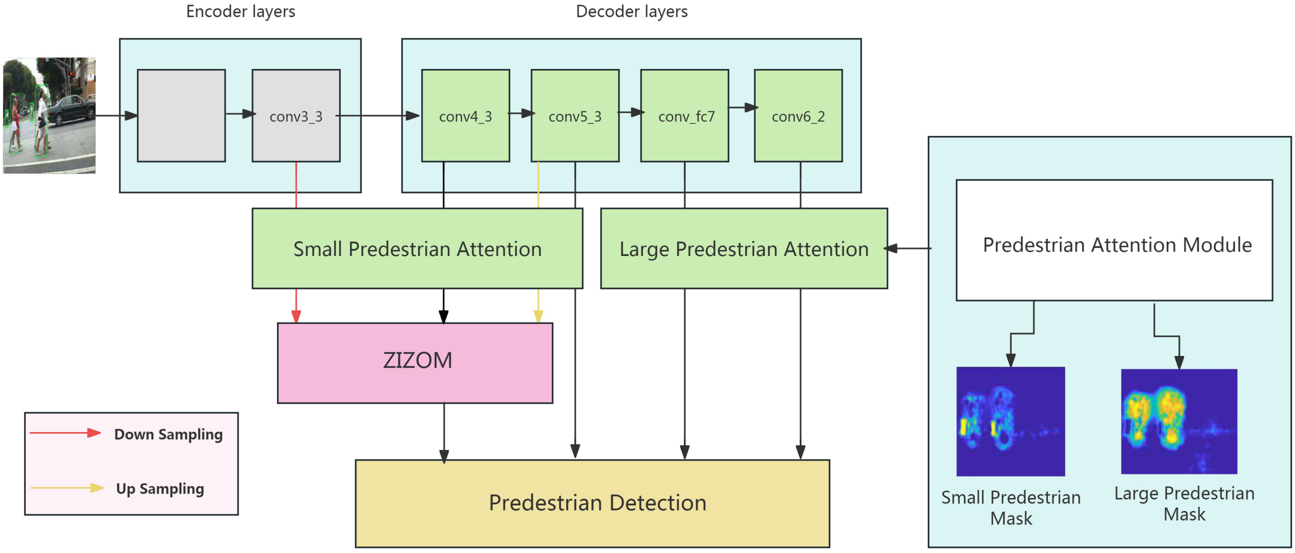
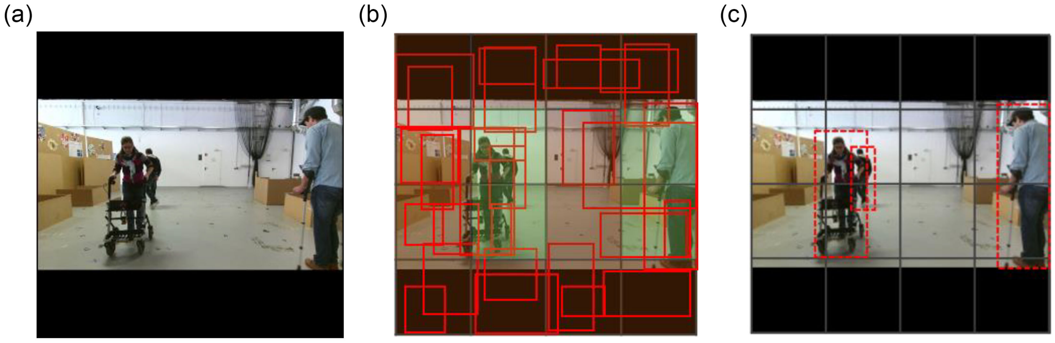


Figure 10
Working of NMS. From Dumitrescu et al. (2022)



demonstrates a padded image, (b) portrays all the bounding box proposals generated, and finally (c) depicts the results after filtering the bounding box proposals through NMS.

Liu et al. (2019) propose an adaptive NMS that increases the threshold value in a crowded scene to preserve overlapping neighboring objects and decreases the threshold value when the pedestrians are separated to disregard false positives. Intersection over union (IoU) is the measure of overlap between the target mask and the proposed bounding box, which is used as a metric to define the object density. The object density "d" of an object is defined as the maximum bounding box IoU with other objects in the ground truth set, which is computed for each proposal and reflects the intensity of crowd occlusion, and using this, the threshold "N" is updated if and only if this object density is greater than the fixed-threshold "F" ($N = \max(F, d)$). Furthermore, a three-convolution layer subnet is incorporated to determine the density of each proposal.

Unlike adaptive NMS illustrated above, Huang et al. (2020) do not change the threshold value according to the conditions; it rather uses a different method of IoU calculation where it computes the IoU between the visible regions of the full body boxes to estimate the overlapping of the two full body boxes. The lower value of IoU indicates the boxes belong to different pedestrians, and a similarly higher value of IoU indicates the same pedestrian; this enables the NMS to eliminate redundant boxes for the same pedestrian and disregard false positives. In addition, to obtain the visible part of a pedestrian, a paired box model is proposed, which simultaneously predicts the full and visible body boxes of the pedestrian.

2.3.4. Other developments

The architecture proposed by Mao et al. (2017) leverages certain channel features, viz. apparent-to-semantic channels, temporal channels, and depth channels, to learn the representations and improve the pedestrian detection performance. It consists of the BN (Simonyan & Zisserman, 2014), the channel feature network (CFN), the RPN, and the FRCNN (Girshick, 2015) for the final detection. All the activation maps from multiple layers in the BN are aggregated and fed to the CFN, which computes the channel feature map. Further, the representations learned by these channel maps help the FRCNN make accurate predictions.

Liu et al. (2019) propose a "center and scale prediction" (CSP) detector, which simplifies the pedestrian detection problem to a straightforward center and scales the prediction task of the pedestrian. The CSP consists of a feature extractor module that extracts aggregated multi-scale features from a fully convolutional backbone network. Further, the detection head computes the center heatmap and the scale map through a series of convolution layers.

To improve the efficiency of the model in Adarsh et al. (2020), Gong et al. (2020) replace the max pooling layer with 2-step convolutions, replace traditional convolutions with anti-residual blocks employing depth-wise separable convolution (Chollet, 2017), and finally they add an additional upsampling layer for a three-scale detection enabling efficient detection of small pedestrians.

The architecture proposed by Brazil et al. (2017) consists of an RPN (Ren et al., 2015) and a binary classification network to perform pedestrian detection on bounding box proposals generated from the RPN (refining the classification scores). A segmentation infusion network is attached to the deepest convolution layers of both sub-networks, and it illuminates the pedestrians in the shared feature maps preceding the classification layer by computing two masks reflecting the likelihood of residing in the pedestrian or background segment.

Zhou and Yuan (2018) leverages full-body and visible-body estimation to perform pedestrian detection. Two separate branches are incorporated, each taking input from the ROI pooling layer of the backbone network and respectively performing bounding box regression and classification (same as Ren et al. (2015)) for the full body and visible body parts of the pedestrian. These complementary classification scores are later fused together to improve detection robustness.

2.4. Autonomous driving

2.4.1. Adoption of asymmetric convolutions

Figure 11 illustrates the distinction between standard convolutions and asymmetric convolutions. It can be seen in the figure that instead of using a 3 × 3 convolution kernel, asymmetric convolution uses 3 × 1 and 1 × 3 convolution kernels, respectively. This adoption enables the usage of fewer parameters, which aids in reducing the computational complexity. Efficient dense modules with asymmetric convolution network (EDA-Net) (Lo et al., 2019) aims to achieve high performance segmentation to guide autonomous driving at a low computational cost.

Figure 12 gives an overview of the architecture and the connection flow in the network. EDA-Net achieves its aim by incorporating EDA blocks that comprise EDA modules. Figure 12 gives an overview of the architecture of EDA Net. It can be seen in the figure that each EDA module subsequently comprises a pointwise convolution layer (which reduces the number of input channels) and two pairs of asymmetric convolution layers (Romera et al., 2017; Szegedy et al., 2016) (which reduces computational complexity) employing dilated convolution (Zhao et al., 2017) at the second pair to obtain more contextual information.

However, some modules are fully dilated EDA modules, and the dilation rate gradually increases (the receptive field increases) throughout the modules in the network. Leveraging dense connectivity from Huang et al. (2017), each module is densely connected to each other in EDA-Net, which helps increase processing efficiency and combine multi-scale features as each EDA module has feature maps at a different receptive field.

Similar to EDA-Net mentioned above, Li et al. (2019) aim to obtain high efficiency at a low computational cost and employ asymmetric convolutions and dilated convolutions for the same. The architecture mainly comprises depthwise asymmetric bottleneck modules, which have a two-branch structure where the two branches are responsible for extracting local and contextual information, respectively, which are further concatenated.

The first branch consists of a 3 × 3 convolution layer followed by an asymmetric pair of convolution layers, whereas the second branch applies dilated convolution to depth-wise asymmetric convolution layers to extract better contextual information at a low computational cost. The outputs from these two branches are further fused and fed to a pointwise convolution layer, which is further concatenated with the input.

2.4.2. Encoder–decoder networks

Orsic et al. (2019) use Res-Net (He et al., 2016) as the encoder network and a symmetric decoder network, and it proposes two distinct ways to increase the receptive field, resulting in an

Figure 11
Working of asymmetric convolutions. (a) Demonstrates a simple 3 × 3 convolution operation and (b) demonstrates an asymmetric convolution operation. From Kim et al. (2019)

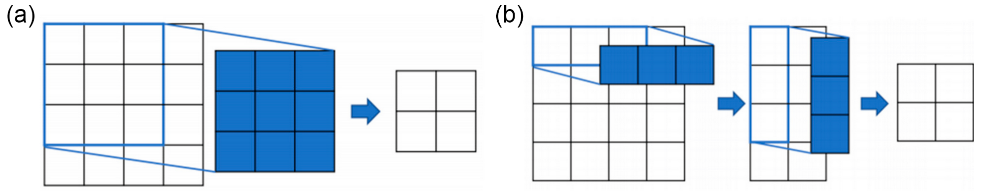


Figure 12
EDA-Net. From Papadeas et al. (2021)

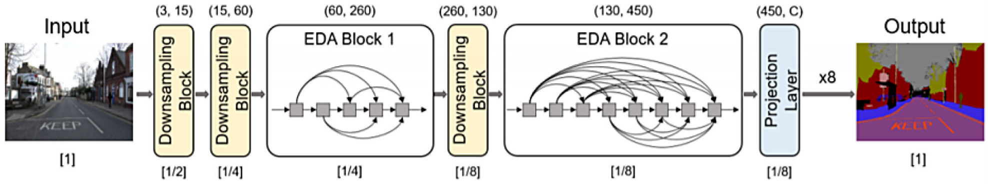
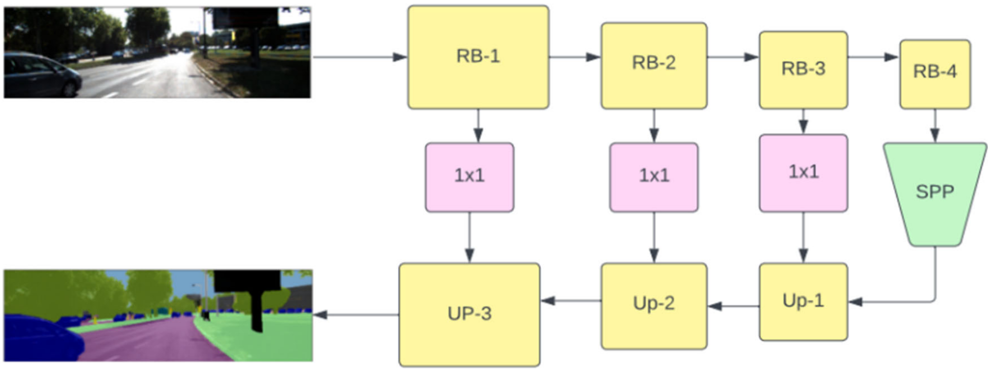


Figure 13
Single-scale model described by Li et al. (2019). “RB” refers to residual convolutional blocks; “Up” refers to upsampling units; “SPP” denotes the spatial pyramid pooling block



increase in model capacity. The first method is elaborated precisely in Figure 13, where a spatial pyramid pooling block (Zhao et al., 2017) is integrated after the last block of the encoder and combines features of the encoder at different pooling levels. The second method differs from the first in the following way: the pyramid fusion model consists of two encoder instances at different resolutions, and these feature maps are concatenated from each encoder at the same scale.

The model proposed in Chaurasia and Culurciello (2017) uses fewer parameters as compared to the approach mentioned above by linking the encoder–decoder connections in a novel way. Rather than passing only skip connections from the encoder to the respective decoder through a pointwise convolution, the proposed architecture bypasses the input to the encoder to the output of the respective decoder to recover spatial information at each level, which also enables the decoder to learn fewer parameters.

Different from the method illustrated above, Treml et al. (2016) use a modified Squeeze-Net 1.1 architecture (Iandola et al., 2016) as its encoder to reduce computation and maintain efficiency. Furthermore, the decoder employs parallel dilated convolutions (Chen et al., 2014), which uses four dilated convolutions at different dilation factors to combine one output of the encoder at four different resolutions, which increases the receptive field and regards multi-scale spatial information. Refinement modules similar to Pinheiro et al. (2016) are incorporated, which take the output of the respective encoder before pooling and the previous upsampler as input and learn to assign weights to these inputs before feeding them to the next upsampler.

Distinct from all three encoder–decoder networks explained above, Nirkin et al. (2021) employ a hypernetwork (Ha et al., 2016) kind of architecture for the encoder, which additionally generates weights for the decoder. In the hypernetwork, there is a backbone architecture (Tan & Le, 2019) where the head in the backbone is replaced by a context-head employing a nested U-Net architecture (Qin et al., 2020) that generates signals and helps calculate weights for the decoder. Each decoder layer consists of meta-blocks that employ the inverted residual architecture of Sandler et al. (2018) and use dynamic patch-wise convolutions; there is a weight mapping network attached to each block. This weight mapping network computes the weight just before usage (minimizing memory usage) through the input of the signal from the context head.

2.4.3. Instance-level segmentation

Initially, the CNN model (Schwing & Urtasun, 2015) pre-trained on ImageNet is used to compute a pixel-level map of instances at the patch level (Zhang et al., 2016). Furthermore, a densely connected pixel-wise Markov random field is incorporated, which combines all the predictions at the local level (patch level) and determines a globally consistent classification.

The module aims to minimize an energy function that consists of pairwise smoothness, local prediction, and an interconnected component term. The smoothness term aims to assign the same labels to pixels having similar features. The local prediction terms learn a potential through a series of Gaussian potentials to encode that pixel prediction at the local level, which should match the pixel predictions at the global level. Finally, the interconnected component term encodes the fact that one instance cannot belong to two distinct components.

De Brabandere et al (2017) propose a discriminative loss function that first calculates pixel embeddings for each pixel, where pixel embeddings of the same instance end up close together in the vector space. It mainly consists of two terms: variance, which pulls the embeddings at a certain distance apart closer to the mean embedding of their cluster, and distance, which pushes the cluster centers within a certain distance away from each other. As the loss converges, no embedding is closer to an embedding from a different cluster as compared to any embedding inside its cluster.

2.4.4. Other developments

Chen et al. (2017) propose a novel importance-aware loss function that assigns higher weights to important objects, forcing the model to pay more attention to these important objects and segment them with higher precision for safe driving. It operates in a hierarchical structure where each object is assigned to a group on each level, where each level indicates the importance, and accordingly, the weights of the objects at higher levels are multiplied by larger important factors, thereby contributing more to the loss function.

Hong et al. (2021) propose a novel network that mainly consists of a deep dual-resolution network that consists of two separate branches where each block is a residual basic block (He et al., 2016) extracting local information and global information, respectively, which goes under bilateral fusion (fusing local and global information) viz. fusing the high-resolution into the low resolution and vice-versa. Further, it consists of a deep aggregation pyramid pooling module to further extract global information from the last low-resolution feature map which fuses feature maps at different scales in a hierarchical-residual way (Gao et al., 2019).

3. Discussion and Analysis

In this survey, a variety of deep learning models to perform image segmentation on distinct applications were introduced and evaluated. Synapse multi-organ CT (Fu et al., 2020), Ubiris.v2 (Proença et al., 2009), Caltech (Dollar et al., 2011), and Camvid (Brostow et al., 2009) datasets were used to evaluate the models in the fields of medical imaging, iris recognition, pedestrian detection, and autonomous driving, respectively. Common metrics that are used to evaluate these models are stated as follows:

1. Medical imaging – Dice coefficient score (DSC) and Hausdorff distance (HD).
2. Iris recognition – Nice1 score and F score, these metrics are taken from the NICE1 contest (Proença & Alexandre, 2007).
3. Pedestrian detection – Average log miss rate on both reasonable occlusion (RO) and heavy occlusion (HO).
4. Autonomous driving – Mean intersection over union (mIoU).
5. Looking at the results at a high level, models employing attention mechanisms achieved the highest performance scores in all applications except “autonomous driving.” To be specific, as can be seen in Table 1, image transformer models were seen to be the most efficient for medical image segmentation applications. The inherent ability of attention mechanisms to capture intricate patterns and localize relevant features makes them well-suited for segmenting structures and abnormalities in medical images accurately.
6. For iris recognition, attention models and image transformers can effectively capture fine-grained details and local variations in the iris texture, leading to improved iris recognition accuracy. These models can focus on informative regions, such as the iris texture patterns, while disregarding irrelevant areas, enhancing the recognition performance.
7. In pedestrian detection, attention models and image transformers excel at capturing global and local spatial dependencies, enabling precise localization of pedestrians within complex scenes. By attending to discriminative regions, such as the human body parts, these models can effectively detect pedestrians with higher accuracy and robustness.
8. In autonomous driving, it was observed that improving pooling modules and convolution methods helped the model obtain a high-performance score with low parameters. Improving pooling modules and convolution methods in autonomous driving models resulted in higher performance scores with fewer parameters compared to self-attention models due to their computational and parameter efficiency. Pooling and convolution operations excel at extracting local features and capturing spatial relationships, which are crucial for autonomous driving tasks such as object detection and localization. Their ability to process information efficiently and capture relevant details contributed to the improved performance achieved in a resource-constrained environment.
9. While performance was high among the entirety of the models reviewed, some of them stood out with better performance in

Table 1
Illustration of high-performing models from each application. The signs (↑, ↓) beside each metric name indicate the relationship between the metric and the performance. “↑” sign indicates higher the metric score, better the performance. “↓” sign indicates lower the metric score, better the performance

Application	Model evaluated	Dataset	Performance score	
Medical imaging		Synapse multi-organ CT	DSC (↑)	HD (↓)
	U-Net (Ronneberger et al., 2015)		76.85	39.70
	TransUNet (Chen et al., 2021)		77.48	31.69
	SwinUNet (Cao et al., 2022)		79.13	21.55
Iris recognition		Ubiris.v2	Nice 1(%) (↓)	F(%) (↑)
	IrisParseNet (Wang et al., 2020)		0.84	91.78
	BiTrans (Wei et al., 2021)		0.85	92.11
	RAG-Net (Wei et al., 2022)		0.83	93.02
	FD-UNet (Zhang et al., 2019)		0.4	94.81
Pedestrian detection		Caltech	Miss rate (RO) (%) (↓)	Miss rate (HO) (%) (↓)
	FasterRCNN +visible-box attention (Zhang et al., 2018)		10.33	45.18
	MS-CNN (Cai et al., 2016)		10	49.1
	(Zhou and Yuan, 2018)		9.4	43.9
	MGAN (Pang et al., 2019)		6.83	38.16
	CSP (Liu et al., 2019)		3.8	36.5
Autonomous driving		Camvid	mIoU (%) (↑)	
	Pyramid model with Res-Net backbone (He et al., 2016)		65.70	
	EDANet (Lo et al., 2019)		66.4	
	DABNet (Li et al., 2019)		66.4	
	LinkNet (Chaurasia and Culurciello, 2017)		68.3	
	DDR-Net (Hong et al., 2021)		76.3	
	HyperSeg (Nirkin et al., 2021)		79.1	

Figure 14
Performance score comparison for models in medical imaging

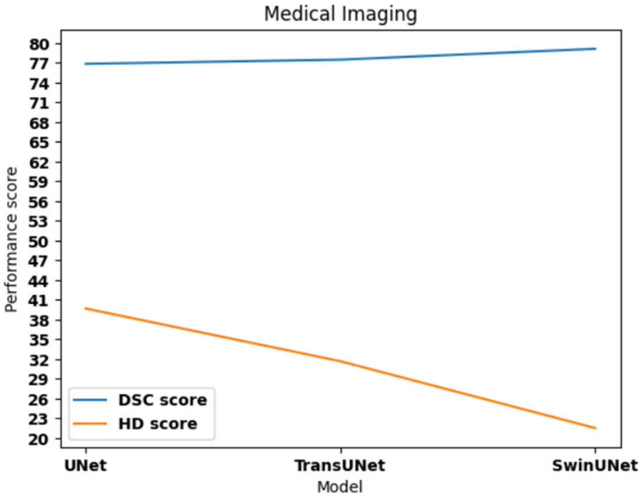
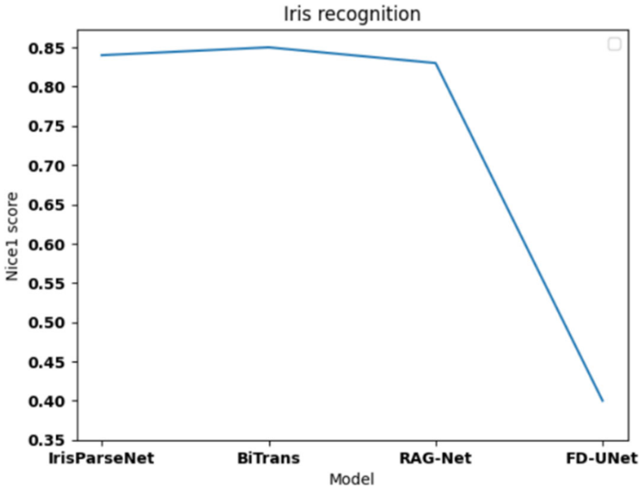


Figure 15
Performance score comparison for models in iris recognition



each domain. Table 1 illustrates the best-performing models for each application with its respective dataset and the common metric used to evaluate the accuracy of the model. These results display the effectiveness of deep learning models to tackle the complex task of image segmentation.

10. Prominent performance can be observed in Figures 14, 15, 16, and 17, where the models SwinUnet, FD-UNet, CSP, and HyperSeg excel in the domains of medical imaging, iris recognition, pedestrian detection, and autonomous driving, respectively. These models demonstrate superior capabilities in their respective applications, highlighting their effectiveness and suitability for addressing the specific challenges in each domain.

4. Limitations

Deep-learning-based algorithms remain effective for image segmentation purposes, but they do come with significant drawbacks. Some of the typical limitations are stated as follows:

Figure 16
Performance score comparison for models in pedestrian detection

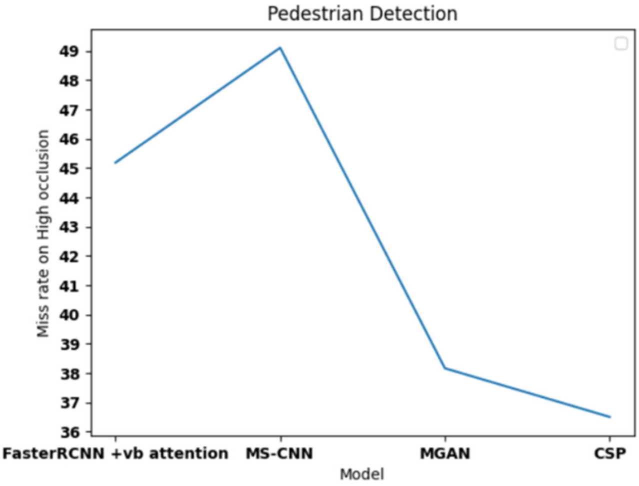
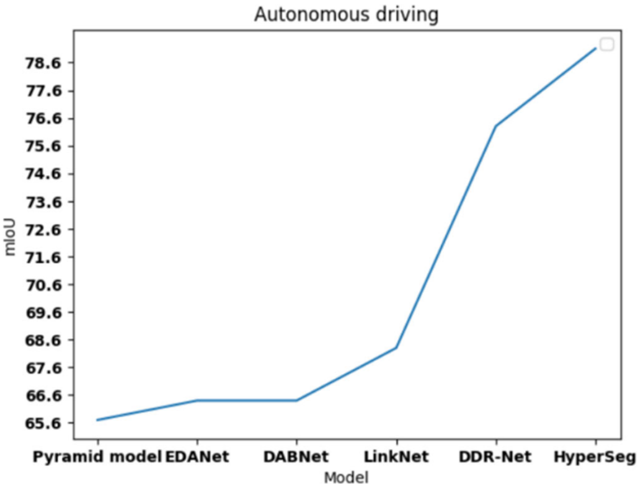


Figure 17
Performance score comparison for models in autonomous driving



- One of the primary impediments to deep learning's widespread deployment techniques in clinical practice is often the fluctuation in the data itself (resolution, contrast). These fluctuations in the data restrict the model's ability to be widely deployed, as inaccurate results might be predicted by the model when exposed to these altered feature elements.
- Deep learning models often have a weak generalization ability to other datasets or applications, which disallows the researchers to often apply a high-performing model of one scenario to other situations as the accuracy of the model will drop in varied situations.
- The huge parameter count and the computational complexity of the models make it difficult to integrate these models into real-time applications that are used on embedded devices such as mobile handheld devices. Mobile handheld devices are the most common source of usage by the common public, and integrating these applications into these devices is critical to aiding the public maximally.

- The paucity of annotated image datasets for training purposes. This scarcity restricts the volume of training data and hence degrades the learnability of the model. Therefore, researchers are forced to annotate the images manually, which is not a practical approach every time.
- Degradation in the performance of the model due to occlusion in images. Occlusion leads to a lesser extraction of relevant and important information from the image. This in turn leads to a decline in the model's performance.

In addition to the limitations mentioned above, the application of "iris recognition" faces further challenges in the segmentation of the iris region. Challenges such as poor illumination and reflection in the iris images are commonly observed in the data. Furthermore, the model finds it challenging to segment the iris in unconstrained scenes, particularly when using low-quality, low-resolution iris images.

5. Conclusion and Future Scope

This work presents a comprehensive overview of distinct deep learning techniques used to tackle the critical task of image segmentation in various applications. This collection of possible methods to solve the task of image segmentation aids researchers in finding and using the most suitable techniques required to ameliorate the performance of novel models discovered in the future. Furthermore, a table listing the most accurate and efficient state-of-the-art models for each application is provided to gain a deeper understanding of the various models and their performance. It was observed that with respect to all the applications studied, attention-based segmentation methods and image transformers had a superior performance than other methods. In addition, certain limitations in the domain of image segmentation have been analyzed and described, which cover the entirety of each application. As part of further exploration of the domain of image segmentation, promising techniques that can be adopted are listed as follows:

- Self-supervised learning: This technique was developed to learn representations from large datasets without the need for manually tagged data. This prevents any human intervention, which lowers the operational cost.
- Weakly supervised learning: This method is associated with learning strategies that are characterized by imprecise or coarse-grained labeling. These labels are often far less expensive to obtain than fine-grained labels for supervised algorithms.
- Pre-training of image transformer models using BERT (Devlin et al., 2018). This will aid the model in producing accurate segmentation maps.

Looking ahead, there are several potential recommendations and new directions to explore in the domain of image segmentation:

- Semi-supervised learning: Image segmentation models may perform better if strategies that incorporate labeled and unlabeled data are studied. To benefit from both labeled and unlabeled data, semi-supervised learning techniques such as consistency regularization and self-training might be investigated.
- Real-time segmentation: Developing lightweight models that can perform real-time segmentation on resource-constrained devices, such as mobile phones, is an important direction to explore. Techniques like network pruning and architecture design optimization can help achieve real-time and resource efficient segmentation

Image segmentation techniques hold vast potential beyond the domains discussed in this work. Promising domains for exploration include:

- Remote sensing: Applying segmentation in remote sensing can assist in land cover classification, change detection, and environmental monitoring, contributing to fields like climate studies and disaster management.
- Video surveillance: Image segmentation is used in video surveillance to detect and track objects of interest, analyze activities, and identify anomalies. Accurate segmentation helps enhance security systems and enable proactive monitoring.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

Adarsh, P., Rathi, P., & Kumar, M. (2020). YOLO v3-Tiny: Object Detection and Recognition using one stage improved model. In *2020 6th International Conference on Advanced Computing and Communication Systems*, 687–694.

Al-Amri, S. S., Kalyankar, N. V., & Khamitkar, S. D. (2010). Image segmentation by using edge detection. *International Journal on Computer Science and Engineering*, *2*(3), 804–807.

Al-Fahoum, A. S. (2003). Adaptive edge localisation approach for quantitative coronary analysis. *Medical and Biological Engineering and Computing*, *41*, 425–431.

Al-Fahoum, A. S., & Reza, A. M. (2004). Perceptually tuned JPEG coder for echocardiac image compression. *IEEE Transactions on Information Technology in Biomedicine*, *8*(3), 313–320.

Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., & Asari, V. K. (2018). Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. *arXiv preprint:1802.06955*.

Apostolopoulos, S., Zanet, S. D., Ciller, C., Wolf, S., & Sznitman, R. (2017). Pathological OCT retinal layer segmentation using branch residual u-shape networks. In *MICCAI 2017: 20th International Conference*, 294–301.

Arora, A., Jayal, A., Gupta, M., Mittal, P., & Satapathy, S. C. (2021). Brain tumor segmentation of MRI images using processed image driven u-net architecture. *Computers*, *10*(11), 139.

Arsalan, M., Naqvi, R. A., Kim, D. S., Nguyen, P. H., Owais, M., & Park, K. R. (2018). IrisDenseNet: Robust iris segmentation using densely connected fully convolutional networks in the images by visible light and near-infrared light camera sensors. *Sensors, 18*(5).

Al-Fahoum, A., Jaber, E. B., & Al-Jarrah, M. A. (2014). Automated detection of lung cancer using statistical and morphological image processing techniques. *Journal of Biomedical Graphics and Computing, 4*(2), 33–42.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(12), 2481–2495.

Bezerra, C. S., Laroca, R., Lucio, D. R., Severo, E., Oliveira, L. F., Britto, A. S., & Menotti, D. (2018). Robust iris segmentation based on fully convolutional networks and generative adversarial networks. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images*, 281–288.

Brazil, G., Yin, X., & Liu, X. (2017). Illuminating pedestrians via simultaneous detection & segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 4950–4959.

Brostow, G. J., Fauqueur, J., & Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters, 30*(2), 88–97.

Cai, Z., Fan, Q., Feris, R. S., & Vasconcelos, N. (2016). A unified multi-scale deep convolutional neural network for fast object detection. In *Computer Vision–ECCV 2016: 14th European Conference*, 354–370.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, 205–218.

Chaurasia, A., & Culurciello, E. (2017). Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing*, 1–4.

Chen, B. K., Gong, C., & Yang, J. (2017). Importance-aware semantic segmentation for autonomous driving system. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1504–1510.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., . . . , & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint:2102.04306*.

Chen, J., Nakashika, T., Takiguchi, T., & Ariki, Y. (2015). Content-based image retrieval using rotation-invariant histograms of oriented gradients. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 443–446.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv preprint:1412.7062*.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint:1706.05587*.

Cherabit, N., Chelali, F. Z., & Djeradi, A. (2012). Circular Hough transform for iris localization. *Science and Technology, 2*(5), 114–121.

Chernov, N., & Lesort, C. (2005). Least squares fitting of circles. *Journal of Mathematical Imaging and Vision, 23*, 239–252.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258.

De Brabandere, B., Neven, D., & Van Gool, L. (2017). Semantic instance segmentation for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 7–9.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Feifei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint:1810.04805*.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761.

Dumitrescu, F., Boiangiu, C. A., & Voncilă, M. L. (2022). Fast and robust people detection in RGB images. *Applied Sciences*, *12*(3), 1225.

Fu, S., Lu, Y., Wang, Y., Zhou, Y., Shen, W., Fishman, E., & Yuille, A. (2020). Domain adaptive relational reasoning for 3d multi-organ segmentation. In *23rd International Conference on Medical Image Computing and Computer-Assisted Intervention*, 656–666.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint:1602.07360*.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059.

Gao, S. H., Cheng, M. M., Zhao, K., Zhang, X. Y., Yang, M. H., & Torr, P. (2019). Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(2), 652–662.

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

Gong, X., Ma, L., & Ouyang, H. (2020). An improved method of Tiny YOLOV3. *IOP Conference Series: Earth and Environmental Science*, *440*(5).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . , & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., . . . , & Liu, J. (2019). CE-Net: Context encoder network for 2D medical image segmentation. *IEEE Transactions on Medical Imaging, 38*(10), 2281–2292.

Ha, D., Dai, A., & Le, Q. V. (2016). Hypernetworks. *arXiv preprint:1609.09106*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hong, Y., Pan, H., Sun, W., & Jia, Y. (2021). Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv preprint:2101.06085*.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., . . . , & Wu, J. (2020). Unet 3+: A full-scale connected Unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1055–1059.

Huang, X., Ge, Z., Jie, Z., & Yoshie, O. (2020). NMS by representative region: Towards crowded pedestrian detection by proposal pairing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10750–10759.

Huang, Z. C., Chan, P. P., Ng, W. W., & Yeung, D. S. (2010). Content-based image retrieval using color moment and Gabor texture feature. In *2010 International Conference on Machine Learning and Cybernetics*, *2*, 719–724.

Idrissa, M., & Acheroy, M. (2002). Texture classification using Gabor filters. *Pattern Recognition Letters*, *23*(9), 1095–1102.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., & Schiele, B. (2016). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Computer Vision–ECCV 2016: 14th European Conference*, 34–50.

Jalilian, E., & Uhl, A. (2017). Iris segmentation using fully convolutional encoder–decoder networks. In B. Bhanu & A. Kumar (Eds.), *Deep learning for biometrics* (pp. 133–155). Cham: Springer.

Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., & Johansen, H. D. (2020). Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd*

International Symposium on Computer-Based Medical Systems, 558–564.

Kaul, C., Manandhar, S., & Pears, N. (2019). Focusnet: An attention-based fully convolutional network for medical image segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging*, 455–458.

Kim, S., Nam, J., & Ko, B. (2019). Fast depth estimation in a single image using lightweight efficient neural network. *Sensors, 19*(20).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Lalaoui, L., & Mohamadi, T. (2013). A comparative study of image region-based segmentation algorithms. *International Journal of Advanced Computer Science and Applications*, *4*(6), 198–206.

Li, G., Yun, I., Kim, J., & Kim, J. (2019). Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. *arXiv preprint:1907.11357*.

Li, Y. H., Huang, P. J., & Juan, Y. (2019). An efficient and robust iris segmentation algorithm using deep learning. *Mobile Information Systems*, *2019*.

Lian, S., Luo, Z., Zhong, Z., Lin, X., Su, S., & Li, S. (2018). Attention guided U-Net for accurate iris segmentation. *Journal of Visual Communication and Image Representation*, *56*, 296–304.

Liang, M., & Hu, X. (2015). Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3367–3375.

Lin, C., Lu, J., Wang, G., & Zhou, J. (2018). Graininess-aware deep feature learning for pedestrian detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 732–747.

Liu, S., Huang, D., & Wang, Y. (2019). Adaptive NMS: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition*, 6459–6468.

Liu, W., Liao, S., Ren, W., Hu, W., & Yu, Y. (2019). High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5187–5196.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., . . . , & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Lo, S. Y., Hang, H. M., Chan, S. W., & Lin, J. J. (2019). Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In *Proceedings of the ACM Multimedia Asia*, 1–6.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*, 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*(7), 674–693.

Mao, J., Xiao, T., Jiang, Y., & Cao, Z. (2017). What can help pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3127–3136.

Nirkin, Y., Wolf, L., & Hassner, T. (2021). Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4061–4070.

Orsic, M., Kreso, I., Bevandic, P., & Segvic, S. (2019). In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12607–12616.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66.

Pan, S., Sun, S., Yang, L., Duan, F., & Guan, A. (2015). Content retrieval algorithm based on improved HOG. In *2015 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence*, 438–441.

Pang, Y., Xie, J., Khan, M. H., Anwer, R. M., Khan, F. S., & Shao, L. (2019). Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4967–4975.

Papadeas, I., Tsochatzidis, L., Amanatiadis, A., & Pratikakis, I. (2021). Real-time semantic image segmentation with deep learning for autonomous driving: A survey. *Applied Sciences*, *11*(19).

Park, J., Woo, S., Lee, J. Y., & Kweon, I. S. (2018). Bam: Bottleneck attention module. *arXiv preprint:1807.06514*.

Pinheiro, P. O., Lin, T. Y., Collobert, R., & Dollár, P. (2016). Learning to refine object segments. In *Computer Vision–ECCV 2016: 14th European Conference*, 75–91.

Proença, H., & Alexandre, L. A. (2007). The NICE. I: Noisy iris challenge evaluation-Part I. In *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 1–4.

Proença, H., Filipe, S., Santos, R., Oliveira, J., & Alexandre, L. A. (2009). The UBIRIS. v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(8), 1529–1535.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, *106*.

Ragab, D. A., Sharkas, M., Marshall, S., & Ren, J. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ, 7*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 28*.

Romera, E., Alvarez, J. M., Bergasa, L. M., & Arroyo, R. (2017). Efficient convnet for real-time semantic segmentation. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, 1789–1794.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, 234–241.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Sardar, M., Banerjee, S., & Mitra, S. (2020). Iris segmentation using interactive deep learning. *IEEE Access*, *8*, 219322–219330.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, *53*, 197–207.

Schwing, A. G., & Urtasun, R. (2015). Fully connected deep structured networks. *arXiv preprint:1503.02351*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint:1409.1556*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-V4, inception-ResNet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, *31*(1).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 32–42.

Treml, M., Arjona-Medina, J., Unterthiner, T., Durgesh, R., Friedmann, F., Schuberth, P., . . . , & Hochreiter, S. (2016). Speeding up semantic segmentation for autonomous driving. In *29th Conference on Neural Information Processing Systems (NIPS 2016)*.

Wang, C., Muhammad, J., Wang, Y., He, Z., & Sun, Z. (2020). Towards complete and accurate iris segmentation using deep multi-task attention network for non-cooperative iris recognition. *IEEE Transactions on Information Forensics and Security*, *15*, 2944–2959.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, *2*, 1398–1402.

Wei, J., Huang, H., Sun, M., Wang, Y., Ren, M., He, R., & Sun, Z. (2021). Toward accurate and reliable iris segmentation using uncertainty learning. *arXiv preprint:2110.10334*.

Wei, Y., Zeng, A., Zhang, X., & Huang, H. (2022). RAG-Net: ResNet-50 attention gate network for accurate iris segmentation. *IET Image Processing*, *16*(11), 3057–3066.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint:1511.07122*.

Zhang, S., Yang, J., & Schiele, B. (2018). Occluded pedestrian detection through guided attention in CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6995–7003.

Zhang, W., Lu, X., Gu, Y., Liu, Y., Meng, X., & Li, J. (2019). A robust iris segmentation scheme based on improved U-net. *IEEE Access*, *7*, 85082–85089.

Zhang, Z., Fidler, S., & Urtasun, R. (2016). Instance-level segmentation for autonomous driving with deep densely connected MRFs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 669–677.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2881–2890.

Zhao, L., Zhou, D., Jin, X., & Zhu, W. (2022). nn-TransUNet: An automatic deep learning pipeline for heart MRI segmentation. *Life*, *12*(10), 1570.

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., & Fan, Y. (2018). A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *Medical Image Analysis, 43*, 98–111.

Zhao, Z., & Kumar, A. (2017). Towards more accurate iris recognition using deeply learned spatially corresponding features. In *Proceedings of the IEEE International Conference on Computer Vision*, 3809–3818.

Zhou, C., & Yuan, J. (2018). Bi-box regression for pedestrian detection and occlusion estimation. In *Proceedings of the European Conference on Computer Vision*, 135–151.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 3–11.

How to Cite: Purohit, J. & Dave, R. (2023). Leveraging Deep Learning Techniques to Obtain Efficacious Segmentation Results. *Archives of Advanced Engineering Science* 1(1), 11–26, https://doi.org/10.47852/bonviewAAES32021220