Fully Convolutional Network with Task Partitioning for Inshore Ship Detection in Optical Remote Sensing Images

Haoning Lin, Zhenwei Shi, *Member, IEEE*, Zhengxia Zou

***Abstract*—Ship detection in optical remote sensing imagery has drawn much attention in recent years, especially with regards to the more challenging inshore ship detection. However, recent work on this subject relies heavily on hand-crafted features that require carefully tuned parameters and on complicated procedures. In this paper, we utilize a fully convolutional network (FCN) to tackle the problem of inshore ship detection and design a ship detection framework that possesses a more simplified procedure and a more robust performance. When tackling the ship detection problem with FCN, there are two major difficulties: 1) the long and thin shape of the ships and their arbitrary direction makes the objects extremely anisotropic and hard to be captured by network features and 2) ships can be closely docked side by side, which makes separating them difficult. Therefore, we implement a task partitioning model in the network, where layers at different depths are assigned different tasks. The deep layer in the network provides detection functionality and the shallow layer supplements with accurate localization. This approach mitigates FCN's trade-off between localization accuracy and feature representative ability, which is of importance in the detection of closely docked ships. The experiments demonstrate that this framework, with the advantages of FCN and the task partitioning model, provides robust and reliable inshore ship detection in complex contexts.**

***Index Terms*—inshore, ship detection, fully convolutional network, optical remote sensing.**

I. Introduction

THE ship detection in remote sensing imagery has been under extensive investigation over the last decades, both in synthetic aperture radar (SAR) imagery and in optical imagery. Recently, ship detection in optical imagery is under more active research because of its high resolution and human eye friendly color presentation.

A considerate amount of research in optical imagery focuses on the detection of different types of objects, such as roads [1], buildings [2], oil tanks [3], vehicles [4] and ships [5], [6]. Aside from detecting scattered objects, the classification of scenes also received a lot of attention recently, such as in [7], where the objective is to classify image patches into different categories, such as buildings, forests, harbors, etc.

Inshore ship detection presents more challenges seeing that the targets have extremely long and thin shapes and are subject to rotation. They are also surrounded by complex contexts such as nearby ships and docks. Consequently, recent research on inshore ship detection has been based on the detection of ship foredecks first and has been focusing on carefully hand-crafted features. Methods in [8] and [9] both use Harris corner detection for ship foredeck detection while [10] relies on line segment detection for preliminary proposal selection. These approaches also rely on procedures such as edge extraction and image binarization that require carefully tuned parameters and do not generalize well when the targets are not presented in ideal quality.

Recently, convolutional neural networks (CNN) have been utilized in a large number of applications on remote sensing images [11]. Fully Convolutional Networks (FCN) is a special kind of CNN that is used to label remote sensing images pixel by pixel [12], [13]. CNNs and FCNs do not rely on hand-crafted features and are able to automatically learn features from labeled data and thus are easy to implement. They also show great generalization ability in wide range of applications. Therefore, we are motivated to replace hand-crafted features with a robust FCN framework.

Neural networks occupy a dominant position in detection and classification in everyday images. However, its trade-off between localization accuracy and representative ability is a rarely mentioned limitation. The size difference between targeted objects in everyday images and remote sensing images is also seldom mentioned. Objects in remote sensing images can be small and closely arrayed, making the accurate localization of targets more important. Recent study shows the increase in the depth of layers increases the representation ability [14]. However, the incurred increased scale of down-sampling also decreases localization accuracy and can inhibit its use in remote sensing images.

In this paper, we partition the detection task among the network layers at different depths to combine the advantages of shallow and deep networks, featuring both localization accuracy and representative ability. In our proposed method, the deep part of the network is used to provide a coarse-positioned detection and also a confined, more managed and thus simpler problem space for the shallow part, which then is able to supplement with accurate localization. The task partitioning model presents similarities to the attention model

This work was supported in part by the National Natural Science Foundation of China under Grant 61671037, in part by the Beijing Natural Science Foundation under Grant 4152031, and in part by the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, under Grant BUAA-VR-16ZZ-03. *(Corresponding author: Zhenwei Shi)*

The authors are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: harrylin@buaa.edu.cn; shizhenwei@buaa.edu.cn; zhengxiazou@buaa.edu.cn).

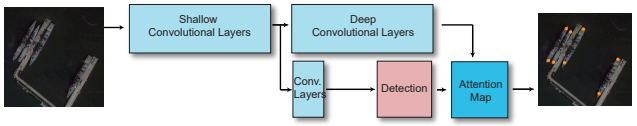

Fig. 1. The structure of the proposed network. All the convolutional layers are marked by light blue blocks and the size of blocks approximately indicates the number of included layers. The pink block denotes the foredeck/stern detection result produced by the shallow path, which has high localization accuracy. The dark blue block denotes the attention map produced by the deep path which has high recall/accuracy performance.

[15] used in neural network community. Nevertheless, the implementation details and the concepts of these two models are quite different. We will provide a brief introduction to the attention model and a comparison for clarity. We also use similar terms, such as attention map, in the following sections. Moreover, the idea of attention map is also similar to saliency detection [16]–[18] and we refer the interested readers to related literature.

The main contributions of our work are as follows,

- Focusing on the problem of inshore ship detection, we replace hand-crafted features with those learned by FCN, which allows unified optimization rather than individually tuned parameters and constitutes a more robust and scalable framework.
- With the task partitioning model, the tasks of localization and detection are partitioned onto different layers of the network, thereby mitigating the localization accuracy/detection ability trade-off common in FCNs and is of vital importance in ship detection tasks in remote sensing imagery.

In Section II we give a brief introduction to FCN and the attention model. Section III describes the proposed FCN with task partitioning. Section IV demonstrates the performance of our framework and Section V concludes the letter.

II. Related Work

CNN is extremely effective in image related tasks, such as object detection and classification [19], [20]. A CNN consists of a cascade of convolutional layers which convolute their inputs with kernels and pass the outputs along in order. The layers are also occasionally interleaved with activation functions, such as ReLU [21], to enhance the network's ability to represent non-linear features, and pooling layers, to reduce computation complexity and improve robustness and generalization ability. Typically, one or two fully connected layers are located at the end to produce a scalar label as the output of the network.

The convolutional layers keep their calculated outputs in accord with the inputs spatially, whereas the fully connected layers produces feature vectors or scalar labels that have no direct spatial information. FCNs are a type of CNN designed to predict a label map rather than a scalar label for an input image, by replacing fully connected layers in CNN with small sized convolutional layers and are often used in pixel-labeling task [22].

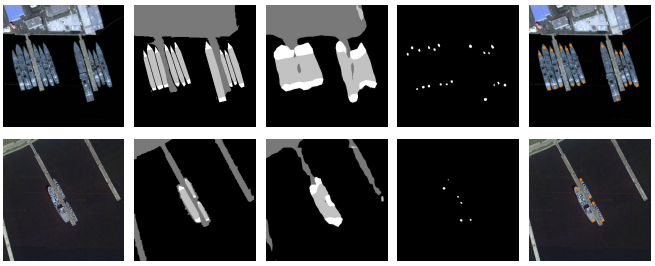

Fig. 2. The images (a) and labels (b) used to train the network. Results from deep and shallow paths are shown in (c,d), respectively. Here different gray levels indicate different categories (sea, land, ship body and foredeck/stern). The first 3 classes are regarded as non-attention and the foredeck/stern (white) is attention in the attention map. Note that the results (d) are the ones that have gone through the attention map so the redundant detection results outside attention are not shown. The final results are shown as composite images (e) for clarity.

The convolutional networks are trained through stochastic gradient descent, which includes iterative forward passes that takes labeled data as input, and backward propagations that update the parameters (i.e. weights) to minimize the difference between the actual and desired outputs, otherwise known as the loss of the network.

The attention mechanism is a design pattern that was recently widely used in neural network community. It first achieved successful implementations in language understanding such as machine translation [23] and later in visual tasks such as [15]. These approaches train the networks to generate both feature maps that encode the information of the input, and the attention maps that reveal regions of the feature maps where the following parts of the network should focus on. In our approach, however, we follow another line of thought, putting back-propagating loss, rather than feature maps, through the attention maps.

III. Proposed Method

A. Motivation

Unlike everyday images where the coverage of an image is always limited and the objects of interests are usually near the center of an image, remote sensing images, if not manually selected and cropped, usually have no specific region of interest. Consequently, to understand a remote sensing image, it is reasonable for machine learning to provide a label for each pixel of the image, rather than a single label for the entire image. In this paper, we use FCN to label every pixel in the remote sensing images.

Recent CNN backbones such as Resnet and VGG are effective in visual tasks such as everyday object classification and segmentation. These networks all have a large number of layers and a high amount of down-sampling with pooling layers (usually 8x, 16x) because the targeted objects often have a radius of over hundreds of pixels. In our dataset, however, with image resolution at 1 meter/pixel, the distance between the center of two ships that are docked side by side is less than 20 pixels, and the down-sampling will merge these two ships into neighboring pixels as can be seen in Fig. 2c. Because

a ship also can have the length at over 200 pixels and can be positioned at any orientation in the image, a shallow and simple network will not be able to effectively detect even small parts such as foredecks or sterns.

We introduce the idea of task partitioning into this detection framework, which mitigates the need for CNN's representative ability by confining the problem space with the deep layers, and acquires accurate localization with the shallow layers.

B. Proposed Network

The structure of our proposed network can be seen in Fig. 1. We base our network on Resnet-50 [14] and modify it into a fully convolutional network by replacing the last fully connected layers with convolution layers (Res-FCN). We nominally split the original network into 2 parts, the shallow layers and the deep layers. We add two convolutional layers after the shallow layers and lead the network into a separate shallow path and the deep layers establish the deep path after the shallow layers. Here the deep path is designed to produce the aforementioned attention maps, and the shallow path is designed to produce the detection results with accurate localization. The combination of attention maps and the detection results from the shallow path allows the network to produce results that have high recall/accuracy and high localization accuracy.

We formulate the problem as the detection of key points of the ships, i.e. foredecks and sterns. The FCN is widely used as segmentation frameworks, but the impact will be minimal if we train FCN to segment the areas that are regarded as the center of the target objects. The segmentation areas can then be clustered into key points in the same ways in detection frameworks, such as non-maximum suppression (NMS). The figures in this paper are all presented as in segmentation problems to conform to the nature of the FCN.

We label the training data with 4 labels, foredeck/stern, ship body, sea and land (see Fig. 2). By providing sea and land labels, we hope to prevent the network from degenerating and thus to enhance the generalization ability of the network, considering that most of the areas in a remote sensing image do not contain ships and yet we keep the network "active" in those areas nonetheless. The deep path, which produces the attention maps, is trained with these labels to discriminate these classes and provide coarse locations and the shallow path, is trained to provide accurate locations of the foredecks/sterns. With the attention map from the deep path, the shallow path only has to locate the accurate positions of the foredecks/sterns within the coarse foredeck/stern candidates.

We additionally include a variation in the training labels for the shallow/deep path to make them more suitable in their tasks. The foredeck/stern areas in the labels given to the deep path are morphologically dilated to increase recall rate of the detection. The opposite is done for shallow path to increase accuracy in localization.

The experiments show that with the simplified problem space, the shallow path, despite its limited depth, is able to locate the foredeck/stern accurately with high recall/precision. With the foredeck/stern candidates, the task of locating the whole ships becomes trivial. We extract rectangle patches that have one foredeck/stern on each end and use a simple convolutional network to valid the candidates.

C. Conventional Attention Model

The conventional attention model exerts focus, i.e. weights on features and thus relieves the following network layers of features that are regarded useless. Here we denote the input features from previous part of the network by $f \in \mathbb{R}^{C\times H\times W}$. First, a summarized feature map is calculated by the attention model as follows,

$$s = \mathrm{g}(\mathbf{W} * f + b) \tag{1}$$

where $*$ denotes convolution operation, W denotes the convolution filter and g the non-linear activation function. This can be viewed the same as a conventional layer, with the exception that $s \in \mathbb{R}^{C\times H\times W}, C = 1$. Next, a spatial softmax operation is applied on s, which can be regarded as a normalization,

$$\sigma(l) = \frac{e^{s(l)}}{\sum_{l'\in L} e^{s(l')}} \tag{2}$$

where l is the spatial index of the features which includes 2 dimensions (x, y) corresponding to the size of s, H and W, respectively. L marks the neighboring locations of l. Here the calculated σ is the attention map and is then applied on features via element-wise production across channels. Features with large corresponding values in the attention map are reserved and those with small ones can be viewed as discarded.

D. Proposed Task Partitioning Model

The task partitioning model in our approach is implemented from another perspective. The network is separated into 2 paths, shallow and deep path. The deep path has more discriminating ability and produces the attention map and is trained explicitly. The attention map is applied on the shallow path and affects the shallow path losses that are taken into consideration, rather than the feature. In the training stage, the deep path helps to provide a simpler problem space for the shallow path, i.e. it uses its attention map to decide which easy part of the problem the shallow path needs to solve and which it does not, allowing the shallow path to focus on finding the accurate location of the targets. In the inference stage, the attention map decides which of the detections from shallow path are valid.

The process of producing and applying the attention map can be formulated as follows,

$$s = \mathrm{g}(\mathbf{W} * f + b) \qquad , s \in \mathbb{R}^{2\times H\times W} \tag{3}$$

$$\sigma_{(l)} = \mathbf{1}(s_{(0,l)} < s_{(1,l)}) \qquad , \sigma \in \mathbb{R}^{1\times H\times W} \tag{4}$$

$$c' = c \odot \sigma \tag{5}$$

where $\mathbf{1}(\bullet)$ denotes the element-wise function that outputs 1 when its corresponding input element is true and otherwise 0. $\odot$ is the element-wise multiplication operation and c is the loss that goes through the attention map. The deep path is trained discriminately and the training follows the conventional object classification scheme in neural network study. Because here,

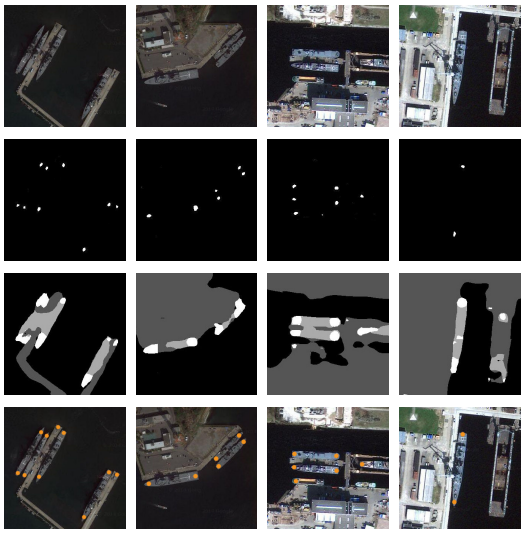

Fig. 3. Outputs of proposed network. Original images (Top), detection results (Middle-Top), attention maps (Middle-Bottom) and Composite results (Bottom) for clarity. Here in the attention maps we show all the classification results (4 classes) for completeness and when acting as attention maps, only the white areas are attention areas.

for simplicity, the classification problem is reduced to two categories (attention and no attention), we simplify the equation into comparing the 1st dimension of s on every spatial location l (intuitively, the two elements on every l denote the scores for non-attention, attention, respectively). Different from the aforementioned conventional attention model, which produces soft scores, our model produces binary attention maps.

In our experiment, we train the deep path to classify pixels into 4 classes (sea, land, ship body, foredeck/stern) and when used to produce attention maps, the 1st three classes are regarded as non-attention.

IV. Experiments

A. Experiment Dataset

We experiment our proposed method on an optical remote sensing dataset collected from Google Earth and GaoFen-2 satellite. This dataset contains 24 images each with above 5000×5000 pixels and with a resolution of 1 meter/pixel. These images feature both large harbor areas and rich land objects, which suits the need to fully test our framework. We set our method to only detecting battleships longer than 100 meters in order to limit the size range of the targets, because we surmise that a multi-scale framework and a higher resolution dataset would be needed for smaller-sized targets. We select 14 images as our training set to train our Res-FCN and the rest as the test set. Due to the limitation of GPU memory, the images are cut into 321×321 patches for training. Moreover, we rotate the patches that include ships to augment the training set in orientations and lay more attention on ship objects.

TABLE I
The precision and recall rate of the methods

Method	tp	fp	fn	Precision	Recall
Method in [8]	7680	480	1120	94.1 %	87.3%
Proposed method	8060	340	740	96.1 %	91.7%

B. Qualitative and Quantitative Performance

We test our model on the test set and select the areas that feature dense distribution of ships to showcase the performance of the method in Fig. 3, which demonstrates the network produces accurate locations for the foredeck/stern candidates. Although the results are initially produced by the shallow network and may produce a large number of redundant candidates, the attention map from the deep network is able to filter out these false targets. With the non-maximum suppression method, the exact location of each foredeck/stern is able to be acquired and enables further validation.

We use the precision and recall rate to quantitatively evaluate our complete ship detection framework on the test set. These are calculated with

$$\text{Precision} = \frac{tp}{tp + fp}$$
$$\text{Recall} = \frac{tp}{tp + fn}$$

where tp, fp, fn represent the number of true positives, false positives and false negatives in the detection results, respectively. To test our method comprehensively, we augment our test dataset via flipping, rotation and slight contrast modification and acquire a test dataset with 8800 targets in total. We reproduce the method in [8] as a baseline (Table I), of which the parameters are hand-selected in regard to the training dataset. The method in [8] has a few limitations despite its good performance under ideal circumstances. The method includes sea/land segmentation by minimizing an energy function with the split Bregman method, ship foredeck detection with Harris corner detection and validation with the shape of the foredeck, ship width and length. The detection and validation procedures both rely on the sea/land segmentation result and requires hand-tuned parameters, which compromise the robustness of the method. Our method outperforms the baseline in that its procedure is more concise, requires no empirical selected parameters and is robust in complex context.

C. Discussion on the Task Partitioning Model

To demonstrate the ability and necessity of our task partitioning model, we present an experiment with a multi-scale network. The widely used multi-scale structure concatenates the features from layers of different depths to compensate the down-sampling in the deep layers and to utilize the details from shallow layers. The network is able to outline the fine borders of the ships, however, some of the detections of neighboring foredeck/stern occasionally join each other (see Fig. 4) making the separation difficult. Only the shallow network is used to obtain the results in Fig. 4. Even though we accomplish high localization accuracy, the limited representative ability of the shallow network causes it to produce many false positives.

Fig. 4. Top: the detection results with the multi-scale network. Bottom: the detection results with the shallow network. The green areas are those detected as positive. Notice the positive pixels from different ships can join together and makes separation difficult.

TABLE II
THE COMPARISON BETWEEN DIFFERENT IMPLEMENTATIONS OF THE NETWORK

	Layer 12	Layer 25	Layer 12 (hole)	Layer 25 (hole)
Precision	87.4 %	85.4 %	89.1 %	86.0 %
Recall	92.1 %	89.1 %	93.2 %	89.3 %

D. Network Structure Options

We also experiment with the depth of shallow layers and the hole settings [22] with the network. The original Resnet-50 features 50 convolutional layers and down-samples at layers 1,2,12,25,41. We remove the down-sampling at layer 12 to decrease the down-sampling factor to 16x and experiment with the depth of shallow layers at 12 and 25, which correspond to outputs with down-sampling factor 4x and 8x, respectively. Although layer 25 has more depth, its down-sampling negatively affects the performance in accurate localization. The detections of neighboring foredecks/sterns start to join each other. We also add hole settings on layers 20 and 23, increasing their effective kernel size to 5, which enlarge the receptive field of the shallow layers. The experiment shows that the hole setting slightly increases the performance of the shallow layer. The comparison results can be seen in Tab. II. Here, we only show the performance of foredeck/stern detection instead of ship detection for a more direct presentation.

V. Conclusion

With the feature learning power of Res-FCN we manage to replace carefully hand-crafted features with machine-learned ones, increasing the generalization ability and scalability of the model. We use the task partitioning model to mitigate the limitations of CNNs and partition the task into detection and accurate localization. This combination demonstrates its ability to solve the inshore ship detection problem successfully. The network is currently unable to provide predictions on the direction of the proposed candidates. For future work, we aim to train the network to distinguish the direction of foredecks/sterns and thus make the framework more informative and robust.

References

[1] J. Wang, J. Song, M. Chen, and Z. Yang, "Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine," *International Journal of Remote Sensing*, vol. 36, no. 12, pp. 3144–3169, 2015.

[2] K. Stankov and D.-C. He, "Detection of buildings in multispectral very high spatial resolution images using the percentage occupancy hit-or-miss transform," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 10, pp. 4069–4080, 2014.

[3] L. Zhang, Z. Shi, and J. Wu, "A hierarchical oil tank detector with deep surrounding features for high-resolution optical satellite imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4895–4909, 2015.

[4] X. Yu and Z. Shi, "Vehicle detection in remote sensing imagery based on salient information and local shape feature," *Optik-International Journal for Light and Electron Optics*, vol. 126, no. 20, pp. 2485–2490, 2015.

[5] Z. Shi, X. Yu, Z. Jiang, and B. Li, "Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 4511–4523, 2014.

[6] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with svd networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–14, 2016.

[7] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 98, pp. 119–132, 2014.

[8] G. Liu, Y. Zhang, X. Zheng, X. Sun, K. Fu, and H. Wang, "A new method on inshore ship detection in high-resolution satellite images using shape and context information," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 3, pp. 617–621, 2014.

[9] X. Ren, L. Jiang, and X.-a. Tang, "In-shore ship extraction from hr optical remote sensing image via salience structure and gis information," in *Ninth International Symposium on Multispectral Image Processing and Pattern Recognition (MIPPR2015)*. International Society for Optics and Photonics, 2015, pp. 98 150U–98 150U.

[10] R. Zhang, J. Yao, K. Zhang, C. Feng, and J. Zhang, "S-cnn ship detection from high-resolution remote sensing images," *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 423–430, 2016.

[11] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, 2016.

[12] S. Paisitkriangkrai, J. Sherrah, P. Janney, V.-D. Hengel *et al.*, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 36–43.

[13] H. Lin, Z. Shi, and Z. Zou, "Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network," *Remote Sensing*, vol. 9, no. 5, p. 480, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[15] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[16] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE transactions on cybernetics*, vol. 43, no. 2, pp. 660–672, 2013.

[17] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 6, pp. 1279–1289, 2016.

[18] Q. Wang, Y. Yuan, and P. Yan, "Visual saliency by selective contrast," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1150–1155, 2013.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.