ISSN: 0067-2904

Efficient Hybrid DCT-Wiener Algorithm Based Deep Learning Approach For Semantic Shape Segmentation

Kaustubh V. Sakhare*, Vibha Vyas
Department of Electronics and Telecommunication, College of Engineering, Pune, India

Received: 28/9/2020 Accepted: 13/4/2021

Abstract
Semantic segmentation is effective in numerous object classification tasks such as autonomous vehicles and scene understanding. With the advent in the deep learning domain, lots of efforts are seen in applying deep learning algorithms for semantic segmentation. Most of the algorithms gain the required accuracy while compromising on their storage and computational requirements. The work showcases the implementation of Convolutional Neural Network (CNN) using Discrete Cosine Transform (DCT), where DCT exhibit exceptional energy compaction properties. The proposed Adaptive Weight Wiener Filter (AWWF)rearranges the DCT coefficients by truncating the high frequency coefficients. AWWF-DCT model reinstate the convolutional layers giving modularity in the design using multi scale convolution block. The impact of selection of DCT coefficients in the proposed model is validated on the benchmark database as City Spaces. The same level of accuracy compared to the conventional algorithm is achieved using only 40 % of the DCT coefficients. Extensive experiments validate the advantages of adaptive DCT modeling of CNN in semantic segmentation and image classification.

Keywords: Semantic Segmentation, Convolutional Neural Network, Discrete Cosine Transform, Weiner Filter, Multi scale convolution block.

1. Introduction

Segmentation is one of the crucial steps in image analysis. Semantic segmentation is one of the popular exploitations of segmentation. Performing semantic labeling to partition objects present in the images [1]. Pixel level labeling is carried out for set of object categories. Thus, making it more reliable solution to the complex problems such as scene understanding, autonomous vehicles, and medical image analysis [2]. Over the years, conventional segmentation techniques like shape-based segmentation, K Means clustering, contour-based techniques, sparsity-based methods were overtaken by deep learning-based approaches. Deep learning in semantic labeling has shown remarkable results with extensive performance improvements on benchmark datasets such as PASCAL VOC, City space [3, 4]. Deep Learning for semantic segmentation is formulated as: Fully Connected Networks [4, 5], Encoder-based models, dilated convolution models, Convolutional models with graphical models, Multiscale and pyramidal network-based models [4]. Most of the methods listed execute the convolutional neural network in spatial domain. The work proposed in this paper, uses implementation of CNN in the DCT domain [6].

DCT gives higher level of energy compaction; at the same time, sparser representation of the features comprehends the diversified objects present in the image. The deep learning network

*Email: kvsakhare@pict.edu

give better feature representation, as higher levels of the neural architecture comprehend the major objects as well as layered architecture and achieve similar dimensionality reduction. Motivated with these common attributes between DCT and deep neural networks to separate the frequencies and energy compaction [7]. The proposed architecture engages an Adaptive Wiener Weight Filter (AWWF) for selection of the DCT coefficients. The pretrained network as image net is expected to enrich the encoding phase. As the AWWF-DCT results in a reduced feature resolution by one eighth of the image, we reinstate down sampling operations in Image net. Various Multiscale convolution blocks are tried out to define appropriate feature concatenation. Extensive experiments on City Space database validate the proposed model compared to the existing frameworks in semantic segmentation.DCT speed up the training of fully connected sparse feature extraction, which has resulted in sparser weight matrices trained over the data.

The major contributions of this paper are:

1. Modular approach of feature extraction and decoding is adopted to keep best formulations in practice and improve on the resolution at higher layers.
2. The DCT operation performed on the images gives sparser representation of the weight matrices.
3. DCT wiener filter is proposed as adaptive weight filter, used to achieve significant convergence speedup and case specific accuracy. Improvement is seen at early stage learned feature maps.
4. Multiscale convolution block is developed to perform convolution operation with **'n'** convolution kernels in a parallel mode. The Multiscale feature fusion is achieved by combining/averaging results of Multiscale convolution.

The layout of the paper follows as: Section 2 presents the related work on the semantic segmentation and Harmonic Convolutional neural networks. Section 3 presents the DCT used to construct Adaptive Weiner Weight Filter for semantic segmentation. Section 4 illustrates the proposed Hybrid DCT-AWWF based deep learning approach for image analysis using semantic labels. Section 5 validates the proposed model with the help of methodical experiment followed by conclusions and list of references.

2. Related work

Deep learning is the most contributed to research topic in artificial intelligence field recently. These are effectively used to create a model for perceiving and understanding large quantities of data, such as images and sound [8, 9]. The contemporary literature with respect to semantic segmentation problem, see numerous deep neural network approaches proposed. The prevailing discussion tries to find out some of the unique approaches in semantic segmentation and their common features to narrow down the literature findings.

Fully connected networks ;(shown in Figure 1) can be seen as one of the initial efforts in applying deep learning for semantic segmentation. The architecture reinstated all the fully connected layers by the convolutional layers from the pretrained architectures such as VGG16. Thus, enabling it to handle the arbitrary sized input and giving segmentation output of the same size [10,11]. Skip connections as shown in Figure 2up sample the features from the final layer while combining those with the previous layers. Deep layers represent the semantic information, which is combined with contextual information conveyed by the shallow layers. Validation of the model is done on PASCAL VOC, NYUDv2 datasets, comparable performance is observed on those benchmark datasets. [4].

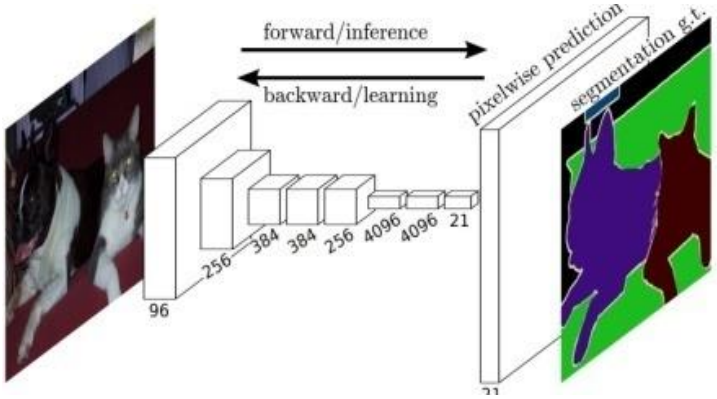


Figure 1-Image Segmentation Network using Fully Convolution Block.

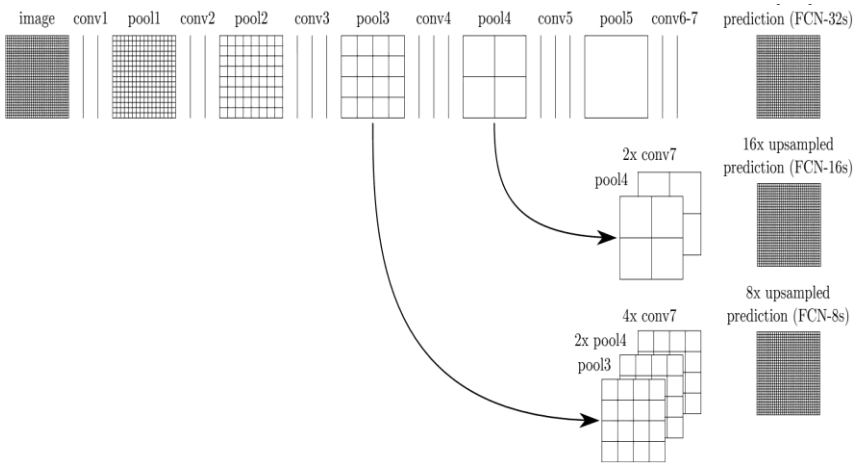


Figure 2-Skip connections Network selectively adding the high-low level information

FCN has performed better on variable sized images; it has some constraints when using it in real time implementations. FCN has a loophole of underperforming on global context information. To overcome these challenges, ParseNet is proposed to handle the global context information. FCN are derived by adding global context to FCNs as layer wise averaging the feature to enhance features at each location. The context vector is formed by pooling the feature map. Similar unpooling method is followed to normalize the context vector and generate the new feature space of similar with same aspect ratio. In essence, FCN with added global context is the main motive behind forming ParseNet [12, 13, and 14].Figure 3 shows these feature maps grouped together.

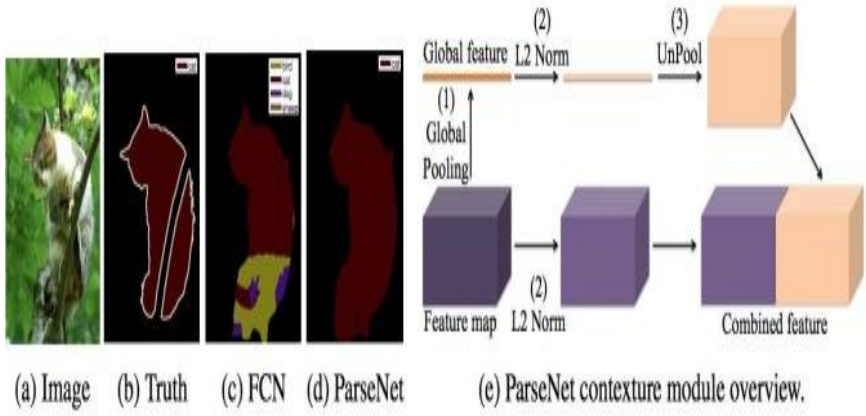


Figure 3-ParseNet, using extra global context (d) than an FCN (c) results in smoother segmentation.

Generic application areas of FCN are iris segmentation, brain tumor segmentation, skin lesion segmentation [11, 15]

Encoder –based models: encoder-decoder model based on convolutional neural network is one of the popular models for image segmentation. Convolutional layers of VGG-16 model will be used as input in the Encoder while feature vector acts as input to the deconvolution network. The encoder gives the class probabilities. The deconvolutionalstage of the framework has

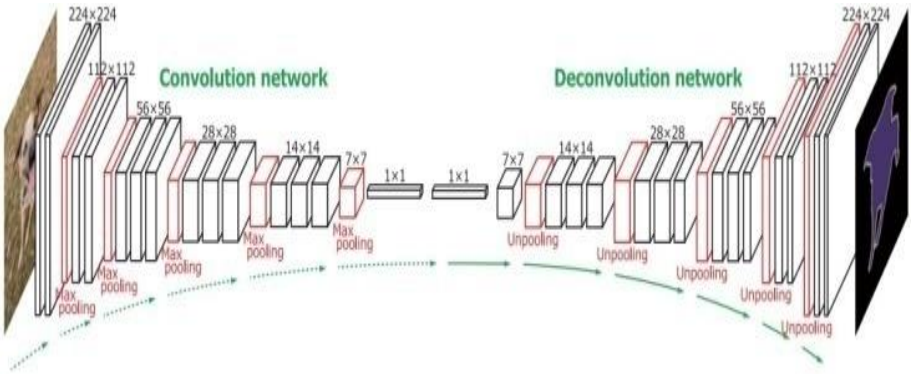


Figure 4- Semantic Segmentation using Deconvolutionalmodule.

deconvolution layers and unpooling layers associated with it as shown in Figure 4. It determines the pixel-wise class labels and segmentation masking. The efforts done at the initial level attracted 72.5 % accuracy on PASCAL VOC 2012 dataset.

Biomedical Image segmentation using Encoder-Decoder Models: The Unet architecture [16, 17] proposed for segmenting the biological images. The data augmentation improvised the model accuracy. In the Unet architecture the contracting path is used to acquire the context, and localization is achieved in the expanding path. The similarity to FCN architecture is observed in the contraction phase, as it extracts the features with convolutions. While expander phase up sample the results with the deconvolution. Feature maps are maintained from decoder to the encoder phase to avoid loss of the pattern information. A stride convolution categorizes each pixel to the relevant class. Numerous developments are seen in U-Net architecture as shown in Figure 5and can be used to extend its utility to 3D Images up to complex problems like road segmentation [17, 18].

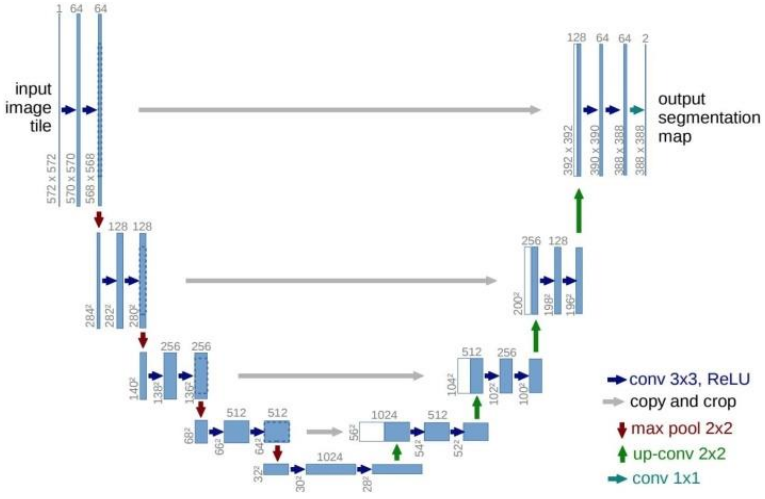


Figure 5-U-net model

V Net modules shown in Figure 6drew attention due to its use for 3D medical image

segmentation and new proposed objective function as Dice Coefficient. It enables the model to handle the diversified and imbalanced data modeling use cases. The proposition made it effective in segmenting the whole data object at one time for challenging problems like MRI image segmentations [18, 19, 20].

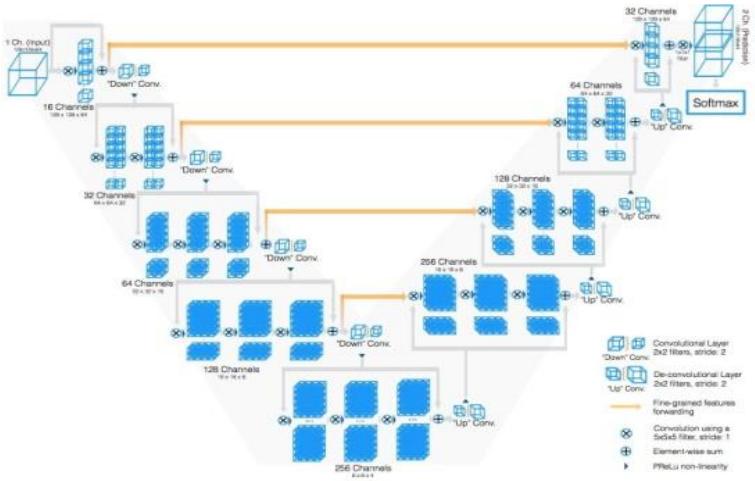


Figure 6-V-net model for 3D image segmentation.

Dilated convolution models: This model is based on the dilation rate. As shown in Figure 7, a 3*3 having dilation rate 2, offers the same receptive field as one generated by kernel 5*5. The number of parameters on the other side remains the same. Improvisation in the receptive field without compromising on the computational cost. For the input signal x(i), the dilated convolution output will be given as $y_i = K * x[i + rk]w[k]$; where r is the dilation rate. A lot of recent work is identified in dilated convolution as Deep Lab [21], multi scale context aggregation [22] Atrous Spatial Pyramid Pooling (DenseASPP) [23].

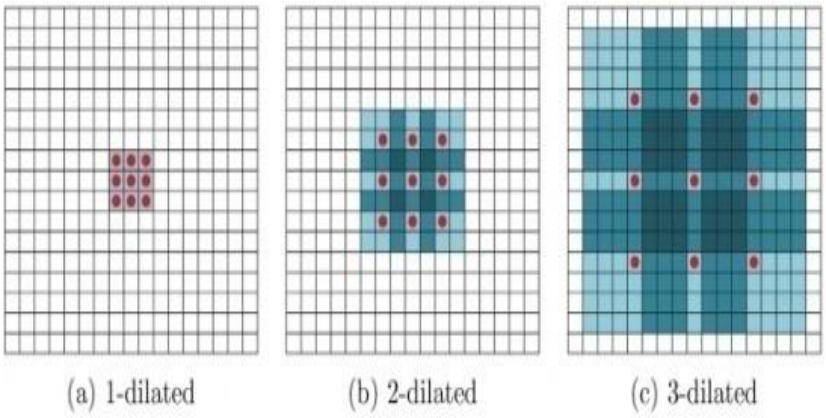


Figure 7-Dilated convolutions. A 3 × 3 kernel at different dilation rates.

DeepLabv1 [21] and DeepLabv2 [23] are one of the prominent choices for semantic segmentation. The dilated convolution used in that, resolve the issue of decreasing resolution generated due to max pooling and stride factor. Spatial Pyramidal pooling based on Atrous convolution is used to address the convolutional feature layer. Better object localization is achieved using Deep CNNs and locating the object boundaries using probabilistic model. ResNet-101 based Deep Lab yields 79.3 % mean Intersection of Union (mIoU) on the PASCAL VOC 2012 dataset. 70.4 % mIoU score for City space database.

Several works considered combining spectral information with CNNs. CNNs especially is trained for detection of multiple compressed images. A common practice in various works

[23] is to arrange histograms of pre-selected DCT coefficients by 1-dimensional CNN. In another work [24] a multi-branch 2-dimensional CNN was trained on feature vector spanned by the first 20 AC coefficients (which exhibits nonzero frequencies in DCT) extracted from compressed images. Use of spectral representation of the images is popular for object recognition.

Training of shallow fully connected neural networks [25] and fully connected sparse auto encoders [7, 26] with DCT on low resolution images and truncation of the coefficients helped speed up the performance. Similarly Radial Basis Function with DCT were used in face recognition [27, 28].

3. DCT- Adaptive Weight Wiener Filter (DCT-AWWF)

DCT –Weiner filter design to pull the high frequency components are required for convolution layer. The proposed Efficient DCT-AWWF model is shown in Figure 8. Spatial domain implementation of CNN may miss the manipulating features. Processing the input in DCT inherently de-correlate the features using sparse representation. $\tilde{X}$ represent the DCT coefficients. The encoders in semantic segmentation when dealing with large feature maps get additional advantage of dimensionality reduction in DCT domain. Wiener filter will be employed to adopt the important DCT features to learn importance of each DCT coefficient. Multiscale convolution block is used along with a deep convolutional stream to acquire more unique and sparse features [29, 30, 36, 37, 38].

Combination of these features in the encoder gets the optimized features in the context driven way. Layered architecture is employed with convolution performed at different scales. The concatenation of the convolution here gives sparser representation and further achieves the down sampling.

The architecture is detailed in the next section.

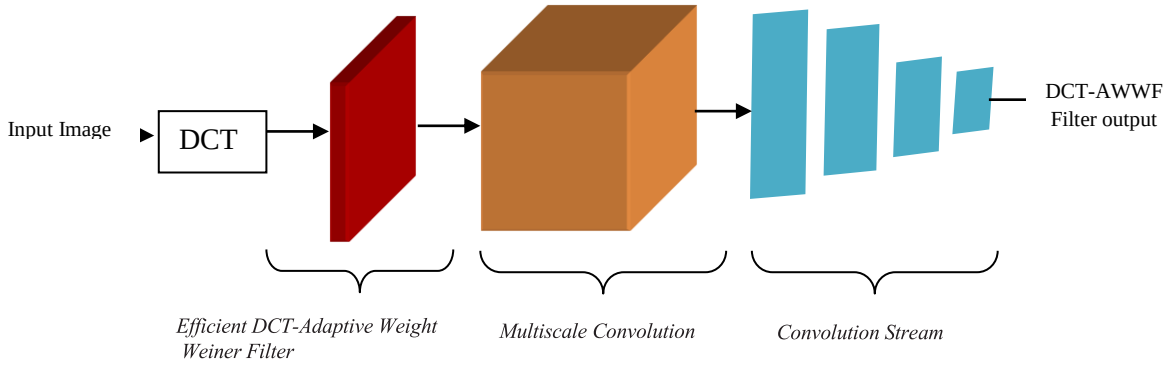


Figure 8-Efficient DCT-AWW Filter

4. Hybrid DCT-Wiener based deep learning approach for semantic shape segmentation

DCT-AWWF Encoder Framework:

The encoder for semantic segmentation comprises of Wiener based adaptive weight filter, convolution layer and max pooling and BN layer as shown in Figure 8.

As discussed in section 3, the important DCT coefficients are identified using W weight matrix to represent semantic segmentation [31, 40]. The DCT coefficients are given by the matrix$X$. The output matrix will be generated as given in Equation [1].

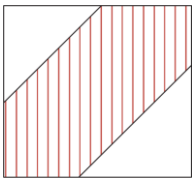
$$\tilde{X} = X \cdot W \text{---------- (1)}$$

Here"$\cdot$"characterizes array multiplication, while $\tilde{X}^{(1)}$ holds the output of the first layer. The matrix Wave elements 1 and 0, corresponding to the high frequency components related to the semantic segmentation. High-frequency band in the image, houses the semantic features of animage.As the DCT operation is carried out, the spectral domain of an image is sliced to

three sub bands as given in Figure 9 (a). b1 and b2 are used as thresholds for the frequencies to sort the DCT coefficients (i.e., frequencies). Equation [2] gives the weight matrix $\dot{W}$

$$\dot{W} = I(\tilde{X} - b1 \geq 0) - I(\tilde{X} - b2 \geq 0) \text{---------- (2)}$$

$\tilde{X}$ is the matrix of DCT indices formed in thecrisscrossmannershowninFigure9(b) and I (·) gives a matrix having the elements 1and0, indicating the element-by-element argument of the function if it istrueornot.$b_1$and$b_2$are determined in the context-driven way.



1	2	6	7	15
3	5	8	14	16
4	9	13	17	22
10	12	18	21	23
11	19	20	24	25

Figure 9 (a)- Frequency domain slicing and **Figure 3 (b)**-zigzag scan table for the 5 × 5 block

Convolutional Layers

Apart from dimensionality reduction, convolution layers extract deep features. In a convolutional stage, the convolution and activation operation are given in the Equation [3]

$$\tilde{X}^{(1)} = \eta \sum_{i=1}^{C_j} \dot{W}^{(\ell)} * \tilde{X}^{(\ell-1)} + b_j^l \text{---------- (3)}$$

For ith layer

Convolution operation is performed by * Activation function is given byη(·): Sigmoid, ReLU, Bounded ReLU can be applied as activation function.$C_j$is the channel number of this layer.

The detailed architecture is shown in Figure 10.

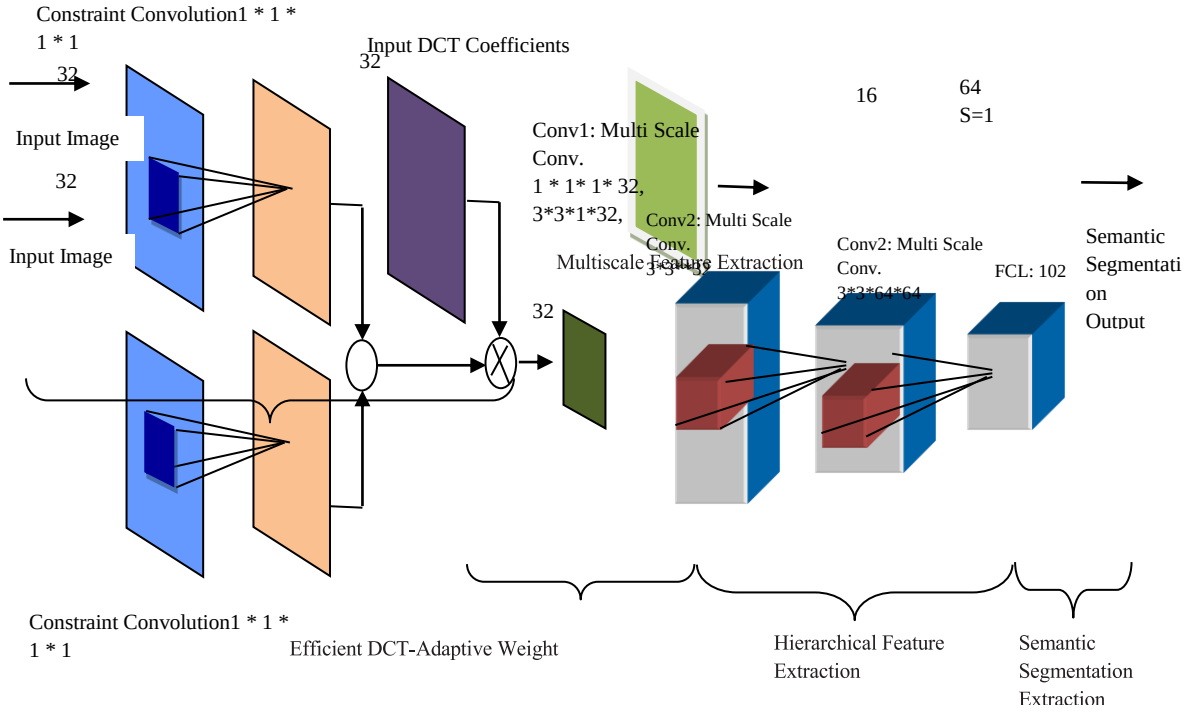


Figure 10-Proposed Architecture

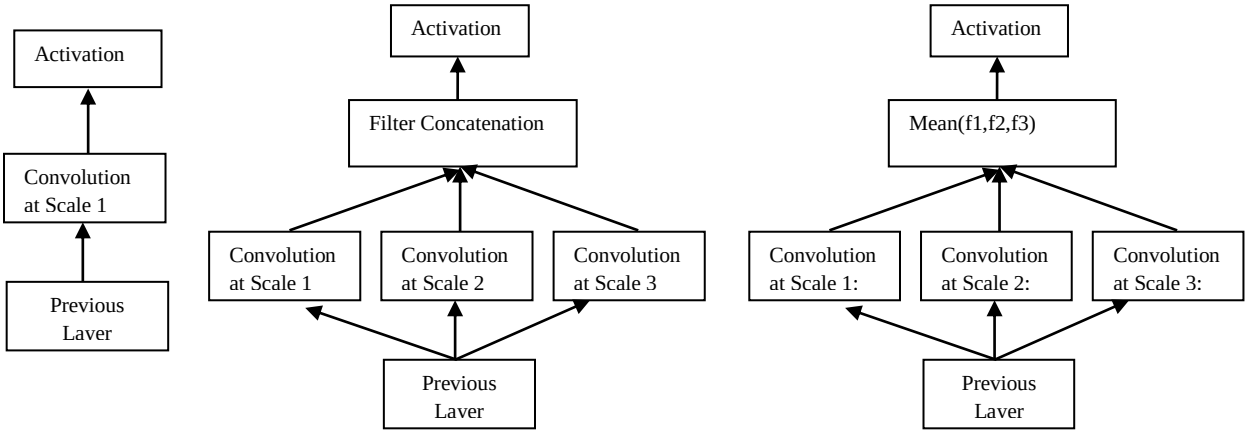
Multiscale convolutions are formed with various kernel sizes and concatenated to get a high-level feature map [32]. The concatenation layer performs the feature fusion by combining the output of multi-scale convolution into an element wise average function μ (),given as

$\mu(x_1,x_2,\ldots,x_n)$=mean$(x_1,x_2,\ldots,x_n)$.

The concatenation of the features is performed at the fusion level. It helps in understanding the high frequency features in the initial layers. To find the best of the feature selection module, The Multiscale convolution designs are compared with various convolution layer architecture. As a single-scale convolutional layer Figure 11(a), inception module Figure 11 (b) and proposed convolution module at different scales. Figure 11 (c) proposed in the paper.

Pooling Layers and Batch Normalization

Batch Normalization will follow each convolutional layer. This design minimizes the covariance during the training phase while fastening the training process. Inserting the pooling layer after the normalization layer, decreasing the training time and additionally reducing dimensionality is achieved in this stage. Over fitting even will be overcome with the help of pooling layer [6].



(a) (b) (c)

Figure 11-(a) Single scale convolution, (b) inception module (c) Proposed Convolution module at different scales

Encoder Layer

Fully-connected layer and layer with softmax function will be part of classification layer. FC layer ensures that all the features will be used for the classification. The softmax function will take the input towards last stage of semantic encoder [6, 36, 39, 41].

Network Architecture

The CNN architecture with proposed DCT Adaptive weight wiener filter is shown in figure 4.The design of the architecture is illustrated as given below. Figure 10 depict the output at every stage for input image 224*224.

The Efficient DCT-AWWF filter with relevant operations is mentioned in Equations [1] and [2]. Equation [3] is rounded off using constrained convolutions. Constrained convolution layers are built up using stride of 1. BReLU is used as activation function, associated with element-wise subtraction. The convolutions have kernel size 1*1, bias updated at the time of training while a fixed weight of 1 is used [6, 33].

Conv1 is generated as multi-scale convolutional block. 32 kernels of different sizes 1*1, 3*3, and 5*5 are applied, generating 32 feature maps at different scales with size 32*32. BN layer follows the Convolution function. Max-pooling layer with window size 2*2 and step size 2 will give the input feature map same with the spatial resolution reduced to 25 % of the original value [6, 33, 34].

In the proposed architecture, the 5^{th} and 8^{th} layer will be convolutional layers Conv2 and Conv3. Convolution 2 have 64 kernels with dimension 3*3*32, convolution 3 have 64 kernels of size 3*3*64. To get the normalization of the output features, both convolution layers will

be followed by Batch Normalization layers and pooling layers. Fully connected layer 1 and 2, FC1 and FC2 are presented after the last pooling layer. FC1 consist of 1024 neurons while FC2 has 2 neurons. At the output layer, the output of FC2 is given to the softmax layer, giving prediction of the object class and class label [44, 45]**.**

5. Experimental Results and Analysis

The Adam optimization [19, 42] was used in training the network. The optimizer accelerated the convergence on every dataset. The code was implemented Google colab environment. The implementation was done in two stages: the encoder was trained to categorize down sampled section of the input image, the decoder was appended and trained on the network to perform up sampling and classification of every pixel at the later stage. Setting the learning rate to 5e - 4 and weight change to 2e- 4 and keeping the batch processing of 10 performed consistently well [19, 33,43].

A model is evaluated using quantitative accuracy, speed, and storage requirements. Pixel Accuracy: it is a ratio of properly classified pixels with total pixels count in the image. When K is foreground classes, for k+1 class, is represented in the Equation [4].

$$\text{Pixel Accuracy} = \frac{\sum_{i=0}^{K} P_{ii}}{\sum_{i=0}^{K} \sum_{j=0}^{K} P_{ij}} \text{---------- (4)}$$

Intersection of Union (IoU) as given in Equation [5] is popularly used in semantic segmentation. It is calculated as intersection between the estimated segmentation vector and ground truth proportioned with union between the estimated segmentation vector and ground truth.

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \text{---------- (5)}$$

The performance of the recent segmentation algorithms is given by Mean –IoU, where it states the average IoU over all classes. The performance of DCT-AWWF encoder on city space dataset is validated. VGG16 is set as base architecture, as its one of the fastest segmentation models. VGG16 has resulted in lesser parameters and memory consumption, compared to FCN. The comparison of the results is done based on class average accuracy and intersection-over-union (IOU). The dataset has total 5000 annotated images, from which 2975 images were selected for training. Validation dataset was maintained using 500 images. Whereas 1525 image were maintained in the testing dataset. [35,44]. Cityscapes exhibit dynamically varying road scenarios, featuring many pedestrians and cyclists. For the model training 7 classes were considered. As reported in Table1, the proposed architecture outperforms Unet in pixel average accuracy and class IoU DCT-AWWF Model will be an outperformer in the City spaces benchmark. Figure 12 presents prediction examples on validation dataset.

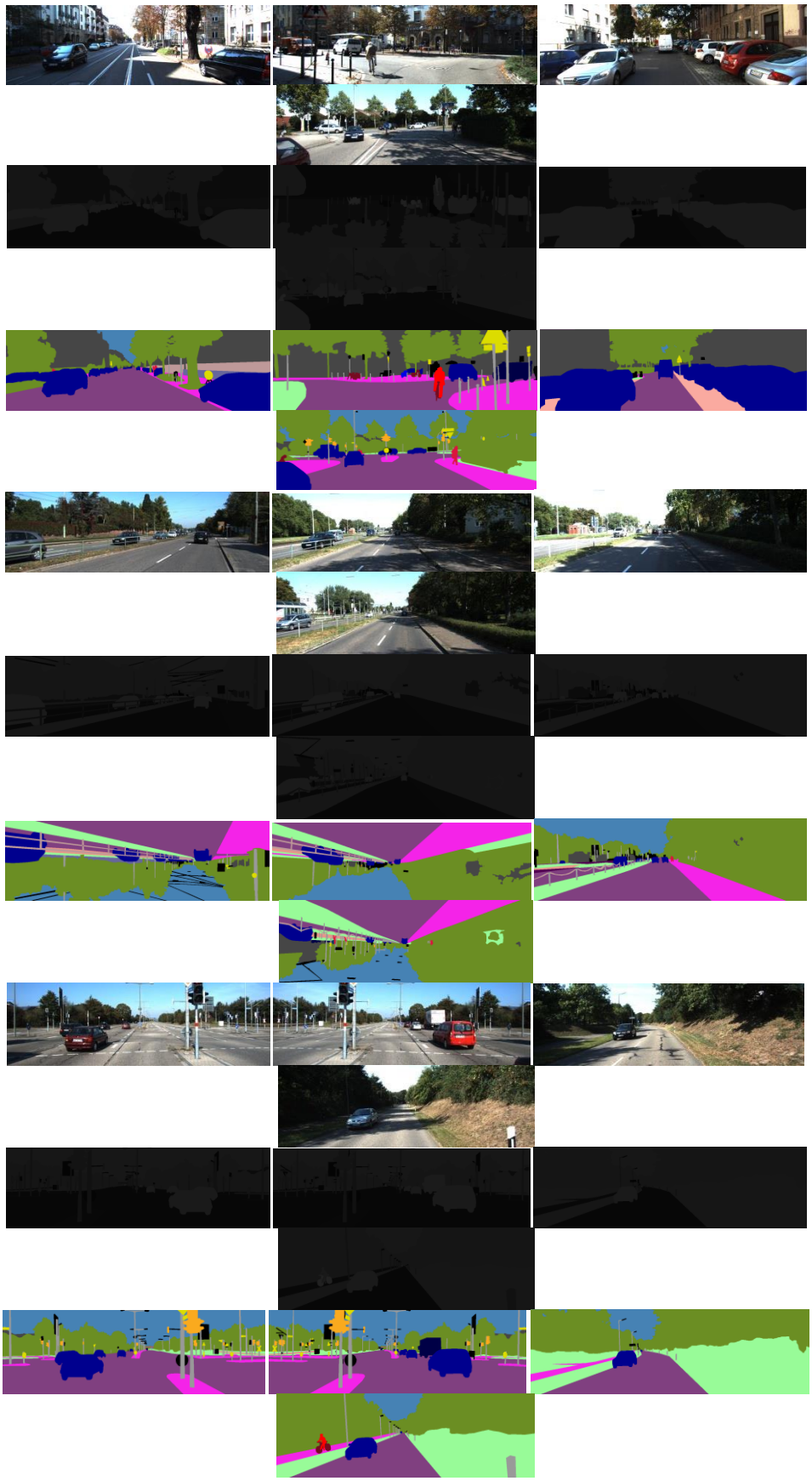


Figure 12-Semantic Segmentation results: row (1) original Image, row (2) Semantic Images row (3) Semantic segmentation images

Table 1-Results of U Net Architecture with Proposed DCT-AWWF Architecture

Object / Model	Camping Car	Car	Road	Person	Pole	Side walk	Cyclist	Class Avg	Class IoU
UNet	84.6	**87.3**	92.3	55.0	**47.5**	74.1	26.0	66.7	**55.6**
DCT-AWWF	**88.8**	**91.2**	**95.1**	**67.2**	45.4	**86.7**	**34.1**	**72.6**	52.4

Conclusion:

The work carried out here exploited the semantic segmentation architecture in the DCT domain. Feature representation and manipulation in frequency domain become advantageous. The article proposes and validates a hybrid DCT-Weiner based deep learning approach for semantic shape segmentation. The semantic segmentation layered architecture was used to extract features in the encoder stage. The performance was investigated in frequency domain. The high frequency information in the image has given better feature representation. The hybrid design applies a DCT-Wiener filter to pull the high frequency components required for convolution layer. Multi scale convolution model was developed for sparse feature representation. The model proposed in the paper achieved remarkable results compared to the existing encoders used in semantic segmentation. The class average accuracy achieved more than the conventional U Net architecture. The semantic labeling with reduced computational complexity can be seen as solution in some of the applications such as autonomous vehicles, scene understanding, and augmented reality.

References

[1] L.-C.Chen, G.Papandreou, F.Schroff,andH.Adam,"Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv: 1706.05587*,2017.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprintarXiv: 1409.1556*,2014.

[3] V.Badrinarayanan,A.Handa,andR.Cipolla,"Segnet:Adeepconvolutionalencoder-decoderarchitecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv: 1505.07293*,2015.

[4] M.Cordts,M.Omran,S.Ramos,T.Rehfeld,M.Enzweiler,R.Benenson,U.Franke,S.Roth,andB .Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,2016.

[5] V. Goel, J. Weng, and P. Poupart, "Unsupervised video object segmentation for deep reinforcement learning," in *Advances in Neural Information Processing Systems*, 2018, pp.5683–5694.

[6] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor.

[7] A. Liu, Z. Zhao, C. Zhang, and Y. Su, "Median filtering forensics in digital images based on frequency-domain features," Multimedia Tools Appl., vol. 76, no. 6, pp. 22 119–22 132, 2017.

[8] S. Minaee and Y. Wang, "An admm approach to masked signal decomposition using subspace representation," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3192–3204,2019.

[9] 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.

[10] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner*et al.*, "Gradient-based learning applied to document recognition,"*ProceedingsoftheIEEE*, vol. 86, no. 11, pp. 2278–2324,1998.

[11] J.Fu,J.Liu,Y.Wang,J.Zhou,C.Wang,andH.Lu,"Stackeddeconvolutionalnetworkforsemantic segmentation,"*IEEETransactions on Image Processing*,2019.

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W.Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[13] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance- aware semantic segmentation," in *Proceedings of the IEEEConference on Computer Vision and Pattern Recognition*, 2017, pp.2359–2367.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp.2881–2890.

[15] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE transactions on medical imaging*, vol. 36, no. 9, pp. 1876–1886,2017.

[16] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv: 1511.06434*,2015.

[17] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp.565–571.

[18] Olaf Ronneberger, Philipp Fischer,"U-Net: Convolutional Networks for Biomedical Image Segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, pp 234-241

[19] Z.Zhou,M.M.R.Siddiquee,N.Tajbakhsh,andJ.Liang,"Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp.3–11.

[20] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang, and H. Lu, "Adaptive context network for scene parsing," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp.6748–6757.

[21] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang, and H. Lu, "Adaptive context network for scene parsing," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp.6748–6757.

[22] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015.

[23] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cot- trell, "Understanding convolution for semantic segmentation," in winter conference on applications of computer vision. IEEE, 2018, pp. 1451–1460.

[24] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic

depth," *arXivpreprint arXiv:1603.09382*,2016

[25] S. Kant, P.Kumar,A.Gupta,andR.Gupta, "Leukonet: Dct-based cnn architecture for the classification of normal versus leukemic blasts in b-all cancer," arXiv:1810.07961,2018.

[26] A. Tuama, F. Comby, and M. Chaumont, "Camera model identification with the use of deep convolutional neural networks," in Proc. IEEE Int. Workshop Inf. Forensics Secur., Seattle, WA, USA, Dec. 2016, pp. 1–6

[27] S. Liew, M. Khalil-Hani, and R. Bakhteri, "Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems," Neurocomputing, vol. 216, pp. 718–734, 2016.

[28] L. Gueguen, A. Sergeev, R. Liu, and J. Yosinski. Faster neural networks straight from JPEG. In International Conference on Learning Representations Workshop, 2018.

[29] Q. Wang and R. Zhang. Double JPEG compression forensics based on a convolutional neural network. EURASIP Journal on Information Security, 2016(1):23, Oct 2016

[30] Y. Wang, C. Xu, S. You, D. Tao, and C. Xu. Cnnpack: Packing convolutional neural networks in the frequency domain. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 253–261. Curran Associates, Inc., 2016

[31] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 7168–7177. IEEE, 2017

[32] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor.

[33] Semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3029–3037.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*,2015.

[35] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," IEEE Transactions on Image Processing, 2019.

[36] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp.3562–3572.

[37] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp.2393–2402.

[38] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Multi- scale context intertwining for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp.603–619.

[39] J.He,Z.Deng,L.Zhou,Y.Wang,andY.Qiao,"Adaptive pyramid context network for semantic segmentation," in *Conference on Computer Vision and Pattern Recognition*, 2019, pp.7519–7528.

[40] R.Hu,P.Dollár,K.He,T.Darrell andR.Girshick,"Learning to segment everything," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp.4233–4241.

[41] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang,and H. Adam, "Masklab: Instance segmentation by refining object detection with semantic and direction features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp.4013–4022.

[42] X.Chen,R.Girshick,K.He and P.Dollár,"Tensormask: A foundation for dense object segmentation," *arXiv preprintarXiv: 1903.12174*,2019.

[43] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen,and P. Luo, "Polarmask:

Single shot instance segmentation with polar representation," *arXiv preprint arXiv:1909.13226*, 2019.

[44] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cot- trell, "Understanding convolution for semantic segmentation," in *winter conference on applications of computer vision*. IEEE, 2018, pp. 1451–1460.

[45] A. Hatamizadeh, D. Sengupta, and D. Terzopoulos, "End-to-end deepconvolutionalactivecontoursforimagesegmentation,"*arXivpreprint arXiv: 1909.13359,*2019.