IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE SYSTEMS, MAN AND CYBERNETICS SOCIETY SECTION

Received September 16, 2021, accepted October 20, 2021, date of publication October 26, 2021, date of current version November 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3123066

FTNet: Feature Transverse Network for Thermal Image Semantic Segmentation

KAREN PANETTA[1], (Fellow, IEEE), SHREYAS KAMATH K. M.[1], (Member, IEEE), SRIJITH RAJEEV[1], (Member, IEEE), AND SOS S. AGAIAN[2], (Fellow, IEEE)

[1]Department of Electrical and Computer Engineering, Tufts University, Medford, MA 02155, USA
[2]Department of Computer Science, City University of New York, New York, NY 10016, USA

Corresponding author: Shreyas Kamath K. M. (skamat04@tufts.edu)

ABSTRACT Thermal imaging is a process of using infrared radiation and thermal energy to collect information about objects. It is superior to visible imaging for its ability to operate in darkness and tolerate illumination variations. In addition, it has potential to penetrate smoke, aerosol, dust, and mist, which are critical inhibitors for visible imaging applications, including semantic segmentation. Unfortunately, current state-of-the-art image semantic segmentation methods (i) mainly concentrate on visible spectrum images, which do not adequately capture the context of corresponding pixels, particularly edge details in thermal images, and (ii) accept a trade-off between higher accuracy and lower speed, or vice-versa. Here, a novel end-to-end trainable convolutional neural network architecture, feature transverse network (FTNet), has been proposed to solve the aforementioned problems. FTNet captures and optimizes feature representation at the multi-scale resolution, thereby improving the capability to process high-resolution images and producing quality output with a lower computational cost. Extensive computer experimentations were conducted on publicly available benchmarking thermal datasets, including SODA, MFNet, and SCUT-Seg, to demonstrate the effectiveness of the proposed FTNet compared to state-of-the-art methods. This comparison includes multiple aspects, including the quantitative accuracy and speed of the various approaches. The source code is available at https://github.com/shreyaskamathkm/FTNet.

INDEX TERMS Convolutional neural network, FTNet, semantic segmentation, thermal segmentation, edge-guidance, transverse network.

I. INTRODUCTION

Image segmentation is the process of partitioning images into multiple segments [1], and it is one of the most challenging tasks in computer vision. It paved the way towards scene understanding, whose importance is highlighted by the fact that an increasing number of applications nourish from inferring knowledge from imagery, including autonomous driving [2]–[4], computational photography [5], [6], biomedical analysis [7], [8], and augmented reality [9]–[13].

Semantic image segmentation (SS) is a high-level task formulated as a classification problem of pixels with semantic labels [11]. Semantic segmentation algorithms identify regions of different objects in the scene by grouping parts of the image together based on the same object of interest and assigning a label to each pixel of an input image. In contrast, instance segmentation treats multiple objects of the same class as distinct individual instances. Panoptic segmentation assigns two labels to each pixel of an image, namely a semantic label and an instance id. The identically labeled pixels belong to the same semantic class, and instance their ids distinguish its instances.

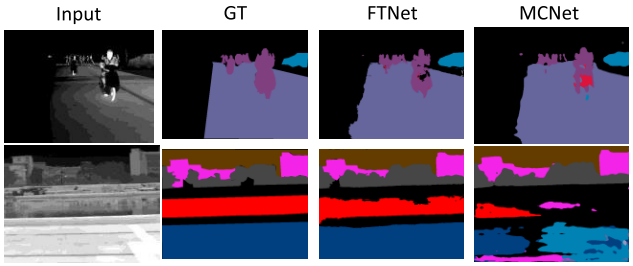


FIGURE 1. Examples demonstrating the effectiveness of the FTNet's ability to reconstruct semantic maps with higher accuracy and crisper edges. Column (a) Input thermal images, (b) Ground truth semantic maps, (c) Results produced by FTNet, and (d) Results generated by the state-of-the-art network- MCNet.

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaojie Su.

Despite significant advancements, semantic image segmentation is still considered a challenging task due to the adverse environmental conditions caused by imaging limitations of the visible spectrum. For instance, visible cameras are susceptible to lighting conditions and become invalid in total darkness. Furthermore, their imaging quality decreases significantly in adverse environmental conditions, such as rain and smog [11].

Thermal imaging is a process of utilizing infrared radiation and thermal energy to gather information about objects. It is superior to visible imaging for its ability to operate in darkness, and across illumination variations. It offers the capability to penetrate smoke, aerosol, dust, and mist [14]. The global thermal imaging market for the mobility industry is expected to reach $3.22 billion by 2025 [15]. The growth of the market can be attributed to growing awareness and lower prices of thermal cameras. This has led to their application in many computer vision tasks, such as detection [16], [17], tracking, segmentation [18]–[21], and individual and emotion identification [22]–[25]. Thermal image-based computer vision will be instrumental in improving driver-assist systems that are increasingly quintessential in consumer car models. These sensors offer additional information to existing autonomous driving sensory systems, which strives to improve performance in identifying objects within a vehicle's surroundings to enhance driving reactions. However, the inter-class variance of objects in thermal images is extremely low, making accurate labeling near boundaries difficult, resulting in a large amount of semantic ambiguity and intensifying the challenge of semantic segmentation. The state-of-the-art (SOTA) semantic segmentation approaches focus on diminishing semantic ambiguity using the rich context information of visible images. However, redundant and noisy semantic information from thermal images may clutter the final semantic maps. An example of ambiguous boundary and the noise-induced from thermal sensors is illustrated in Figure 2.

Convolutional Neural Networks (CNN) are among the most effective and widely used deep learning (DL) architectures in computer vision, including classification, detection, and segmentation. Semantic segmentation models usually follow an encoder-decoder architecture. A deep convolutional neural network (CNN) typically computes a feature hierarchy layer by layer in the encoder stage. It develops an inherent multi-scale pyramid shape. At the decoder end stage, a high-semantic feature map is up-sampled and fused with the previous layer feature map through lateral connections to recover higher spatial dimensions. After extracting spatial details, the network predicts the class for each pixel to complete the segmentation process.

However, this progress has come with a voracious appetite for computing power which will rapidly become technically and economically prohibitive [27].

This article presents a novel end-to-end trainable convolutional neural network architecture, named feature transverse network (FTNet), to address these issues. The proposed

FIGURE 2. Illustration of RGB and thermal images from MFN dataset [26]. The object boundaries in thermal images can be visualized as ambiguous and noisy compared to their RGB counterpart, which will adversely affect segmentation.

FTNet network will be designed and optimized to perform image segmentation of thermal images. FTNet consists of two main components: a high-low feature traversing and an edge guidance part. The architecture is equipped with skip connections between these two networks to use high-resolution image details during the reconstruction. An example of the results obtained using FTNet is illustrated in Figure 1.

Some of the notable contributions of FTNet include:

1) a unified end-to-end trainable network that captures discriminative thermal image features from multiple resolutions and combines them in a fully connected approach;

2) a network that captures and optimizes feature representation at the multi-scale resolution, thereby improving the capability of handling high-resolution images and producing quality output at a lower computational cost;

3) a network whose main representations are shared between the semantic segmentation and edge guidance structures, which means that the FTNet simultaneously achieves semantic segmentation and edge detection without significantly increasing the model complexity;

4) an extensive computer simulation performed on challenging thermal semantic segmentation tasks on benchmarks datasets including SODA [11], MFNet [26], and SCUT-Seg [28], which validate the performance of the proposed model compared with state-of-the-art methods such as MCNet [28], PSPNet [29], DeepLabv3 [30], and HRNet [31];

5) the source code, which will be made available on GitHub for the research community.

The remainder of the paper is organized as follows. In section II, the recent related literature is reviewed. A detailed description of the FTNet architecture and its analysis is provided in section III. Section IV presents the experimental results, including training details, ablation studies, and benchmark results. Finally, a brief discussion and conclusion are provided in sections V and VI, respectively.

II. RELATED WORK

This section provides an overview of some of the most prominent DL architectures in use for the computer vision community for visible and thermal image semantic segmentation.

TABLE 1. Literature review of the state-of-art techniques for image semantic segmentation.

Author	Explanation
FCN-32s [42]	This network employs exclusively locally connected layers, such as convolution, pooling, and upsampling. Avoiding the use of dense layers reduces parameters. Such an architecture also allows a fully convolutional network (FCN) to work for variable image sizes.
SegNet [43]	This method consists of an encoder network, a corresponding decoder network, and a pixel-wise classification layer. The decoder upsamples the lower resolution input feature maps using pooling indices computed in the max-pooling step of the corresponding encoder. The sparse upsampled maps are then convolved with trainable filters to produce dense feature maps.
UNet [44]	This network was developed for biomedical image segmentation. The contraction path (also called the encoder) is used to capture the context in the image. The symmetric expanding path (also called the decoder) enables precise localization using transposed convolutions. Thus, it is an end-to-end network that comprises only convolutional layers because it can accept images of any size.
FPN [45]	This network accepts a single-scale image of arbitrary size and provides proportionally sized feature maps at multiple levels in a fully convolutional fashion. Feature Pyramid Network (FPN) is independent of the convolutional backbone architectures. Therefore, FPN acts as a generic solution for building feature pyramids inside deep convolutional networks.
PSPNet [29]	This network employed a pyramid parsing module that utilized global context information by different region-based context aggregation. The local and global clues together made the final prediction more reliable.
DeepLabv3 [30]	This network aims at solving the problem of segmenting objects at multiple scales. It employed atrous convolution in cascade or in parallel to capture multi-scale context by adopting multiple atrous rates.
ICNet [46]	This method developed a cascaded network and fusion unit that incorporated multi-resolution branches under proper label guidance to reduce a large portion of computation for pixel-wise label inference.
LinkNet [47]	This network aims to reduce the number of parameters by bypassing the spatial information directly from the encoder to the corresponding decoder, improving accuracy and significantly decreasing processing time.
PAN [48]	This method boosts information flow in a proposal-based instance segmentation framework by shortening the information path between lower layers and topmost features. Additionally, adaptive feature pooling is employed to make useful information in each feature level propagate directly to the following proposal subnetworks. A complementary branch capturing different views for each proposal is created to improve mask prediction further.
DANet [49]	This method appends two types of attention modules on top of traditional dilated FCN, which applied the semantic interdependencies in spatial and channel dimensions. This approach adaptively integrates local features with their global dependencies.
UNet++ [50]	This network connected the encoder and decoder sub-networks through a series of nested, dense skip pathways. The re-designed skip pathways aim at reducing the semantic gap between the feature maps of the encoder and decoder sub-networks.
ENCNet [51]	This method introduced a context encoding module, which captured the semantic context of scenes and selectively highlighted class-dependent feature maps.
HRNet [31]	This method overcomes the low-resolution representations by utilizing a high-resolution convolution stream throughout the network. It gradually adds high-to-low resolution convolution streams and connects multi-resolution streams in parallel.
MCNet [28]	This method captured the inter-class and intra-class contextual dependencies using a multi-level correction process. In addition, prior edge knowledge was combined with the context information to obtain the final feature representations.

Some of the earliest segmentation approaches include thresholding [32]–[34], histogram-based bundling, region growing [35], k-means clustering [36], watersheds [37], active contours [38], and graph cuts [39]. Most traditional semantic segmentation algorithms are based on low-order visual information of the images. Therefore, the semantic maps produced by these methods are often not ideal when complex segmentation tasks which require artificial auxiliary information are presented [40]. However, DL architectures have shifted the paradigm in the field of segmentation with remarkable performance improvements on popular benchmarks [41].

Long *et al.* [42] developed one of the first semantic segmentation DL architectures using a fully convolutional network. It was able to produce an output of the corresponding size with arbitrary size input and effective reasoning. However, it did not utilize the global context information efficiently. Noh *et al.* [52] generated dense segmentation masks using a sequence of deconvolution operations. The network consisted of deconvolution and unpooling layers, which alleviated a few of the existing limitations.

Eliminating downsampling may increase resolution; however, it affects the receptive field in subsequent layers, increasing context loss. To overcome this, Chen *et al.* [53] and Yu and Koltun*et al.* [54] used dilated convolution to enlarge the receptive field of neural networks. Chen *et al.* [30] further combined cascaded and parallel modules of dilated convolutions.

Badrinarayanan *et al.* [43] introduced unpooling layers for upsampling as a replacement for transposed convolutions. This network eliminated the parameters required for learned upsampling, thereby achieving a balance between memory and precision. Liu *et al.* [55] proposed a method that models global context directly instead of relying on the largest

receptive field of the network. This method merged the output of the global pooling layer from previous layers with the current map of the posterior layer to generate the final classifier prediction with both having the same size [56].

Multi-scale feature analysis has been extensively studied and has been deployed in various neural network architectures. One of the most prominent multi-scale feature analysis models was FPN [57], [58], which introduced multi-scale feature fusion by setting a top-down pathway. Following this, various feature pyramidal-based architectures have been introduced; for example, PANet [48] suggests an additional bottom-up path augmentation for preserving the local context, and NASFPN [59] was introduced for object detection. This network exploited the neural architecture search framework. Wang *et al.* [31] proposed a multi-branch parallel structure that can efficiently utilize the fine-grained spatial information, which is generally lost in encoder-decoder-based models due to the downsampling and upsampling process. However, it does not consider global context information and boundary information [60]. Zhao *et al.* [29] further developed a method to learn feature representations at different scales. However, it has a sizeable model complexity and computational requirements.

Edge guidance is simple but effective in indicating the semantic separation between different regions [61], [62]. In fact, there exists few traditional high-order conditional random field (CRF) [63], and CNN based [64], [65] semantic segmentation methods utilize superpixels for retaining boundary information. However, superpixel based approaches are unlearnable and not robust [66]. Liu *et al.* [66] addressed this issue using edge loss reinforced structures constructed from encoder and decoder to retain spatial boundary information for remote sensing images.

In the thermal image segmentation domain, Li *et al.* [11] designed a gated feature-wise transform layer to adaptively embed edge information as the guidance of a semantic segmentation network. This network extracted edges utilizing the HED (Holistically nested Edge Detection) network [67] and embedded the edge features into a network proposed by Chen *et al.* [30]. Xiong *et al.* [28] developed a thermal image semantic segmentation method that utilized multi-level edge knowledge to get more edge and shape features. Ha *et al.* [26] incorporated RGB and thermal information to perform segmentation. This network utilized two separate encoders to extract features from visible and thermal images and fuse them in the decoder to produce the probability map for the semantic segmentation results. Other methods that used RGB and thermal images are described in [68]–[70]. Table 1 provides a chronological list of various other image segmentation methods, along with a brief explanation for each method.

III. PROPOSED METHOD

This section presents an end-to-end trainable convolutional neural network architecture called the feature transverse network (FTNet). A high level flow diagram of the proposed system is provided in Figure 3. This paper aims to construct a function $f(I)$ developed specifically to link each pixel in an image, where I is an input image of any arbitrary size (m, n) to a class label with the same dimension. This network combines the low-level layers with poor semantic features and strong resolution with the high-level layers that have rich semantic features and scarce resolution. Following this theme, the novel FTNet comprises an encoder network, a corresponding transverse decoder network, and a final pixel-wise classification layer. This network aims at capturing and optimizing feature representation at the multi-scale resolution, thereby improving the capability of handling high-resolution images and producing quality output at a lower computational cost than the SOTA techniques. Additional details of these components are provided in further subsections.

A. ENCODER NETWORK

Since the main focus of the proposed network is to build a position-sensitive model capable of pixel-level classification, FTNet employs existing SOTA backbones that follow the design rule of LeNet-5 [71]. The spatial size of the features in these classification-based backbones is gradually reduced from a high-level representation to a low-level representation, thereby allowing FTNet to capture features with different representation capabilities.

For simplicity of exposition, consider the case in which ResNet50 is used as a backbone. The basic structure of the encoder network is visualized in Figure 3 (a). The encoder network aims at acquiring features at different resolutions by subsampling at various stages. The convolutions in these networks can be divided into four stages, and the output of each stage's last block from the encoder network can be represented as $\{\mathrm{E_i}|i = 1, 2, 3, 4\}$. This bottom-up pathway extracts and establishes a feature pyramid by incorporating features from the last convolutional layer in each stage. The extension to other backbones is straightforward.

B. FEATURE TRANSVERSE NETWORK (DECODER)

Recent developments have demonstrated the evidence for the necessity of exploring all the multiple-resolution representations for a broad range of vision problems [31]. Following this, a transverse network with a top-down path comprising feature aggregation to produce final semantic features at high-resolution is introduced. An illustration of the proposed decoder network is displayed in Figure 3 (b).

The proposed FTNet considers the features from the stages $\{\mathrm{E_i}|i = 1, 2, 3, 4\}$, which comprises multiple resolutions $\mathfrak{r} = 1/4, 1/8, 1/16$, and $1/32$, respectively. These feature maps are passed through a set of residual units $\mathfrak{U}$ as proposed by He *et al.* [72]. The illustration of this unit is provided in Figure 3 (d). Each residual unit can be defined as formulated in equation (1).

$$\Psi_{l+1} = \mathbb{R}(\omega_s(\mathrm{I}(\Psi_l)) + \Omega(\omega_l * \Psi_l + \mathrm{b_l})| \times \{\omega_l = [\varpi_{l,k} : k = 1 \leq k \leq K]\} \quad (1)$$

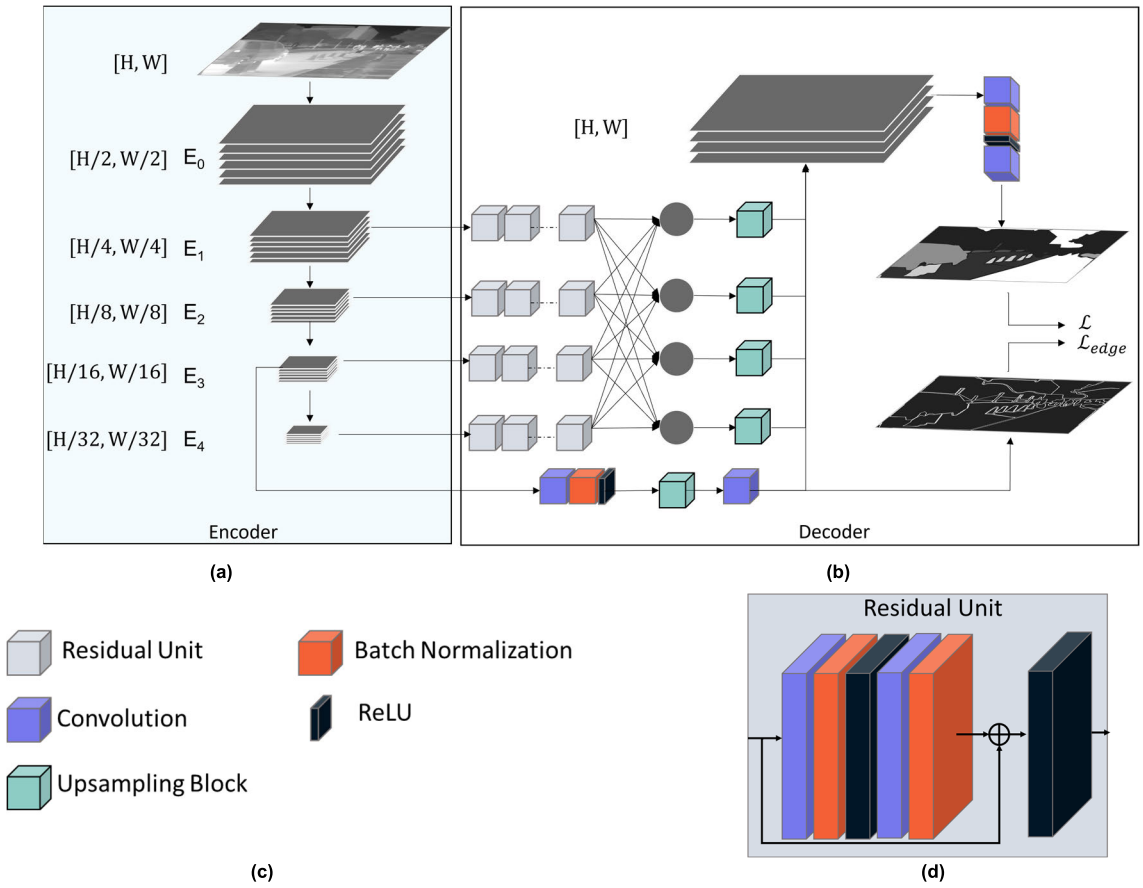


FIGURE 3. The network architecture of the proposed feature transverse network (FTNet) is shown, where (a) provides the encoder structure that extracts features at different resolutions, (b) provides the decoder structure designed in a transverse manner that reconstructs a high-resolution semantic map, (c) provides the legend used in FTNet, and (d) visualizes the residual unit proposed by He *et al.* [72].

where Ψ_l is the input feature map for the l^{th} residual layer, ω_l and b_l are the associated set of weights and biases respectively, K denotes the number of weights, Ω denotes the combination of layers CONV $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ CONV $\rightarrow$ BN, $\mathbb{R}$ denotes the ReLU activation function, and I is the identity map which may comprise of weight ω_s when the features maps do not have the same number of channels.

A set of residual units along each resolution form a residual stream. The output of these residual streams can be formulated as $\{\Phi_i | i = 1, 2, 3, 4\}$. These streams are fused in a fully connected fashion to take advantage of information exchange across multi-resolution representations. The integration of multi-resolution features maps $\mathfrak{r}(m)$ at the i^{th} stage is a summation of different feature maps with a corresponding function f. In certain cases, when dilated convolutions are used for semantic segmentation purposes, the last three layers comprise dilated convolutions. To support this mechanism, the corresponding residual streams contain dilated convolution to maintain the same resolution. A broad formulation of both these cases can be defined as shown in (2)

$$\mathcal{D}_i = \sum_{j=0}^{4} f_{ij}(\Phi_{ij}) | \quad i = 0, 1, 2, 3, 4 \tag{2}$$

The function $f_{ij}(\cdot)$ is dependent on feature resolutions. It can be formulated as shown in III-C.

$$f_{ij}(\cdot) = \begin{cases} \Phi_{ij}, & if\ i = jor \mathfrak{r}_i = \mathfrak{r}_j \\ \downarrow (\Phi_{ij}), & if\ (i < j)\ and\ \mathfrak{r}_i < \mathfrak{r}_j \\ \uparrow (\Phi_{ij}) & if\ (i > j)\ and\ \mathfrak{r}_i > \mathfrak{r}_j \end{cases} \tag{3}$$

To downsample ($\downarrow$) by a factor of 4, two strided convolutions with kernel size 3×3 are utilized. For upsampling ($\uparrow$), a resize convolution with a bilinear kernel and a convolutional layer of kernel size 1×1 are utilized. No function is applied when the input and output feature maps' resolutions are identical and along the same residual stream. When the residual stream is different, dilated convolution is applied to maintain the same resolution. Finally, all the feature maps are upsampled to the original resolution and passed through Θ, which comprises CONV $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ CONV with convolution kernel size 1×1.

C. EDGE GUIDANCE (EG)

Thermal image features are coarse due to low resolution and contrast. The object boundaries are ambiguous due to thermal crossover, and the images are noisy due to the design of

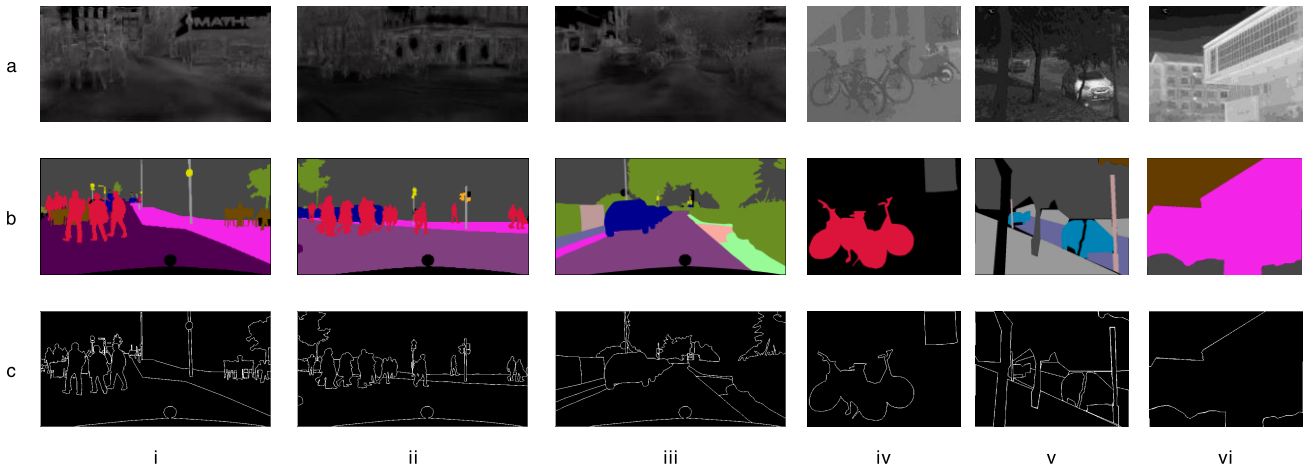


FIGURE 4. Illustration of thermal images with their corresponding masks from the Cityscapes and SODA datasets. Rows [a-c] shows the thermal images with the corresponding semantic maps and edge maps. Columns [i-iii] comprises Cityscapes images converted to the thermal domain, and columns [iv – vi] contain images converted from the SODA dataset.

thermal sensors. Due to these concerns, the reconstruction of the semantic maps generally depends on low-level features and edges details.

Considering these observations, edge map detection is introduced in the decoder section. The edges are extracted from the E_3 layer and passed through CONV $\rightarrow$ BN $\rightarrow$ ReLU. This is upsampled to the original resolution, passed through a CONV layer, and finally appended to the feature maps before applying the function Θ. It should be noted that the edge ground truth is obtained by a simple calculation of the semantic ground truth gradient, which does not require additional labeling effort. A detailed study of edges extracted from various parts of the encoder is provided in further sections.

TABLE 2. Ablation study on the effect of edge guidance applied at different encoder blocks.

Method	Edge Guidance (EG)	mIoU (%)
FTNet ResNet- 50 Filters = 32 Residual units ($\mathfrak{U}$) = 4	No Edge	55.50
	E_1	56.14
	E_2	54.93
	E_3	**56.71**
	E_4	55.99
	E_1, E_4	55.03
	E_2, E_4	54.40
	E_3, E_4	56.10

D. LOSS

An edge-based loss function is employed to ensure the prediction of crisp edges along the boundaries of semantic maps. In edge detection cases, the labels for edges and backgrounds are highly imbalanced. A binary cross-entropy with an adaptive balancing mechanism proposed by Xie and Tu [67] is utilized to overcome this issue. For an image with a ground truth which comprises of Z_+ edge pixels and Z_- background pixels, the prediction $\tilde{p}$ can be formulated as shown in (4).

$$\mathcal{L}_{edge}(\tilde{p}) = -\frac{|Z_-| \sum_{i \in Z_+} \log(\tilde{p}_j)}{|Z_+ \cup Z_-|} - \frac{|Z_+| \sum_{i \in Z_-} \log(1-\tilde{p}_j)}{|Z_+ \cup Z_-|} \tag{4}$$

This loss function handles the class imbalance by providing equal weights irrespective of the ratio of Z_+ and Z_- between the two classes.

A cross-entropy loss defined in (5) is utilized to supervise the semantic maps generated.

$$\mathcal{L} = -\frac{1}{N} \sum_m^N y_m \log(\Gamma_\eta(x_m \mid z) + (1-y_m) \log(1-\Gamma_\eta(x_m \mid z)) \tag{5}$$

where $\Gamma_\eta(x_m \mid z)$ denotes the probability at pixel m with the parameter η, y_m is the ground truth. The total loss adapted to train FTNet is denoted as:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{edge} + \beta \mathcal{L} \tag{6}$$

TABLE 3. Ablation study about the impact of various encoder stage extractions. The number of params and flops are for input size 640 × 480.

Method	Stages	Φ_{ij}	Params (M)	FLOPS (G)	mIoU (%)
FTNet ResNet- 50 Filters = 32 Residual units ($\mathfrak{U}$) = 4	I + $\{E_i \mid i = 0,1,2,3,4\}$	C	25.72	64.58	56.34
		D - C	25.75	161.72	56.88
	$\{E_i \mid i = 0,1,2,3,4\}$	C	25.31	33.98	54.95
		D - C	25.5	134.82	57.11
	$\{E_i \mid i = 1,2,3,4\}$	C	25.31	28.71	55.50
		D - C	25.31	125.68	56.83

Note: I: image, D: Dilation, C: Convolution, and Id: Identity. E_i denotes the features extracted from the encoder stages.

where α and β are continuous hyper-parameters and denote the weights for edges and semantic loss respectively. For experimental results, β was fixed to 1, while the α were varied. More discussions are provided in the later sections. This configuration helps in obtaining refined, spatially consistent, and crisp boundary-located semantic maps.

IV. EXPERIMENTAL RESULTS

This section provides the performance evaluation of the FTNet. After describing the experimental settings, datasets, and training details, the performance comparisons with SOTA methods are provided to demonstrate the effectiveness and generalization ability of the proposed FTNet.

TABLE 4. Impact of utilizing different encoders, filter size, and the number of residual units $\mathfrak{U}$ with the proposed FTNet. the flops are calculated for images of input size 640 × 480.

Encoder	Filters	Residual Units ($\mathfrak{U}$)	Params (M)	FLOPs (G)	mIoU (%)
DS ResNet - 50	32	4	25.44	41.39	55.99
	64	4	27.77	57.9	57.45
ResNest - 50	32	4	27.25	40.96	55.16
	64	4	29.57	57.47	57.98
ResNet – 50 [72]	32	4	25.32	33.04	56.71
	64	4	27.65	49.55	58.82
	128	2	32.13	93.63	59.40
ResNeXt – 50 [77]	32	4	26.64	33.97	57.07
	64	4	28.96	50.48	59.45
	128	2	33.44	94.55	59.94
DS ResNet – 101*	128	2	44.44	59.92	59.89
ResNest – 101	128	2	48.04	66.37	58.89
ResNet – 101 [72]	128	2	44.31	51.57	60.84
ResNeXt – 101 [77]	128	2	90.60	104.58	62.67

*DS denotes deep stem

A. DATASET

For training and testing purposes, the SODA dataset [11] was employed. It comprises 2,168 annotated images among which, 1,168 images are used for training, and 1,000 images are used for testing. Due to the scale of this dataset, synthetic images were utilized for pretraining the network. These synthetic images were obtained by translating the Cityscapes [73] dataset from RGB space to thermal space. The original Cityscapes dataset contained 5,000 images, including 2,975 training images, 500 validation images, and 1,525 test images. Following the training protocol described in [11], all the training, validation, and testing images were combined for pretraining purposes. As the dataset does not comprise the ground truth edge maps, they were generated following the strategy used in [74]. In this protocol, the ground truth semantic maps are utilized to generate edges for each class. However, as FTNet incorporates only binary information, the protocol was adapted to produce binary edge maps instead of the generating edges for each class. An example of set of images utilized for training are provided in Figure 4.

B. TRAINING DETAILS

For training the network, a progressive learning algorithm is adopted. Initially, the encoder parameters are loaded with pretrained ImageNet weights. For pretraining the model with thermal features, synthetic thermal Cityscapes images were utilized. The thermal input images were augmented by performing random cropping (from 640 × 480 to 480 × 480), and random scaling in the range of (0.5, 2), and performing random horizontal flipping along with the corresponding masks. The SGD optimizer [75] with base learning rate 0.01, momentum 0.9, and weight decay 0.0001 is employed to train the model. As the decoder section is randomly initialized, the learning rate of this network is increased by a factor of 10. The batch size was set to 16. The network was trained for 100 epochs, and poly learning rate policy with the power of 0.9 is used to drop the learning rate.

As the Cityscapes dataset labels are different from the SODA dataset, the last layer of the network is discarded after pretraining and adjusted to match the number of classes from the SODA. The protocol for training remains the same except for the initial learning rate, which is dropped by a factor of 10.

The experiments were conducted on the PyTorch platform [76]. The models were trained on 2 V100 GPUs, and it takes around 12 hours to complete FTNet training (Cityscapes pretrain and SODA training). To evaluate FTNet, Intersection over Union (IoU), also known as the Jaccard Index, is utilized. It provides the ratio of the intersection of the pixel-wise classification results with the ground truth to their union. A higher percentage depicts how close the predicted class maps are to the ground truth.

C. ABLATION STUDIES

Extensive ablation studies on the various architectural components of FTNet architecture were performed. The ResNet 50 architecture was utilized as the encoder for baseline results. The number of filters in the decoder was set to 32, and the number of residual units $\mathfrak{U}$ for each residual stream Φ_i was set to 4. Equal weights ($\alpha = \beta = 1$) were provided to both edge and semantic maps for all the ablation studies except for loss weight analysis.

1) IMPACT OF ENCODER STAGES EXTRACTIONS AND DILATION

This ablation study comprises evaluating the model's performance when features are extracted from various encoder stages and an analysis of dilation. The results are provided in Table 3. Initially, all the resolution scales, including the image space was utilized to reconstruct semantic maps. Gradually, the image space was discarded, and further analysis led to the elimination of features from E_0 stage. For semantic segmentation techniques, down-sampling may cause loss of spatial information; however, it is required to understand the scenes and reconstruct the semantic maps with finer details.

Excluding down-sampling may increase resolution; however, it affects the receptive field in subsequent layers, increasing context loss. To overcome this, dilated convolutions were employed to adjust receptive fields of feature points without decreasing the resolution of feature maps [30].

Replacing strided convolutions with dilation-based convolution in FTNet provided superior results for all three cases. However, there exists a trade-off between the accuracy and the number of FLOPs. The models in all three cases have the same number of parameters for tuning, but FTNet with dilated convolutions have approximately 100 G extra FLOPS in all cases. As indicated by Minaee *et al.* [41], an ideal model should consider multiple aspects, which include quantitative accuracy, speed (inference time), and storage requirements (memory footprint). Following this suggestion, FTNet aims at decreasing the FLOPS, thereby decreasing inference time while achieving higher accuracy. As FTNet with $\{E_i|i = 1, 2, 3, 4\}$ without dilation has lower FLOPs with acceptable accuracy, the rest of the ablation study utilizes the strided convolution in the model.

2) IMPACT OF EDGE GUIDANCE (EG)

Considering that thermal images are generally blurry and lack color information when compared to the visible domain, low-level features and edge details are crucial for generating semantic maps. In deep CNNs, there is a trade-off between semantics and resolution at the low-level and high-level layers. This trade-off is quantitatively shown in Table 2. Edges were extracted from each stage $\{E_i|i = 1, 2, 3, 4\}$ and various combinations of E_i were investigated. It can be verified that extracting edges from E_3 the stage provides the best mIoU. Furthermore, the accuracy increased by 1.21% from 55.5% to 56.71%. This confirms that high-level semantics with sufficient resolution provides better edges. This significant improvement proves that the edge guidance increases the learning effectiveness of a neural network by capturing varying structures to encode meaningful features.

On the contrary, the combination of edges from various encoder stages does not perform as expected for thermal imagery. This is because initial encoder stages have poor semantic and high-level layers, especially E_4, which has rich semantics, but low resolution. This was validated by examining the E_1 and E_4 combination. The edge extracted from E_1 provided 56.14% and E_4 provided 55.99%, while the combination of both resulted in 55.03%, which is subpar.

3) IMPACT OF ENCODER TOPOLOGIES

The successful use of CNNs in image classification tasks has accelerated the research in architectural design. Since then, numerous network architectures have been proposed to address this task. Typically, these networks are used as encoders for complex tasks such as object detection, classification, and semantic segmentation. This section aims at evaluating the performance of the FTNet decoder with various encoder architectures on the SODA dataset. For a fair comparison, all the decoder components of the FTNet network were fixed, and only the encoder was replaced. The results of this study are provided in Table 4. It can be seen that the mIoU scores of deep stem ResNet-based architecture underperformed while ResNet and ResNeXt provided superior mIoU results when the filter sizes were set to 32 and 64. The ResNet and ResNeXt models were further investigated with 128 filter sizes and two residual units. The ResNeXt model using this setting outperformed other architectures from the ResNet family. It indicates that the inclusion of cardinality is of paramount importance to achieve better semantic maps. This feature of the ResNeXt model is more effective and of central importance, in addition to the dimensions of width and depth.

4) IMPACT OF EDGE LOSS WEIGHTS

As the defined loss from equation IV) comprises of two components, namely, semantic loss and edge guidance loss, it is necessary to adapt them to the same order of magnitude to obtain optimum results. A small edge loss weight may lead to a failure of edge supervision, while a large weight may dominate the semantic loss. It is necessary to optimize them as semantic loss is always higher than edge loss. In this ablation study, different α values were used to empirically determine the best edge loss weight. Table 5 provides the complete set of variations and their corresponding mIoU scores.

TABLE 5. Analysis of hyperparameter-α weights in the loss function with $\beta = 1$. FTNET - ResNeXt – 50 with 128 filter size and 2 residual units (U) was utilized for this ablation study.

α	mIoU (%)		
	SODA Dataset	MFN Dataset	SCUT-Seg Dataset
1	59.61	46.52	65.43
5	59.77	46.91	66.29
10	60.05	46.90	66.40
15	59.61	46.98	66.59
20	**60.08**	**47.12**	**66.73**
30	60.04	47.08	66.8

When setting $\alpha = \beta = 1$, the boundaries are not crisp when compared to the results obtained with $\alpha = 20$ and $\beta = 1$. This discrepancy explains that setting the magnitude of different losses is very crucial to gain better accuracy. From the table, it can be determined that $\alpha = 20$ provides superior accuracy.

D. BENCHMARK RESULTS

To show the effectiveness of the proposed FTNet, it is compared with SOTA methods such as FCN [42], UNet [44] with ResNet based encoder, PSPNet [29], HRNet [31], ICNet [46], UNet++ [50], PAN [48], LinkNet [47], FPN [45], ENCNet [51], DANet [49], PSPNet [29], PSANet [78], SegNet [43], and MCNet [28]. For all these comparisons, FTNet with ResNeXt backbone was utilized. The hyperparameters

TABLE 6. Performance comparison in terms of per class IOU with state-of-the-art methods on the soda test dataset. the proposed ftnet uses resnext – 50, filters = 128, α =20, and $\mathfrak{U}$-2.

Classes	HRNet	UNet++	DeepLabv3	PSPNet	MCNet	FTNet
background	52.31%	51.79%	53.72%	54.30%	47.81%	53.71%
person	71.84%	72.33%	70.22%	71.20%	66.02%	73.83%
building	74.88%	71.97%	74.79%	74.05%	68.15%	73.32%
tree	72.40%	71.83%	71.92%	72.73%	69.46%	72.96%
road	79.19%	78.26%	80.58%	79.24%	77.41%	80.42%
pole	35.78%	35.17%	31.87%	33.40%	27.77%	37.88%
grass	47.44%	46.84%	47.99%	47.60%	43.85%	49.04%
door	32.66%	25.51%	36.59%	35.08%	26.84%	36.61%
table	11.50%	3.20%	17.04%	20.86%	10.36%	14.17%
chair	39.35%	34.27%	42.79%	39.66%	32.40%	43.35%
car	80.93%	81.17%	80.93%	80.35%	73.16%	81.38%
bicycle	50.55%	51.85%	49.95%	47.32%	39.71%	55.12%
lamp	67.01%	42.22%	62.29%	65.89%	34.98%	68.57%
monitor	64.34%	61.28%	62.02%	62.39%	54.44%	68.72%
traffic cone	54.78%	53.25%	50.65%	54.01%	38.70%	59.49%
trash can	62.63%	44.29%	67.54%	64.47%	53.17%	63.46%
animal	58.82%	45.16%	55.83%	61.91%	47.91%	60.73%
fence	58.33%	50.93%	56.36%	57.94%	48.79%	55.74%
sky	86.88%	84.07%	85.50%	86.93%	81.55%	85.56%
river	74.01%	62.45%	75.65%	77.76%	67.42%	75.05%
sidewalk	49.31%	48.86%	51.45%	49.24%	46.76%	52.76%
mIoU	58.33%	53.18%	58.37%	58.87%	50.32%	60.09%

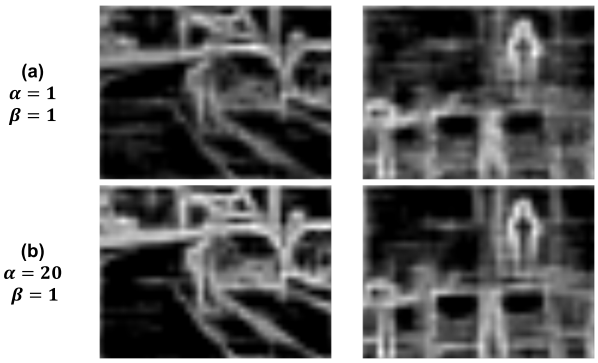


FIGURE 5. Illustration of edges obtained under different α and β conditions. Edges obtained in panel [b] have higher intensity with sharper boundaries when compared to panel [a]. This difference enables better segmentation with crisp boundaries.

used were filters = 128, residual units = 2, α = 20, and $\beta = 1$.

The performance was tested on the MFNet dataset [26], SODA dataset [11], and SCUT-Seg dataset [28] to demonstrate the generalization ability of FTNet. MFNet dataset is a public semantic segmentation dataset based on the driving scene with RGB-T images. It comprises 1569 images divided into 784, 392, and 393 images for training, validation, and testing. The SCUT-Seg dataset is a set of thermal images collected with different driving scenarios. This was an extension of the SCUT FIR Pedestrian Dataset [79]. It consists of 1345 training and 665 testing images. Furthermore, pretrained weights of cityscapes data were utilized as initial parameters. The rest of the training details remain the same.

For better understanding, the per-class IoU is provided in Table 6. The complete set of mIoU for all datasets is provided in Table 7. These results demonstrate the performance of FTNet when compared to the SOTA methods. FTNet achieves 60.09% accuracy on the SODA dataset while the top four SOTA methods were MCNet, HRNet, PSPNet, and DeepLabV3 with 50.32%, 58.33%, 58.87%, and 58.37%, respectively. Furthermore, FTNet's performance on MFN and SCUT-Seg datasets is notable when compared to SOTA.

Qualitative evaluation is essential in image segmentation assessment, and the segmentation maps of FTNet and SOTA are assessed based on the human visual system. This analysis is done using humans as an observer [80]–[84]. Human visual analysis is critical in identifying characteristics of algorithms that quantitative metrics may not identify correctly. For instance, in a less detailed or incorrectly labeled dataset, quantitative metrics will automatically penalize segmentation algorithms for correctly segmenting the object.

The qualitative comparisons are provided in Figure 6 and Figure 7. These examples show complex indoor and outdoor scenarios with numerous object instances in multiple scales and partial occlusion. These scenes were also captured with diverse lighting conditions, including day and night. These

TABLE 7. Semantic segmentation results on the different datasets. The number of params and flops are calculated for input size 640 x 480. blue text indicates the best and red text indicates the second-best performance.

Model	Backbone	Params (M)	FLOPS (G)	mIoU (%)		
				SODA Dataset	MFN Dataset	SCUT-Seg Dataset
LinkNet [47]	ResNet-18	11.59	14.94	25.94	24.80	36.69
FPN [45]	ResNet -50	26.11	36.76	50.24	41.20	53.67
ICNet [46]	Dilated-ResNet- 50	26.33	40.71	51.02	40.53	55.47
PAN [48]	Dilated-ResNet- 50	24.26	40.94	37.72	29.34	40.13
UNet* [44]	ResNet -50	32.52	50.99	50.96	39.03	57.25
HRNet [31]	HRNet-32	29.54	53.33	58.33	46.48	**65.23**
DeepLabv3 ** [30]	Dilated-ResNet- 50	39.84	56.96	58.37	45.74	59.97
FCN-32s** [42]	VGG-16	15.30	94.34	46.57	37.12	42.86
ENCNet [51]	Dilated-ResNet- 50	33.71	163.87	56.00	44.17	60.62
SegNet [43]	VGG-16	29.49	203.1	48.90	43.59	55.28
PSPNet [29]	Dilated-ResNet- 50	46.79	207.99	**58.87**	**46.51**	59.94
DANet [49]	Dilated-ResNet- 50	47.66	232.25	50.61	39.26	50.53
UNet++ [50]	ResNet - 50	48.98	270.57	53.18	43.59	58.04
MCNet [28] ***	Dilated-ResNet- 50	35.67	131.87	50.32	42.28	56.87
FTNet –Filters 128, $\mathfrak{U}$=2, α = 20, β =1	ResNeXt – 50 [77]	33.44	94.55	**60.08**	**47.12**	**66.73**

* indicates that a different encoder was utilized when compared to the original, ** indicates that the methods were re-implemented, *** indicates the network was specifically developed for thermal images.

outdoor examples show challenging scenarios with various objects, such as cars and pedestrians, in close proximity to each other and far apart. FTNet effectively addresses these challenges and yields reliable semantic maps. Figure 6 panels [a,c] and Figure 7 panels [a,b] show that the person class and other objects have well-defined edges compared to SOTA methods. Figure 6 panel c shows that the two pedestrians near the car have crisp boundaries with finer labels representing the input.

Similarly, in Figure 6 panel b, the car has better edges, and FTNet and PSPNet, DANet, and UNet++ could detect the pole. However, FTNet was able to detect the pole with higher similarity to the input image. In Figure 7 panel b, the shape of the person and monitors were more clearly defined when compared to others. In Figure 7 panel c, most of the SOTA results were erroneous, but FTNet had a clear semantic map of chairs and tables. Overall, FTNet yielded acceptable results even though the SODA dataset had few indoor data representations. The proposed network reconstructs semantic maps with a higher correlation to the ground truth despite the poor quality of the thermal images. These examples illustrate the ability of the FTNet to perform better in circumstances where ambiguous object boundaries are introduced by thermal crossover compared to the other models.

Since the processing time of CNN-based semantic segmentation tasks is crucial, the inference speed of the network is computed and tabulated. A 640×480 image was run through the network 300 times, and then the average of the results was considered a single run to calculate the computation time. This experiment was repeated ten different times, and the average time of these ten runs is provided as the inference time in Table 8. This table provides the runtime result of different approaches without any optimizations and provides the number of parameters. Runtime was measured on an Intel i9-9900K 3.60GHz CPU system and an Nvidia RTX 2080 Ti GPU. The simulation results show that the FTNet's runtime performance is comparable to other SOTA methods. In terms of the number of parameters and FLOPS, this model has less memory overhead (-2.23M) and calculation (-37.42G Flops) when compared to MCNet. Even though the number of parameters is slightly higher than HRNet by 3.9M, the inference time is comparable and provides better accuracy. These observations demonstrate the potential of FTNet for application on edge devices and intelligent systems such as automated driving and video surveillance applications.

E. DISCUSSION

To the best of the authors' knowledge, the most similar works to the FTNet are MCNet [28] and HRNet [31].

MCNet introduced multiple structures to preserve boundary information rather than post-fine-tuning the semantic segmentation results. They utilize two feature representations E_1 and E_4 (see Figure 3 (a)) from a dilated encoder network. It employs a loss function that spans across multiple levels of a correlation matrix correction module. However, FTNet does not use dilated networks, thus reducing the number of parameters. FTNet exploits all the feature representations $\{E_i | i = 1, 2, 3, 4\}$ in a transverse structure to aid the network in producing high-quality semantic maps. Furthermore, a novel edge guidance mechanism is developed to produce crisp boundaries. Finally, a weighted loss function is explored to ensure that the edge and semantic losses have the same

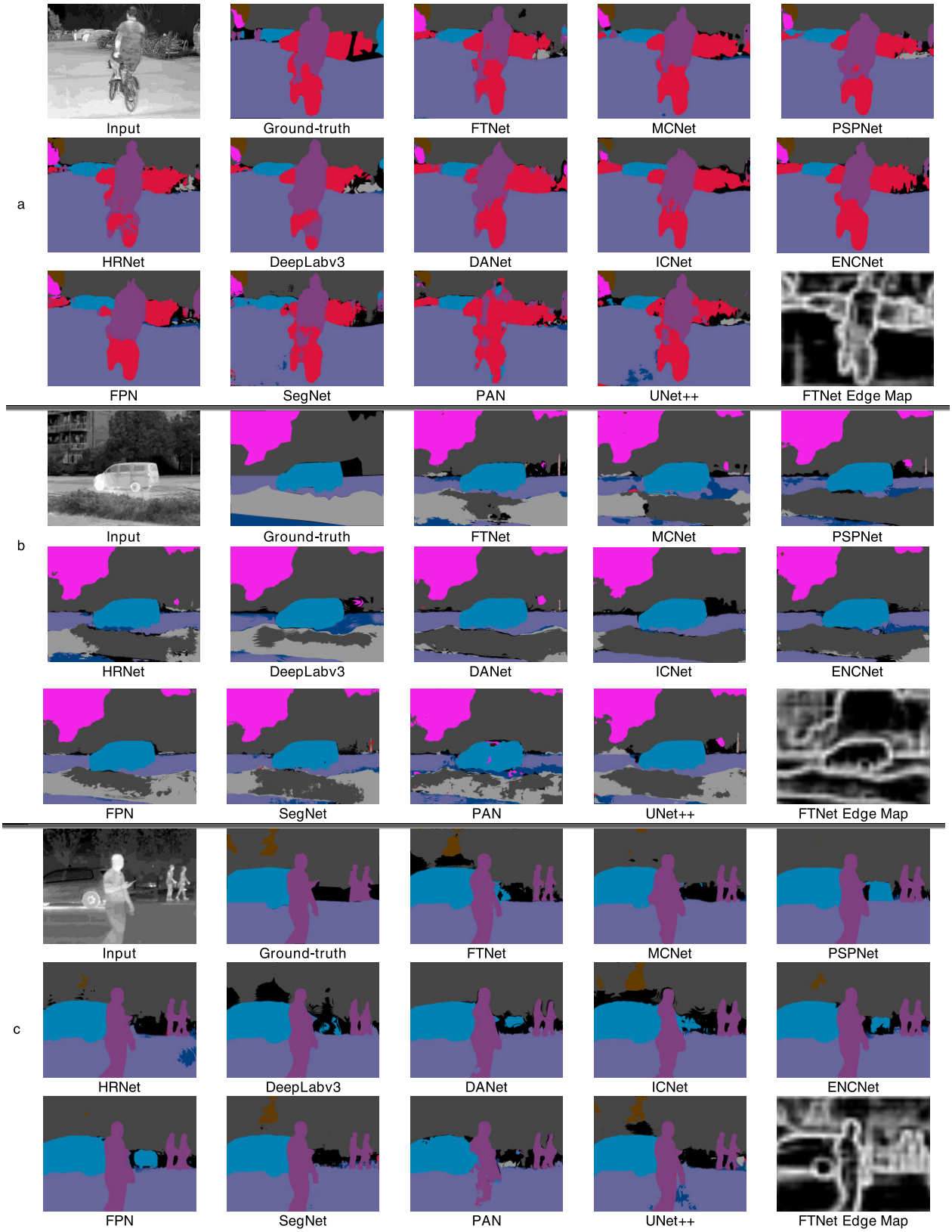


FIGURE 6. Qualitative comparison of thermal image semantic segmentation of outdoor environments with SOTA methods. In addition to the segmentation output, the edge map reconstructed from FTNet is provided. It can be seen from the images that FTNet has provided clear boundaries when compared to SOTA methods. In panel (b), the car has finer boundaries near the tires and the pole was detected with clear distinction even though they were missing in the ground truth. In panels (a) and (c), the person class was segmented more finely while SOTA methods had ambiguous maps.

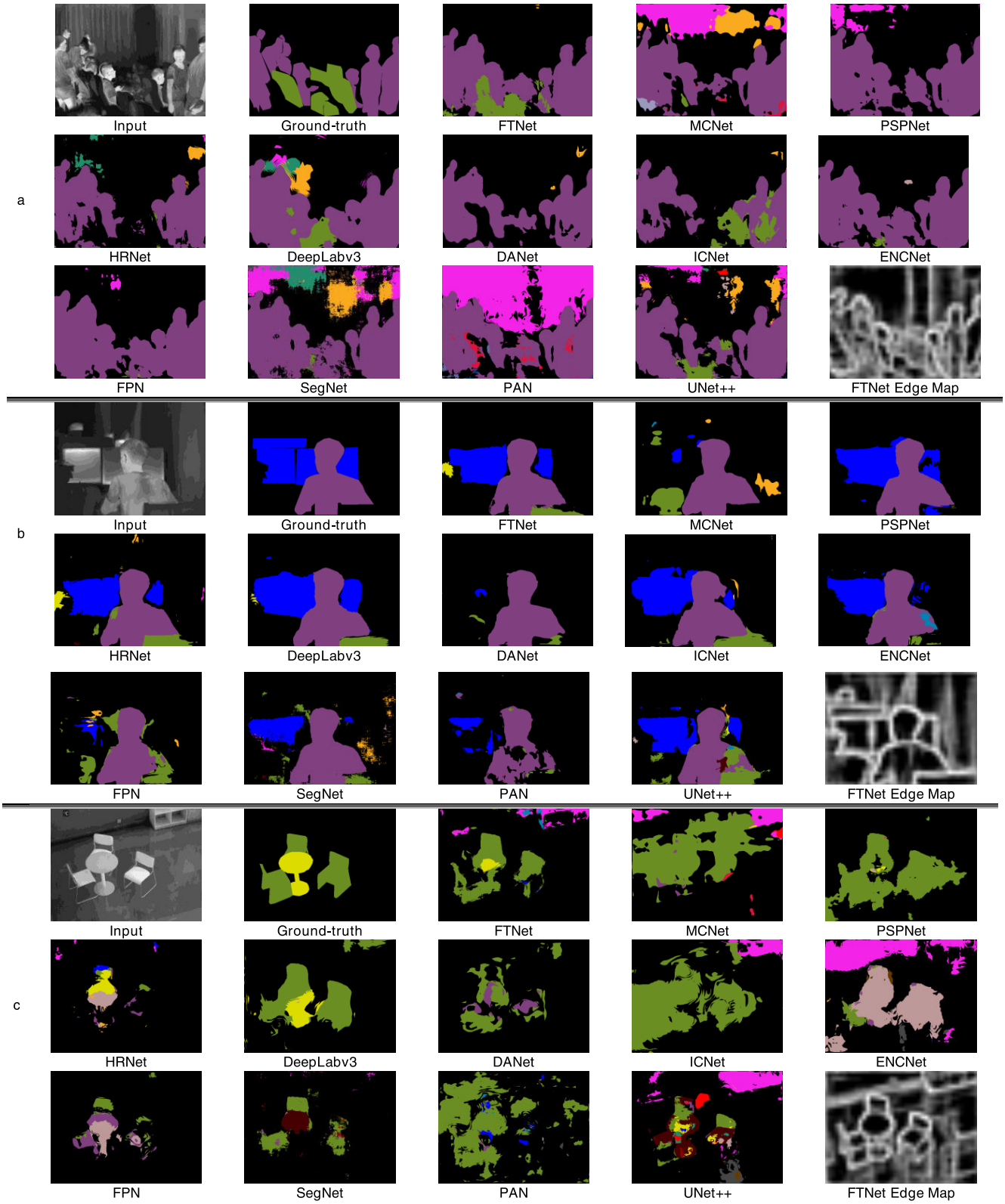


FIGURE 7. Qualitative comparison of thermal image semantic segmentation of indoor environments with SOTA methods. In addition to the segmentation output, the edge map reconstructed from FTNet is provided. In panel (a), the person and chair class are close to the ground truth, while other SOTA methods except DeepLabv3 fail to detect the chair. In panel (b), FTNet could detect monitors with much better boundaries while SOTA methods lack distinction and clarity. Panel (c) FTNet predicted the chairs more accurately while SOTA methods have erroneous results.

TABLE 8. Average execution time of 300 simulations on a 640 × 480 image.

Model	Params (M)	FLOPS (G)	Time (in ms)
LinkNet [47]	11.59	14.94	5.69
FPN [45]	26.11	36.76	12.80
UNet* [44]	32.52	50.99	15.86
FCN-32s** [42]	15.30	94.34	17.41
DeepLabv3 ** [30]	39.84	56.96	18.45
PAN [48]	24.26	40.94	18.87
ICNet [46]	26.33	40.71	20.51
SegNet [43]	29.49	203.1	39.68
HRNet [31]	29.54	53.33	42.51
ENCNet [51]	33.71	163.87	47.92
UNet++ [50]	48.98	270.57	50.64
PSPNet [29]	46.79	207.99	51.14
DANet [49]	47.66	232.25	59.12
MCNet [28]	35.67	131.87	153.57
FTNet- ResNeXt – 50, F-128, $\mathcal{U}$ -2	33.44	94.55	42.78

order of magnitude. This loss function preserves the boundary information along with accurate semantic maps.

HRNet extracts features from high-resolution feature maps in parallel with the low-resolution feature maps. The extracted feature maps from multiple parallel streams are fused to obtain high-resolution representations. However, the encoder network is not interchangeable with existing encoder backbones such as VGG, ResNet, and ResNeXt.

On the contrary, FTNet is carefully designed to transverse through multiple streams of existing serially connected encoder networks. This mechanism exploits all the feature maps at various resolutions, including the low-level feature maps containing high semantic information. The introduction of the edge guidance counterpart into this network has an immense impact on the performance shown in both quantitative and qualitative analysis. Existing SOTA encoder networks such as Xception [85], DenseNet [86], and MobileNet [87] can be repurposed with the decoder of FTNet for various applications, including image denoising and recoloring. Furthermore, the decoder in FTNet can be readily extended to incorporate the outputs of dilated convolution in applications where it is necessary to preserve the resolution.

The benchmarking datasets used in this article [11], [26], [28] have promoted the research of semantic segmentation using thermal images. However, the annotations provided in these datasets are coarse and less detailed compared to RGB datasets. This is most likely due to the extremely low inter-class variance of objects in thermal images, making accurate labeling near boundaries difficult. Additionally, some of the labels in the datasets were misclassified in SODA Dataset. The challenges mentioned above can be visualized in Figure 6 and Figure 7 (input and ground truth). Moreover, the distribution of the semantic labels is highly imbalanced among these benchmark datasets. For example, the SODA dataset comprises 1,304 road images, whereas the number of monitor images is 75. These issues lead to a negative impact on the performance of the proposed and SOTA methods.

V. CONCLUSION

In this work, a novel deep learning-based semantic segmentation network, FTNet was presented. This network aims at exploring the multi-resolution representation to perform pixel-wise classification accurately. The proposed FTNet is an end-to-end trainable architecture with ResNeXt encoder and employs a novel transverse-based decoder network, efficient in terms of parameters/operations and computation time. This transverse-based network captures discriminative features from multiple resolutions and combines them in a fully connected fashion to achieve semantic maps close to the ground truth. An edge guidance mechanism is proposed to overcome the poor quality, single-channel, and blurry object boundary attributes of thermal images. The introduction of weighted loss further improves spatial boundary information and reduces semantic ambiguity. Extensive quantitative analysis demonstrated that FTNet achieved mIoU of 60.08%, 47.12%, and 66.73% on SODA, MFN, and SCUT-Seg Dataset with 33.44M parameters and 94.55G FLOPS. Furthermore, the qualitative analysis showed that FTNet reconstructed rich semantic maps with crisp boundaries. These results show that FTNet can potentially optimize thermal image perception in intelligent systems such as automated driving and video surveillance applications, computational photography, biomedical analysis, and augmented reality.

As a part of future work, the authors intend to explore dilated convolution with reduced parameters and check the system's performance on RGB datasets. Furthermore, FTNet will be tested on other position-sensitive vision applications, such as facial landmark detection, image super-resolution, image recoloring, and image denoising. Another promising future work will be applying the proposed model in other domains such as hyperspectral imaging.

REFERENCES

[1] R. Szeliski, *Computer Vision: Algorithms and Applications*. New York, NY, USA: Springer, 2010.
[2] H. Kim, J. Koo, D. Kim, B. Park, Y. Jo, H. Myung, and D. Lee, "Vision-based real-time obstacle segmentation algorithm for autonomous surface vehicle," *IEEE Access*, vol. 7, pp. 179420–179428, 2019.
[3] B. Olimov, J. Kim, and A. Paul, "REF-Net: Robust, efficient, and fast network for semantic segmentation applications using devices with limited computational resources," *IEEE Access*, vol. 9, pp. 15084–15098, 2021.
[4] W. Wang, Y. Fu, Z. Pan, X. Li, and Y. Zhuang, "Real-time driving scene semantic segmentation," *IEEE Access*, vol. 8, pp. 36776–36788, 2020.
[5] B. Li, Y. Shi, Z. Qi, and Z. Chen, "A survey on semantic segmentation," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2018, pp. 1233–1240.
[6] N. Salamati, D. Larlus, G. Csurka, and S. Süsstrunk, "Semantic image segmentation using visible and near-infrared channels," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 461–471.
[7] P. Yin, R. Yuan, Y. Cheng, and Q. Wu, "Deep guidance network for biomedical image segmentation," *IEEE Access*, vol. 8, pp. 116106–116116, 2020.

[8] F. Meng, L. Guo, Q. Wu, and H. Li, "A new deep segmentation quality assessment network for refining bounding box based segmentation," *IEEE Access*, vol. 7, pp. 59514–59523, 2019.

[9] Y. Xu, S. Arai, F. Tokuda, and K. Kosuge, "A convolutional neural network for point cloud instance segmentation in cluttered scene trained by synthetic data without color," *IEEE Access*, vol. 8, pp. 70262–70269, 2020.

[10] I. Cabrilo, P. Bijlenga, and K. Schaller, "Augmented reality in the surgery of cerebral aneurysms: A technical report," *Oper. Neurosurg.*, vol. 10, no. 2, pp. 252–261, Jun. 2014.

[11] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3069–3082, Jul. 2021.

[12] T. Singha, D.-S. Pham, and A. Krishna, "FANet: Feature aggregation network for semantic segmentation," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov./Dec. 2020, pp. 1–8.

[13] S. Rajeev, Q. Wan, K. Yau, K. Panetta, and S. Agaian, "Augmented reality-based vision-aid indoor navigation system in GPS denied environment," *Proc. SPIE Mobile Multimedia/Image Process., Secur., Appl.*, vol. 10993, May 2019, Art. no. 109930P.

[14] K. J. Havens and E. J. Sharp, "Imager selection," in *Thermal Imaging Techniques to Survey and Monitor Animals in the Wild*, K. J. Havens and E. J. Sharp Eds. Boston, MA, USA: Academic, 2016, pp. 121–141.

[15] Research and Market. (Apr. 2021). *Global Thermal Imaging Market for the Mobility Industry 2020-2025: Analysis of Applications, Products and Countries*. [Online]. Available: https://www.researchandmarkets.com/reports/5315057/

[16] E. M. Zaihidee, K. H. Ghazali, and A. A. Almisreb, "Comparison of human segmentation using thermal and color image in outdoor environment," in *Proc. IEEE Conf. Syst., Process Control (ICSPC)*, Dec. 2015, pp. 152–156.

[17] S. Rajeev, S. K. KM, Q. Wan, K. Panetta, and S. S. Agaian, "Illumination invariant NIR face recognition using directional visibility," *Electron. Imag.*, vol. 2019, no. 11, pp. 273-1–273-7, 2019.

[18] S. K. Biswas and P. Milanfar, "Linear support tensor machine with LSK channels: Pedestrian detection in thermal infrared images," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4229–4242, Sep. 2017.

[19] J. Ge, Y. Luo, and G. Tei, "Real-time pedestrian detection and tracking at nighttime for driver-assistance systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 2, pp. 283–298, Jun. 2009.

[20] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5743–5756, Dec. 2016.

[21] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 808–823.

[22] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, and X. Yuan, "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 509–520, Mar. 2020.

[23] L. Kezebou, V. Oludare, K. Panetta, and S. Agaian, "TR-GAN: Thermal to RGB face synthesis with generative adversarial network for cross-modal face recognition," *Proc. SPIE Mobile Multimedia/Image Process., Secur., Appl.*, vol. 11399, Apr. 2020, Art. no. 113990P.

[24] S. Kamath K. M., R. Rajendran, Q. Wan, K. Panetta, and S. Agaian, "TERNet: A deep learning approach for thermal face emotion recognition," *Proc. SPIE*, vol. 10993, May 2019, Art. no. 1099309.

[25] Q. Wan, S. P. Rao, A. Kaszowska, V. Voronin, K. Panetta, H. A. Taylor, and S. Agaian, "Face description using anisotropic gradient: Thermal infrared to visible face recognition," *Proc. SPIE Mobile Multimedia/Image Process., Secur., Appl.*, vol. 10668, May 2018, Art. no. 106680V.

[26] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5108–5115.

[27] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," 2020, *arXiv:2007.05558*.

[28] H. Xiong, W. Cai, and Q. Liu, "MCNet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene," *Infr. Phys. Technol.*, vol. 113, Mar. 2021, Art. no. 103628.

[29] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

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

[31] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, and Y. Mu, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2020.

[32] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[33] R. Rajendran, K. Panetta, and S. Agaian, "A human visual based binarization technique for histological images," *Proc. SPIE Mobile Multimedia/Image Process., Secur., Appl.*, vol. 10221, May 2017, Art. no. 102210Q.

[34] K. Piniarski and P. Pawlowski, "Segmentation of pedestrians in thermal imaging," in *Proc. Baltic URSI Symp. (URSI)*, May 2018, pp. 210–211.

[35] R. Nock and F. Nielsen, "Statistical region merging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1452–1458, Nov. 2004.

[36] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm," in *Procedia Computer Science*, vol. 54. Amsterdam, The Netherlands: Elsevier, 2015, pp. 764–771.

[37] L. Najman and M. Schmitt, "Watershed of a continuous function," *Signal Process.*, vol. 38, no. 1, pp. 99–112, Jul. 1994.

[38] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.

[39] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[40] W. Song, N. Zheng, R. Zheng, X. Zhao, and A. Wang, "Digital image semantic segmentation algorithms: A survey," *J. Inf. Hiding Multim. Signal Process.*, vol. 10, no. 1, pp. 196–211, 2019.

[41] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 17, 2021, doi: 10.1109/TPAMI.2021.3059968.

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

[43] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

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

[45] A. Kirillov, K. He, R. Girshick, and P. Dollar. (2017). *A Unified Architecture for Instance and Semantic Segmentation*. [Online]. Available: http://presentations.cocodataset.org/COCO17-Stuff-FAIR.pdf

[46] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 405–420.

[47] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[48] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*.

[49] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[50] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multi-Modal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[51] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[52] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

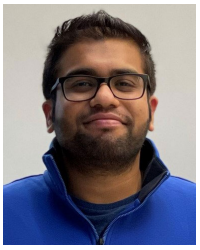
[53] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

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

[55] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*.

[56] B. Artacho and A. Savakis, "Waterfall atrous spatial pooling architecture for efficient semantic segmentation," *Sensors*, vol. 19, no. 24, p. 5361, Dec. 2019.
[57] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[58] S. Seferbekov, V. Iglovikov, A. Buslaev, and A. Shvets, "Feature pyramid network for multi-class land segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 272–275.
[59] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.
[60] Z. Xu, W. Zhang, T. Zhang, and J. Li, "HRCNet: high-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 13, no. 1, p. 71, Dec. 2020.
[61] K. A. Panetta, E. J. Wharton, and S. S. Agaian, "Logarithmic edge detection with applications," *J. Comput.*, vol. 3, no. 9, pp. 11–19, Sep. 2008.
[62] S. S. Agaian, K. A. Panetta, S. C. Nercessian, and E. E. Danahy, "Boolean derivatives with application to edge detection for imaging systems," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 40, no. 2, pp. 371–382, Apr. 2010.
[63] P. Kohli, L. Ladický, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 302–324, May 2009.
[64] A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr, "Higher order conditional random fields in deep neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 524–540.
[65] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3376–3385.
[66] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li, "ERN: Edge loss reinforced semantic segmentation network for remote sensing images," *Remote Sens.*, vol. 10, no. 9, p. 1339, Aug. 2018.
[67] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.
[68] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, pp. 161–171, Jan. 2019.
[69] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, Jul. 2019.
[70] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "PST900: RGB-thermal calibration, dataset and segmentation network," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 9441–9447.
[71] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[73] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[74] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam, "CASENet: Deep category-aware semantic edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5964–5973.
[75] N. Ketkar, "Stochastic gradient descent," in *Deep Learning With Python*. Springer, 2017, pp. 113–132.
[76] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, and Z. Lin, "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.
[77] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.
[78] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 267–283.
[79] Z. Xu, J. Zhuang, Q. Liu, J. Zhou, and S. Peng, "Benchmarking a large-scale FIR dataset for on-road pedestrian detection," *Infr. Phys. Technol.*, vol. 96, pp. 199–208, Jan. 2019.
[80] S. Z. Syed Zaini, N. N. Sofia, M. Marzuki, M. F. Abdullah, K. A. Ahmad, I. S. Isa, and S. N. Sulaiman, "Image quality assessment for image segmentation algorithms: Qualitative and quantitative analyses," in *Proc. 9th IEEE Int. Conf. Control Syst., Comput. Eng. (ICCSCE)*, Nov. 2019, pp. 66–71.
[81] X. Zhang, P. Ye, and G. Xiao, "VIFB: A visible and infrared image fusion benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 104–105.
[82] G. Dong, Y. Yan, C. Shen, and H. Wang, "Real-time high-performance semantic image segmentation of urban street scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3258–3274, Jun. 2021.
[83] D. Nie and D. Shen, "Adversarial confidence learning for medical image segmentation and synthesis," *Int. J. Comput. Vis.*, vol. 128, nos. 10–11, pp. 2494–2513, Nov. 2020.
[84] S. Qiu, Y. Zhao, J. Jiao, Y. Wei, and S. Wei, "Referring image segmentation by generative adversarial learning," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1333–1344, May 2020.
[85] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[86] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[87] 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," 2017, *arXiv:1704.04861*.

KAREN PANETTA (Fellow, IEEE) received the B.S. degree in computer engineering from Boston University, Boston, MA, USA, and the M.S. and Ph.D. degrees in electrical engineering from Northeastern University, Boston. She is currently the Dean of graduate engineering education and a Professor with the Department of Electrical and Computer Engineering. She is also an Adjunct Professor of computer science with Tufts University, Medford, MA, USA, and the Director of the Dr. Panetta's Vision and Sensing System Laboratory. Her research interests include developing efficient algorithms for simulation, modeling, signal, and image processing for biomedical and security applications. She was a recipient of the 2012 IEEE Ethical Practices Award and the Harriet B. Rigas Award for Outstanding Educator. In 2011, she was awarded the Presidential Award for Engineering and Science Education and Mentoring by U.S. President Obama. She is the Vice President of SMC, Membership, and Student Activities. She was the President of the 2019 IEEE-HKN. She is the Editor-in-Chief of the *IEEE Women in Engineering Magazine*. She was the IEEE-USA Vice-President of communications and public affairs. From 2007 to 2009, she served as the Worldwide Director for the IEEE Women in Engineering, overseeing the world's largest professional organization supporting women in engineering and science.



SHREYAS KAMATH K. M. (Member, IEEE) received the bachelor's degree (B.E.) in electronics and communication engineering from Visvesvaraya Technological University, Belgaum, India, in 2014, and the M.S. degree in electronic and computer engineering from The University of Texas at San Antonio, USA. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Tufts University, USA. He is working as a Graduate Research Assistant with the Visual and Sensing Laboratory, Tufts. His main areas of research interests include signal/image processing, deep learning, computer vision, 3D scanning, and automated biometric technologies, particularly focusing on fingerprints and their applications.

SRIJITH RAJEEV (Member, IEEE) received the bachelor's degree (B.E.) degree in electronics and communication engineering from Visvesvaraya Technological University, India, in 2014, and the M.S. degree in electrical and computer engineering from The University of Texas at San Antonio, USA, in 2016. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Tufts University, USA. His current research interests include artificial intelligence, eye-tracking, signal/image processing, machine learning, robotics, 3D sensors and modeling, digital forensic, and biomedical applications.

SOS S. AGAIAN (Fellow, IEEE) is currently a Distinguished Professor with The City University of New York/CSI. His research interests include computational vision and machine learning, large-scale data analytic analytics, multi-modal data fusion, biologically inspired signal/image processing modeling, multi-modal biometric and digital forensics, 3D imaging sensors, information processing, security, and biomedical and health informatics. He has authored more than 650 technical articles and ten books in these areas. He is also listed as a co-inventor on 44 patents/disclosures. The technologies that he invented have been adopted by multiple institutions, including the U.S. Government, and commercialized by industry. He is a fellow of SPIE, IS&T, and AAAS. He also serves as a Foreign Member for the Armenian National Academy. He received the Maestro Educator of the year, sponsored by the Society of Mexican American Engineers. He received the Distinguished Research Award at The University of Texas at San Antonio. He was a recipient of the Innovator of the Year Award, in 2014, the Tech Flash Titans-Top Researcher-Award (*San Antonio Business Journal*), in 2014, the Entrepreneurship Award (UTSA-2013 and 2016), and the Excellence in Teaching Award, in 2015. He is an Editorial Board Member for the journal of *Pattern Recognition and Image Analysis* and an Associate Editor for several journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS, *Journal of Electrical and Computer Engineering* (Hindawi Publishing Corporation), *International Journal of Digital Multimedia Broadcasting* (Hindawi Publishing Corporation), and *Journal of Electronic Imaging* (IS&T and SPIE).