Journal of Imaging Science and Technology® 69(4): 1–10, 2025.

ExtremeMETA: High-speed Lightweight Image Segmentation Model by Remodeling Multi-channel Metamaterial Imagers

Quan Liu and Brandon T. Swartz
Vanderbilt University, Nashville, TN 37212

Ivan Kravchenko
Oak Ridge National Laboratory, Oak Ridge, TN 37830

Jason G. Valentine and Yuankai Huo
Vanderbilt University, Nashville, TN 37212
E-mail: yuankai.huo@vanderbilt.edu

Abstract. *Deep neural networks (DNNs) have heavily relied on traditional computational units, such as CPUs and GPUs. However, this conventional approach brings significant computational burden, latency issues, and high power consumption, limiting their effectiveness. This has sparked the need for lightweight networks such as ExtremeC3Net. Meanwhile, there have been notable advancements in optical computational units, particularly with metamaterials, offering the exciting prospect of energy-efficient neural networks operating at the speed of light. Yet, the digital design of metamaterial neural networks (MNNs) faces precision, noise, and bandwidth challenges, limiting their application to intuitive tasks and low-resolution images. In this study, we proposed a large kernel lightweight segmentation model, ExtremeMETA. Based on ExtremeC3Net, our proposed model, ExtremeMETA maximized the ability of the first convolution layer by exploring a larger convolution kernel and multiple processing paths. With the large kernel convolution model, we extended the optic neural network application boundary to the segmentation task. To further lighten the computation burden of the digital processing part, a set of model compression methods was applied to improve model efficiency in the inference stage. The experimental results on three publicly available datasets demonstrated that the optimized efficient design improved segmentation performance from 92.45 to 95.97 on mIoU while reducing computational FLOPs from 461.07 MMacs to 166.03 MMacs. The large kernel lightweight model ExtremeMETA showcased the hybrid design's ability on complex tasks.*
Keywords: *large convolution kernel, model compression, segmentation, meta-material*

[DOI: 10.2352/J.ImagingSci.Technol.2025.69.4.040403]

1. INTRODUCTION

In the realm of modern computer vision, digital neural networks play a pivotal role. Arguably, convolutional neural network (CNN) stands out as the most extensively employed AI approach, particularly in tasks like image classification, segmentation, and detection. Traditional CNNs face several challenges when deployed in resource-constrained environments, such as those found in IoT devices, edge computing systems, and drone operations. These applications demand real-time performance with minimal power consumption, low latency, and efficient processing capabilities, which are difficult to achieve with standard CNN architectures due to their computational complexity and large memory requirements. As IoT and edge computing continue to expand in fields like smart cities, autonomous vehicles, and drone-based surveillance, it is critical to develop CNN models that operate effectively in these environments. Addressing these challenges not only enhances the scalability and adaptability of CNN-based solutions but also enables more efficient and reliable system operations in real-time applications. Despite the advent of vision transformer-based models, convolution remains integral for extracting local image features. Presently, CNNs are typically implemented on computational units like CPUs and GPUs. However, this conventional design approach brings forth substantial challenges, including a formidable computational load, notable latency issues, and heightened power consumption. These limitations are prominent in drone operations, Internet of Things (IoT), and edge computing applications, which emphasize the need for a lightweight model to analyze efficiently. Recognizing the critical need for DNN models with reduced energy consumption and lower latency, the AI community has embarked on a quest for more efficient solutions. Despite these efforts, achieving DNNs that are light with low power consumption in the current research trends is an elusive goal.

Recent breakthroughs in optical computational units, including metamaterials (refer to Figure 1), have brought to light the potential for neural networks that operate without energy consumption and at unprecedented speeds. The current cutting-edge metamaterial neural network (MNN) takes on a hybrid form, leveraging optical processors as a lightspeed and energy-free front-end convolutional operator alongside a digital feature aggregator. This novel approach significantly reduces computational latency. By assigning the convolution operations to optical units, more than 90 percent of the floating- point operations (FLOPs) inherent in conventional CNN backbones like VGG and ResNet are

Received June 4, 2024; accepted for publication Nov. 24, 2024; published online Jan. 31, 2025. Associate Editor: Henry Y.T. Ngan.

1062-3701/2025/69(4)/040403/10/$25.00

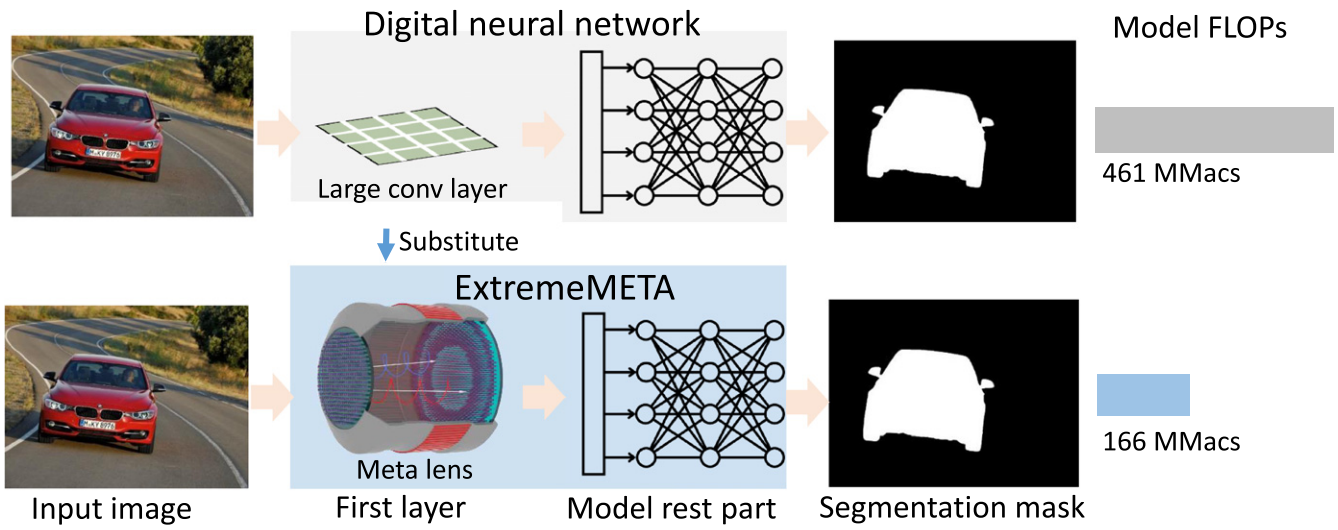

Figure 1. This study provides a hybrid pipeline for designing and optimizing a large kernel digital neural network. The proposed ExtremeMETA is efficient for segmentation tasks with less FLOPs in computation.

effectively off-loaded. This marks a noteworthy departure from traditional architectures, opening up new avenues for efficient and high-performance neural network designs. However, the hybrid design is fundamentally influenced by the physical structure including the limited kernel size and channel number. Moreover, the hybrid system is also limited by what can be fabricated as the first optical layer of the neural network.

Based on our proposed LMNN (large kernel metamaterial neural network) model, the hybrid design achieved promising performance on the classification task. However, LMNN has a few limitations, namely: (1) this model can only perform image classification tasks instead of model complex tasks like image segmentation and object detection; (2) input images are in low resolution (28×28), and (3) leverages the computation burden to the optical part, the digital part requires efficiency improvement operation like model compression in the inference stage. While the LMNN reduces computational complexity by offloading much of the burden to the first layer using a metaoptic lens, it faces limitations in segmentation and object detection, where fine-grained spatial understanding is required and the loss of this capability may be due to the early emphasis on feature extraction in LMNN. In practical applications, such as autonomous driving or medical imaging, this limitation can affect the network's ability to deliver accurate pixel-level segmentation or precise object localization.

In this study, we propose a novel large kernel lightweight segmentation model, ExtremeMETA, which maximizes the efficiency advantages of optic signal computation, while compressing the digital processing model to further improve the model segmentation efficiency. To adapt the segmentation task on large images, the proposed lightweight large kernel model achieves larger receptive fields, the ability to analyze larger images, and covers general vision tasks, image classification segmentation, and detection. Furthermore, the complexity of the model digital processing part is explicitly addressed via a set of model compression methods. We evaluated our design on image segmentation tasks using three public datasets: the portrait dataset, the Stanford dataset, and KITTI dataset. The proposed lightweight large kernel model achieved superior segmentation accuracy as compared with the state-of-the-art (SOTA) segmentation model. Overall, the system's contributions are as follows:

- We propose a new large convolution kernel CNN network to achieve a large reception field, lower energy consumption, and less latency.
- We introduce model reparameterization to improve large convolution kernel performance and sparse convolution kernel compression mechanism to compress the multi-branch sparse-convolution design to a single layer for the hybrid system implementation. The model compression mechanism improves the model efficiency for digital processing.
- The task limitations of large convolution hybrid models are explicitly addressed via performing segmentation tasks on multiple datasets from different categories.

The rest of the article is organized as follows. In Section 2, we present the background and related research relevant to large kernel convolution, model compression, and ONNs on image processing tasks. In Section 3, our proposed lightweight lightspeed model is presented. It includes the large kernel reparameterization, sparse convolution compression, and multipath model compression. Section 4 details the dataset and experiment implementation details. Section 5 analyzes the experimental results and ablation study. Then, in Sections 6 and 7, we provide the discussion and conclude our work.

2. RELATED WORK

2.1 *Large Kernel Convolution Design*

In the realm of CNNs, the design and utilization of large kernel convolutions have garnered significant attention in recent years. Numerous studies have explored the benefits of using larger convolutional kernels, such as 7×7 or 11×11, to capture broader spatial contexts and more intricate patterns within images [1, 2]. Early research efforts focused on understanding the impact of kernel size on model performance, with findings suggesting that larger kernels can lead to improved feature extraction and recognition accuracy, especially for complex visual tasks [3].

Building on these findings, subsequent studies proposed various strategies to incorporate large kernel convolutions into CNN architectures effectively. These strategies often involved modifying network architectures, adjusting kernel sizes, or integrating multi-scale features to enhance the robustness and versatility of CNN models [4, 5]. Additionally, advancements in hardware acceleration and parallel processing have facilitated the efficient implementation of large kernel convolutions, enabling their widespread adoption across diverse computer vision applications [6, 7].

Overall, the related work on large kernel convolution design underscores its pivotal role in advancing the capabilities of CNNs for tackling increasingly complex and demanding visual recognition tasks [8, 9].

2.2 *Optic Neural Network*

Optic neural networks (ONNs) have emerged as a promising paradigm for accelerating neural network computations by leveraging the unique properties of optical computing. Inspired by the principles of light-based signal processing, ONNs exploit the parallelism, high bandwidth, and low energy consumption inherent in optical systems to achieve significant computational efficiency gains compared to traditional electronic implementations. A considerable body of research has focused on exploring various aspects of ONNs, including optical device design, system architectures, and algorithmic frameworks tailored to optical computing platforms [10–12].

Early studies laid the groundwork for ONNs by demonstrating their potential for accelerating matrix-vector multiplications, a fundamental operation in neural network inference [13, 14]. Subsequent works have extended ONN capabilities to encompass more complex neural network layers and architectures, paving the way for practical applications in tasks such as image classification, object detection, and natural language processing [15, 16].

Key challenges in ONN research include addressing optical noise, device nonlinearity, and scalability issues, which require interdisciplinary efforts spanning optics, photonics, and machine learning [17, 18]. Despite these challenges, ONNs hold great promise for enabling ultra-fast and energy-efficient neural network computations, with the potential to revolutionize various domains of artificial intelligence and computing [19, 20].

2.3 *Segmentation Model*

Recent advancements in segmentation techniques have introduced novel methods that improve accuracy and robustness in challenging tasks. For instance, the use of a topological loss function based on persistent homology has shown promise in improving the structural integrity of segmentation outputs, particularly in applications where shape preservation is critical [21]. Additionally, the boundary-enhanced dual-stream network has demonstrated significant improvements in semantic segmentation, particularly in high-resolution remote sensing images where fine boundary details are crucial [22]. These models offer innovative solutions for specific segmentation challenges, complementing the growing body of research on improving segmentation accuracy. Our proposed model, ExtremeMETA builds on this foundation by providing a model that is both computationally efficient and highly accurate, making it suitable for a wide range of applications, from general-purpose segmentation to more domain-specific tasks.

2.4 *Convolution Neural Network Model Compression*

In the field of CNNs, model compression techniques have garnered significant attention as a means to reduce the computational complexity and memory footprint of deep learning models without sacrificing performance. A diverse range of methods has been proposed to compress CNNs, including pruning, quantization, low-rank approximation, knowledge distillation, and weight sharing. Pruning techniques aim to remove redundant or less important parameters from the network, thereby reducing its size and computational cost [23, 24]. Quantization methods reduce the precision of network parameters, often by representing weights and activations with fewer bits, to decrease memory requirements and improve inference speed [25]. Low-rank approximation techniques exploit the underlying structure of weight matrices to factorize them into smaller, more computationally efficient components [26]. Knowledge distillation involves training a compact "student" network to mimic the predictions of a larger "teacher" network, transferring knowledge from the latter to the former [27]. Additionally, weight sharing approaches reduce redundancy by sharing parameters across different parts of the network [28].

Collectively, these model compression techniques offer effective strategies for deploying CNNs on resource-constrained devices or accelerating inference in large-scale deployment scenarios. Ongoing research in this area continues to explore novel compression algorithms, optimization strategies, and application-specific considerations to further improve the efficiency and effectiveness of compressed CNN models.

2.5 *Model Efficiency Improvement*

Recent studies have focused on improving the efficiency and performance of models in various signal processing and communication-related tasks, which closely align with the objectives of our work. For example, [29] introduced a manifold regularization-based deep convolutional autoencoder

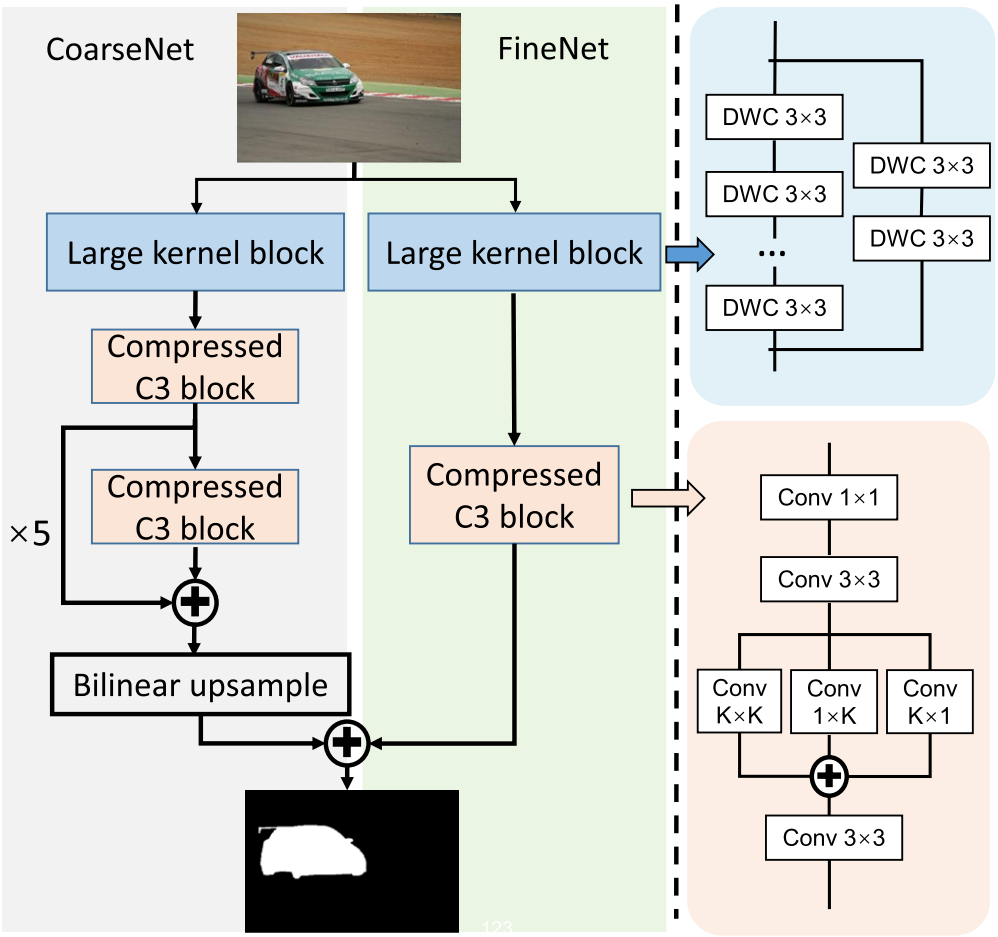

Figure 2. Lightweight segmentation model with hybrid metaoptics design. The model has two parts: CoarseNet and FineNet. The large kernel block is composed of depthwise convolution layers.

for unauthorized broadcasting identification, addressing a critical challenge in signal security and classification. Additionally, [30] and the multi-scale radio transformer method have advanced the field of lightweight automatic modulation classification, particularly in resource-constrained environments like drone communication systems. Similarly, [31] demonstrated how lightweight networks can achieve real-time classification of wireless communication signals, making them highly applicable for low- power devices. Furthermore, CNN-LSTM-driven methods have been proposed for real-time transformer discharge pattern recognition, showcasing the potential of combining CNNs with temporal models in complex pattern recognition tasks.

3. METHOD

3.1 *Problem Statement*

We extensively study the trainability of large kernels on MNNs and unveil three main observations: (i) traditional convolution kernel shows limited improvement on large images; (ii) the MNN is only available on classification task; (iii) metamaterial implementation limited the computation ratio on segmentation model which is typical in a complex structure. Model is shown in Figure 2.

3.2 *Large Convolution Design with Multiple Path Design*

Limited by the image size and the task for the model, our previous proposed model, LMNN achieved the prediction performance with kernel size 9×9. Two major limitations exist when applying the large kernel design to the MNN: (1) the metamaterial implementation limits the image size to a small range; (2) only the classification task is available to be validated on the MNN model when the segmentation task and detection task are too difficult to be implemented under the optic implementation limitation. To address the challenges, we proposed our model from two perspectives: (1) from kernel design, we employ the large convolution kernel with parameterization design to construct the convolution layer (larger than 9×9); (2) from model design, our proposed lightweight segmentation model based on the multipath model structure composed of a course segmentation path and a light refinement path proposed by [32].

3.3 *Model Compression with Sparse Convolution*

Model compression is a crucial technique aimed at enhancing the efficiency of deep learning models by reducing their size and computational demands while maintaining their

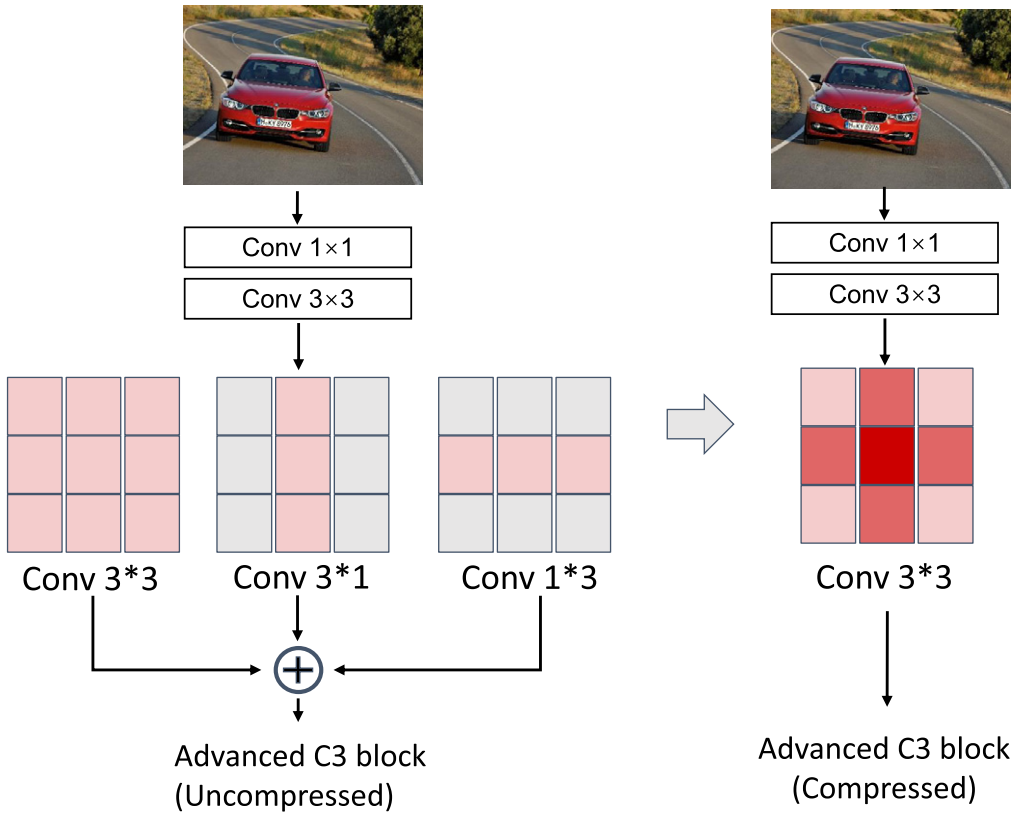

Figure 3. Model compression on segmentation model digital processing part. The left panel shows the multipath structure of the advanced C3 block. The right panel shows the compression mechanism.

performance standards. Among various strategies employed for model compression, pruning, and quantization stand out as widely adopted methodologies. Pruning, a prominent model compression technique, involves the systematic removal of redundant or unnecessary parameters from neural networks. By identifying and eliminating connections that contribute minimally to the model's performance, pruning effectively reduces the model's size and computational requirements. This process permits a more streamlined network architecture without sacrificing accuracy, making it particularly valuable for resource-constrained environments or deployment on edge devices.

We applied model compression and parameterization together for the sparse convolution kernel which is shown in Figure 3. Sparse convolution refers to a convolution operation where the kernel (filter) contains mostly zero values, resulting in a sparse structure. When using a kernel size of 1×3 (1 row and 3 columns), the convolution operation typically involves sliding this kernel over the input data and performing element-wise multiplication followed by summation along the spatial dimensions.

$$O_{h,w,c'} = \sum_{i=0}^{2} \sum_{j=0}^{C-1} I_{h,w+i,j} \times K_{0,i,j,c'}, \tag{1}$$

where, I is the input tensor, K is the kernel tensor, O is the output tensor, and $\times$ is the convolution operation.

For the ExtremeC3 block, we have three convolution paths with kernel size $k \times k$, $1 \times k$, and $k \times 1$. Denoting the individual kernels as $k_{1\times k}$, $k_{k\times k}$, and $k_{k\times 1}$; the compressed convolution kernel is expressed as follows:

$$K_{\text{combined}}(i,j) = w_{1xk} \times K_{1xk}(i,j) + w_{kxk} \times K_{kxk}(i,j) + w_{kx1} \times K_{kx1}(i,j). \tag{2}$$

The compressed multipath convolution block saves computation complexity in the inference stage.

The use of sparse convolution compression in ExtremeMETA significantly improves efficiency by reducing the number of unnecessary computations, particularly in non-critical areas of the network. This technique compresses the model by introducing sparsity convolution and multipath in the convolutional layers, which leads to lower memory usage and faster inference times. In practical deployment, especially in resource-constrained environments such as edge devices or IoT systems, this results in reduced computational load, lower power consumption, and faster real-time performance without compromising model accuracy.

4. DATA AND EXPERIMENTAL DESIGN

4.1 *Data Description*

Three public datasets, EG1800 [33], Stanford Car dataset [34], and KITTI dataset [35], were used to evaluate

the lightweight large kernel model on segmentation tasks. For the EG1800 dataset, we employed 1887 images in 600 × 800 resolution with semantic segmentation masks. The EG1800 dataset was collected from Flickr with the manually annotated mask of the portrait. The Stanford Car dataset is composed of 16,185 RGB images of cars with the point coordinate of the car's location in the images. The KITTI dataset is popular in mobile robotics and autonomous driving and features diverse traffic scenarios captured using high-resolution RGB, grayscale stereo cameras, and a 3D laser scanner. However, it lacks inherent ground truth annotations for semantic segmentation. To adapt to the segmentation task, both the Stanford Car dataset and the KITTI dataset need to address the annotation limitation.

4.2 *Data Generation with Foundation Model*

Regarding the lack of segmentation annotation in Stanford Car and KITTI datasets, we employed the Segment Anything Model (SAM) [36] to generate the object mask based on the prompts of object location. SAM is a foundation model that has a zero-shot ability to segment objects on new image distributions. The RGB image of Stanford Car and KITTI datasets and bounding box coordinate is provided to SAM, which generates the object masks. With the help of the SAM, the RGB images with object mask annotations are available for model training.

4.3 *Large Kernel Digital Design on Segmentation Model*

The large kernel design was applied to the segmentation network's first convolution layer design. Since the first layer was designed to be substituted by the metaoptic lens in the inference stage, our large kernel design was under physical limitation. On the other hand, the optic lens provided lightspeed computation which we took advantage of. Based on the multipath segmentation network, the first convolution layers of the CoarseNet and FineNet parts were redesigned with the large convolution kernel with parameterization following the strategy in our previous work LMNN [37]. Since the image was large compared with FashionMNIST previously used, our kernel size increased from 9 × 9 to 15 × 15. The channel number was expanded from 12 to 48. The Larger convolution kernel and channel number provided the capability of the first layers and handled the complex situation.

4.4 *Model Design with Optic Constrain*

Constrained by fabrication issues, the metaoptic layer has limitations on both channel number and input size. The trade-off in model performance between input size and channel number is discussed. The size-first design uses the largest input image size under fabrication constraint. Channel-first design prefers more channel numbers under the fabrication limitation.

4.5 *Model Compression Efficiency*

Besides enlarging the capability of the first layer, our proposed lightweight segmentation network is compressed in the digital part. Since compression affects the model's complexity and efficiency, we evaluated if the compressed model loses accuracy. To test the efficiency of the model compression strategy, the model FLOPs, parameters, and FLOPs ratio of the first convolution layer.

Table I. Segmentation performance on EG1800.

Model	Kernel size	1st Conv FLOPs (%)	Model FLOPs	Digital FLOPs	Test (mIoU)
ExtremeC3	3 × 3	10.87	199.4	199.4	0.9249
	11 × 11	62.11	469.14	469.14	0.9323
	15 × 15	75.30	719.62	719.62	0.9301
Digital	N/A	N/A	174.10	174.10	0.9086
Ours	1 × 1	2.80	182.06	174.10	0.9137
	3 × 3	10.87	199.40	174.10	0.9234
	11 × 11	59.68	431.81	174.10	0.9415
	15 × 15	63.36	475.16	174.10	**0.9418**

Model FLOPs and digital FLOPs unit is MMacs.

5. RESULT

In this section, we evaluate our proposed lightweight segmentation network with a simple model structure, using the EG1800 dataset, Stanford Car dataset, and KITTI dataset. Since the Stanford Car dataset and KITTI dataset are car images, we train the model and test the two datasets together.

5.1 *Segmentation Performance on Portrait Dataset*

We evaluate the lightweight segmentation model on EG1800 dataset together with model parameters and first convolution FLOPs ratio. As shown in Table I, the original ExtremeC3 model cannot take advantage of the large convolution kernel on the first layer, 15 × 15 kernel showed even lower performance than 11 × 11. The model performance without the first convolution layer showed a 2% drop compared with the ExtremeC3 model with 3 × 3 kernel size. Our proposed hybrid lightweight segmentation model achieved the best performance with 15 × 15 convolution kernel which had the same digital computation FLOPs.

Besides improving the model performance with advanced design on the first convolution layer, we evaluate the model efficiency improvement by model compression. Following the experiment setting in Table I, we applied model compression, including sparse convolution kernel compression and multipath parameterization, to each model design and show the efficiency evaluation matrix in Table II. The compression method showed efficient computation on digital FLOPs without affecting model performance (mIoU).

In comparison to traditional CNN architectures, ExtremeMETA achieved lower computational complexity by employing sparse convolution compression and metaoptic lens techniques, which reduced redundant operations in early layers. This resulted in faster processing and reduced memory requirements. However, like most efficient models, MobileNets [38] and EfficientNet [7], there is a trade-off

Table II. Segmentation performance on EG1800 after model compression.

Model	Kernel size	1st Conv FLOPs (%)	Model FLOPs	Digital FLOPs	Test (mIoU)
ExtremeC3	3 × 3	11.33	191.32	191.32	0.9233
	11 × 11	63.21	461.07	461.07	0.9315
	15 × 15	76.16	711.55	711.55	0.9289
Digital	N/A	N/A	166.03	166.03	0.9031
Ours	1 × 1	3.17	174.25	166.03	0.9121
	3 × 3	11.33	191.32	166.03	0.9217
	11 × 11	60.81	423.74	166.03	0.9404
	15 × 15	64.45	467.09	166.03	**0.9420**

Model FLOPs and digital FLOPs unit is MMacs.

Table III. Segmentation performance on car dataset.

Model	Kernel size	Train (KITTI+Stanford)	Test (mIoU)	KITTI	Stanford
ExtremeC3	3*3	95.02	92.51	84.45	95.23
	11*11	95.12	92.09	84.37	95.39
	15*15	76.09	70.25	22.69	95.22
Digital	N/A	93.31	89.11	78.47	94.27
Ours	1*1	94.13	90.94	82.68	93.15
	3*3	94.97	92.01	85.05	94.77
	11*11	95.79	92.91	85.33	95.97
	15*15	96.05	93.17	87.41	95.19

Model FLOPs and digital FLOPs unit is MMacs.

between computational efficiency and segmentation accuracy. In practical applications, such as real-time image segmentation on edge devices, ExtremeMETA demonstrated improved processing speed and reduced power consumption while maintaining competitive segmentation accuracy. The trade-off is most noticeable in tasks requiring extremely fine-grained segmentation, where traditional CNNs may offer marginally better accuracy at the cost of significantly higher computational demands.

5.2 *Segmentation Performance on Car Dataset*

To validate our lightweight segmentation model with more datasets, we conducted experiments on the car dataset, including the Stanford Car dataset and KITTI dataset with semantic segmentation mask as ground truth. Both the Stanford Car dataset and the KITTI datasets were used for model training, even though the resolutions were different. Using the same experimental setup described in Table III, we applied model compression and multipath parameterization to model design, and present the resulting efficiency evaluation matrix in Table IV.

5.3 *Model Robustness*

To evaluate the generalization ability of ExtremeMETA, we conducted experiments on datasets beyond those used for training, specifically the Portrait and Pet datasets. As shown

Table IV. Segmentation performance on car dataset after model compression.

Model	Kernel size	1st Conv (%) FLOPs (%)	Model FLOPs	Digital FLOPs	Test (mIoU)
ExtremeC3	3*3	11.33	191.32	191.32	91.36
	11*11	63.21	461.07	461.07	92.45
	15*15	76.16	711.55	711.55	70.01
Digital	N/A	N/A	166.03	166.03	88.97
Ours	1*1	3.17	174.25	166.03	90.94
	3*3	11.33	191.32	166.03	94.25
	11*11	60.81	423.74	166.03	95.32
	15*15	64.45	467.09	166.03	93.05

Model FLOPs and digital FLOPs unit is MMacs.

Table V. Comparison of model performance on Portrait and Pet datasets.

Model	Portrait (mIoU)	Pet (mIoU)
YOLO	**92.67**	70.48
Ours	91.84	**73.87**

in Table V, ExtremeMETA achieved an mIoU of 91.8439 on the Portrait dataset and 73.8717 on the Pet dataset, significantly outperforming YOLO on both. These results demonstrate that ExtremeMETA generalizes well across different types of images, even in tasks that involve varying levels of complexity, such as fine-grained segmentation in the Pet dataset. The model's architecture, including sparse convolution compression and metaoptic lens techniques, allows it to adapt to different domains with minimal loss in performance, making it a versatile solution for various practical applications.

To provide a visual comparison of the segmentation results, we present qualitative examples from the Portrait and Pet datasets in Figure 4. The figure shows the original images, the segmentation outputs generated by ExtremeMETA, and the corresponding ground truth. As demonstrated, ExtremeMETA accurately segmented both human portraits and animal shapes, closely matching the ground truth in each case. These visual results further validate the effectiveness of ExtremeMETA in diverse segmentation tasks, showing that it can generalize well across different image types and maintain high segmentation accuracy.

5.4 *Ablation Studies*

Due to the fabrication limitation of the metalens array, the priority of channel number and input image size need to be fixed. The results of the experiment are shown in Figure 5. The left panel illustrates how increasing the input image size enhances performance compared to expanding the number of channels in a convolution layer. The gray area depicts the performance disparity expressed as mIoU. Increasing the input image size enhances the model's ability to capture finer spatial details, which improves performance in tasks like semantic segmentation. However, it also increases

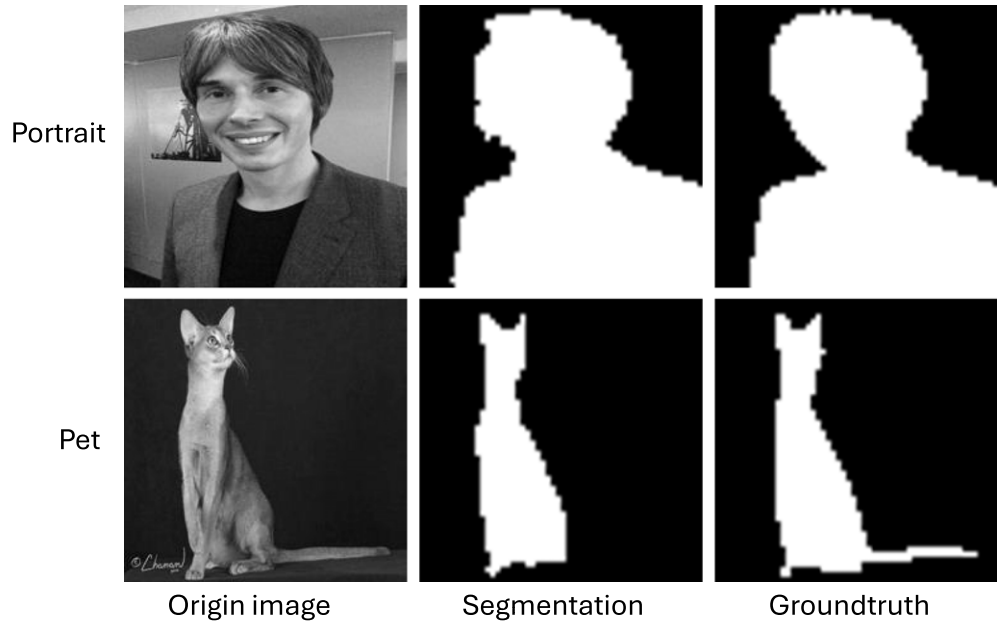

Figure 4. Segmentation results on the Portrait and Pet datasets. The first column shows the original images, the second column presents the segmentation results from ExtremeMETA, and the third column displays the ground truth. The results show that ExtremeMETA effectively captured the boundaries and shapes of objects with high accuracy.

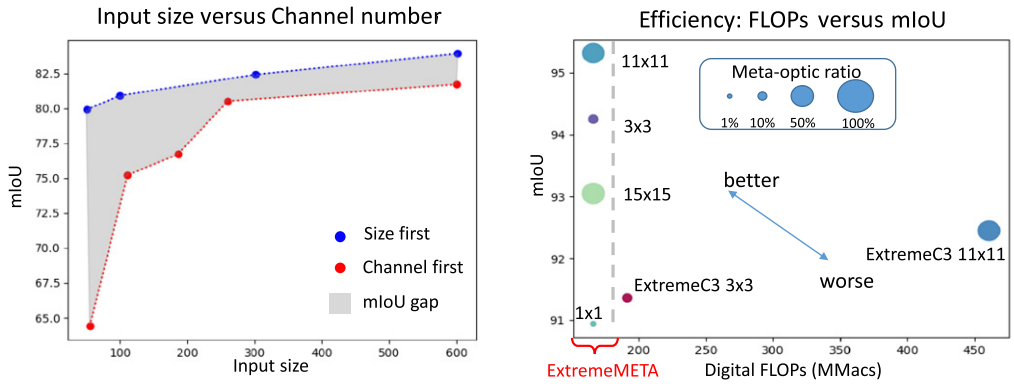

Figure 5. Model ablation study. Left panel: trade-off between input image size and channel number of convolution layer. Right panel: model efficiency visualization comparing model FLOPs and mIoU.

computational cost. Expanding the number of channels, while boosting the model's capacity to extract complex features, raises the risk of overfitting and computational load. The gray area in the left panel highlights that, in this case, increasing the input size led to a greater improvement in mIoU than expanding the number of channels, indicating that capturing spatial details was more impactful for performance. On the right panel, the effectiveness of utilizing large convolution kernels is shown. Circles of various colors represent different convolution layer architectures, with the area of each circle indicating the ratio of FLOPs for the layer when implemented using metaoptic materials. The x-axis represents the model's FLOPs, excluding the layer intended for fabrication.

5.5 *Model Compression*

Figure 6 demonstrates that the compressed model achieves a reduction of 8 MMacs in FLOPs, decreasing from 174.10 MMacs to 166.03 MMacs. The right panel indicates that the compressed model maintains equivalent performance to the original model. This consistency in performance establishes that ExtremeMETA not only enhances the efficiency of the digital components but also contributes to the overall optimization of the hybrid system.

6. DISCUSSION

Given the demonstrated superior performance of large convolution kernels in tasks such as image classification and segmentation, there exists substantial potential for their application in a wider array of complex computer vision tasks. Large convolution kernels have shown remarkable effectiveness in tasks like image classification and segmentation, primarily due to their ability to capture more extensive spatial information and intricate patterns within images. This success suggests that employing large convolution

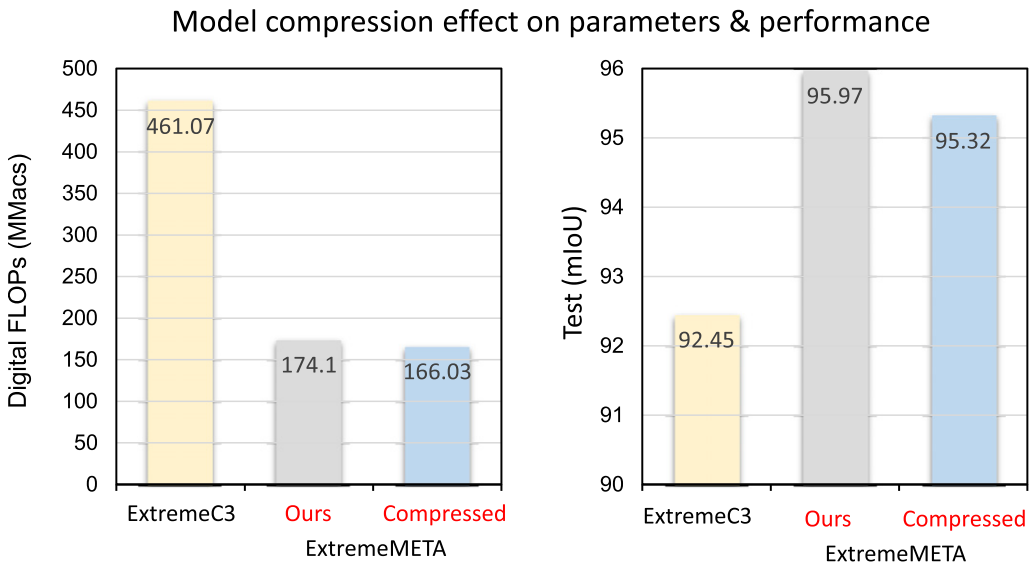

Figure 6. Model compression performance. Left panel: origin model, ExtremeMETA, and compressed model parameters comparison; right panel: model performance after compression.

kernels in other computer vision tasks could yield significant improvements.

One such task is object detection, where accurately identifying and localizing objects within images is crucial. By utilizing large convolution kernels, the model can better discern the detailed features of objects, leading to more precise detection results. This can be particularly beneficial in scenarios with small or occluded objects, where finer details are essential for accurate recognition as the results shown in the experiments on the car dataset.

Furthermore, in tasks involving image generation or synthesis, such as style transfer or super- resolution, large convolution kernels can enhance the model's ability to capture intricate textures and details, resulting in more realistic and high-fidelity output images. These kernels can effectively extract and preserve fine-grained features, which are instrumental in faithfully replicating the characteristics of the input images.

The application can be extended to video processing tasks such as action recognition or video segmentation, where large convolution kernels can enhance the model's capability to analyze temporal and spatial dependencies across frames. By incorporating information from a broader context, these kernels enable a more robust understanding of dynamic scenes, leading to improved performance in tasks requiring temporal coherence and contextual understanding.

The adoption of large convolution kernels holds promise for advancing various complex computer vision tasks beyond traditional image classification and segmentation. Their ability to capture intricate details and spatial relationships makes them a valuable tool for enhancing the performance and capabilities of computer vision models across diverse applications.

7. CONCLUSION

In this study, we presented a novel large kernel lightweight segmentation model that harnesses the efficiency advantages of optical signal computation while integrating digital processing model compression techniques to further enhance segmentation efficiency. Our model offers larger receptive fields tailored for segmentation tasks on large images, extending its applicability to various vision tasks including image classification, segmentation, and detection. Through extensive evaluations on diverse datasets, including the portrait, Stanford, and KITTI datasets, our proposed approach has demonstrated superior segmentation accuracy compared to state-of-the-art models. Our contributions encompass the introduction of a novel large convolution kernel CNN network for larger reception fields, reduced energy consumption, and lower latency, alongside the introduction of model reparameterization and sparse convolution kernel compression mechanisms to enhance model performance and efficiency in digital processing. By explicitly addressing task limitations and conducting segmentation tasks on multiple datasets from different categories, our work represents a significant step forward in the development of efficient and effective segmentation models for a wide range of computer vision applications.

Summary of Contributions: Our work offers key advancements in computer vision by addressing the computational and practical challenges in deploying CNN-based models in real-world scenarios. We introduced an architecture that not only improves segmentation accuracy but also reduces computational complexity, making it highly suitable for resource-constrained environments such as IoT devices and edge computing. The proposed model compression techniques further contribute to lower energy consumption and faster processing times, highlighting the potential for widespread adoption across various industries, from autonomous systems to medical imaging. Our findings

push the boundaries of segmentation model efficiency and performance, paving the way for future innovations in various fields.

ACKNOWLEDGMENT

Y.H. and Q.L. acknowledge support from NIH under contract R01DK135597. Y.H. is the corresponding author. B.T.S. and J.G.V. acknowledge support from DARPA under contract HR001118C0015, NAVAIR under contract N6893622C0030 and ONR under contract N000142112468. Metaoptic devices were manufactured as part of a user project at the Center for Nanophase Materials Sciences (CNMS), which is a US Department of Energy, Office of Science User Facility, Oak Ridge National Laboratory.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Preprint, arXiv:1409.1556 (2014).

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2015), pp. 1–9.

[3] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," Preprint, arXiv:1311.2901 (2014).

[4] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Rethinking the inception architecture for computer vision," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 2818–2826.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 770–778.

[6] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Efficient and accurate approximations of nonlinear convolutional networks," *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2018), pp. 1984–1992.

[7] M. Sun, Z. Liu, X. Wang, W. Qiao, and K. Lin, "Efficientnet: Rethinking model scaling for convolutional neural networks," *Int'l. Con. on Machine Learning* (PMLR, Stockholm, Sweden, 2019), pp. 6105–6114.

[8] M. Lin, Q. Chen, and S. Yan, "Network in network," Preprint, arXiv:1312.4400 (2013).

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2017), pp. 4700–4708.

[10] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, and D. Englund, "Deep learning with coherent nanophotonic circuits," Nature Photonics **11**, 441–446 (2017).

[11] X. Lin, Y. Rivenson, D. Teng, L. Wei, H. Günaydın, Y. Zhang, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," Science **361**, 1004–1008 (2018).

[12] T. W. Hughes, M. Minkov, I. A. Williamson, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement: supplementary material," Optica **Part F127** (2018).

[13] A. N. Tait, M. A. Nahmias, B. J. Shastri, P. R. Prucnal, and J. S. Harris, "The physics of optical neural networks," Appl. Phys. Rev. **4**, 021105 (2017).

[14] A. N. Tait, M. A. Nahmias, B. J. Shastri, P. R. Prucnal, and J. S. Harris, "Optical implementation of deep networks," Appl. Optics **55**, A71–A82 (2016).

[15] M. Miscuglio, J. Dambre, and P. Bienstman, "All-optical nonlinear activation function for photonic neural networks [invited]," Opt. Mater. Express **8**, 3851–3863 (2018).

[16] L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutiérrez, I. Fischer, and C. R. Mirasso, "Photonic information processing beyond turing: an optoelectronic implementation of reservoir computing," Opt. Express **20**, 3241–3249 (2012).

[17] S. Jutamulia and F. T. S. Yu, "Overview of hybrid optical neural networks," Opt. Laser Technol. **28**, 85–97 (1996).

[18] K. M. Boehm, P. Khosravi, R. Vanguri, J. Gao, and S. P. Shah, "Harnessing multimodal data integration to advance precision oncology," Nature Rev. Cancer **22**, 71–88 (2022).

[19] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleido-bert: Vision-language pre-training on fashion domain," *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2021), pp. 12642–12652.

[20] Y. B. Ovchinnikov, J. Müller, M. Doery, E. Vredenbregt, K. Helmerson, S. Rolston, and W. Phillips, "Diffraction of a released bose-einstein condensate by a pulsed standing light wave," Phys. Rev. Lett. **83**, 284 (1999).

[21] J. R. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. P. King, "A topological loss function for deep-learning based image segmentation using persistent homology," IEEE Trans. Pattern Anal. Mach. Intell. **44**, 8766–8778 (2020).

[22] X. Li, L. Xie, C. Wang, J. Miao, H. Shen, and L. Zhang, "Boundary-enhanced dual-stream network for semantic segmentation of high-resolution remote sensing images," GIScience Remote Sens. **61**, 2356355 (2024).

[23] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Montréal, Canada, 2015), pp. 1135–1143.

[24] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," Int'l. Conf. on Learning Representations (OpenReview, Toulon, France, 2016).

[25] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," Preprint, arXiv:1609.07061 (2017).

[26] E. L. Denton, W. Zaremba, J. Bruna, Y. Lecun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Montréal, Canada, 2014), pp. 1269–1277.

[27] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," Preprint, arXiv:1503.02531 (2015).

[28] W. Chen, J. Wilson, S. Tyree, K. Weinberger, and Y. Chen, "Compressing neural networks with the hashing trick," *Int'l. Conf. on Machine Learning* (JMLR, Lille, France, 2015), pp. 2285–2294.

[29] Q. Zheng, P. Zhao, D. Zhang, and H. Wang, "MR-DCAE: Manifold regularization-based deep convolutional autoencoder for unauthorized broadcasting identification," Int. J. Intell. Syst. **36**, 7204–7238 (2021).

[30] Q. Zheng, X. Tian, Z. Yu, Y. Ding, A. Elhanashi, S. Saponara, and K. Kpalma, "Mobilerat: A lightweight radio transformer method for automatic modulation classification in drone communication systems," Drones **7**, 596 (2023).

[31] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," Preprint, arXiv:2110.02178 (2021).

[32] H. Park, L. L. Sjösund, Y. Yoo, J. Bang, and N. Kwak, "Extremec3net: Extreme lightweight portrait segmentation networks using advanced c3-modules," Preprint, arXiv:1908.03093 (2019).

[33] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, and I. Sachs, "Automatic portrait segmentation for image stylization," *Computer Graphics Forum* (Wiley Online Library, Hoboken, NJ, 2016), Vol. 35, pp. 93–102.

[34] J. Krause, J. Deng, M. Stark, and L. Fei-Fei, "Collecting a large-scale dataset of fine-grained cars," *Proc. 1st IEEE Workshop on Fine-Grained Visual Classification (FGVC) in Conjunction with the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2013).

[35] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," *Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2012), pp. 3354–3361.

[36] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," *Proc. IEEE/CVF Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2023), pp. 4015–4026.

[37] Q. Liu, H. Zheng, B. T. Swartz, Z. Asad, I. Kravchenko, J. G. Valentine, and Y. Huo, "Digital modeling on large kernel metamaterial neural network," J. Imaging Sci. Technol. **67** (2023).

[38] A. G. Howard, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," Preprint, arXiv:1704.04861 (2017).