IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 October 2022, accepted 20 November 2022, date of publication 24 November 2022,
date of current version 30 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3224588

RESEARCH ARTICLE

CM-UNet: ConvMixer UNet for Segmentation of Unknown Objects in Cluttered Scenes

XIAOQIAN HUANG[1], RANA AZZAM[2], SAJID JAVED[2], DONGMING GAN[3], LAKMAL SENEVIRATNE[2], ABDELQADER ABUSAFIEH[4], AND YAHYA ZWEIRI[1,2], (Member, IEEE)

[1]Advanced Research and Innovation Center (ARIC), Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[2]Khalifa University Center for Autonomous Robotic Systems (KUCARS), Khalifa University, Abu Dhabi, United Arab Emirates
[3]School of Engineering Technology, Purdue University, West Lafayette, IN 47907, USA
[4]SVP Technology and Advanced Materials, Strata Manufacturing PJSC, Al Ain, United Arab Emirates

Corresponding author: Xiaoqian Huang (xiaoqian.huang@ku.ac.ae)

This work was supported in part by the Advanced Research and Innovation Center (ARIC), which is funded in part by STRATA Manufacturing PJSC (a Mubadala company); and in part by the Khalifa University of Science and Technology (Khalifa University Center for Autonomous Robotic Systems) under Grant RC1-2018-KUCARS.

ABSTRACT Object segmentation in cluttered environments is a fundamental pre-processing step for many perception-related tasks such as vision-based robotic grasping. Most of the existing object segmentation methods are incapable of precisely segmenting unknown objects, particularly in scenarios exhibiting significant occlusion. In this paper, we propose a novel approach for refining the segmentation of unknown objects in cluttered scenes. More specifically, a ConvMixer-based UNet model is designed to enhance the segmentation mask and boundary of unknown objects appearing in cluttered scenes. In our model, we leverage the object's semantic and localization information, which are essential for successful segmentation, using a ConvMixer-based Cross Fusion (CMCF) module. Furthermore, we propose to use patch embedding as a pre-processing step, where input data is rearranged to expedite processing and improve the efficiency of the system. CM-UNet was trained and extensively tested on various challenging publicly available datasets, including unknown objects in un-structured scenes. Thorough evaluations, in terms of segmentation accuracy and processing efficiency, were conducted against state-of-the-art solutions, where the superiority of our model was proven. CM-UNet has shown its ability to efficiently improve the segmentation accuracy of unknown objects in cluttered scenes, even in presence of occlusion.

INDEX TERMS ConvMixer-based network, UNet, object segmentation, cluttered scene, unknown objects, robotic grasping.

I. INTRODUCTION

Robotic grippers have enabled the automation of key manufacturing processes in the industry and hence have gained immense importance over the past years, especially with assist of perception such as vision-based tactile sensing [1], grasping slip detection [2] and robotic sorting applications [3]. Robotic grasping is one of the tasks that robotic grippers have excelled in the industrial field, where they have been shown to expedite manufacturing while improving throughput. Robotic grasping is a complex task by which a robotic gripper grasps a particular object from its surroundings, after attentively perceiving the environment, identifying and locating the object of interest, and finally planning the kinematics of the system. The success of robotic grasping is directly affected by the quality of the segmentation technique used to locate the object prior to grasping it. Several segmentation approaches have been proposed in the literature, yet the majority assumes structured task environments that do not resemble the actual industrial production line. This makes the segmentation approaches prone to high errors and hence hinders the full automation of the system. Therefore, there is

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

an absolute necessity to improve the accuracy and efficiency of object segmentation approaches, particularly in the case of unknown object geometries in cluttered environments.

State-of-the-art object segmentation approaches can be classified into two main categories; model-based and learning-based. Model-based segmentation approaches assume that objects exhibit a particular geometry, without considering their structural shape variations. In the past decades, a multitude of model-based segmentation methods were proposed based on the Active Contour Model (ACM) [4], whose efficiency for object segmentation was verified through various approaches [5], [6], [7]. Model-based segmentation approaches were also developed based on the Gaussian mixture model (GMM) [8], [9], which is a statistical model that can well describe the spatial distribution and the characteristics of the data in the parameter space. However, the recognition model is generally developed based on various parameters and priors, the selection of which is very significant yet challenging. Moreover, such models suffer from imaging noise caused by the intensity inhomogeneity and the local characteristics of image gradients. This has motivated the introduction of learning-based approaches to achieve more efficient segmentation that is resilient to variations in object geometries across various environmental scenarios.

Learning-based object segmentation approaches are developed based on various deep learning models that have demonstrated unprecedented performance and have achieved significant results. For example, the authors in [10] proposed a new hybrid method for switching between linear and nonlinear spectral unmixing of hyper-spectral data based on neural networks as a possible way to achieve segmentation. Convolutional neural network (CNN) is the most widely used class of neural networks for such applications. The problem of object detection, which constitutes a significant component of object segmentation, involves processing an image to identify and locate instances of objects of interest in a particular scene. Locating an object implies estimating its location and size to facilitate defining its bounding box. In view of the fact that multiple objects may appear in a single scene and that objects may be at various locations in different sizes, the object detector is presented with endless possibilities to work out the problem. In other words, the object detector has to process a huge amount of ''regions'' in the image to correctly pinpoint the location of the object of interest. Alternatively, Region CNN (R-CNN) was proposed in [11] to mitigate this issue by reducing the amount of regions that have to be examined by the neural network. A selective search [12] is used to select a fixed number of regions, referred to as region proposals, to pass to the CNN to carry out object detection. Nevertheless, having to perform feature extraction for every region proposal is computationally expensive, rendering the performance inefficient for the target applications. To further alleviate the shortcomings of this approach, Fast R-CNN was proposed in [13] where the convolution operation is carried out only once per image instead of region proposal. A feature map is generated and processed by a set of fully connected layers that generate the detection results. In addition to the bounding box, Mask R-CNN was proposed in [14] where an object mask is also generated. Transformers neural networks, which currently lead the trend in computer vision [15], were shown to successfully perform object segmentation with high precision. This is attributed to their ability to process global features in the image by means of their self-attention module [16].

Despite their ability to achieve high segmentation accuracy in simple environments, the majority of the existing approaches to object segmentation suffer to correctly segment objects of interest in cluttered environments, under various illumination conditions, an in presence of occlusions [17]. Depending on the environment, the size of the object in the observed scene, and the camera's field of view, different parts of the object of interest may be occluded [18]. Hence, for successful object segmentation, it is necessary to recover or at least predict the occluded part of the object prior to object detection, which is challenging to achieve. Another major limitation is that of the assumption of prior knowledge of the objects of interest, since segmentation approaches depend on object detection networks for known objects. However, the majority of the practical scenarios involve target objects that are unknown. This has motivated the emergence of approaches for refining the predicted object segmentation, which in combination with existing segmentation approaches is capable of improving the accuracy of prediction.

In this work, we address object segmentation refinement for unknown objects in cluttered environments. More particularly, we present a novel ConvMixer-based U-Net (CM-UNet) for segmentation refinement of unknown objects, as illustrated in Fig. 1. The model consists of an encoder (conv block and ResNet block) and a decoder (ConvMixer Cross Fusion (CMCF) module) for contextual feature extraction and spatial information fusion, respectively. We developed the CMCF module to leverage the semantic and localization information while filtering out unrelated features using a light-weight architecture. Compared to the state-of-the-art transformer-based U-Net, our CM-UNet with CMCF shows a huge advantage of around 50% reduction on time efficiency, in addition to better refinement accuracy. Through ablation study and experiments, we provide evidences that patch embedding (PE) is of great importance to both transformer and ConvMixer based architectures due to its ability of locality preservation. Furthermore, our developed CMCF module has the ability to filter out non-semantic features to achieve more accurate segmentation.

II. RELATED WORK

A. OBJECT SEGMENTATION

Object segmentation approaches in the literature can be classified into recognition/model-based approaches and learning-based approaches. The latter are more prevalent among the state-of-the-art approaches and heavily depend on feature

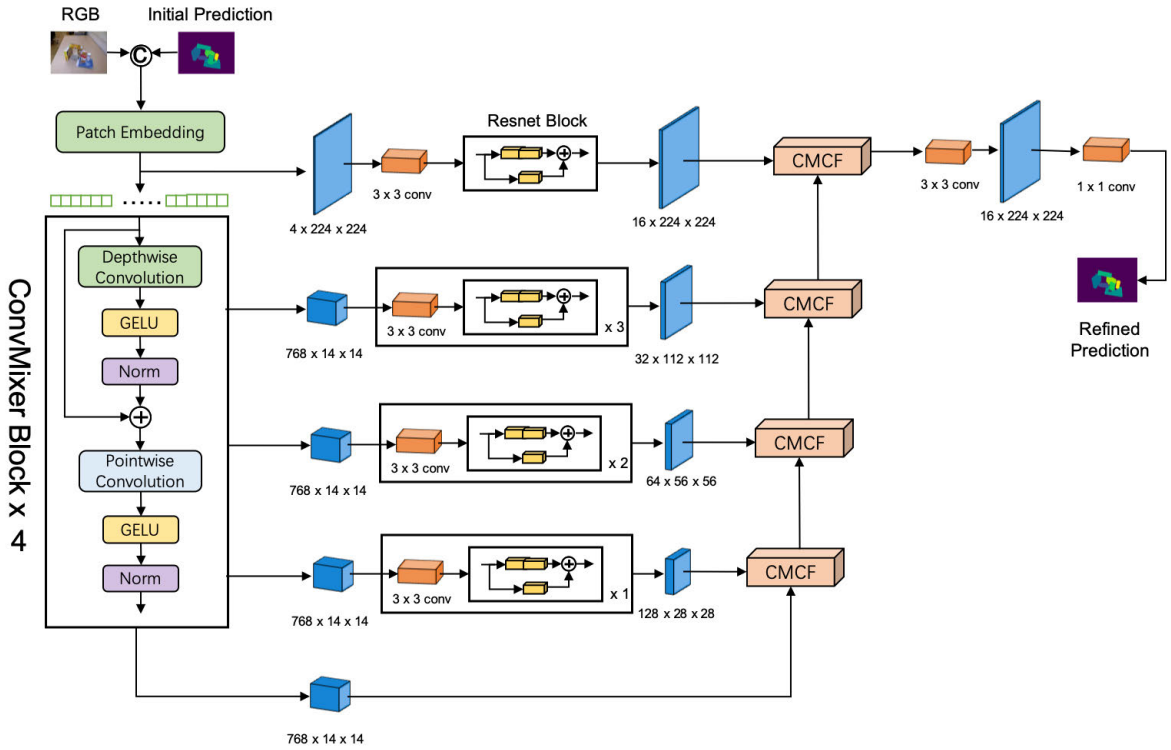


FIGURE 1. Proposed CM-UNet which is a fully convolutional network with developed CMCF module using spatial-locations mix mechanism. The model consists of encoder (conv block and ResNet block) and decoder (ConvMixer Cross Fusion (CMCF) module) for contextual feature extraction and spatial information fusion, respectively.

extraction. Convolutional neural networks (CNNs) are known as a powerful tool for extracting representative features in images. Nevertheless, it suffers from the loss of spatial information, which is attributed to the convolution operation that downsamples the features in every convolutional layer [19].

On the other hand, standard transformers neural networks can capture long-range correlations between feature elements by means of the self-attention mechanism. Besides the self-attention mechanism, U-Net transformer networks make use of cross attention in skip connections to further filter out non-semantic information from the spatial information and hence obtain the correlations between elements [20]. Accordingly, U-Net transformer networks outperform U-Net and attention U-Net in terms of segmentation accuracy when applied on a small dataset.

The transformer's self-attention module is limited to explore intra-sample correlation. To contemplate intra- and inter-correlation, researchers have developed Mixed Transformer U-NET (MT-Unet) with Mixed Transformer Module (MTM) as presented in [21]. MTM consists of two parts; Local-Global Gaussian -Weighted Self-Attention (LGG-SA) with lower computation cost, and External Attention (EA) for inter-correlation learning. Experimentation results have shown that MT-Unet surpasses other state-of-the-art methods without pre-training.

For unknown object segmentation, some examples demonstrate that two-stage prediction, initial segmentation and refinement, can work well [22]. A two-stage Fully Convolutional Neural Network (FCNN) pipeline was proposed in [23] to predict and refine the segmentation of human hairs. Specifically, the second stage was designed as a border refinement with a symmetric encoder-decoder FCNN architecture to refine the hair boundary. In [24], Progressive Boundary Refinement Network (PBRNet) whose structure is similar to that of U-Net, was firstly applied into temporal action detection problem. The network structure is designed for multiple tasks including coarse pyramidal detection and refined pyramidal detection, then the output goes through the fine-grained detection module to localize the action boundary and segment action instances precisely. A two-stage cascaded U-Net was developed in [25], which fine segments objects in the second refining stage based on the coarse segmentation in the first stage, thanks to the automatic context from the original input. Inspired by the two-stage learning-based segmentation, we developed CM-UNet which exhibits a symmetric architecture with skip connections to refine the initially predicted segmentation. Segmentation refinement benefits from such architecture due to its ability to enrich the semantic information associated with the object of interest.

B. ConvMixer

When using the standard Vision Transformer (ViT) model [26], the first step is to embed the input images as patches then pass them as inputs to transformer encoders.

However, its performance highly relies on and is sensitive to the training hyper-parameters such as the optimizer and learning rate. By comparing the performance of standard ViT on ImageNet, the authors in [27] observed that it underperforms state-of-the-art CNNs. Moreover, the self-attention mechanism in transformer networks has a quadratic time computation complexity $O(n^2)$ and requires $O(n^2)$ memory, that could be computationally expensive applied on the dataset of large-size images [20]. However, the complexities of a point-wise and a depth-wise convolutions are $O(patches * channels^2)$ and $O(patches * channels * kernel\_size^2)$, respectively. By comparison, the ConvMixer network is more suitable for large-size image datasets. Then, the authors replace the patch stem with a standard convolutional stem, which demonstrates a better performance on ImageNet with faster convergence and greater in-sensitiveness to hyper-parameters. Enlightened by the idea of mixer's MLPs blocks [28] and the direct processing of embedding patches [26], the MLP-mixer is developed based on multi-layer perceptrons (MLPs) without any convolutions or self attention [29]. Two types of MLPs are used; the channel-mixing MLPs and the token-mixing MLPs, allowing the communication between channels and spatial locations. It shows competitive performance to the image classification benchmarks. Besides, the order of patches in images and pixels in patches does not affect the MLP-mixer's performance.

Then, the question that whether transformer benefits more from its architecture or input patches was explored in [30]. The experimental results indicate that the patch representation probably leads to the great performance of ViT and other new architectures. Building on this discovery, ConvMixer was developed based on a simple architecture that consists of a patch embedding layer and repeated fully-convolution block [30]. Different from ViT and MLP-Mixers, ConvMixer only uses standard convolutions, yet it outperforms both ViT and MLP-Mixers and is competitive with the standard vision models such as ResNet with sub-optimal hyper-parameters.

In this work, we developed CM-UNet for object segmentation refinement based on learning long-term spatial and contextual features, but with less computational complexity and higher efficiency compared to transformer-based U-Net. In addition, we explored that patch embedding does play an important role in the segmentation refinement task.

III. PROPOSED APPROACH

In this section, the overall architecture of the proposed ConvMixer UNet (CM-UNet), illustrated in Fig. 1, will be presented in detail. The system consists of three main components; patch embedding, spatial-locations mix mechanism, and CMCF module as will be discussed in Section III-A, Section III-B, and Section III-C, respectively.

As mentioned in Section I, U-Net is capable of fusing both contextual and positional information. However, it cannot perform complex segmentation and refinement tasks with traditional convolution layers due to the lack of global features. ConvMixer exhibits a transformer-like structure that directly operates on embedded patches to allow a larger receptive field, but with less complexity and parameters compared to a transformer. As the patches go through more encoder and decoder layers in the network, the size of the feature map decreases and hence the amount of contextual features reduces. To that end, we developed ConvMixer-based Cross Fusion module to filter out the non-semantic information and to enrich the semantic information in the deep network layers.

A. PATCH EMBEDDING

While transformer neural networks show outstanding performance in natural language processing (NLP), their computational complexity is very high when used for image processing due to the high execution time and memory consumption requirements. For an image with height H and width W, the computational cost can reach $(H \times W)^2$ which cannot be executed on general hardware. The use of transformer networks in computer vision was pioneered by A. Dosovitskiy [26] upon the proposal of ViT with patch embedding module.

Patch embedding is the process by which an image $x \subseteq R^{H\times W\times C}$ is first split into patches of the same height and width P, analogous to word tokens in NLP, then re-arranged into a flattened 2D sequence, while preserving their locality. The patch embedding process is described in Equation (1).

$$x \subseteq R^{H\times W\times C} \rightarrow x \subseteq R^{N\times H/P\times W/P} \tag{1}$$

where $N = C \times H \times W/P^2$ is the output channel. H, W, C and P represent the height, width, image channels, and patch size [26]. Such rearrangement drastically reduces the computational complexity of the algorithm as compared to processing the whole image directly. The remaining modules in the processing pipeline are applied to the patches directly.

In this work, input images in both training and testing datasets are resized as $224 \times 224 \times 3$. After applying patch partitioning with patch size 16 and depth 4, a sequence of 784 patches will be obtained as expressed in Equation (1). Then a $786 \times 14 \times 14$ image is obtained for the subsequent convolutional operations of ConvMixer as shown in Equation (2).

$$z_{pe} = BN(\sigma \cdot Conv(z_{in}, stride = P, kernel\_size = P))\} \tag{2}$$

where $Conv$, $sigma$ and BN represent convolution operation, activation function and batch normalization, respectively, and z_{in} represents the inputs.

B. SPATIAL-LOCATIONS MIX MECHANISM

The key function of the self-attention mechanism in transformer networks is to extract the long-term and global features [28]. In the field of computer vision, this technique is widely used to enhance feature discrimination [31]. In ConvMixer, convolutions are employed with a large kernel to mix spatial locations from different distances. As illustrated in Fig. 2, the ConvMixer layer is a residual structure that utilizes pure convolutions. To achieve the goal of mixing spatial

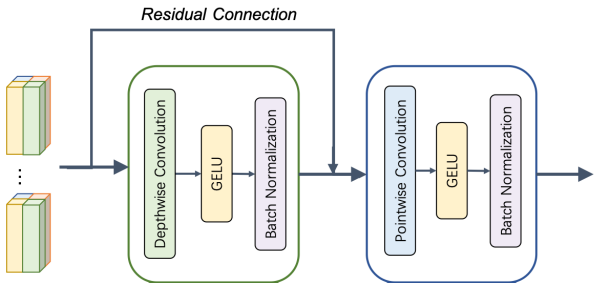


FIGURE 2. ConvMixer layer: residual connection of depth-wise convolution and normalization, followed by point-wise convolution.

locations, the result of residual connection of the inputs and outputs processed by depth-wise convolution goes through the point-wise convolution as calculated in Equation (3).

$$z_r = BN(\sigma \cdot DepthConv(z_{pe})) + z_{pe}$$
$$z_c = BN(\sigma \cdot PointConv(z_r)) \quad (3)$$

The difference among standard, depth-wise, and point-wise convolutions is as depicted in Fig. 3. For the case of normal convolution, illustrated in Fig. 3 (a), the filter is applied to the input to mix channel information. More particularly, as the input image and the filter have the same depth C, the channel information will be convolved and hence the output's depth will be equal to 1. As for the depth-wise convolution, shown in Fig. 3 (b), the filter has the same depth as the input, yet channel-wise convolution is carried out. Hence, the number of the channels in the output remains unchanged and the channel information is reserved and inherited throughout the convolution operation. Point-wise convolution, on the other hand, refers to the standard convolution with $1 \times 1 \times C$ filter as indicated in Fig. 3 (c), and hence considers information of individual input elements.

C. ConvMixer CROSS FUSION MODULE (CMCF)

Inspired by the self-attention mechanism and ConvMixer's structure, we designed the ConvMixer Cross Fusion Module (CMCF) to enrich the semantic information of the low-resolution maps obtained from the ResNet blocks. The global dependencies and relationships between the contextual and spatial information can be learned explicitly. Moreover, it allows to filter out the non-semantic features and to obtain a fine spatial resolution.

As illustrated in Fig. 4, after position embedding, point-wise convolution is applied to the higher-resolution feature maps x_{enc} with d channels and $2H \times 2W$ resolution to mix spatial locations and output x_p. Similarly, the depth-wise convolution is applied to the lower-resolution feature maps y_{dec} with $2d$ channels and $H \times W$ resolution to mix channel locations and output y_p. Inspired by the transformer architecture, the x_{pe}, z_u, and y_u are analogous to the key, value, and query in the self-attention mechanism. After aggregating spatial and channel information, z_u is up-sampled and multiplied by x_{pe} to get the attention map from key and query. Finally, we apply

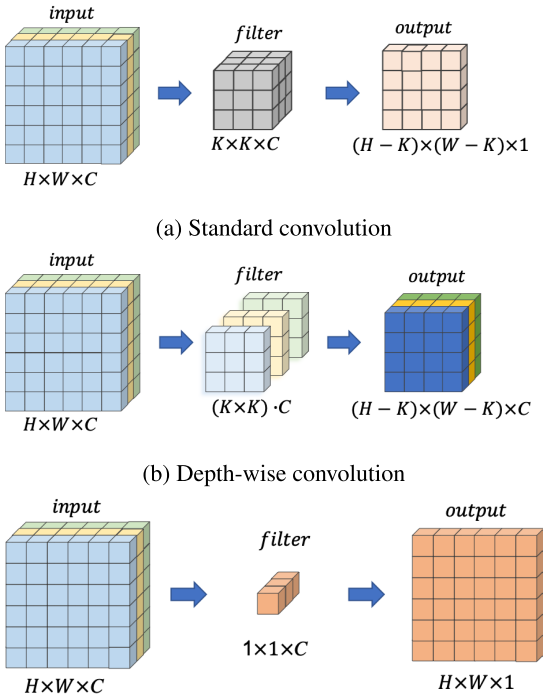


FIGURE 3. Working principle and comparison of standard convolution, depth-wise convolution and point-wise convolution.

the dot-product attention to output the feature maps as in Equation (4), where $Conv2D$ represents the 2D convolution block in Fig. 4:

$$out_j = Conv2D(\sum_{i=1}^{N}[x_{pei} \cdot z_{uj}, y_{u\{i,j\}}]) \quad (4)$$

IV. PERFORMANCE EVALUATION

A. DATASETS

To train the proposed segmentation refinement method for unknown objects for robotic grasping tasks, the Tabletop Object Dataset (TOD) [32] is used. TOD is a synthetic, large-scale dataset consisting of 20k cluttered scenes and a total of 100k images of objects in an indoor environment. For evaluations and testing, the OCID and OSD public datasets [33] were employed. For each scene in the OCID dataset, a 640 × 480 organized XYZRGBL point cloud, depth image, RGB image and 2d-label-masks with unique integer-label for each object are provided. A total of 89 representative objects are selected from the Autonomous Robot Indoor (ARID) and YCB Object and Model Set (YCB) subsets. Such objects were placed in various arrangements where they appeared separated from each other, physically touching each other, or occluded. As for the OSD, the RGB image, depth image, and the segmentation ground truth are provided. The OCID contains 2346 images labeled semi-automatically and the OSD contains 111 images labeled manually. Such images

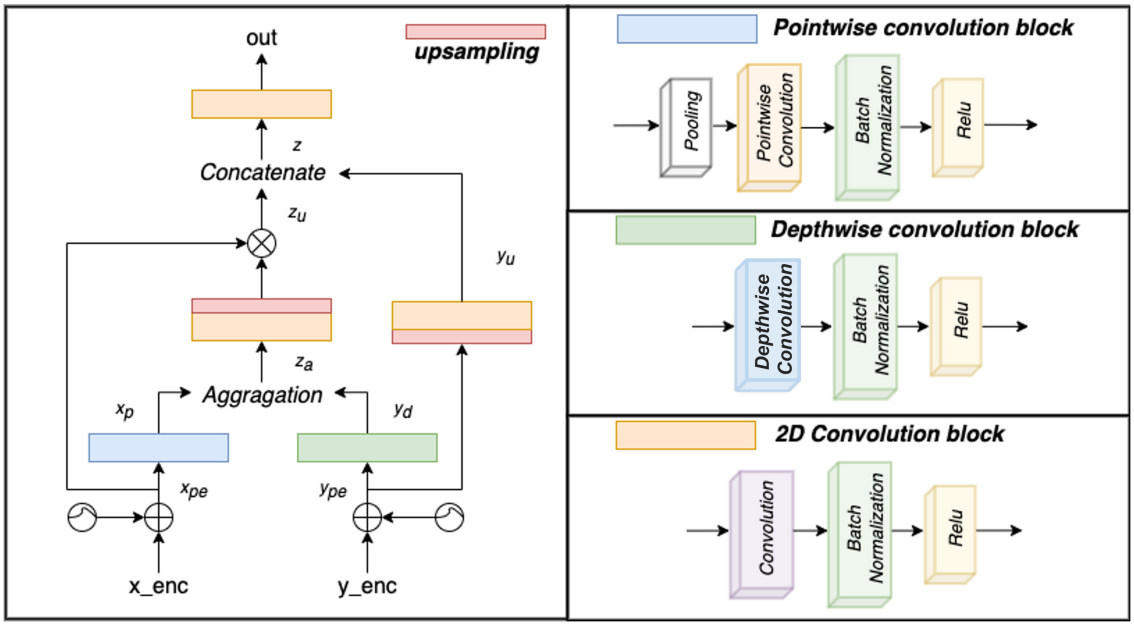


FIGURE 4. Architecture of ConvMixer Cross Fusion Module (CMCF), which is inspired by the self-attention module but with full convolution operations.

include objects placed on a table or on the floor in cluttered and real scenes.

B. IMPLEMENTATION DETAILS

Our proposed segmentation refinement network consists of a U-shape four encoder blocks and four decoder blocks, with a four-channel input concatenating the initial predicted segmentation mask and the original RGB image. We trained CM-UNet in 30 training epochs and used a batch size of 16 on TOD datasets using ADAM optimizer with 10e-4 learning rate. During training, we used the weighted BCE loss L_{seg} to calculate the difference between the predicted masks $y = \{y_1, y_2, \ldots, y_n\}$ and the ground truth masks $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ as:

$$L_{seg} = \frac{\sum_i^N (L_n \cdot Mask_w)}{\sum_i^N Mask_w} \tag{5}$$

where $Mask_w$ is the weighted mask, and $L_n = -\frac{1}{N}[\hat{y}_n \cdot log\sigma(y_n) + (1-\hat{y}_n) \cdot log(1-\sigma(y_n))]$ is the BCE loss to measure the predicted mask error of a single batch.

We evaluate the performance from two aspects; (1) the overlapping area between the segmented mask and the corresponding ground truth, which will be referred to as overlap hereafter, and (2) the overlap between the detected and ground truth boundaries that outline the objects, which will be referred to as boundary. The *F_score*, *Precision* and *Recall* are used to evaluate the matching degree of the predicted segmentation and the corresponding ground truth [34]. Particularly, *Precision* indicates the quality of segmentation calculated as the percentage of correctly labeled pixels. *Recall* represents the ratio of correctly segmented pixels to the total of pixels in the ground truth scene. Consequently, the *F_score* is computed as the harmonic mean of *Precision* and *Recall*.

$$Precision = \frac{\sum_{i=1}^N ob_i \cap gt_i}{\sum_{i=1}^N ob_i} \tag{6}$$

$$Recall = \frac{\sum_{i=1}^N ob_i \cap gt_i}{\sum_{i=1}^N gt_i} \tag{7}$$

$$F\_score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

where $ob = \{ob_1, ob_2, \ldots, ob_i\}$ represents the segmentation results, and $gt = \{gt_1, gt_2, \ldots, gt_i\}$ represents the corresponding ground truth. $\sum_{i=1}^N ob_i \cap gt_i$ indicates the number of pixels in the overlapping area between the predicted segmentation and the matched reference object.

C. EXPERIMENTAL RESULTS

We employed our CM-UNet to refine the segmentation predicted by DSN [35] on OCID and OSD datasets, and evaluated the performance using *Precision*, *Recall*, and *F_score*, as described in Section IV-B. Tables 1 shows the evaluation results of segmentation refinement on OCID dataset. Table 2 shows the quantitative evaluation of segmentation refinement on OSD dataset. It is clear from both tables that the proposed ConvMixer-UNet improves the segmentation done by DSN with around 8% higher F_score on mask overlap. Besides, Fig. 5 illustrates sample scenes from the testing set, along with the corresponding segmentation ground truth, predicted segmentation, and refined segmentation. It can be noticed that the proposed segmentation refinement method was able to improve the initial segmentation output and resulted in a higher number of correctly segmented objects. In addition, the refined masks obtained from the proposed method are

TABLE 1. Quantitative evaluation of segmentation refinement on OCID dataset. *Precision*, *Recall*, and *F_score* (as described in Sec IV-B) are utilized to compare the initial segmentation predicted by DSN [35] and its refined segmentation by our model.

Terms		DSN	Our model
Overlap	F_score	0.76482	0.85298
	Precision	0.85346	0.85299
	Recall	0.80708	0.85714
Boundary	F_score	0.71467	0.72916
	Precision	0.77474	0.74529
	Recall	0.74460	0.72048

TABLE 2. Quantitative evaluation of segmentation refinement on OSD dataset. *Precision*, *Recall*, and *F_score* (as described in Sec IV-B) are utilized to compare the initial segmentation predicted by DSN [35] and its refined segmentation by our model.

Terms		DSN	Our model
Overlap	F_score	0.74686	0.82932
	Precision	0.70175	0.85711
	Recall	0.81863	0.82121
Boundary	F_score	0.55861	0.67047
	Precision	0.53187	0.67569
	Recall	0.65181	0.69002

closer to the instant contours observed from real RGB images. It is also worth noting that our proposed model can refine the predicted mask of OCID to more accurately resemble the scene than the ground truth mask. This is attributed to the possible errors of ground truth resulting from the semi-automatic labeling of the dataset.

D. COMPARISON TO STATE-OF-THE-ART METHODS

In this section, the same experimental protocols described in Section IV-B will be carried out to demonstrate the performance improvement achieved by our proposed method as compared to state-of-the-art segmentation approaches; particularly mask-RCNN [14], UCN [36], PointGroup [37], and DSN [35]. Testing scenarios are taken from the OCID and OSD datasets [38] and segmentation performance is quantitatively evaluated using the normalized metrics, because the values obtained using the unnormalized metrics in Equations (6)-(8) are heavily affected by the size of the objects in the scene. For instance, if there are two objects in the scene; one is much bigger than the other and the individual accuracies of their segmentation are drastically different, the overall accuracy will be closer to the segmentation accuracy of the large object. To circumvent this issue, we utilized the normalized metrics to make them independent of the object sizes in the scene, as listed in Equations (9) and (10), where m, n are the labels of prediction and ground truth of individual objects, and E represents the Hungary assignment between the predicted and ground truth instance masks I_m^M and I_n^N. Besides, $P_{m,n}$, $R_{m,n}$, and $F_{m,n}$ represent *Precision*, *Recall*, and F_score of I_m^M and I_n^N. M, N are the number of segmented objects and the true number of objects in the scene, respectively.

$$F_{m,n} = \frac{2 * P_{m,n} * R_{m,n}}{P_{m,n} + R_{m,n}} \tag{9}$$

$$F\_score' = \frac{\sum_{(m,n)\subset E} F_{m,n}}{max(N, M)} \tag{10}$$

TABLE 3. Results of segmentation refinement performance quantified using overlap *F_score* and *F_score′* on both OCID and OSD datasets as compared to state-of-the-art segmentation approaches.

MODELS	F_score		F_score'	
	OCID	OSD	OCID	OSD
MASK-RCNN (2018) [14]	0.6314	0.6411	0.6324	0.5971
MASK-RCNN + **our model**	**0.6671**	**0.7172**	**0.6914**	**0.6912**
UCN (2020) [36]	0.7599	0.7931	0.7324	0.7136
UCN + **our model**	**0.8071**	**0.8021**	**0.7613**	**0.7295**
PointGroup (2020) [37]	0.6320	0.6659	0.6145	0.6011
PointGroup + **our model**	**0.6564**	**0.7110**	**0.6757**	**0.6244**
DSN (2021) [35]	0.7648	0.7469	0.7147	0.7469
DSN + **our model**	**0.8530**	**0.8293**	**0.8176**	**0.8057**

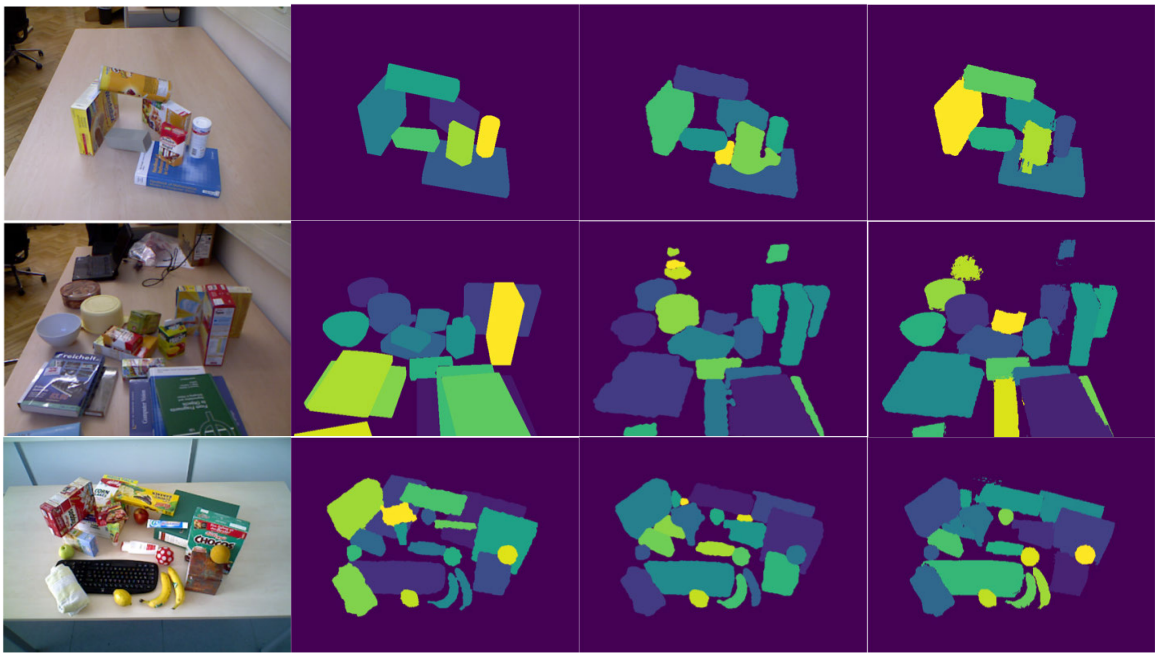
Table 3 lists the F_score and F_score' of the predicted segmentation results obtained by state-of-the-art segmentation methods with and without our proposed refinement approach on two testing datasets. The table shows the refinement improvement achieved by our approach as compared to the listed state-of-the-art segmentation techniques. Based on the predicted segmentation by mask-RCNN [14], UCN [36], PointGroup [37], and DSN [35], our model improves the overlap F_score' by 9.33%, 3.95%, 9.96%, and 14.40% on OCID dataset, respectively. Besides, our model improves the overlap F_score' by 15.76%, 2.23%, 3.88%, and 7.87% on OSD dataset, respectively. Similarly, our model improves the overlap F_score by 5.65%, 6.21%, 3.86%, and 11.53% on OCID dataset, respectively. Besides, our model improves the overlap F_score by 11.87%, 1.13%, 6.77%, and 11.03% on OSD dataset, respectively. Therefore, these tests have proven the effectiveness of our model and have demonstrated its ability to enhance the segmentation accuracy on challenging datasets comprising unknown objects in cluttered scenes.

E. ABLATION STUDY

In this section, the selection of (1) the network architecture, i.e. transformer network or ConvMixer, and (2) the modules along the processing pipeline, i.e. patch embedding, CMCF, and input modes, in the proposed segmentation refinement method will be justified through an ablation study. The segmentation results will be compared for different models as listed in Table 4 and as illustrated in Fig. 6. Similarly, F_score, *Precision* and *Recall* in Equation (6)-(8) are utilized to quantify the accuracy of refinement. Moreover, Frames Per Second (FPS) is computed to indicate the rate at which images are being processed. We computed FPS on NVIDIA V100 Tensor Core GPU. Giga floating-point operations (GFLOPs) required for one single pass, are also calculated to evaluate the time efficiency.

1) ConvMixer VS. TRANSFORMER NETWORKS

In this section, a comparison between convolution-based segmentation refinement and attention-based segmentation refinement will be conducted, while maintaining the input and output dimensions of the system. More particularly, a Convmixer network with a convolution-based decoder, CMCF, will be compared to a transformer network with an



(i) RGB images (j) Ground truth (k) DSN predicted Initial mask (l) Refined segmentation

FIGURE 5. Visualization of initial predicted segmentation by DSN on the testing set and the refined segmentation by our model.

attention-based decoder, namely Multi-Head Cross Attention (MHCA) module [20]. Fig.6 depicts the segmentation refinement results achieved by various structures. In this section, the performance of the models referred to as (a), (b), (c), and (f) will be evaluated.

By analyzing the results obtained by model (a), (b), (c), and (f), it was observed that the segmentation refinement was comparable in terms of improving the mask overlap and boundary of the segmented objects, yet the ConvMixer-based models achieved a slightly higher accuracy. However, the image processing time of the ConvMixer-based architectures, models (c) and (f), is reduced by approximately 55.8% compared to the transformer-based architectures, models (a) and (b). Hence, it is concluded that the ConvMixer-based architecture outperforms the transformer-based architectures in terms of computation and time efficiency. This is attributed to the simple architecture exhibited by the ConvMixer-based models which only consists of a patch embedding layer and repeated fully-convolutional layers block.

2) IMPACT OF CMCF

In this section, the effectiveness of the CMCF module will be investigated by comparing the performance of the ConvMixer-based architecture with (model (f)) and without (model (e)) the CMCF module. The same training parameters, experimental protocols, and evaluation metrics were used. As listed in Table 4 and depicted in Fig.6, both models demonstrate outstanding performance on the evaluation set, however, model (f) which contains the CMCF module achieves higher accuracy on overlap and boundary estimation of the predicted segmentation. This is attributed to the fact that CMCF is capable of filtering out the non-semantic features, and hence can provide a more accurate refinement result when mixing the spatial and contextual information. Besides, it is noteworthy that the time efficiency is improved by 18%, due to the separable convolution structure of CMCF module. The number of parameters of a single separable convolution is only $3 \times 3 + 1 \times 3 \times 4 = 39$, which is much less than the one of the traditional convolution block $4 \times 3 \times 3 \times 3 = 108$.

3) IMPACT OF PATCH EMBEDDING

So far, the best performing model in terms of segmentation refinement accuracy and time and memory complexity is the ConvMixer-based model with CMCF. In this section, the impact of adding a patch embedding module, introduced in Section III-A, will be studied. The segmentation refinement results obtained by model (c), ConvMixer + CMCF without Patch Embedding, and model (f), ConvMixer + CMCF + Patch Embedding, are shown in Fig. 6 and the evaluation metrics are listed in Table 4. It is clear from the results that the model with patch embedding performs better than the model without patch embedding in terms of F_score, precision, and recall of both overlap and boundary prediction. Particularly, the F_score of overlap is improved by approximately 2% due to the use patch embedding. The improvement that patch embedding adds to segmentation refinement can also be seen

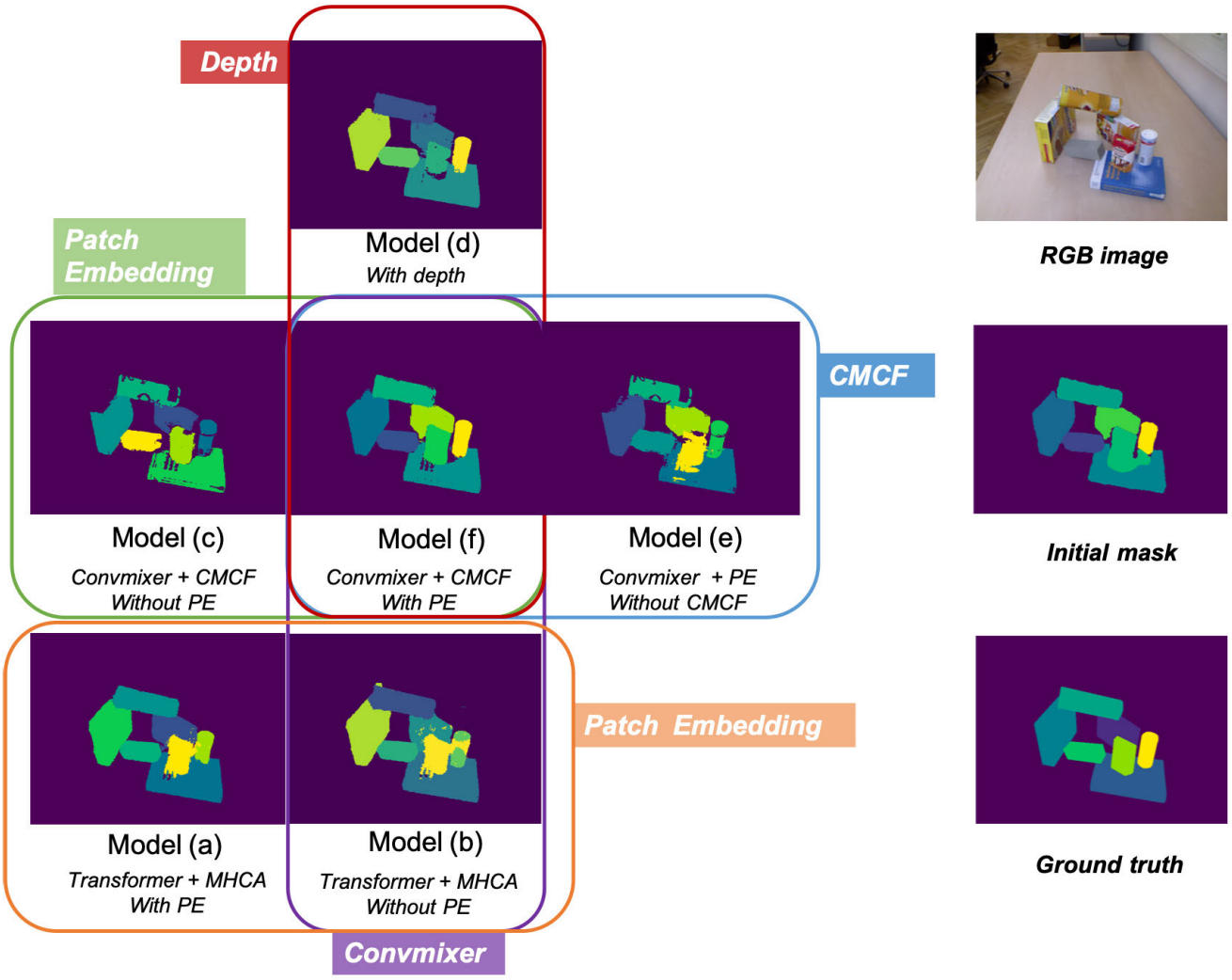


FIGURE 6. Ablation study results - Input RGB image, initial segmentation mask predicted by DSN, and the corresponding segmentation ground truth are depicted on the right side of the figure. The segmentation refinement results achieved by the transformer-based and ConvMixer-based models that were considered in the ablation study are shown and labeled accordingly.

TABLE 4. Ablation study on the effect on ConvMixer, inputs mode, PE and CMCF module by the quantitative asset on accuracy and efficiency. The FPS is computed on NVIDIA V100 Tensor Core GPU.

Terms	Overlap			Boundary			FPS	GFLOPs
	F_score	Precision	Recall	F_score	Precision	Recall		
(a) Transformer + MHCA	0.8409	0.8479	0.8517	0.7172	0.7006	0.7228	0.1797	6.3120
(b) Transformer + PE+ MHCA	0.8503	0.8472	0.8572	0.7228	0.7286	0.7246	0.1791	6.2091
(c) CM+ CMCF	0.8324	0.8342	0.8330	0.7248	0.7381	0.7224	0.2798	7.6215
(d) CM + PE + CMCF+ depth	0.8528	0.8472	0.8526	0.7281	0.7438	0.7203	0.2847	6.5187
(e) CM + PE	0.8482	0.8525	0.8469	0.7222	0.7291	0.7240	0.2869	3.3259
(f) CM+ PE + CMCF	0.8530	0.8530	0.8571	0.7292	0.7453	0.7205	0.3387	6.4721

in the transformer-based models, models (a) and (b). These results justify our choice of this module in the system.

4) IMPACT OF THE INPUT MODE

In this section, the choice of the input to the proposed segmentation refinement method will be justified. As described in the Section IV-B, the input to our system consists of 4 channels; the RBG image and the corresponding initial mask predicted by any segmentation technique, DSN in our case. A fifth dimension, referring to the depth of the scene, was added to the input to test if such information is needed to further refine the segmentation. This model is referred to as model (d) in Fig. 6 and Table 4 and is compared to model (f) which exhibits the exact same architecture and modules, but does not use the depth information in the input. The results obtained from both models appear to have the same

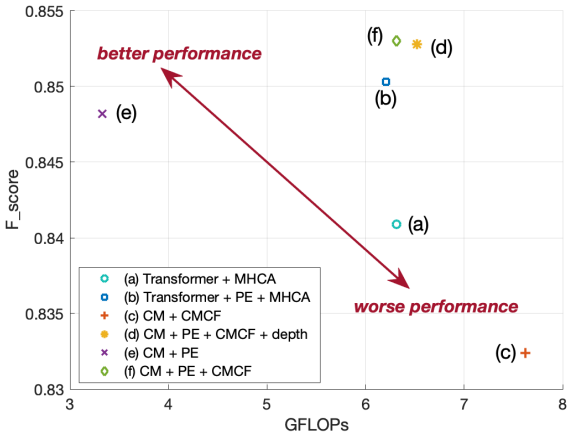


FIGURE 7. Mask overlap F_score and processing speed achieved by the models listed in Table 4.

refinement performance in terms of quantitative evaluation. This is due to that depth information is already incorporated in the initial mask obtained by DSN, and hence having it as an input adds nothing to the accuracy of segmentation. Rather, it negatively affects the computational efficiency of the algorithm by increasing the number of operations needed to carry out the prediction.

The *F_score* of the overlap prediction and the computation complexity for all the models discussed so far are depicted in Fig. 7. Higher *F_score* and lower GFLOPs indicate better capacity on both segmentation refinement accuracy and time efficiency and hence better overall performance. It can be noticed from Fig. 7 that the ConvMixer-based models (d) and (f) with both PE and CMCF present prominent performance where they scored the highest *F_score* and low GFLOPs. Also, it can be seen in Table 4 that model (f) has the highest frame processing rate. In addition, PE and CMCF modules can bring significant improvement on segmentation refinement accuracy when compared to model (c) and (e). On the contrary, transformer-based refinement networks (Model (a), (b)) show poorer performance than ConvMixer based models, both in terms of *F_score* and GFLOPs.

V. CONCLUSION

In this paper, we proposed a CM-UNet for segmentation refinement of unknown objects in cluttered scenarios. The ConvMixer-based Cross Fusion module was developed to fuse the large-scale contextual features and spatial information through encoding the inter-dependent information. When applying the proposed method on the predicted segmentation of the state-of-the-art methods on unseen objects, the average overlapping accuracy is improved by 8.42% compared to the initial prediction by DSN [35] on OCID and OSD datasets [33]. We conducted a thorough ablation study to justify the choice of the ConvMixer-based UNet architecture and we have shown that our proposed method performs better than transformer-based UNet in terms of accuracy and time efficiency in refinement tasks. In addition, we have proven that patch embedding and CMCF modules do bring positive effect on segmentation accuracy.

In the future, we will conduct real robotic grasping experiments for unknown objects using segmentation refinement network. Besides object segmentation, this work can also be applied to semantic segmentation and object localization refinement by transfer learning.

REFERENCES

[1] F. B. Naeini, A. M. AlAli, R. Al-Husari, A. Rigi, M. K. Al-Sharman, D. Makris, and Y. Zweiri, "A novel dynamic-vision-based approach for tactile sensing applications," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 1881–1893, May 2019.

[2] A. Rigi, F. B. Naeini, D. Makris, and Y. Zweiri, "A novel event-based incipient slip detection using dynamic active-pixel vision sensor (DAVIS)," *Sensors*, vol. 18, no. 2, p. 333, 2018.

[3] X. Huang, M. Halwani, R. Muthusamy, A. Ayyad, D. Swart, L. Seneviratne, D. Gan, and Y. Zweiri, "Real-time grasping strategies using event camera," *J. Intell. Manuf.*, vol. 33, no. 2, pp. 593–615, Feb. 2022.

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

[5] H. Ali, N. Badshah, K. Chen, and G. A. Khan, "A variational model with hybrid images data fitting energies for segmentation of images with intensity inhomogeneity," *Pattern Recognit.*, vol. 51, pp. 27–42, Mar. 2016.

[6] T. Chan and L. Vese, "An active contour model without edges," in *Proc. Int. Conf. Scale-Space Theories Comput. Vis.* Cham, Switzerland: Springer, 1999, pp. 141–151.

[7] C. He, Y. Wang, and Q. Chen, "Active contours driven by weighted region-scalable fitting energy based on local entropy," *Signal Process.*, vol. 92, no. 2, pp. 587–600, Feb. 2012.

[8] R. Farnoush and P. B. Zar, "Image segmentation using Gaussian mixture model," Tech. Rep., 2008.

[9] Y. Qi, G. Zhang, Y. Qi, and Y. Li, "Object segmentation based on Gaussian mixture model and conditional random fields," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Aug. 2016, pp. 900–904.

[10] A. Ahmed, O. Duran, Y. Zweiri, and M. Smith, "Hybrid spectral unmixing: Using artificial neural networks for linear/non-linear switching," *Remote Sens.*, vol. 9, no. 8, p. 775, Jul. 2017. [Online]. Available: https://www.mdpi.com/2072-4292/9/8/775

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2014, pp. 580–587.

[12] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.

[13] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[14] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[15] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surveys*, vol. 54, no. 10, pp. 1–41, Jan. 2022.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[17] H. H. Hoang and B. L. Tran, "Accurate instance-based segmentation for boundary detection in robot grasping application," *Appl. Sci.*, vol. 11, no. 9, p. 4248, May 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/9/4248

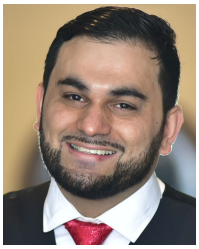
[18] K. Saleh, S. Szenasi, and Z. Vamossy, "Occlusion handling in generic object detection: A review," in *Proc. IEEE 19th World Symp. Appl. Mach. Intell. Inform. (SAMI)*, Jan. 2021, pp. 000477–000484.

[19] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Ann. Transl. Med.*, vol. 8, no. 11, p. 713, Jun. 2020.

[20] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-Net transformer: Self and cross attention for medical image segmentation," in *Proc. IEEE 19th World Symp. Appl. Mach. Intell. Inform. (SAMI)*. Cham, Switzerland: Springer, Sep. 2021, pp. 267–276.
[21] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong, "Mixed transformer U-Net for medical image segmentation," 2021, *arXiv:2111.04734*.
[22] W. Gu, S. Bai, and L. Kong, "A review on 2D instance segmentation based on deep neural networks," *Image Vis. Comput.*, vol. 120, Apr. 2022, Art. no. 104401.
[23] Y. Yan, S. Duffner, X. Naturel, A. Berthelier, C. Garcia, C. Blanc, and T. Chateau, "Two-stage human hair segmentation in the wild using deep shape prior," *Pattern Recognit. Lett.*, vol. 136, pp. 293–300, Aug. 2020.
[24] Q. Liu and Z. Wang, "Progressive boundary refinement network for temporal action detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11612–11619.
[25] X. Huang, Z. Lin, Y. Jiao, M.-T. Chan, S. Huang, and L. Wang, "Two-stage segmentation framework based on distance transformation," *Sensors*, vol. 22, no. 1, p. 250, Dec. 2021.
[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[27] T. Xiao, P. Dollar, M. Singh, E. Mintun, T. Darrell, and R. Girshick, "Early convolutions help transformers see better," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 30392–30400.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[29] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-mixer: An all-MLP architecture for vision," 2021, *arXiv:2105.01601*.
[30] A. Trockman and J. Z. Kolter, "Patches are all you need?" 2022, *arXiv:2201.09792*.
[31] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13756–13765.
[32] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "The best of both modes: Separately leveraging RGB and depth for unseen object instance segmentation," in *Proc. Conf. Robot Learn.*, 2020, pp. 1369–1378.
[33] M. Suchi, T. Patten, D. Fischinger, and M. Vincze, "EasyLabel: A semi-automatic pixel-wise object annotation tool for creating robotic RGB-D datasets," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 6678–6684.
[34] X. Zhang, X. Feng, P. Xiao, G. He, and L. Zhu, "Segmentation quality evaluation using region-based precision and recall measures for remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 102, pp. 73–84, Apr. 2015.
[35] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen object instance segmentation for robotic environments," *IEEE Trans. Robot.*, vol. 37, no. 5, pp. 1343–1359, Oct. 2021.
[36] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, "Learning RGB-D feature embeddings for unseen object instance segmentation," 2020, *arXiv:2007.15157*.
[37] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "PointGroup: Dual-set point grouping for 3D instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4867–4876.
[38] A. Richtsfeld, T. Morwald, J. Prankl, M. Zillich, and M. Vincze, "Segmentation of unknown objects in indoor environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 4791–4796.

XIAOQIAN HUANG received the M.Sc. degree in mechanical engineering from Khalifa University, Abu Dhabi, United Arab Emirates, in 2017. She is currently pursuing the Ph.D. degree in robotics with Khalifa University. During her master's research, she focused on UAV navigation and control based on vision. Her current research interests include robotics and autonomous systems, neuromorphic vision, robotic vision, and artificial intelligence.

RANA AZZAM received the B.Sc. degree in computer engineering, the M.Sc. degree by research in electrical and computer engineering, and the Ph.D. degree in engineering with a focus on robotics from Khalifa University, in 2014, 2016, and 2020, respectively. She is currently a Postdoctoral Fellow with the Department of Aerospace Engineering, Khalifa University. Her research interests include machine learning, reinforcement learning, navigation, and simultaneous localization and mapping.



SAJID JAVED received the B.Sc. degree in computer science from the University of Hertfordshire, U.K., in 2010, and the master's and Ph.D. degrees in computer science from Kyungpook National University, Republic of Korea, in 2017. He is currently an Assistant Professor of computer vision with the Department of Electrical and Computer Engineering (ECE), Khalifa University of Science and Technology, United Arab Emirates. Prior to that, he was a Research Scientist at Khalifa University Center for Autonomous Robotics System (KUCARS), from 2019 to 2021. Before joining Khalifa University, he was a Research Fellow at the University of Warwick, U.K., from 2017 to 2018, where he worked on histopathological landscapes for better cancer grading and prognostication. His research interests include visual object tracking in the wild, multi-object tracking, background-foreground modeling from video sequences, moving object detection from complex scenes, cancer image analytics including tissue phenotyping, nucleus detection, and nucleus classification problems. His research themes involve developing deep neural networks, subspace learning models, and graph neural networks.

DONGMING GAN received the Ph.D. degree in robotics and mechanical engineering from King's College London, Beijing University of Posts and Telecommunications. He is currently an Assistant Professor with the School of Engineering Technology, Purdue University, West Lafayette, IN, USA. His main research interests include robotics, mechanism, and machine theory with a focus on design, modeling, control, and development of intelligent reconfigurable robotic systems, compliant robot manipulators, and flexible light-weight wearable devices for assisting humans in manufacturing, healthcare, and domestic human–robot co-existing scenarios with safe physical interactions and collaborations. He is an Associate Editor of *Mechanism and Machine Theory*, an Associate Editor of *IMechE, Part C: Journal of Mechanical Engineering Science*, a Topic Editor of *Mechanical Sciences*, a Review Editor on the Editorial Board of *Bionics and Biomimetics in Frontiers journal*, and a Guest Editor of *Applied Bionics and Biomechanics*. He is an Elected Committee Member of the ASME DED Mechanism and Robotics Committee and a member of ASME. He has served on program committees and symposiums of several international conferences including IEEE Cyber 2019, IEEE/ASME ReMAR 2012-2021, Parallel 2014/2020, and ASME IDETC 2012-2020.

LAKMAL SENEVIRATNE is currently a Professor of mechanical engineering and the Founding Director of the Centre for Autonomous Robotic Systems (KUCARS), Khalifa University, United Arab Emirates (UAE). He has also served as an Associate Provost for Research and Graduate Studies and an Associate VP Research at Khalifa University. He has published over 400 peer-reviewed publications on these topics. He is a member of the Mohammed Bin Rashid Academy of Scientists in the UAE.

ABDELQADER ABUSAFIEH received the master's degree in mechanical engineering from Villanova University and the Ph.D. degree in materials engineering from Drexel University. He is currently an SVP for Technology and Advanced Materials at STRATA Manufacturing PJSC (a Mubadala Company) where he is responsible for driving research and development strategy and technology development programs within Mubadala Aerospace assets including collaboration initiatives with OEMs, technology partners, and academia. Prior to this role, he worked as a Technical Advisor at Mubadala Aerospace and Defense Unit for investment activities and technology and training initiatives. He has several patents and numerous publications and invited seminars. He sits on a number of senior management boards in academia and industry.

YAHYA ZWEIRI (Member, IEEE) received the Ph.D. degree from the King's College London, in 2003. He is currently an Associate Professor with the Department of Aerospace Engineering and the Deputy Director of Advanced Research and Innovation Center, Khalifa University, United Arab Emirates. He was involved in defense and security research projects in the last 20 years at the Defense Science and Technology Laboratory, King's College London, and the King Abdullah II Design and Development Bureau, Jordan. He has published over 130 refereed journals and conference papers and filed ten patents, USA and U.K. His main research interest includes robotic systems for extreme conditions with particular emphasis on applied AI aspects and neuromorphic vision systems.