Location-Aware Encoding for Lesion Detection in ^{68}Ga-DOTATATE Positron Emission Tomography Images

Fuyong Xing, *Member, IEEE*, Michael Silosky, Debashis Ghosh, and Bennett B. Chin

***Abstract*— *Objective:* Lesion detection with positron emission tomography (PET) imaging is critical for tumor staging, treatment planning, and advancing novel therapies to improve patient outcomes, especially for neuroendocrine tumors (NETs). Current lesion detection methods often require manual cropping of regions/volumes of interest (ROIs/VOIs) a priori, or rely on multi-stage, cascaded models, or use multi-modality imaging to detect lesions in PET images. This leads to significant inefficiency, high variability and/or potential accumulative errors in lesion quantification. To tackle this issue, we propose a novel single-stage lesion detection method using only PET images. *Methods:* We design and incorporate a new, plug-and-play codebook learning module into a U-Net-like neural network and promote lesion location-specific feature learning at multiple scales. We explicitly regularize the codebook learning with direct supervision at the network's multi-level hidden layers and enforce the network to learn multi-scale discriminative features with respect to predicting lesion positions. The network automatically combines the predictions from the codebook learning module and other layers via a learnable fusion layer. *Results:* We evaluate the proposed method on a real-world clinical ^{68}Ga-DOTATATE PET image dataset, and our method produces significantly better lesion detection performance than recent state-of-the-art approaches. *Conclusion:* We present a novel deep learning method for single-stage lesion detection in PET imaging data, with no ROI/VOI cropping in advance, no multi-stage modeling and no multi-modality data. *Significance:* This study provides a new perspective for effective and efficient lesion identification in PET, potentially accelerating novel therapeutic regimen development for NETs and ultimately improving patient outcomes including survival.**

***Index Terms*— Lesion detection, PET, neuroendocrine tumors, deep neural networks, location-aware encoding**

I. Introduction

GASTROENTEROPANCREATIC neuroendocrine tumors (GEP-NETs) are rare, difficult-to-detect tumors which commonly present at advanced stages, with the liver as the most common site of metastases [1]. ^{68}Ga- and ^{64}Cu-DOTATATE positron emission tomography-computed tomography (PET/CT) are widely used molecular imaging techniques for NETs [2]–[4] and show very promising results for accurate staging of GEP-NETs [5], [6]. In order to develop effective treatments, it is critical to correctly identify lesions in PET images. Lesion detection with a high positive predictive value can dramatically accelerate the process of clinical interpretation. Manual lesion identification in ^{68}Ga- and ^{64}Cu-DOTATATE PET images is very labor-intensive, time-consuming and prone to intra-/inter-observer variation in image interpretation. An automated method for accurate lesion localization can assist with detection of NETs and treatment planning, and thus potentially improve patient outcomes including survival.

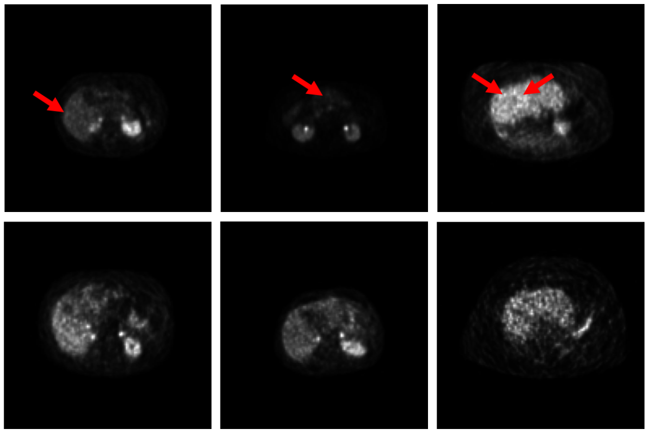


Fig. 1. Some example ^{68}Ga-DOTATATE PET images of livers. Row 1 denotes three different abnormal subjects with each having one or more hepatic lesions (pointed out by red arrows), and row 2 represents three normal subjects without liver lesions.

The complex nature of PET images, however, poses significant challenges for automated lesion detection [7], [8], as shown in Fig. 1. First, PET images usually exhibit low spatial resolution and image contrast such that the boundaries between lesions and surrounding normal regions are not clear. Second, noise is inherently high in PET images compared with anatomical imaging modalities such as CT and magnetic resonance imaging (MRI). In addition, ^{68}Ga-DOTATATE PET imaging typically uses lower administered dose and faster radionuclide decay than the prevailing ^{18}F-fluorodeoxyglucose (FDG)-PET [9] diagnostic tool, so that ^{68}Ga-DOTATATE PET usually has higher image noise and a lower signal-to-noise ratio, thus significantly affecting the lesion detectability. Finally, lesions often show large variability in the shape, size,

F. Xing and D. Ghosh are with Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, CO 80045, USA (e-mail: fuyong.xing@cuanschutz.edu).

M. Silosky and B. B. Chin are with Department of Radiology, University of Colorado Anschutz Medical Campus, CO 80045, USA.

texture, intensity inhomogeneity and other features, and this further challenges lesion detection or identification algorithms for PET images.

Thresholding-based methods are commonly used for lesion identification in PET images at an early stage, but the assumptions on which these methods rely rarely hold in real-world practice [7], [10]. Afterwards, more advanced imaging processing and statistical or machine learning techniques have been applied to automated lesion detection, delivering relatively higher accuracy than thresholding [11]–[17]. Recently, deep learning that shows great success in medical imaging [18]–[21] has been applied to PET image analysis, often leading to improved lesion/tumor identification performance [22], [23]. Many previous approaches take as input a region/volume of interest (ROI/VOI), which is often manually cropped and contains objects of interest only (e.g., lesions or specific organs containing lesions) [7]. These methods use ROIs/VOIs to constrain their outputs and reduce the noise effects outside the ROIs/VOIs, but they require considerable human interaction to isolate the tumors in advance. Some recent deep learning-based lesion detection methods [23] take the entire PET image as input, but there is still much room for improvement. Another large population of approaches use multi-modality images, such as PET/CT or PET/MRI, for lesion quantification [7], [8]. However, these approaches typically require different imaging modalities to be properly aligned or registered, and this may be difficult to achieve in actual practice [7]. In addition, certain tumor boundaries may not be present in CT or MRI images but appear in PET images only, such as liver lesions for GEP-NETs with ^{68}Ga-DOTATATE PET imaging.

In this paper, we propose a novel deep neural network with location-aware feature encoding for single-stage hepatic lesion detection using only PET images (see Fig. 2). Specifically, we design a discriminative codebook learning module and incorporate it into a residual learning-based U-Net-like neural network to enhance feature discriminativeness for lesion detection. We use lesion location labels as auxiliary supervision at hidden layers to directly regularize the training of the codebook, which is thus enforced to encode features that are semantically discriminative with respect to lesion locations. In addition, we introduce a learnable fusion layer to automatically combine the hidden-layer and last-layer output predictions for lesion detection. The entire network is end-to-end trainable and performs lesion identification in a single-pass manner. It requires neither a preprocessing step to crop an ROI/VOI region as model input nor other imaging modalities such as CT or MRI. The proposed method is extensively evaluated on a set of 3D real clinical ^{68}Ga-DOTATATE PET images from 125 subjects and compared with several recent state-of-the-art deep learning approaches. In summary, the contributions are three-fold:

- We design a novel codebook learning module for discriminative feature learning. We incorporate this plug-and-play module into a deep neural network at multiple levels and use direct supervision to encourage multi-scale discriminative representation learning. The network automatically combines the side-output predictions from the codebook learning module and the prediction from the network's last layer via a learnable fusion layer for lesion detection enhancement.
- We introduce a novel single-stage framework for lesion detection using only PET images. This is different from many other studies that require pre-defined regions/volumes of interest (ROIs/VOIs), multi-stage/cascaded modeling, or multi-modality training image data.
- The proposed method significantly outperforms multiple recent deep learning models for hepatic lesion detection with PET imaging, including those specifically designed for tumor identification in PET images.

II. Related Work

Lesion or tumor detection in PET images is very critical for accurate diagnosis of NETs and assessment of the response to therapy. Early-stage methods mainly rely on digital image processing and/or computer vision techniques to design specific algorithms for automated lesion identification [7], [10], [24], [25]. Afterwards, statistical or machine learning can infer the image processing rules from example data for lesion detection and have attracted increasing attention [12], [13], [16], [26]–[32]. However, traditional machine learning requires manual feature engineering for data representation, which is a non-trivial task, especially for complex PET images. Meanwhile, it gives inferior tumor detection performance in PET images compared with end-to-end deep learning in some recent studies [22], [23], [33].

Deep neural networks, particularly convolutional neural networks (CNNs) [34], [35] and their variants [36]–[39], have recently exhibited great power in medical image computing and achieved state-of-the-art performance in various tasks including lesion detection [18]–[21]. Chen *et al.* [40] have exploited a CNN model to identify initial cervical tumors in ^{18}F-FDG PET images and then applied complex post-processing to tumor refinement. Pfaehler *et al.* [41] have presented a U-Net-based neural network for tumor segmentation in a lung cancer PET image dataset and have achieved good performance. Our previous study [42] has adopted a 2D residual fully convolutional network (FCN) to locate liver lesions in individual 2D PET slices followed by using manually annotated liver masks to refine network predictions. However, it does not consider the context information between adjacent slices for model training, and it may increase false positives when applied to 3D image data. In addition, our prior work [42] as well as the aforementioned methods requires a pre-defined ROI/VOI, e.g., the liver region, to isolate lesion regions from the image background. These ROIs/VOIs are often manually determined and thus need additional human effort for data annotation. In this paper, we have introduced a novel 3D neural network, which takes as input the entire 3D PET volume and does not require pre-defined VOIs.

Some recent deep models take advantage of end-to-end algorithm design and do not need ROIs/VOIs to be cropped or determined in advance. Leung *et al.* [43] have used a U-Net-like architecture [37] to detect lung tumors with a multi-stage training pipeline, Lu *et al.* [44] have modified the U-Net with a dropblock technique to segment tumors in lung

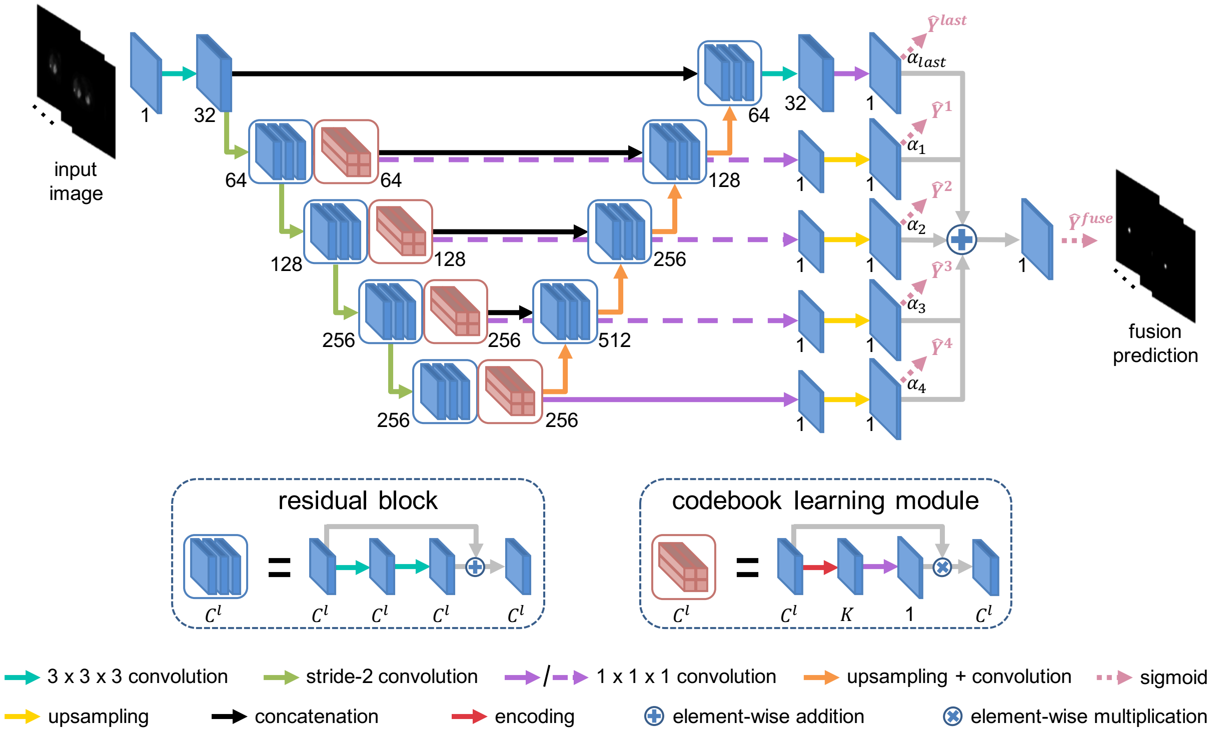


Fig. 2. The proposed neural network with location-aware encoding for single-stage lesion detection in PET images. The blue boxes represent the feature maps with the number of channels presenting below. The arrows with different colors denote distinct operations. To avoid cross connections, we use dashed lines to represent the $\mathbf{1} \times \mathbf{1} \times \mathbf{1}$ convolutions on some codebook learning modules for side-output predictions. The encoding layer (red arrow) in the codebook learning module takes as input a $\boldsymbol{C^l}$-channel feature map and produces a $\boldsymbol{K}$-channel output feature map, where $\boldsymbol{K}$ is the number of codewords in the codebook. The $\hat{\boldsymbol{Y}}^{last}$, $\{\hat{\boldsymbol{Y}}^{l}\}_{l=1}^{4}$ and $\hat{\boldsymbol{Y}}^{fuse}$ represent prediction maps from the last layer, the codebook learning modules and the fusion layer, respectively.

cancer FDG PET images, and Liu *et al.* [45] have applied a Bayesian encoder-decoder neural network to oncological PET segmentation. By considering the information between adjacent slices, Blanc-Durand *et al.* [46] have built a 3D U-Net model for gliomas segmentation in ^{18}F-fluoro-ethyl-L-tyrosine (^{18}F-FET) PET images. Iantsen *et al.* [47] have incorporated residual learning blocks [48] and squeeze-excitation blocks [49] into a 3D U-Net [38] for automated tumor uptake segmentation in cervical cancer PET images and have obtained better performance than conventional thresholding and standard U-Net models. Despite promising performance of these methods, there is still large room for improvement, especially for those multi-stage methods that introduce additional variability in tumor identification. In addition, none of these approaches are designed and evaluated on DOTATATE PET image data, which typically has higher image noise and lower lesion detectability than FDG PET [9], [42].

A popular line of research for lesion identification in PET images is to use multimodal imaging data, such as PET/CT or PET/MRI [7], [50]–[55]. Xue *et al.* [56] have introduced a multimodal neural network for lesion segmentation, which uses shared down-sampling blocks between the PET and CT encoding branches for feature co-learning. Jin *et al.* [57] have built a two-streamed neural network for tumor detection in esophageal cancer and conducted both early and late feature fusion for PET and CT images. Kumar *et al.* [58] have used a co-learning CNN model to quantify the importance of each modality's features and fuse complementary information from multimodal image data for tumor detection. Guo *et al.* [59] have employed different fusion networks to identify tumors using a mixture of PET, CT and MRI images, and other multimodal information fusion approaches have also been applied to lesion detection [60]–[64]. Learning with multimodal images assumes an appropriate registration between different modalities, but this assumption might not always hold in reality [7]. Meanwhile, for some diseases, lesions may not be present in the anatomical modality and thus there is no correspondence between tumor boundaries in PET and CT (or MRI) images, such as GEP-NETs with ^{68}Ga-DOTATATE PET imaging. In this paper, we will focus on lesion detection using only PET images.

III. Single-Stage Lesion Detection in DOTATATE PET

Our end-to-end lesion detection neural network is built on a 3D encoder-decoder architecture with long-range skip connections, as shown in Fig. 2. The network learns an inherent codebook on-the-fly, which consists of a set of visual codewords to model input data distribution, with multi-scale auxiliary supervision for lesion location-aware feature encoding. The auxiliary supervision is directly linked to multi-level hidden layers of the neural network and enhances the discriminativeness of learned features to facilitate lesion identification. In addition, the network adopts a learnable fusion layer to combine the hidden-layer and last-layer outputs for model training, which allows the network to automatically adjust the contribution of each output prediction and optimize a weighted fusion for lesion detection.

Fig. 2 shows our U-Net-like network architecture, which mainly consists of one contracting path (encoder) and one expanding path (decoder), with each containing four residual learning blocks [65]. The contracting path uses stride-2 convolutional layers to stack the residual blocks, while the expanding path links up its residual blocks with interpolation-based upsampling followed by convolutional operations. One codebook learning module is added on top of each residual block in the contracting path, and the output of each codebook learning module is copied and concatenated with feature maps of the corresponding residual block in the expanding path via a long-range skip connection. A learnable fusion layer is applied to information aggregation of the outputs from the hidden layers and the last layer. All convolutional layers in the residual blocks use a 3D kernel of $3 \times 3 \times 3$, and each is followed by an instance normalization layer [66] and an exponential linear unit [67].

A. Lesion Location-Aware Encoding

Incorporating codebook learning into deep neural networks can enhance expressive power of the networks' feature representations and has produced improved performance in different computer vision tasks, compared with the counterpart without codebook learning [68]–[70]. Inspired by [68], we construct a novel lesion location-aware codebook learning module to encode multi-scale spatial information within input images for lesion localization. Specifically, we tailor the codebook learning technique in [68] and make the following significant improvements: 1) we extend the module to learn rich hierarchical features from 3D volumes for object detection instead of 2D image classification, 2) we use auxiliary, side-output supervision to directly regularize the module learning such that the lesion location-relevant information can be encoded in feature learning, and 3) we insert this module into multiple hidden layers of the neural network to extract multi-scale location-aware features, instead of placing it on only the penultimate layer that learns much coarse-scale features, which may contain limited local details of lesions. Our codebook learning module is also different from [69], [70], which aggregate codebook-encoded features across the entire image such that spatial information is lost; instead, our module learns to capture spatial context and highlight salient regions with auxiliary supervision.

Our 3D codebook learning module mainly consists of an encoding layer, a $1 \times 1 \times 1$ convolutional layer followed by an instance normalization layer [66], and a sigmoid activation function. Specifically, we modify the encoding layer in [68] by changing 2D to 3D operators, removing the aggregation operation for encoded features to avoid losing of image spatial information, and outputting the coding coefficients to directly highlight target regions. Formally, let $\boldsymbol{Z} \in \mathbb{R}^{C^l \times D^l \times H^l \times W^l}$ denote the input feature map of our improved encoding layer, where C^l, D^l, H^l and W^l represent the channel, depth, height and width of the feature map, respectively. The goal of the encoding layer is to learn a visual codebook and use it to encode discriminative features for lesion detection. Specifically, the encoding layer first interprets the feature map $\boldsymbol{Z}$ as a set of C^l-dimensional, voxel-level visual descriptors $\{\boldsymbol{z}_i \in \mathbb{R}^{C^l}\}_{i=1}^{M^l}$, where $M^l = D^l \times H^l \times W^l$. Then, it simultaneously learns an inherent codebook $\boldsymbol{B}$ composed of K codewords, $\boldsymbol{B} = \{\boldsymbol{b}_k \in \mathbb{R}^{C^l}\}_{k=1}^{K}$, and produces an output feature map $\boldsymbol{U} \in \mathbb{R}^{K \times D^l \times H^l \times W^l}$, which contains a group of K-dimensional coding coefficient vectors $\{\boldsymbol{u}_i \in \mathbb{R}^K\}_{i=1}^{M^l}$, one for each input visual descriptor. Instead of relying on hard-assignment coding that is widely used in the traditional bag-of-visual-words (BoVW) model [71], [72], we adopt a soft-assignment coding strategy [73], [74] to address codeword ambiguity and make the codebook learning module differentiable, so that the entire neural network can be trained with standard backpropagation in an end-to-end manner. Specifically, the j-th component of the i-th coding coefficient $\boldsymbol{u}_i$ is

$$u_{ij} = \frac{e^{-s_j ||\boldsymbol{z}_i - \boldsymbol{b}_j||_2^2}}{\sum_{k=1}^{K} e^{-s_k ||\boldsymbol{z}_i - \boldsymbol{b}_k||_2^2}}, \tag{1}$$

where $\{s_k\}_{k=1}^{K}$ are scalar-valued smoothing factors for the assignment, one for each codeword. These factors are automatically learned during model training so as to allow for a finer modeling of the distribution of input descriptors $\{\boldsymbol{z}_i\}_{i=1}^{M^l}$. The $||\cdot||_2$ is an l_2 norm to measure the distance between each pair of input descriptor and codeword. The u_{ij} denotes the degree of membership of descriptor $\boldsymbol{z}_i$ to codeword $\boldsymbol{b}_j$, i.e., soft assignment, and a higher value of u_{ij} means that $\boldsymbol{z}_i$ is closer to $\boldsymbol{b}_j$.

Given the coding coefficients $\{\boldsymbol{u}_i\}_{i=1}^{M^l}$ for the voxel-level input descriptors $\{\boldsymbol{z}_i\}_{i=1}^{M^l}$, it is common to perform an aggregation operation, e.g., $\sum_{i=1}^{M^l} u_{ij}$ summing over all the voxels, to obtain an image-level representation for different visual tasks [68]–[70], [74]. However, this aggregation operation removes the spatial information about the locations of target objects in the input feature map and thus may pose challenges for object localization, such as lesion detection in PET images. Thus, instead of conducting an aggregation operation, we propose to use the voxel-wise coding coefficients $\{\boldsymbol{u}_i\}_{i=1}^{M^l}$ to directly highlight the target lesion regions and suppress irrelevant activations within the feature map. To this end, we add a $1 \times 1 \times 1$ 3D convolutional layer followed by instance normalization on top of the encoding layer, and use a sigmoid function as the activation to produce a voxel-wise scaling feature map $\boldsymbol{V} \in \mathbb{R}^{1 \times D^l \times H^l \times W^l}$, which contains M^l scaling factors $\{v_i \in \mathbb{R}\}_{i=1}^{M^l}$ (see Fig. 2). Then, we apply a voxel-wise multiplication to $\boldsymbol{V}$ and $\boldsymbol{Z}$, and output a scaled feature map $\boldsymbol{Z}' \in \mathbb{R}^{C^l \times D^l \times H^l \times W^l}$ to emphasize the lesion locations and prune irrelevant responses in other regions. Formally, this computation is formulated as

$$\boldsymbol{V} = \sigma(g(\boldsymbol{U})), \tag{2}$$

$$\boldsymbol{Z}' = \boldsymbol{V} \otimes \boldsymbol{Z}, \tag{3}$$

where $g(\cdot)$ represents the convolutional operation followed by instance normalization, $\sigma(\cdot)$ denotes the sigmoid activation function, and $\otimes$ means the voxel-wise multiplication.

With voxel-level lesion labels (e.g., 3D binary images) on the last layer of a neural network and an appropriate loss function for lesion detection, we can train the network including the codebook learning module using the standard backpropagation algorithm [75]. However, the supervision from only the last layer may not provide sufficient support to the codebook for learning discriminative features in (early) hidden layers [76], potentially leading to performance degradation of lesion detection. Therefore, we propose to directly add auxiliary supervision, i.e., lesion labels, on top of the codebook learning module and enforce it to understand spatial information about lesion locations for enhancement of feature discriminativeness. Specifically, we place another $1 \times 1 \times 1$ 3D convolutional layer on the module for a side-output prediction of lesion locations and introduce an auxiliary lesion detection loss to explicitly regularize the module training (see Fig. 2). In this way, the codebook learning module can directly receive gradients from this side-output loss, in addition to the supervision backpropagated from the last layer. Thus, the codebook significantly improves the discriminativeness of encoded features with respect to predicting lesion locations and is specifically optimized for the lesion detection task. Note that our codebook learning module simultaneously builds the codebook $\boldsymbol{B}$ and encodes the features $\boldsymbol{Z}'$ in an end-to-end, supervised manner, by taking advantage of the readily available lesion annotations from training data. This is different from the traditional BoVW model [71], [72], which conducts codebook learning and feature encoding in a separate, unsupervised mode and thus may not be appropriate for supervised-learning downstream tasks.

B. Learnable Fusion of Multi-Scale Predictions

NETs typically exhibit significant spatial scale variation in PET images with volumes ranging from a few to hundreds of voxels. This may pose a great challenge for deep neural networks to learn effective feature representations for different-sized lesions. In particular, the high layers in neural networks extract coarse-scale features and ignore local details, and thus may have difficulty in capturing information for small lesions after conducting several downsampling operations. Inspired by [77], [78], we introduce multiple side-output predictions to multi-level layers of the neural network (see Fig. 2), enforcing it to learn multi-scale feature representations for lesion detection. Specifically, we insert multiple codebook learning modules to the network, with each linking to a residual block in the contracting path, so that each module is responsible for encoding discriminative feature maps at a certain scale. With lesion location-aware feature encoding, lesion position-relevant activations are merged via skip connections at each scale and upsampled back to the high-resolution space for lesion detection.

In order to directly take advantage of side-output predictions from the codebook learning modules, we incorporate an additional learnable fusion layer into the network so that the hidden-layer, side-output predictions and the last-layer output prediction are fused via a weighted sum to produce a fused prediction. Because the fusion weights are automatically learned during training, the network can dynamically adjust the relative importance of each prediction for lesion detection. Formally, let $\boldsymbol{A}^l \in \mathbb{R}^{D^l \times H^l \times W^l}$ be the output prediction map, before applying a sigmoid activation function, of the l-th codebook learning module, where $l = 1, 2, ..., L$. Similarly, denote $\boldsymbol{A}^{last} \in \mathbb{R}^{D \times H \times W}$ as the output prediction map of the network's last layer, before using the sigmoid function. The prediction $\hat{\boldsymbol{Y}}^{fuse} \in [0,1]^{D \times H \times W}$ of our fusion layer is

$$\hat{\boldsymbol{Y}}^{fuse} = \sigma(\sum_{l=1}^{L} \alpha_l \cdot f(\boldsymbol{A}^l) + \alpha_{last} \boldsymbol{A}^{last}), \tag{4}$$

where $\{\alpha_l\}_{l=1}^L$ and α_{last} are the learnable fusion weights for side-output and last-layer predictions, respectively. The $f(\cdot)$ denotes an interpolation-based upsampling operation to resize the side-output predictions to the original scale. In our modeling, we incorporate $L = 4$ codebook learning modules into the neural network (see Fig. 2). We do not place a codebook learning module to the first convolutional layer, i.e., on the first skip connection, because it does not capture sufficient semantic context for lesion localization.

C. Loss Function

We formulate lesion detection as a binary voxel-wise classification problem, i.e., lesion voxels versus non-lesion voxels, and optimize the neural network using a weighted binary cross-entropy loss. Because lesions account for a lower proportion of each PET image than non-lesion regions, we assign a higher weight value to the lesion voxels in the loss function for addressing the data imbalance. Note that we add this lesion detection loss to the network's last layer, each codebook learning module, and the fusion layer. Let $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^N$ denote the training data set of N 3D PET images, where $\boldsymbol{X}_i \in \mathbb{R}^{C \times D \times H \times W}$ and $\boldsymbol{Y}_i \in \{0,1\}^{D \times H \times W}$ respectively represent the i-th training image and its associated gold-standard label, which is a binary 3D image with 1's for lesion voxels and 0's for the others. Denote y_{ij}, $\hat{y}_{ij}^{last}$, $\hat{y}_{ij}^{l}$, and $\hat{y}_{ij}^{fuse}$ the j-th voxel value of $\boldsymbol{Y}_i$, $\hat{\boldsymbol{Y}}_i^{last}$, $\hat{\boldsymbol{Y}}_i^{l}$, and $\hat{\boldsymbol{Y}}_i^{fuse}$, respectively. Note that the side-output predictions from the codebook learning modules are upsampled to the original scale $\hat{\boldsymbol{Y}}_i^{l} = \sigma(f(\boldsymbol{A}_i^l)) \in [0,1]^{D \times H \times W}$ for $l = 1, 2, ..., L$, and $\hat{\boldsymbol{Y}}_i^{last} = \sigma(\boldsymbol{A}_i^{last}) \in [0,1]^{D \times H \times W}$. The full loss function $\mathcal{L}$ of lesion detection for the i-th training image $\boldsymbol{X}_i$ is

$$\mathcal{L} = \mathcal{L}^{last} + \sum_{l=1}^{L} \mathcal{L}^l + \mathcal{L}^{fuse}, \tag{5}$$

$$\mathcal{L}^{last} = \frac{-1}{|\boldsymbol{Y}_i|} \sum_{j=1}^{|\boldsymbol{Y}_i|} (\beta y_{ij} \log \hat{y}_{ij}^{last} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}^{last})), \tag{6}$$

$$\mathcal{L}^{l} = \frac{-1}{|\boldsymbol{Y}_i|} \sum_{j=1}^{|\boldsymbol{Y}_i|} (\beta y_{ij} \log \hat{y}_{ij}^{l} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}^{l})), \tag{7}$$

$$\mathcal{L}^{fuse} = \frac{-1}{|\boldsymbol{Y}_i|} \sum_{j=1}^{|\boldsymbol{Y}_i|} (\beta y_{ij} \log \hat{y}_{ij}^{fuse} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}^{fuse})), \tag{8}$$

where $\mathcal{L}^{last}$, $\mathcal{L}^{l}$, and $\mathcal{L}^{fuse}$ are the losses for the network's last layer, the l-th codebook learning module, and the fusion layer, respectively. The $|\boldsymbol{Y}_i|$ represents the cardinality of $\boldsymbol{Y}_i$, and β denotes the weighting parameter to control the relative importance between lesions and non-lesion regions. Applying the loss in Eq. (5) to all the training images, we train the entire neural network including the codebook learning module using both standard supervision from the network's last layer and auxiliary supervision from side-output layers and the fusion layer.

During the testing stage, for each new input image $\boldsymbol{X}$, we have multiple 3D prediction maps from the last year ($\hat{\boldsymbol{Y}}^{last}$), the codebook learning modules ($\hat{\boldsymbol{Y}}^{l}$, $l = 1, 2, ..., L$) and the fusion layer ($\hat{\boldsymbol{Y}}^{fuse}$). Considering that the computation in the last and fusion layers takes into account multi-scale feature representations, we apply an average aggregation operation to the predictions from these two layers and obtain a final prediction map for lesion detection as follows

$$\hat{\boldsymbol{Y}}^{final} = \frac{1}{2}(\hat{\boldsymbol{Y}}^{last} + \hat{\boldsymbol{Y}}^{fuse}). \tag{9}$$

To reduce the effects of noisy predictions, we remove the responses with low values in the final prediction map, i.e., those voxels with values not greater than a threshold τ are suppressed, and then apply connected component analysis to individual lesion identification.

IV. Experiments and Discussion

A. Experimental Setup

1) Dataset: We acquire a real clinical ^{68}Ga-DOTATATE PET liver image dataset using a photomultiplier tube-based PET scanner. The dataset has 125 subjects with 58 abnormal (i.e., patients with hepatic lesions) and 67 normal cases. Each subject has one 3D PET volume consisting of a certain number of 128 × 128 transaxial slices, and the number of slices in the liver volume varies from 23 to 71 for different subjects. Each abnormal PET volume has one or more hepatic lesions. Following [42], we randomly split the dataset into training, validation and test sets with a ratio of 6:2:2. This study is determined to be exempt from IRB review by the Colorado Multiple Institutional Review Board at University of Colorado Anschutz Medical Campus.

2) Implementation Details: We set $\beta = 10$ in Eqs. (6) $\sim$ (8) by using the validation set to search for the best value in the set of $\{0.1, 1, 10, 100\}$. We follow [70] to choose $K = 16$ for codebook learning, because it shows impressive performance in object detection. We train our neural network using stochastic gradient descent with Nesterov momentum [79] and set the parameter values as: momentum = 0.99, learning rate = 10^{-3}, weight decay = 10^{-6}, batch size = 2, and maximum number of iterations = 10^5. For each data batch during training, we load 64 slices for each subject and use zero-valued slice padding for subjects with less than 64 slices, i.e., $C = 1,\ D = 64,\ H = 128$ and $W = 128$. We apply data augmentation to model training, including random rotation within $(-10^o,\ 10^o)$, random horizontal and vertical translation with a displacement in $(-0.125W,\ 0.125W)$ and $(-0.125H,$ $0.125H)$ respectively, and random scaling with a factor in $[0.8, 1.2]$. We stop the training process if the performance on the validation set does not improve for successive 2×10^4 iterations. During testing, we use $\tau = 0.1$ to suppress low-valued predictions for lesion detection.

3) Evaluation Metrics: We follow [42], [54], [60] to use precision, recall and F$_1$ score as the evaluation metrics for lesion detection. We associate automatedly detected lesions with the corresponding gold-standard 3D lesion annotations using the Hungarian algorithm [80]. One detected lesion can correspond to at most one gold-standard annotation, and vice versa. An automatic detection is defined as true positive (TP) if the intersection over union (IoU) between this detection and its associated gold-standard lesion is greater than a threshold, otherwise false positive (FP). In our experiments, we follow [42] to select a 5% IoU threshold to define the TP. The results using a 5% IoU are not significantly different from those obtained with a 10% IoU. In addition, automated detections are considered false positive (FP) if they are not matched with any gold-standard lesions, and gold-standard annotations are viewed as false negative (FN) if they do not have associated automated detections. Based on these definitions, we can quantify the metric values as: $precision = TP/(TP + FP)$, $recall = TP/(TP + FN)$, and $F_1\ score = 2 \cdot precision \cdot recall/(precision + recall)$.

B. Model Evaluation on Lesion Detection

1) Comparison with State of The Art: We compare the proposed method with several recent state-of-the-art deep models, including 3D U-Net [38], V-Net [39], residual pre-activation-based U-Net (RPAU-Net) [47], attention U-Net (ATTU-Net) [81], 3D U-Net with concurrent spatial and channel attention (SCU-Net) [82], project-excite FCN (PE-FCN) [83], residual FCN (RES-FCN) [42], squeeze-excitation normalized network (SEN-Net) [84], hybrid CNN-Transformer U-Net (TransUNet) [85] and volumetric Transformer network (VT-Net) [86]. The RPAU-Net, RES-FCN and SEN-Net are specifically designed for lesion identification in PET images. Table I shows the experimental results of different models. Each model is run 5 times with different random seeds, and the mean and standard deviation of each metric are reported.

We see that our method outperforms other competitors by a large margin in all the three metrics, with a range

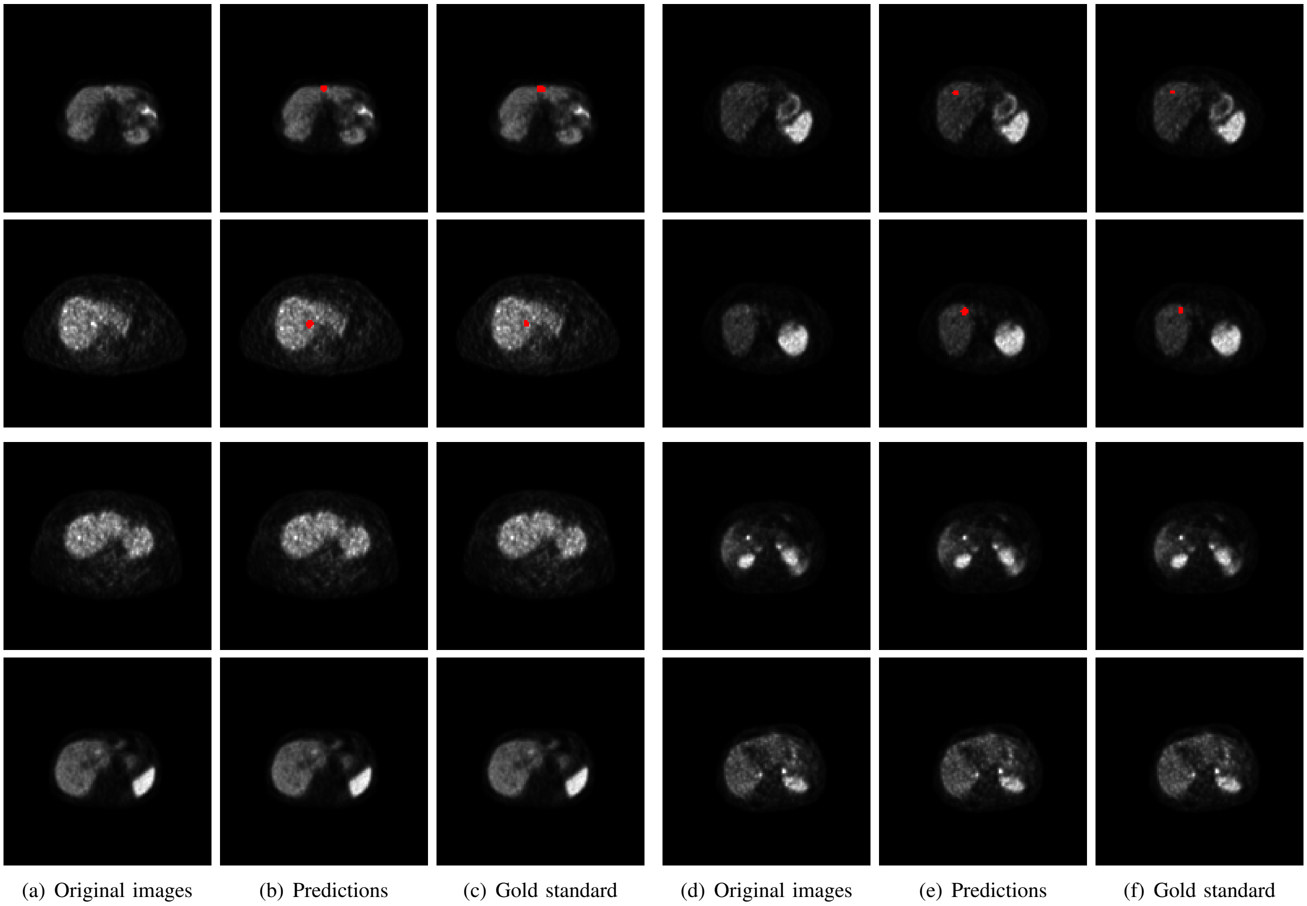


Fig. 3. Qualitative lesion detection results using our method. Rows $\mathbf{1} \sim \mathbf{2}$ represent model predictions on multiple abnormal subjects (lesions marked with red color), and rows $\mathbf{3} \sim \mathbf{4}$ denote predictions on several normal subjects without hepatic lesions. Columns (a)/(d), (b)/(e) and (c)/(f) represents the original images, model predictions and gold standard annotations, respectively.

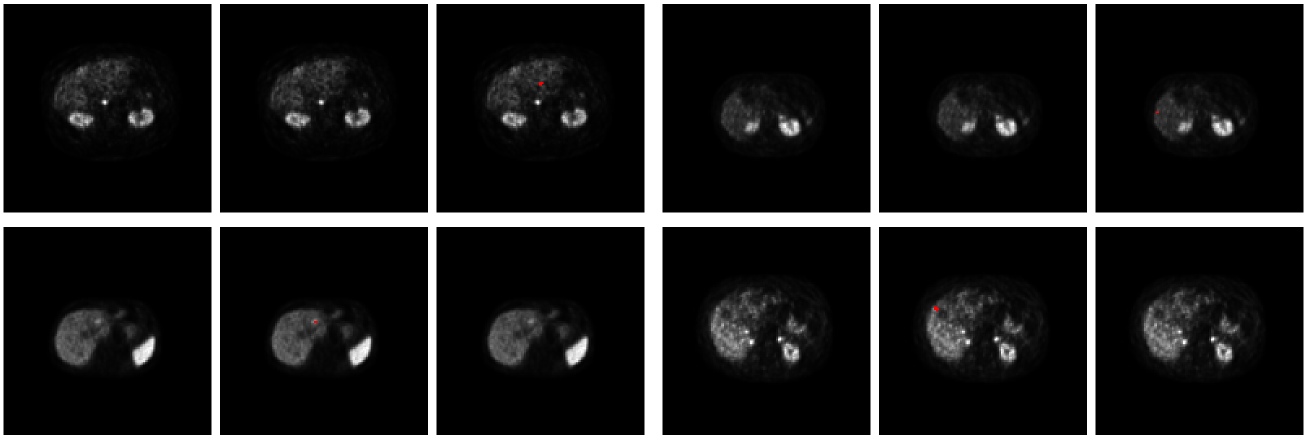
TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS IN LESION DETECTION IN TERMS OF THE MEAN AND STANDARD DEVIATION (STD) OF EACH METRIC: $\boldsymbol{mean} \pm \boldsymbol{std}$. THE HIGHEST VALUE OF EACH METRIC IS HIGHLIGHTED WITH BOLD, AND THE $*$ INDICATES THERE IS A STATISTICALLY SIGNIFICANT DIFFERENCE ($\boldsymbol{p}$-VALUE $<$ 0.05) BETWEEN OUR METHOD AND OTHERS IN TERMS OF F_1 SCORE.

	Precision ($mean \pm std$)	Recall ($mean \pm std$)	F_1 score ($mean \pm std$)
3D U-Net [38]	43.46 ± 1.72	56.41 ± 6.49	$48.97 \pm 3.05^*$
V-Net [39]	92.44 ± 5.75	50.77 ± 5.71	$65.15 \pm 4.24^*$
RPAU-Net [47]	79.70 ± 6.03	45.64 ± 6.36	$57.79 \pm 5.75^*$
ATTU-Net [81]	84.61 ± 2.44	58.97 ± 2.29	$69.47 \pm 1.92^*$
SCU-Net [82]	42.96 ± 4.97	61.54 ± 2.29	$50.32 \pm 2.65^*$
PE-FCN [83]	57.47 ± 11.92	57.95 ± 4.47	$56.72 \pm 3.73^*$
RES-FCN [42]	53.05 ± 7.65	78.97 ± 3.40	$63.17 \pm 6.32^*$
SEN-Net [84]	65.23 ± 3.90	62.56 ± 5.76	$63.80 \pm 4.51^*$
TransUNet [85]	65.20 ± 5.19	55.38 ± 6.80	$59.70 \pm 5.25^*$
VT-Net [86]	53.95 ± 16.67	35.9 ± 7.78	$40.76 \pm 4.48^*$
Ours	$\mathbf{95.54} \pm 3.79$	$\mathbf{73.85} \pm 2.51$	$\mathbf{83.24} \pm 1.93$

of $3.10\% \sim 52.58\%$ for precision, $12.31\% \sim 28.21\%$ for recall, and $13.77\% \sim 34.27\%$ for F_1 score. In particular, our method provides significantly better performance than the others with p-value < 0.05 in Student's t-test in terms of F_1 score. Although V-Net gives a high precision, it misdetects a number of lesions and produces a very low recall, thus leading to a 65.15% F_1 score. Compared with V-Net, RES-FCN significantly improves the recall but has a dramatic decrease in the precision, perhaps because of an increased amount of false positives. The SEN-Net presents a similar F_1 score to V-Net and a relatively higher recall, probably due to the usage of squeeze-excitation normalization layers. The TransUNet gives a comparable precision to SEN-Net but a decreased recall. The 3D VT-Net provides a low F_1 score possibly because the high-complexity model overfits the training data. The ATTU-Net uses an attention mechanism to suppress irrelevant responses in feature maps and thus facilitate lesion localization, delivering a relatively better performance than V-Net, RES-FCN, SEN-Net and others. However, all these approaches are significantly outperformed by our method that produces an average F_1 score of 83.24% (with a maximum F_1 of 86.96%), demonstrating the effectiveness of our method on lesion detection in PET images. Fig. 3 shows qualitative lesion detection results using our method on several example PET images, and Fig. 4 lists some examples of false positives and false negatives for lesion detection.

2) Ablation Study: In order to evaluate the effectiveness of each component of our method, we conduct an ablation study to report the lesion detection performance of the following model variants: 1) $Baseline$: train a lesion detection model



(a) Original images (b) Predictions (c) Gold standard (d) Original images (e) Predictions (f) Gold standard

Fig. 4. Examples of false negatives (row 1) and false positives (row 2) using our method for lesion detection in PET images from different subjects. Columns (a)/(d), (b)/(e) and (c)/(f) represents the original images, model predictions and gold standard annotations, respectively. The red regions are gold-standard annotated lesions (row 1) or predicted lesions (row 2).

TABLE II

ABLATION STUDY OF LESION DETECTION IN TERMS OF THE MEAN AND STANDARD DEVIATION (STD) OF EACH METRIC: $\boldsymbol{mean \pm std}$. THE HIGHEST VALUE OF EACH METRIC IS HIGHLIGHTED WITH BOLD, AND THE $*$ INDICATES THERE IS A STATISTICALLY SIGNIFICANT DIFFERENCE ($\boldsymbol{p}$-VALUE < 0.05) BETWEEN OUR METHOD AND OTHERS IN TERMS OF F_1 SCORE.

	Precision $(mean \pm std)$	Recall $(mean \pm std)$	F_1 score $(mean \pm std)$
Baseline	69.82 ± 7.45	64.62 ± 4.1	$66.97 \pm 5.04^*$
CL	92.08 ± 3.45	65.13 ± 2.05	$76.28 \pm 2.37^*$
PF	85.44 ± 3.86	$\mathbf{73.85} \pm 2.51$	$79.13 \pm 1.75^*$
Ours	$\mathbf{95.54} \pm 3.79$	$\mathbf{73.85} \pm 2.51$	$\mathbf{83.24} \pm 1.93$

with neither codebook learning nor prediction fusion, i.e., using the loss $\mathcal{L}^{last}$ only; 2) CL : train a model with codebook learning but without prediction fusion, i.e., using the loss $\mathcal{L}^{last} + \sum_{l=1}^{L} \mathcal{L}^{l}$; 3) PF : train a model with prediction fusion but without codebook learning , i.e., using the loss $\mathcal{L}^{last} + \mathcal{L}^{fuse}$; 4) $Ours$: the proposed method that trains a model with both codebook learning and prediction fusion, i.e., using the loss $\mathcal{L}^{last} + \sum_{l=1}^{L} \mathcal{L}^{l} + \mathcal{L}^{fuse}$. We run each model 5 times with different random seeds, and report the mean and standard deviation for each metric.

Table II lists the experimental results of the ablation study. Both CL and PF outperform the $Baseline$ model, indicating that incorporating either codebook learning or prediction fusion into model learning is beneficial to lesion detection. We note that the CL model increases the F_1 score from 66.97% to 76.28% compared with the $Baseline$, suggesting that codebook learning can encourage the neural network to learn discriminative feature representations for lesion identification. Combining codebook learning and prediction fusion, our method further significantly improves the F_1 score to 83.24%, and this confirms the effectiveness of our method.

3) Effects of Model Parameters: Our method has an important hyperparameter, the number of codewords (K) in each codebook, which controls the expressive power of the codebook. The left panel of Fig. 5 shows the lesion detection performance of our method using different K values. We see that the F_1 score is relatively low when $K = 4$ or 8, compared with $K >= 16$. This suggests that a small K value may not be sufficient for the codebook to capture the input data distribution and thus leads to poor lesion detection. When $K > 16$, the performance improvement gets saturated. This demonstrates that our codebook learning technique is very effective for feature representation encoding such that we do not need a large codebook to model data distribution.

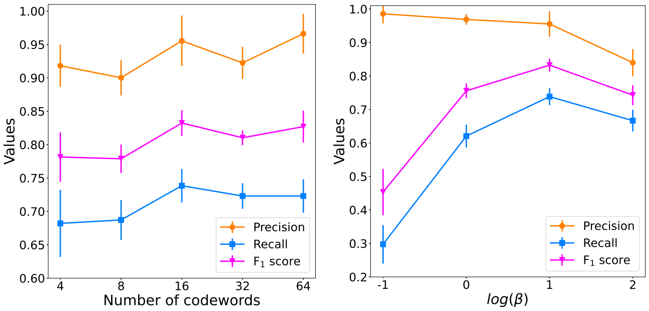


Fig. 5. Lesion detection performance of our method using different numbers of codewords for codebook learning (left) and different $\boldsymbol{\beta}$ values in Eqs. (6) $\sim$ (8) (right). Each curve represents the mean value of 5 runs with different random seeds, and the vertical lines in each curve denote the standard deviation. Note that the $\boldsymbol{x}$-axis in the right plot is $\mathbf{log}(\boldsymbol{\beta})$.

The β in Eqs. (6) $\sim$ (8) is another critical parameter that is used to highlight lesions during model training. The right panel of Fig. 5 shows the experimental results of our method using different β values. As we can see, a small β probably misses many true lesions and gives a very low recall (and thus a low F_1 score), especially for the case $\beta < 1$ which de-emphasizes the lesions for model training. A higher value such as $\beta = 10$ produces much better lesion detection performance with an F_1 score of over 83%. However, a too large β, e.g., 100, may lead to more false positives and thus a lower precision and F_1 score.

TABLE III

Image Segmentation of Different Methods in Terms of the Mean and Standard Deviation (std) of Dice Similarity Coefficient and IoU: $mean \pm std$. The Highest Value of Each Metric Is Highlighted With Bold, and the $*$ Indicates There Is a Statistically Significant Difference (p-value < 0.05) Between Our Method and Others.

	Dice	IoU
	$(mean \pm std)$	$(mean \pm std)$
3D U-Net [38]	57.90 ± 0.89*	55.27 ± 0.89*
V-Net [39]	88.51 ± 0.90*	85.70 ± 0.61*
RPAU-Net [47]	82.29 ± 2.32*	80.14 ± 2.64*
ATTU-Net [81]	89.74 ± 1.31	86.39 ± 1.36
SCU-Net [82]	57.98 ± 1.87*	55.32 ± 1.79*
PE-FCN [83]	79.76 ± 5.51*	76.70 ± 5.65*
RES-FCN [42]	82.24 ± 2.65*	79.06 ± 2.65*
SEN-Net [84]	79.43 ± 2.30*	77.06 ± 2.29*
TransUNet [85]	65.95 ± 7.78*	62.97 ± 7.75*
VT-Net [86]	70.46 ± 5.55*	69.31 ± 5.87*
Baseline	81.87 ± 3.10*	78.49 ± 3.05*
CL	85.07 ± 1.15*	81.74 ± 1.06*
PF	84.44 ± 1.50*	81.01 ± 1.50*
Ours	**91.72** ± 0.37	**88.27** ± 0.36

C. Image Segmentation

We also evaluate our method on PET image segmentation in terms of Dice similarity coefficient and IoU. The top panel of Table III shows the comparison between the proposed method and recent state of the art. As we can see, our method consistently outperforms all the other competitors in terms of both Dice and IoU metrics. In addition, it produces significantly better results (p-value < 0.05) than almost all the others except the ATTU-Net, which gives slightly lower Dice and IoU scores than ours.

The bottom panel of Table III presents the image segmentation results of different variants of our method. We note that either codebook learning or multi-scale prediction fusion can help improve the segmentation compared with the $Baseline$ model, and a combination of these two components can significantly boost the performance. This is consistent with the observation in Table II, further suggesting the superiority of the proposed method.

V. Conclusion

In this paper, we propose a novel U-Net-like neural network for single-stage lesion detection in PET images. It introduces a newly designed codebook learning module into the encoder for multi-scale discriminative feature encoding, and then applies a learnable fusion layer to multi-scale prediction aggregation for lesion identification. The proposed neural network supports single-stage model training and inference, and does not require manual cropping or cascaded models to select ROIs/VOIs as model inputs, which are required by many existing lesion detection methods with PET imaging. In addition, our model is trained with only PET images and does not need other imaging modalities such as CT or MRI, thus eliminating the non-trivial image registration between different modalities. More importantly, this property makes it well suitable for lesion identification in diseases like GEP-NETs, which typically do not have lesion boundaries present in other modalities but PET imaging. Compared with previous studies, this work provides an efficient alternative for effective lesion detection in PET images. It has great potential to expedite new treatment planning and ultimately improve patient outcomes including increased survival rates, especially for NETs.

The experimental results demonstrate that the proposed method significantly outperforms recent state-of-the-art deep learning models in lesion detection, with a p-value < 0.05 in statistical tests. We note that the codebook learning module can effectively boost the performance with a small number of codewords (e.g., 16), compared with the baseline model, and this indicates the great ability of discriminative feature encoding. The multi-scale prediction fusion can also improve the baseline model, demonstrating its importance of addressing scale variation of lesions. The experiments also show that it is necessary to tackle the data imbalance issue, i.e., lesions occupy only a very small proportion of each PET image, and appropriately highlight the lesions for model training.

References

[1] M. Riihimäki *et al.*, "The epidemiology of metastases in neuroendocrine tumors," *International Journal of Cancer*, vol. 139, no. 12, pp. 2679–2686, 2016.

[2] M. Sandström *et al.*, "Comparative biodistribution and radiation dosimetry of ^{68}Ga-DOTATOC and ^{68}Ga-DOTATATE in patients with neuroendocrine tumors," *Journal of Nuclear Medicine*, vol. 54, no. 10, pp. 1755–1759, 2013.

[3] D. Wild *et al.*, "Comparison of ^{68}Ga-DOTATOC and ^{68}Ga-DOTATATE PET/CT within patients with gastroenteropancreatic neuroendocrine tumors," *Journal of Nuclear Medicine*, vol. 54, no. 3, pp. 364–372, 2013.

[4] A. Pfeifer *et al.*, "^{64}Cu-DOTATATE PET for neuroendocrine tumors: A prospective head-to-head comparison with ^{111}In-DTPA-Octreotide in 112 patients," *Journal of Nuclear Medicine*, vol. 56, no. 6, pp. 847–854, 2015.

[5] S. M. Sadowski *et al.*, "Prospective study of ^{68}Ga-DOTATATE positron emission tomography/computed tomography for detecting gastro-entero-pancreatic neuroendocrine tumors and unknown primary sites," *Journal of Clinical Oncology*, vol. 34, no. 6, pp. 588–596, 2016.

[6] S. A. Deppen *et al.*, "Safety and efficacy of ^{68}Ga-DOTATATE PET/CT for diagnosis, staging, and treatment management of neuroendocrine tumors," *Journal of Nuclear Medicine*, vol. 57, no. 5, pp. 708–714, 2016.

[7] M. Hatt *et al.*, "Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group no. 211," *Medical Physics*, vol. 44, no. 6, pp. e1–e42, 2017.

[8] B. Foster *et al.*, "A review on segmentation of positron emission tomography images," *Computers in Biology and Medicine*, vol. 50, pp. 76–96, 2014.

[9] M. S. Silosky *et al.*, "Physical characteristics of ^{68}Ga DOTATATE PET/CT affecting small lesion detectability," *American Journal of Nuclear Medicine and Molecular Imaging*, vol. 11, pp. 27–39, 2021.

[10] J. A. Lee, "Segmentation of positron emission tomography images: Some recommendations for target delineation in radiation oncology," *Radiotherapy and Oncology*, vol. 96, no. 3, pp. 302–307, 2010.

[11] A. Boudraa and H. Zaidi, *Image Segmentation Techniques in Nuclear Medicine Imaging*. Berlin: Springer, 2006.

[12] S. Belhassen and H. Zaidi, "A novel fuzzy c-means algorithm for unsupervised heterogeneous tumor quantification in PET," *Medical Physics*, vol. 37, no. 3, pp. 1309–1324, 2010.

[13] J. Lapuyade-Lahorgue *et al.*, "SPEQTACLE: An automated generalized fuzzy c-means algorithm for tumor delineation in PET," *Medical Physics*, vol. 42, no. 10, pp. 5720–5734, 2015.

[14] M. Aristophanous *et al.*, "A Gaussian mixture model for definition of lung tumor volumes in positron emission tomography," *Medical Physics*, vol. 34, no. 11, pp. 4223–4235, 2007.

[15] T. Layer *et al.*, "Pet image segmentation using a Gaussian mixture model and markov random fields," *EJNMMI Physics*, vol. 2, no. 9, 2015.

[16] M. Hatt *et al.*, "A fuzzy locally adaptive bayesian segmentation approach for volume determination in PET," *IEEE Transactions on Medical Imaging*, vol. 28, no. 6, pp. 881–893, 2009.

[17] M. Hatt *et al.*, “Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications,” *International Journal of Radiation Oncology, Biology, Physics*, vol. 77, no. 1, pp. 301–308, 2010.

[18] H. Greenspan *et al.*, “Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[19] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60 – 88, 2017.

[20] D. Shen *et al.*, “Deep learning in medical image analysis,” *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[21] F. Xing *et al.*, “Deep learning in microscopy image analysis: A survey,” *IEEE Transactions on Neural Network and Learning Systems*, vol. 20, no. 10, pp. 4550–4568, 2018.

[22] M. Hatt *et al.*, “The first MICCAI challenge on PET tumor segmentation,” *Medical Image Analysis*, vol. 44, pp. 177–195, 2018.

[23] H. Zaidi and I. El Naqa, “Quantitative molecular positron emission tomography imaging using advanced deep learning techniques,” *Annual Review of Biomedical Engineering*, vol. 23, no. 1, pp. 249–276, 2021.

[24] H. Li *et al.*, “A novel PET tumor delineation method based on adaptive region-growing and dual-front active contours,” *Medical Physics*, vol. 35, no. 8, pp. 3711–3721, 2008.

[25] M. Abdoli *et al.*, “Contourlet-based active contour model for PET image segmentation,” *Medical Physics*, vol. 40, no. 8, p. 082507, 2013.

[26] A.-S. Dewalle-Vignion *et al.*, “A new method for volume segmentation of PET images, based on possibility theory,” *IEEE Transactions on Medical Imaging*, vol. 30, no. 2, pp. 409–423, 2011.

[27] Z. Li *et al.*, “Lesion detection in dynamic FDG-PET using matched subspace detection,” *IEEE Transactions on Medical Imaging*, vol. 28, no. 2, pp. 230–240, 2009.

[28] Z. Xu *et al.*, “Joint solution for PET image segmentation, denoising, and partial volume correction,” *Medical Image Analysis*, vol. 46, pp. 229–243, 2018.

[29] H. Mi *et al.*, “Joint tumor growth prediction and tumor segmentation on therapeutic follow-up PET images,” *Medical Image Analysis*, vol. 23, no. 1, pp. 84–91, 2015.

[30] B. Foster *et al.*, “Segmentation of PET images for computer-aided functional quantification of tuberculosis in small animal models,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 711–724, 2014.

[31] C. Ballangan *et al.*, “Automated delineation of lung tumors in PET images based on monotonicity and a tumor-customized criterion,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 5, pp. 691–702, 2011.

[32] R. Cui *et al.*, “A multiprocessing scheme for PET image pre-screening, noise reduction, segmentation and lesion partitioning,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1699–1711, 2021.

[33] A. J. Weisman *et al.*, “Comparison of 11 automated PET segmentation methods in lymphoma,” *Physics in Medicine & Biology*, vol. 65, no. 23, p. 235019, 2020.

[34] Y. LeCun *et al.*, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[35] Y. LeCun *et al.*, “Deep learning,” *Nature*, vol. 521, no. 28, pp. 436–444, May 2015.

[36] E. Shelhamer *et al.*, “Fully convolutional networks for semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[37] O. Ronneberger *et al.*, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015, pp. 234–241.

[38] Ö. Çiçek *et al.*, “3D U-Net: Learning dense volumetric segmentation from sparse annotation,” in *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention*, 2016, pp. 424–432.

[39] F. Milletari *et al.*, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proceedings of International Conference on 3D Vision*, 2016, pp. 565–571.

[40] L. Chen *et al.*, “Automatic PET cervical tumor segmentation by combining deep learning and anatomic prior,” *Physics in Medicine & Biology*, vol. 64, no. 8, p. 085019, 2019.

[41] E. Pfaehler *et al.*, “Repeatability of two semi-automatic artificial intelligence approaches for tumor segmentation in PET,” *EJNMMI Research*, vol. 11, no. 1, p. 4, 2021.

[42] J. Wehrend *et al.*, “Automated liver lesion detection in ^{68}Ga DOTATATE PET/CT using a deep fully convolutional neural network,” *EJNMMI Research*, vol. 11, no. 1, p. 98, 2021.

[43] K. H. Leung *et al.*, “A physics-guided modular deep-learning based automated framework for tumor segmentation in PET,” *Physics in Medicine & Biology*, vol. 65, no. 24, p. 245032, 2020.

[44] Y. Lu *et al.*, “Automatic tumor segmentation by means of deep convolutional u-net with pre-trained encoder in PET images,” *IEEE Access*, vol. 8, pp. 113 636–113 648, 2020.

[45] Z. Liu *et al.*, “A Bayesian approach to tissue-fraction estimation for oncological PET segmentation,” *Physics in Medicine & Biology*, vol. 66, no. 12, p. 124002, 2021.

[46] P. Blanc-Durand *et al.*, “Automatic lesion detection and segmentation of 18F-FET PET in gliomas: A full 3D u-net convolutional neural network study,” *PLoS One*, vol. 13, no. 4, p. e0195798, 2018.

[47] A. Iantsen *et al.*, “Convolutional neural networks for PET functional volume fully automatic segmentation: development and validation in a multi-center setting,” *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 48, p. 3444–3456, 2021.

[48] K. He *et al.*, “Identity mappings in deep residual networks,” in *Proceedings of European Conference on Computer Vision*, 2016, pp. 630–645.

[49] J. Hu *et al.*, “Squeeze-and-excitation networks,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[50] V. Oreiller *et al.*, “Head and neck tumor segmentation in PET/CT: The HECKTOR challenge,” *Medical Image Analysis*, vol. 77, p. 102336, 2022.

[51] U. Bagci *et al.*, “Joint segmentation of anatomical and functional images: Applications in quantification of lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT images,” *Medical Image Analysis*, vol. 17, no. 8, pp. 929–945, 2013.

[52] Y. Song *et al.*, “Lesion detection and characterization with context driven approximation in thoracic FDG PET-CT images of NSCLC studies,” *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 408–421, 2014.

[53] C. Lian *et al.*, “Joint tumor segmentation in PET-CT images using co-clustering and fusion based on belief functions,” *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 755–766, 2019.

[54] Y. Zhao *et al.*, “Deep neural network for automatic characterization of lesions on 68ga-psma pet/ct images,” in *Proceedings of The 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2019, pp. 951–954.

[55] L. Bi *et al.*, “Recurrent feature fusion learning for multi-modality pet-ct tumor segmentation,” *Computer Methods and Programs in Biomedicine*, vol. 203, p. 106043, 2021.

[56] Z. Xue *et al.*, “Multi-modal co-learning for liver lesion segmentation on PET-CT images,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3531–3542, 2021.

[57] D. Jin *et al.*, “Deeptarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy,” *Medical Image Analysis*, vol. 68, p. 101909, 2021.

[58] A. Kumar *et al.*, “Co-learning feature fusion maps from PET-CT images of lung cancer,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 204–217, 2020.

[59] Z. Guo *et al.*, “Deep learning-based image segmentation on multimodal medical imaging,” *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, 2019.

[60] Y. Zhao *et al.*, “Deep neural network for automatic characterization of lesions on ^{68}Ga-PSMA-11 PET/CT,” *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 47, pp. 603–613, 2020.

[61] X. Hu *et al.*, “Coarse-to-fine adversarial networks and zone-based uncertainty analysis for NK/T-cell lymphoma segmentation in CT/PET images,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2599–2608, 2020.

[62] P. Borrelli *et al.*, “Ai-based detection of lung lesions in [18F]FDG PET-CT from lung cancer patients,” *EJNMMI Physics*, vol. 8, p. 32, 2021.

[63] L. Sibille *et al.*, “18F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks,” *Radiology*, vol. 294, no. 2, pp. 445–452, 2020.

[64] E. A. Carlsen *et al.*, “A convolutional neural network for total tumor segmentation in [64Cu] Cu-DOTATATE PET/CT of patients with neuroendocrine neoplasms,” *EJNMMI research*, vol. 12, no. 1, pp. 1–10, 2022.

[65] K. He *et al.*, “Deep residual learning for image recognition,” in *IEEE Conf. Comput. Vis. Pattern Recognit*, 2016, pp. 770–778.

[66] D. Ulyanov *et al.*, “Instance normalization: The missing ingredient for fast stylization,” *arXiv:1607.08022*, 2016.

[67] D. A. Clevert *et al.*, “Fast and accurate deep network learning by exponential linear units (ELUs),” in *Proceedings of the International Conference on Learning Representations*, 2016, pp. 1–14.

[68] H. Zhang *et al.*, “Deep TEN: Texture encoding network,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2896–2905.

[69] H. Zhang *et al.*, “Context encoding for semantic segmentation,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[70] T. Bai *et al.*, “Context-aware learning for cancer cell nucleus recognition in pathology images,” *Bioinformatics*, vol. 38, no. 10, pp. 2892–2898, 2022.

[71] J. Sivic and A. Zisserman, “Video google: a text retrieval approach to object matching in videos,” in *Proceedings of IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[72] G. Csurka *et al.*, “Visual categorization with bags of keypoints,” in *Proceedings of ECCV International Workshop on Statistical Learning in Computer Vision*, 2004, pp. 1–16.

[73] J. C. van Gemert *et al.*, “Visual word ambiguity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[74] L. Liu *et al.*, “In defense of soft-assignment coding,” in *Proceedings of International Conference on Computer Vision*, 2011, pp. 2486–2493.

[75] Y. LeCun *et al.*, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[76] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Proceedings of European Conference on Computer Vision*, 2014, pp. 818–833.

[77] C.-Y. Lee *et al.*, “Deeply-supervised nets,” in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, vol. 38, 2015, pp. 562–570.

[78] S. Xie and Z. Tu, “Holistically-nested edge detection,” in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 1395–1403.

[79] I. Sutskever *et al.*, “On the importance of initialization and momentum in deep learning,” in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, no. 3, 2013, pp. 1139–1147.

[80] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[81] O. Oktay *et al.*, “Attention u-net: Learning where to look for the pancreas,” in *Proceedings of the International Conference on Medical Imaging with Deep Learning*, 2018, pp. 1–10.

[82] A. G. Roy *et al.*, “Recalibrating fully convolutional networks with spatial and channel “squeeze and excitation” blocks,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 540–549, 2019.

[83] A.-M. Rickmann *et al.*, “Recalibrating 3D convnets with project & excite,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2461–2471, 2020.

[84] A. Iantsen *et al.*, “Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined pet and ct images,” in *Head and Neck Tumor Segmentation*, 2021, pp. 37–43.

[85] J. Chen *et al.*, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv:2102.04306 [cs.CV]*, pp. 1–13, 2021.

[86] H. Peiris *et al.*, “A robust volumetric transformer for accurate 3d tumor segmentation,” in *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention*, 2022, pp. 162–172.