IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 February 2024, accepted 24 February 2024, date of publication 1 March 2024, date of current version 7 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3372146

RESEARCH ARTICLE

MSGFormer: A DeepLabv3+ Like Semantically Masked and Pixel Contrast Transformer for MouseHole Segmentation

PENG YANG, CHUNMEI LI, CHENGWU FANG, SHASHA KONG, YUNPENG JIN, KAI LI, HAIYANG LI, XIANGJIE HUANG, AND YAOSHENG HAN

Department of Computer Technology and Application, Qinghai University, Xining 810016, China

Corresponding author: Chunmei Li (li_chm0422@sina.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62166033, and in part by the Applied Basic Research Project of Qinghai Province under Grant 2024-ZJ-788.

ABSTRACT In semantic segmentation, the efficient representation of multi-scale context is of paramount importance. Inspired by the remarkable performance of Vision Transformers (ViT) in image classification, subsequent researchers have proposed some Semantic Segmentation ViTs, most of which have achieved impressive results. However, these models often struggle to effectively utilizing multi-scale context, disregarding intra-image semantic context, and neglecting the global context of training data, i.e., the semantic relationships among pixels across different images. In this paper, we introduce the Sliding Window Dilated Attention and combine it with the Spatial Pyramid Pooling (SPP) to form a novel decoder called Sliding window dilated attention spatial pyramid pooling(SwinASPP). By adjusting the sliding window dilation rates, this decoder is capable of capturing multi-scale contextual information from different granularities. Additionally, we propose the Semantic Attention Block, which integrates semantic attention operations into the encoder. And adopt our proposed supervised pixel-wise contrastive learning algorithm, we shift the current training strategy to inter-image for semantic segmentation. Our experiments demonstrate that these methods lead to performance improvements on the SanJiangYuan MouseHole dataset and Cityscapes.

INDEX TERMS Deep learning, semantic segmentation, multi-scale context, semantic context, pixel contrast.

I. INTRODUCTION

Semantic segmentation is one of the fundamental tasks in computer vision, which aims to assign a semantic label to each pixel in an image. Fully Convolutional Networks (FCNs) [1] are the pioneering work that treats semantic segmentation as a pixel-level prediction task, and since then, many subsequent works have been inspired by FCNs.

Following the tremendous success of transformers in the natural language processing(NLP), many scholars have proposed incorporating transformers into visual tasks. Dosovitskiy et al. [2] proposed vision Transformer (ViT) for image classification has achieved remarkable performance. Subsequently, in order to demonstrate the effectiveness of transformer in semantic segmentation, Zheng et al. [3] proposed SETR, had achieved state of the art on ADE20K and Pascal Context. Currently, the mainstream method employs a transformer backbone pretrained on ImageNet [4] as the encoder, in conjunction with a decoder based on CNN for finetuning on semantic segmentation task. CNN-based decoder designs primarily focus on addressing the issue of utilizing multi-scale contextual representations. In order to integrate multi-scale contextual information, most of these works incorporate atrous convolution [5] or pooling operations into the Spatial Pyramid Pooling(SPP) module [6], [7], [8]. Segformer [9] designs a lightweight MLP as the decoder.

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

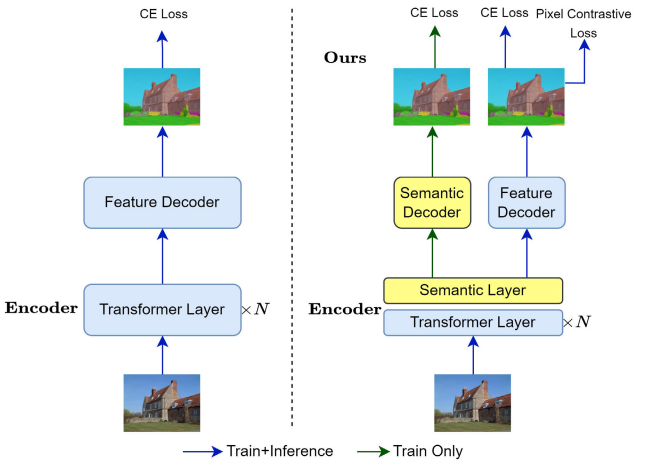


FIGURE 1. Comparison between current transformer-based segmentation model(left) and ours(right). MSGFormer by incorporating an additional semantic layer within the encoder backbone to utilize semantic priors and by employing a pixel-wise contrastive learning algorithm for semantic segmentation. These simple modifications eventually yield performance improvements.

However, CNN possess only a limited receptive field, which only model the local dependencies relationships of pixels when used as decoder, thereby neglecting long-range dependencies to some extent. Although atrous convolution can expand the receptive field, its scope is still limited. The segmentation performance of Segformer [9] is excessively reliant on the capacity of the encoder, which may compromise the upper bound of the model's performance. Therefore, by analyzing the above situation, we argue that a major issue with the current ViT for semantic segmentation is its inadequate utilization of multi-scale contextual information, thereby affecting performance. In order to overcome this limitation, we propose a novel attention mechanism, termed *sliding window dilated attention*. By setting varying dilation rates r, we can capture global contextual information at different granularities. Coupled with sliding window dilated attention, a SPP module evolves into a *sliding window dilated attention spatial pyramid pooling*(SwinASPP), similar to the Atrous Spatial Pyramid Pooling(ASPP) [7] and Pyramid Pooling Module(PPM) [10], which can utilize multi-scale representations for semantic segmentation.

The method employing a pretrained transformer backbone as the encoder and SwinASPP as the decoder for finetuning still has limitations, as it is unable to utilize the semantic-level contextual information within the images. Jin et al. [11] proposed ISNet, which enhances pixel representation by aggregating image-level contextual information and semantic-level contextual information in the decoder structure. However, ISNet is a CNN-based method that only incorporates semantic-level context in the decoder, while the encoder remains unchanged. To address the issues, we use the *SeMask Attention Block*, which integrates semantic information into the hierarchical vision transformer architecture and utilizes semantic context to enhance the global feature information captured by the transformer backone. We insert a semantic layer after each stage of the transformer in the backbone, and employ a lightweight semantic decoder to accumulate semantic features from all stages and utilize our SwinASPP decoder for the main pixel-level prediction.

Although we employ semantic-level context and image-level multi-scale context information, they only consider the local dependency relationships within a single image, neglecting the "global" context of the entire dataset, i.e. the semantic relationships between pixels across images. With the remarkable success of contrastive learning in the field of unsupervised representation learning [12], [13], it has been demonstrated the effectiveness of utilizing global context within training data to enhance performance. Motivated by this, we propose a pixel-wise contrastive algorithm for supervised semantic segmentation. Specifically, in addition to utilizing pixel-wise cross entropy loss for solving pixel classification, we employ *pixel-wise contrastive loss* to calculate pixel-to-pixel contrast, enforcing the embedding of pixels to be close to positive samples while pushing away negative samples. As the pixel-level classification information is provided during training, positive samples are pixels belonging to a same class, while negative samples are from different classes. In this way, global attributes in the embedding space can be captured to better reflect the inherent structure of the training data and achieve more accurate segmentation predictions.

The contributions of this paper include three fold:

- We propose a *Sliding Window Dilated Attention*. By setting different dilation rates and combining with the SPP module, a transformer-based decoder is designed to explore multi-scale context information for semantic segmentation.
- We propose a *SeMask Attention Block*, which can incorporate semantic prior information into pretrained vision transformer backbone, providing semantic context to the encoder.
- We propose a supervised, pixel-wise contrastive learning approach for semantic segmentation. This method exploits the global context of the training data by expanding the training strategy beyond individual images to encompass multiple images. By calculating pixel-to-pixel contrast, our method leverages semantic relationships among pixels and between pixels and semantic regions.

II. RELATED WORK

A. SEMANTIC SEGMENTATION

Semantic segmentation can be regarded as the extension of image classification from image level to pixel level. FCN [1] is a representative work in semantic segmentation, which is a fully convolutional network capable of performing pixel-level classification end-to-end. After that, in order to achieve precise segmentation, researchers have continuously improved the semantic segmentation CNN from various aspects, such as: enlarging the receptive field [5], [7],

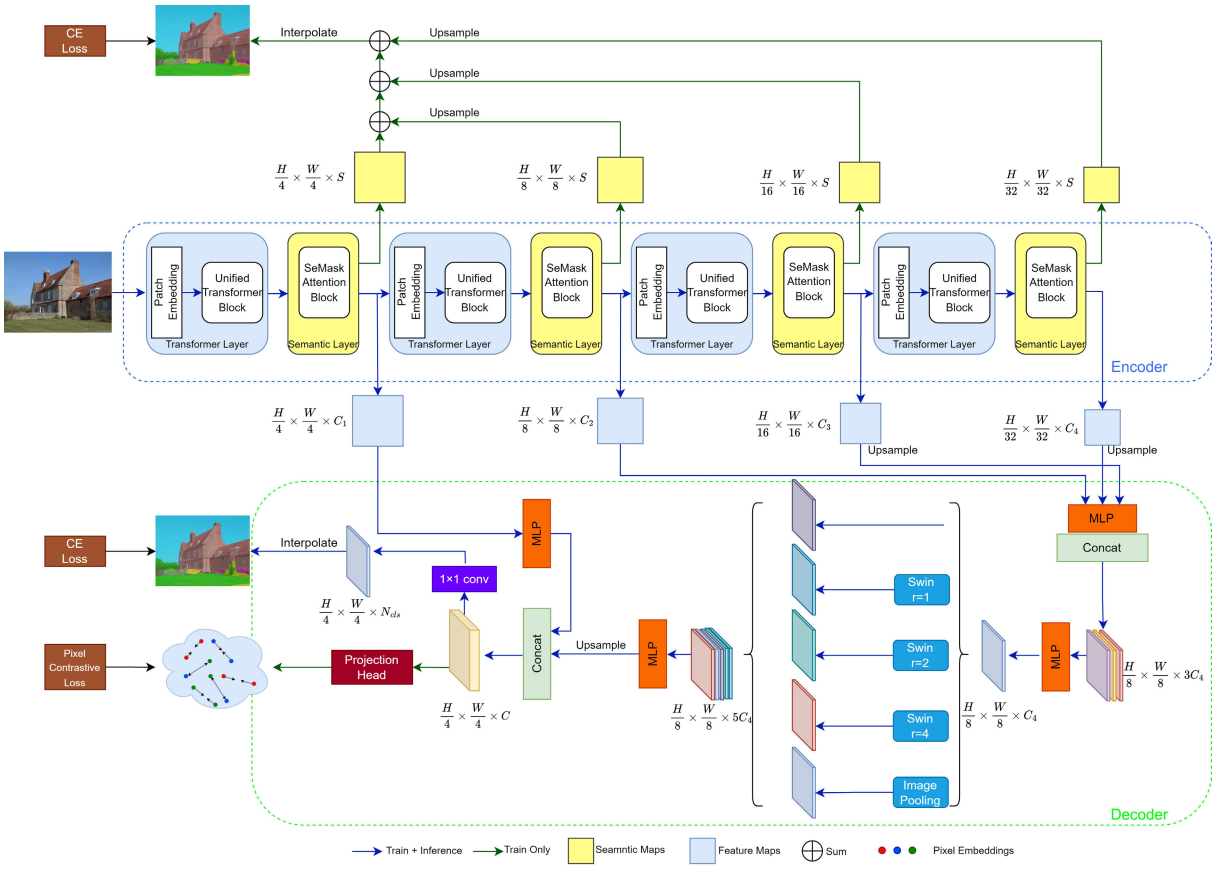


FIGURE 2. The overall structure of network. After the unified transformer layer, we introduce a semantic layer with SeMask blocks Fig. 3(b) to capture semantic context. Semantic maps at each stage are aggregated using a simple upsample + sum operation and supervised for semantic context using weighted CE Loss. In the ASPP decoder, sliding-window dilated attention with dilation rates of r=1, 2, 4 is applied to capture features at different granularities. The final output feature maps undergo two distinct processes: i) They are subjected to a 1 × 1 convolution to reduce the channel dimension to N_{cls}, and supervised by CE loss for the network's main prediction, and ii) They are passed through a projection head to map high-dimensional pixel embeddings into 256-dimensional ℓ_2-norm feature vectors, which are utilized for calculating the contrastive loss.

[10], [14], [15]; supplementing boundary information [16], [17], [18], [19], [20]; refining contextual information [17], [21], [22], [23], [24], [25]; and incorporating attention mechanisms in the decoder [26], [27], [28]. Recent methods have demonstrated the effectiveness of Transformer-based architectures for semantic segmentation [3], [9], [29], [30]. However, these methods still have shortcomings.

B. MULTI-SCALE CONTEXT INFORMATION

In order to enhance the pixel representation in semantic segmentation networks, designing a reasonable context information aggregation scheme is a common approach. To obtain multi-scale contextual information, DeepLab [7], [15], [31] proposes adopt pyramid dilated convolutions with different dilation rates, PSPNet [10] performs spatial pooling at different scales. References [17], [32], and [33] utilizing the image pyramid method. SegNeXt [34] leverages multi-branch deep atrous convolution to capture multi-scale context, and employs Hamburger to further extract global contextual information. MaskFormer [35] transforms the problem of per-pixel classification into mask classification, thus unifying semantic segmentation and instance segmentation tasks. Mask2Former [30] introduces masked attention to reduce computational costs and memory usage, and utilizes multi-scale high-resolution features to segment small objects. By enhancing the subtask framework of MaskFormer [35], it evolves into a versatile image segmentation framework. Recently, most Transformer-based semantic segmentation methods adopt popular Semantic-FPN [36] and UperNet [37] as the fundamental frameworks. They incorporate hierarchical vision Transformers such as [38], [39], [40], and [41] as feature extraction encoder. In this work, we designed a Transformer-based decoder with global attention to explore multi-scale contextual information for semantic segmentation.

C. SEMANTIC CONTEXT INFORMATION

Zhang et al. [21] proposed a context encoding module to capture and utilize the semantic contextual information in images, which selectively emphasizes feature map related to the category. OCRNet [24], ACFNet [42], and SCARF [43], EMANet [44] model contextual relationships

within specific semantic class region based on coarse segmentation. References [11] and [45] proposed specially designed modules that aggregate image-level and semantic-level contextual information in the decoder to enhance pixel representation. More recently, IDRNet [46] employs an intervention-driven approach to transform pixel-level representations into semantic-level representations. Subsequently, it executes deletion diagnostics [47] procedure to model the relationships between semantic-level representations. These works are CNN-based, which captures the semantic context in the decoder. In this work, we argue that the approach mentioned above could lead to a potential loss of semantic information during the encoding stage. Therefore, we propose capturing semantic context during the encoding stage of the Segmentation Vision Transformer.

D. GLOBAL CONTEXT INFORMATION OF TRAINING DATA

Recently, unsupervised contrastive learning [12], [13], [48], [49] has been the most widely used method for learning representations without labels. It only requires learning to distinguish data in the abstract semantic feature space, making the model not only more simplified but also more generalizable. Subsequently, [50], [51], [52], [53] have also demonstrated that label information can help contrastive learning in image-level pattern pre-training. Although some works [54], [55], [56] have addressed the contrastive learning problem in dense prediction tasks, they typically consider contrastive learning as a pre-training step for dense image embedding, and calculate the contrast among pixels using augmented versions of a same image, simply utilizing local context within a single image. References [57] and [58] proposes to mine the contextual information beyond individual images to further augment the pixel representations. In recent study, [59] have aggregated dataset-level contextual information beyond the input images using a memory module. We propose a pixel-to-pixel contrastive learning method for supervised semantic segmentation, which explores the global pixel relationships in the training data.

III. METHOD

In this section, we introduce the semantically masked and pixel-wise contrastive transformer in detail. First, we describe the overview of our transformer encoder. Then, we elaborate our SwinASPP decoder. Finally, introduce the loss function we used in our model, especially the contrastive loss will be described in detail.

A. SEMANTICALLY MASKED ENCODER

Encoder has four different stages,each stage consists of two layers: The transformer layer,which is N_f number of Unified Transformer blocks Fig. 3(a) stacked together; The Semantic layer, which is N_c number of Semantic Attention blocks Fig. 3(b). The transformer layer is followed by a semantic layer to form our SeMask layer.

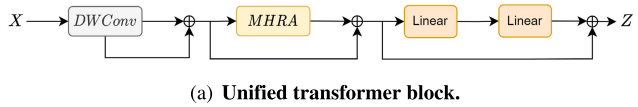
In the process of training, the Transformer Layer outputs is the inputs of Semantic Layer. The intermediate semantic prior features and semantically masked maps Fig. 3(b) will be returned by semantic layer. The semantically masked maps from each stages are aggregated using the SwinASPP decoder for final dense-pixel prediction. The semantic prior features from each stages are aggregated using a lightweight upsample and sum operation-based semantic decoder to predict the semantic-prior for our model.

1) UNIFIED TRANSFORMER LAYER

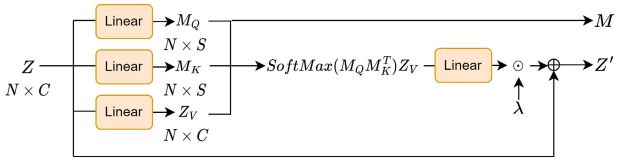
It unifying convolution and self-attention in a concise transformer encoder. In the shallow layers, it use convolution neural networks to decrease computation redundancy. In the deep layers, it use self-attention to capture long-range global dependency. By stacking local and global Unified transformer blocks hierarchically, it can flexibly integrate their cooperative capabilities to promote representation learning while achieving a balance between computational complexity and accuracy.

2) SEMANTIC LAYER

Different from the Unified Transformer layer, the importance of the semantic layer is to model the semantic context, which is used as a prior to calculate the semantic attention weights to update the feature map according to the semantic prior knowledge existing in the image. Within each semantic layer, there are N_c SeMask attention blocks Fig. 3(b). In order to reduce the computational expenditure of SeMask attention block, the method of calculation refers to the window self-attention mechanism of Swin transformer [41]. During the training, the semantic block can use the semantic context provided by the image in the encoding stage to generate semantic prior features to guide the training of the encoder. The feature map Z from the previous transformer layer passes through the semantic block to generate M_Q, M_K, Z_V. We obtain M_K and M_Q by projecting features into the semantic space. The dimension of M_Q and M_K are $N \times S$, where S denotes the number of classes, and the dimension



(a) **Unified transformer block.**



(b) **Semantic Attention Block.**

FIGURE 3. SeMask block. As shown in (a), N_f Uniformer blocks are stacked at each transformer layer, and N_c semantic attention blocks, as shown in (b), are stacked in each semantic layer at every stage Fig. 2. The output Z from the last Uniformer block is fed into the first semantic attention block in the semantic layer.

of Z_V is $N \times C$, where C is the embedding dimension. M_Q generates a semantic graph and calculate the semantic attention matrix utilizing both M_K and M_Q. This matrix is passed through softmax and used to update Z_V, as shown in Fig. 3(b). The Semantic attention equation can be defined as follows:

$$SeAttention(M_Q, M_K, Z_V) = SoftMax(M_Q M_K^T) Z_V \quad (1)$$

We apply matrix multiplication between the feature values and semantic attention weights. The resulting matrix product is subsequently transferred through a linear layer and multiplied by a learnable scalar constant λ for smooth fine-tuning. Following a residual connection, we ultimately obtain the adapted features. These features include abundant semantic information, which we refer to as semantic masking features. Subsequently, the semantic query M_Q is used to optimize the semantic prior graph.

B. SwinASPP DECODER

In order to incorporate multi-scale features, we utilize the architecture of spatial pyramid pooling to combined with sliding window dilated attention and get the novel SPP module called SwinASPP. The structure contains five branches including one shortcut connection, one image pooling branch and three sliding window dilated attention with $r = (1, 2, 4)$. Then the results of the five branches are concatenated together. A MLP layer reduce channel dimension of fused feature map to 512, and then upsample to 1/4 of the image size, fuses with the first stage of the encoder's output. Finally, a 1×1 convolution takes the feature to predict the segmentation logits with $\frac{H}{4} \times \frac{W}{4} \times N_{cls}$ resolution.

1) SLIDING WINDOW DILATED ATTENTION

Conventional self-attention mechanisms possess a global receptive field. However, they incur significant computational cost. To capture multi-scale contextual information from diverse receptive fields while maintaining a balanced computation complexity, we propose a sliding window hole attention with varying dilation rates.

$$\begin{aligned} K_r &= Swin(K, r) \\ V_r &= Swin(V, r) \end{aligned} \quad (2)$$

$$SwinAttention(Q, K_r, V_r) = SoftMax(\frac{QK_r}{\sqrt{d}})V_r \quad (3)$$

where Q refers to query matrix, K_r and V_r refers to the key and value matrix results after applying the sliding window operation with the dilation rate r to K and V, respectively.

C. LOSS FUNCTION

In the training process, both the cross-entropy loss function and pixel contrastive loss function are utilized. The total loss $\mathcal{L}_T$ is calculated as the sum of two pixel-wise cross-entropy losses $\mathcal{L}_1$, $\mathcal{L}_2$ and a pixel-wise contrastive loss $\mathcal{L}_v^{\text{NCE}}$. Loss $\mathcal{L}_1$ is calculated based on the primary prediction from the SwinASPP decoder, while loss $\mathcal{L}_2$ is derived from the semantic prior prediction of our lightweight decoder. As the cross-entropy loss function only explores the relationships between pixels within a single image, it overlooks the global context among the entire training dataset images. Therefore, the pixel contrastive loss function is introduced to enhance the intra-class compactness and inter-class discreteness of all images. During training, pixels within the same class will be continuously pulled closer together, while pixels from different classes will be pushed apart.

1) CROSS-ENTROPY LOSS

The current semantic segmentation task assigning a semantic class to each pixel in an image, treating it as a pixel-level classification task. Specifically, let encoder-decoder produce a dense feature $\mathcal{F} \in \mathbb{R}^{H \times W \times D}$. Then a segmentation head g_{SEG} maps $\mathcal{F}$ into a categorical logits map $O = g_{SEG}(\mathcal{F}) \in \mathbb{R}^{H \times W \times |C|}$. We define our losses on O and $\mathcal{M}$ as follows:

$$\mathcal{L}_1 = \frac{1}{H \times W} \sum_{i,j} \mathcal{L}^{ce} \left(\text{softmax}(O_{[i,j]}), 1_{\bar{c}}^{\top} \left(\mathcal{GT}_{[ij]} \right) \right) \quad (4)$$

$$\mathcal{L}_2 = \frac{1}{H \times W} \sum_{i,j} \mathcal{L}^{ce} \left(\text{softmax}(\mathcal{M}_{[i,j]}), 1_{\bar{c}}^{\top} \left(\mathcal{GT}_{[ij]} \right) \right) \quad (5)$$

where $[i, j]$ denotes the current predicted pixel, $\bar{c}$ denotes the ground-truth label of pixel $[i, j]$, $1_{\bar{c}}$ denotes for converting the class label stored in $\mathcal{GT}$ into a one-hot format. $\mathcal{F}$ denotes the main prediction of the network, and $\mathcal{M}$ denotes the semantic prior prediction.

2) PIXEL-WISE CONTRASTIVE LOSS

a: PIXEL-TO-PIXEL CONTRAST

The cross-entropy loss function treats each pixel independently for prediction, without considering the relationships between pixels within the same image and different images. To address this problem, we employ a pixel contrastive learning approach that regularizes the embedding space and explores the global structure of the training data. Essentially, our contrastive loss computation involves training image pixels as data samples. For pixel v with ground-truth semantic label $\bar{c}$, the positive sample is defined as other pixels belonging to class $\bar{c}$, while the negative sample is defined as the pixels not belonging to class $\bar{c}$. The supervised pixel-level contrastive loss is defined as:

$$\mathcal{L}_v^{\text{NCE}} = \frac{1}{|\mathcal{P}_v|} \sum_{v^+ \in \mathcal{P}_v} - \log \frac{\exp(v \cdot v^+ / \tau)}{\exp(v \cdot v^+ / \tau) + \sum_{v^- \in \mathcal{N}_v} \exp(v \cdot v^- / \tau)} \quad (6)$$

where v^+ is an pixel embedding of positive sample for $\mathcal{P}_v$, $\mathcal{N}_v$ refer to pixel embedding collections of the negative sample, the range of temperature hyper parameters is $\tau > 0$. Note that the positive samples, negative samples and the current pixel v are not limited to a same image.

b: PIXEL-TO-REGION CONTRAST

In contrastive learning, memory bank is a key technique that helps learn good representations by utilizing a large amount of data during training. However, due to the segmentation task setting with a large number of pixel samples, most of them store all the training pixel samples directly, such as traditional memory [60], which will significantly slow down the training process. Maintaining a few latest batches in the queue, such as [61], [62], and [63], is not an optimal solution either, because the most recent batches contain only a limited number of images, reducing the diversity of pixel samples. Hence, we choose to create a pixel queue for each category. For each category, we randomly select a small number of pixels (i.e., U) from each image in the latest mini-batch and add them to the queue with a size of $T \gg U$. In practical use, we find this strategy to be very effective, but the undersampled pixel embeddings are too sparse to utilize only a small amount of information from the image. Therefore, we further construct a region memory bank that stores more representative embeddings absorbed from semantic regions of the image.

In particular, for a segmentation dataset with $|\mathcal{C}|$ semantic classes, our regional memory is constructed with a size of $|\mathcal{C}| \times N \times D$, where D is the dimension of pixel embeddings, and N is the size of the region memory. The $(\bar{c}, n)$-th element in the region memory is obtained by average pooling the D-dimensional feature vectors of all pixel embeddings labeled as class $\bar{c}$ in the current image. Utilizing region memory allows our pixel contrastive loss function to explore the relationship between from pixel to region. When calculating the Eq. 6 for the current pixel v belonging to class $\bar{c}$, the stored region embeddings with the same class $\bar{c}$ are considered as positive samples, while the negative sample is defined as the region embeddings not belonging to class $\bar{c}$. Hence, the overall training objective is:

$$\mathcal{L}_T = \mathcal{L}_1 + \alpha \mathcal{L}_2 + \beta \mathcal{L}_v^{\text{NCE}} \tag{7}$$

where α and β are the coefficient. We empirically set $\alpha = 0.4$ and $\beta = 0.2$.

IV. EXPERIMENTS

A. DATASETS AND METRICS

1) MOUSEHOLE DATASET

In the grasslands of the Sanjiangyuan Region, rodent pest is one of the factors accelerating grassland degradation. The extensive digging, root excavation, and grass consumption by rodents lead to widespread death of pasture. We employ the MouseHole dataset to study the relationship between rodent pest and grassland degradation, this dataset comprises 7,562 finely annotated RGB images captured in the Sanjiangyuan Region of Qinghai Province, covering a total of four semantic classes: eroded grassland around MouseHole, non-eroded grassland around MouseHole, stone, and cow dung. All images in the dataset have a resolution of 512×512. A total of 6,187 images were used as the training set, while 688 images were divided into the validation set, and 687 images served as the test set.

2) CITYSCAPES

CityScapes is one of the most challenging scene parsing datasets containing 5,000 fine-annotated images with 19 categories. The dataset comprises 2,975/500/1,525 images for the training, validation, and test sets, respectively.

3) METRICS

We report mean Intersection-over-Union (mIoU) over all classes for evaluation.

B. IMPLEMENTATION DETAILS

1) TRAINING

All experiments presented in this section are implemented using the MMSegmentation[1] codebase on a server with 8 NVIDIA GeForce GTX 1080Ti. We employ the backbone UniFormer [38] and SwinASPP to comprehensively validate the proposed algorithm. We adhere to the conventions of [9] for training hyper-parameters. To ensure fairness, we initialize backbone with pre-trained weights on ImageNet [4], while the remaining layers being randomly initialized. For data augmentation, we use scaling with a ratio randomly sampled from (0.5,0.75,1.0,1.25,1.5,1.75), color jitter and horizontal flipping. We randomly crop large images and pad small images to a same size of 512×512 for MouseHole dataset and 768×768 for Cityscapes. In order to train the model for semantic segmentation tasks, we employed the AdamW [64] optimizer with a base learning rate γ_0. We adopt the polynomial annealing policy to schedule the learning rate $\gamma = \gamma_0 \left(1 - \frac{N_{iter}}{N_{total}}\right)^{0.9}$. A linear warm-up strategy was used for 1,500 iterations. We set the base learning rate γ_0 to 0.00006, weight decay to 10^{-2} and train for 160K iterations with a batch size of 16 for MouseHole dataset and 8 for Cityscapes.

2) INFERENCE

To deal with varying image sizes during the inference, we maintain the aspect ratio constant and resize the images to the smaller edge resolution, and then rescale to the original dimensions before calculating the evaluation metrics.

C. ABLATION STUDIES

1) SwinASPP DECODER

In order to prove that SwinASPP improves efficiency, we compare it with DeepLabv3+ [31] and SegFormer [9]. To ensure a fair comparison, they utilized Uniformer [38] as the encoder for both. Table 1 presents a comparison of parameters, FLOPs, and mIoU. In Table 2, We investigated the influence of various dilation rates on performance.

[1] https://github.com/open-mmlab/mmsegmentation

TABLE 1. Comparision of SegFormer and DeepLabV3+ with ours. "Ours" refers to the encoder using only Uniformer-S, while the decoder employs SwinASPP.

Method	Backbone	Params(M)	FLOPs(G)	mIoU
DeeplabV3+	ResNet101	62.57	253.93	77.58
SegFormer	UniFormer-S	24.19	58.29	77.46
Ours	UniFormer-S	33.13	68.49	**77.74**

TABLE 2. Ablation on dilation rates. The $r = (1, 2, 4)$ is the optimal setting.

Rate	Params(M)	FLOPs(G)	mIoU
(1,1,1)	33.13	68.49	77.52
(1,2,1)	33.13	68.49	77.61
(1,2,3)	33.13	68.49	77.66
(1,2,4)	33.13	68.49	**77.74**

2) SeMasK BLOCK

We conducted ablation studies on different variants of the SeMask Block. We investigate the impact of semantic attention and the number of SeMask blocks (N_c), reporting results through single-scale inference on the MouseHole val dataset. In Table 3, by replacing the Semantic Attention Block with a simple self-attention block on the Uniformer-S variant, it becomes evident that the simple attention does not contribute to result improvement. This demonstrates the effectiveness of our SeMask Block. In Table 4, we investigate the impact of the number of SeMask attention blocks on performance by varying the N_c values within each semantic layer of the Uniformer-S variant. We observe that $N_c = [1, 1, 1, 1]$ is the optimal setting.

TABLE 3. Ablation on Semantic Attention. A simple self-attention block result in performance degradation.

Encoder + Decoder	SA Block	SeMask Block	mIoU(%)
Uniformer-S + SwinASPP			77.74
Variant Uniformer-S + SwinASPP	✓		77.61
SeMask-S + SwinASPP		✓	**78.02**

TABLE 4. Ablation on N_c. The $N_c = [1, 1, 1, 1]$ is the optimal setting.

Encoder + Decoder	N_c	Params(M)	mIoU(%)
SeMask-S + SwinASPP	[1,1,1,1]	38.04	**78.02**
SeMask-S + SwinASPP	[1,1,1,2]	38.28	77.63
SeMask-S + SwinASPP	[2,2,2,2]	39.02	76.07

D. PIXEL-WISE CONTRASTIVE LOSS

We verify the design of our contrastive loss function. In Table 5, our baseline employs Uniformer as the encoder and SwinASPP as the decoder. We respectively incorporate pixel contrast and region contrast, and observe consistent performance gains (pixel contrast from 77.74% to 77.86%, region contrast from 77.74% to 77.97%). Finally, the combination of both forms of contrast yields improved segmentation performance, highlighting the necessity of jointly considering pixel-to-pixel contrast and pixel-to-region contrast.

TABLE 5. Ablation on pixel contrast and region contrast. They both contribute to performance improvements, but their combination yields even better results.

Encoder+Decoder	Pixel	Region	mIoU(%)
Uniformer-S + SwinASPP			77.74
Uniformer-S + SwinASPP	✓		77.86
Uniformer-S + SwinASPP		✓	77.97
Uniformer-S + SwinASPP	✓	✓	**78.13**

E. MAIN RESULTS

1) MOUSEHOLE DATASET

Utilizing SeMask Uniformer as the encoder and SwinASPP as our primary predictor during training, along with the incorporation of cross-entropy and pixel-level contrast loss functions, we achieved a leading performance of 78.3% on the mIoU metric. The comparison of our results with those of other models is presented in Table 6.

2) CITYSCAPES

We conducted experiments on the Cityscapes dataset and reported the results in Table 7. The results show that our model achieves a competitive performance with mIoU of 81.43%. Therefore, our approach can obtain better feature representations for semantic segmentation.

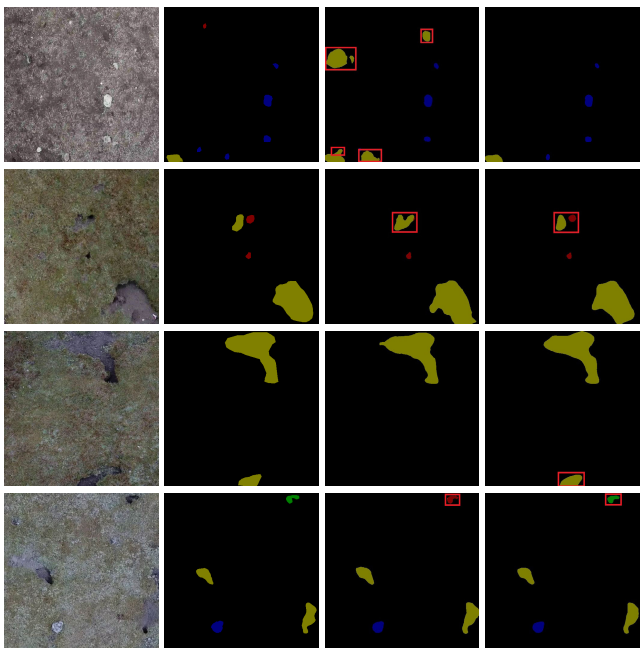


FIGURE 4. Qualitative results on MouseHole-Val. The improved areas are marked with red solid boxes.

3) QUALITATIVE RESULTS

In Fig. 4, we compare the qualitative results of SegFormer and MSGFormer on the Mousehole dataset. The Fig. 4 results

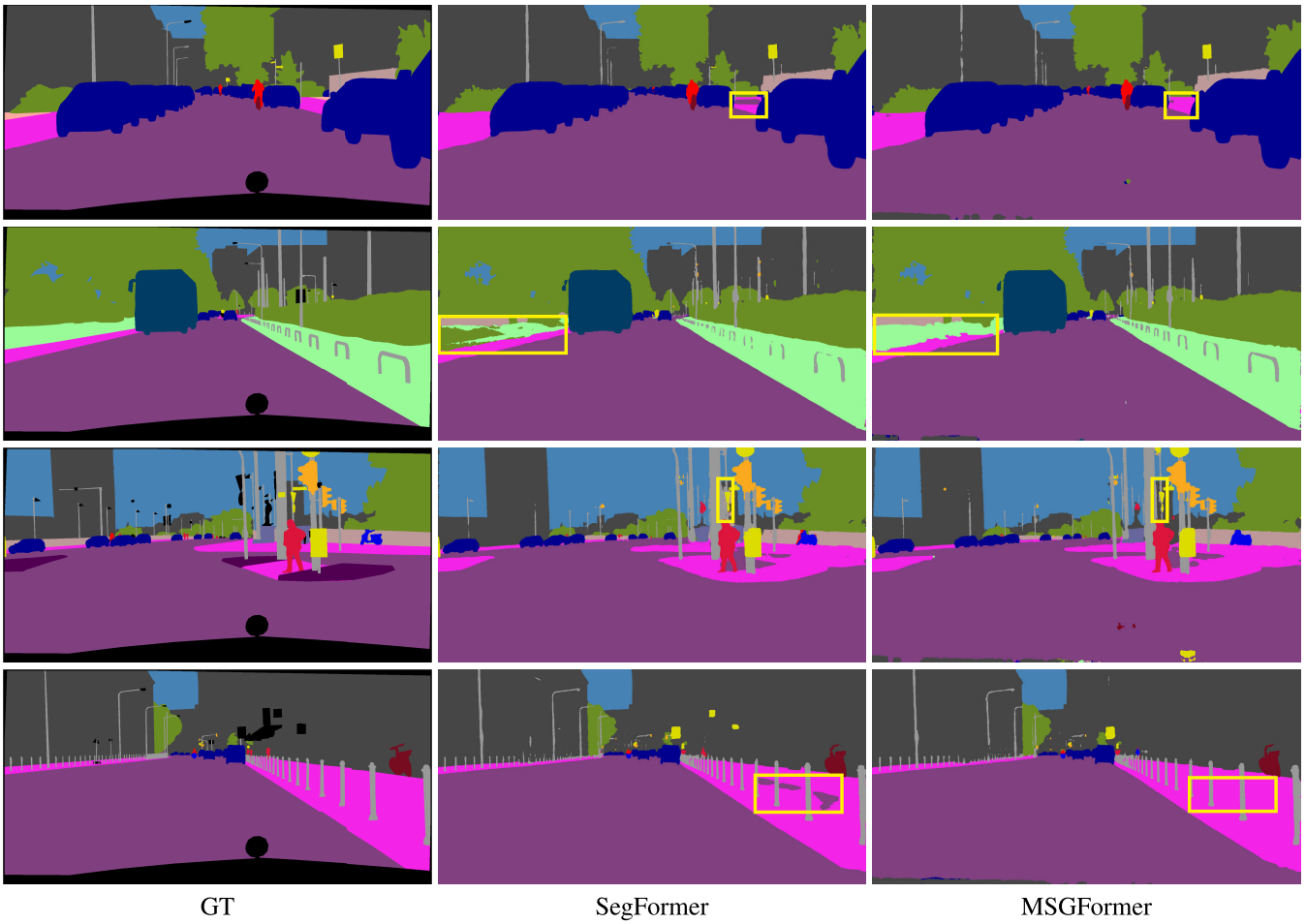


FIGURE 5. Qualitative results on Cityscapes-Val. The improved areas are marked with yellow solid boxes.

TABLE 6. Performance comparison on MouseHole-Val. We report both single-scale(SS) and multi-scale(MS) mIoU on MouseHole validation set. We use the results from the official MMSegmentation trained model.

Model	Backbone	FLOPs(G)	mIoU(SS)	mIoU(MS)
PSPNet [10]	ResNet101	256.14	76.53	77.48
GCNet [65]	ResNet101	275.38	76.61	77.6
PSANet [66]	ResNet101	272.18	76.01	77.2
NonLocal [67]	ResNet101	277.53	76.57	77.56
DeepLabV3 [7]	ResNet101	347.34	76.45	77.71
CCNet [68]	ResNet101	278.07	77.12	77.88
DNLNet [69]	ResNet101	277.53	76.55	77.43
DANet [70]	ResNet101	276.75	77.01	77.65
OCRNet [71]	ResNet101	230.61	77.38	78.36
PIDNet [72]	-	34.45	76.49	-
DeepLabV3+ [31]	ResNet101	253.93	77.58	77.98
SexNeXt-B [34]	MSCAN-B	32.778	77.22	-
SETR [3]	ViT-S	315.84	72.01	72.89
SegFormer [9]	Uniformer-S	58.29	77.46	78.05
Lawin [74]	Uniformer-S	39.8	77.59	78.16
Mask2Former [30]	ResNet101	210.72	77.93	-
PoolFormer-M36 [73]	-	67.74	76.62	-
MSGFormer*(Ours)*	SeMask Uniformer-S	79.33	**78.3**	**78.96**

demonstrate that our MSGFormer generates segments that are both more accurate (as shown in the second and fourth rows) and more complete (as shown in the third row) in complex grassland scenes. Fig. 5 shows that our approach obtain significant improvements in challenging areas, such as small objects and object boundary. This is because

TABLE 7. Performance comparison on Cityscapes-Val. All experimental results are obtained under the input size of 768 × 768. We use the results from the official MMSegmentation trained model.

Model	Backbone	Params(M)	FLOPs(G)	mIoU(SS)
PSANet [66]	ResNet101	75.766	630	77.21
CCNet [68]	ResNet101	66.45	628	77.82
OCRNet [71]	ResNet101	55.52	518	78.46
DeepLabV3+ [31]	ResNet101	60.21	572	78.69
PIDNet [72]	-	37.31	77.53	76.53
SexNeXt-B [34]	MSCAN-B	27.63	73.15	**81.92**
PoolFormer-M36 [73]	-	59.77	152.2	77.05
Mask2Former [30]	ResNet101	63.2	432.26	80.74
SegFormer [9]	Uniformer-S	24.2	131.55	80.36
Lawin [74]	Uniformer-S	27.21	89.03	80.32
MSGFormer*(Ours)*	SeMask Uniformer-S	39.8	178.49	81.43

our Transformer encoder can capture semantic context, while pixel contrast enables discriminative representations, retaining more detailed semantic information. The improved regions are marked with solid boxes.

V. CONCLUSION

In this work, we observe several limitations in the current Semantic Segmentation ViT models, including the lack of an efficient decoder to utilize multi-scale context and the disregard for rich semantic relations among pixels

across different images. Additionally, direct finetuning of the segmentation encoder failed to consider the image's semantic context comprehensively. Therefore, we propose the Sliding Window Dilated Attention, integrated it into the SPP to capture multi-scale contextual information from different granularities efficiently. By means of pixel-wise contrastive learning, we achieved cross-image category-discriminative representations under supervised settings, learning global context from the training data. We propose the Semantic Attention Block, which utilizes semantic attention to capture semantic context and enhance the semantic representation of feature maps. Finally, we conduct experiments on the MouseHole dataset of SanJiangYuan project and the public dataset Cityscapes, our approach demonstrate improvements in semantic segmentation performance. We believe that the Transformer architecture proposed in this work holds important reference value for future research in this field.

REFERENCES

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

[2] 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*.

[3] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

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

[6] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jul. 2006, pp. 2169–2178.

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

[8] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 1458–1465.

[9] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[11] Z. Jin, B. Liu, Q. Chu, and N. Yu, "ISNet: Integrate image-level and semantic-level context for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7169–7178.

[12] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[14] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1743–1751.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[16] X. Li, X. Li, L. Zhang, G. Cheng, J. Shi, Z. Lin, S. Tan, and Y. Tong, "Improving semantic segmentation via decoupled body and edge supervision," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 435–452.

[17] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5168–5177.

[18] M. A. Islam, M. Rochan, N. D. B. Bruce, and Y. Wang, "Gated feedback refinement network for dense image labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4877–4885.

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

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

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

[22] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "OCNet: Object context network for scene parsing," 2018, *arXiv:1809.00916*.

[23] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, and N. Sang, "Context prior for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12413–12422.

[24] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 173–190.

[25] R. P. K. Poudel, U. Bonde, S. Liwicki, and C. Zach, "ContextNet: Exploring context and detail for semantic segmentation in real-time," 2018, *arXiv:1805.04554*.

[26] Q. Song, K. Mei, and R. Huang, "AttaNet: Attention-augmented network for fast and accurate scene parsing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2567–2575.

[27] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," 2020, *arXiv:2005.10821*.

[28] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, and L. J. Latecki, "LEDNet: A lightweight encoder–decoder network for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1860–1864.

[29] E. Xie, W. Wang, W. Wang, P. Sun, H. Xu, D. Liang, and P. Luo, "Segmenting transparent object in the wild with transformer," 2021, *arXiv:2101.08461*.

[30] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1280–1289.

[31] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[32] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[33] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3194–3203.

[34] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 1140–1156.

[35] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17864–17875.

[36] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6392–6401.

[37] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.

[38] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "UniFormer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, pp. 12581–12600, 2023.
[39] J. Jiao, Y.-M. Tang, K.-Y. Lin, Y. Gao, J. Ma, Y. Wang, and W.-S. Zheng, "DilateFormer: Multi-scale dilated transformer for visual recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 8906–8919, 2023.
[40] W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, X. He, and W. Liu, "CrossFormer++: A versatile vision transformer hinging on cross-scale attention," 2023, *arXiv:2303.06908*.
[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[42] F. Zhang, Y. Chen, Z. Li, Z. Hong, J. Liu, F. Ma, J. Han, and E. Ding, "ACFNet: Attentional class feature network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6797–6806.
[43] X. Ding, C. Shen, Z. Che, T. Zeng, and Y. Peng, "SCARF: A semantic constrained attention refinement network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3002–3011.
[44] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9166–9175.
[45] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.
[46] Z. Jin, X. Hu, L. Zhu, L. Song, L. Yuan, and L. Yu, "IDRNet: Intervention-driven relation network for semantic segmentation," 2023, *arXiv:2310.10755*.
[47] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, 1977.
[48] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.
[49] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[50] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.
[51] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4182–4192.
[52] X. Zhao, R. Vemulapalli, P. A. Mansfield, B. Gong, B. Green, L. Shapira, and Y. Wu, "Contrastive learning for label efficient semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10603–10613.
[53] L. Wei, L. Xie, J. He, X. Zhang, and Q. Tian, "Can semantic labels assist self-supervised visual representation learning?" in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 2642–2650.
[54] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3023–3032.
[55] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12546–12558.
[56] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16679–16688.
[57] Z. Jin, T. Gong, D. Yu, Q. Chu, J. Wang, C. Wang, and J. Shao, "Mining contextual information beyond image for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7211–7221.
[58] H. Hu, J. Cui, and L. Wang, "Region-aware contrastive learning for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16271–16281.
[59] Z. Jin, D. Yu, Z. Yuan, and L. Yu, "MCIBI++: Soft mining contextual information beyond image for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5988–6005, May 2023.
[60] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.
[61] X. Wang, H. Zhang, W. Huang, and M. R. Scott, "Cross-batch memory for embedding learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6387–6396.
[62] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," in *Proc. Comput. Vis. and Pattern Recognit.*, Mar. 2020.
[63] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.
[64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[65] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.
[66] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 267–283.
[67] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[68] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[69] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu, "Disentangled non-local neural networks," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 191–207.
[70] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.
[71] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," 2019, *arXiv:1909.11065*.
[72] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19529–19539.
[73] H. Yan, C. Zhang, and M. Wu, "Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention," 2022, *arXiv:2201.01615*.
[74] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10809–10819.

PENG YANG was born in 2000. He received the B.S. degree in software engineering from Jiangxi Normal University, in 2022. He is currently pursuing the M.S. degree with Qinghai University, China. His research interests include computer vision and grassland degradation assessment. He is engaged in researching semantic segmentation.

CHUNMEI LI was born in 1972. She is currently an Associate Professor and a Master's Supervisor with the Department of Computer Science, Qinghai University. She has led the National-Level Natural Science Foundation Project and three provincial-level projects. She has authored over 40 teaching and research articles. Her research interests include computer vision, deep learning, and data mining. She is a Professional Member of China Computer Federation. She holds the position of a Supervisor with Qinghai Province Computer Society.

CHENGWU FANG was born in 1999. He received the B.S. degree in software engineering from Qingdao University, in 2022. He is currently pursuing the M.S. degree with Qinghai University, China. His research interests include computer vision and calculating grassland carrying capacity. He is dedicated to researching object detection.

SHASHA KONG was born in 2000. She received the B.S. degree in data science and big data technology from Hubei University of Economics, in 2022. She is currently pursuing the M.S. degree with Qinghai University. Her research interests include speech recognition and knowledge graphs. She is focused on studying low-resource speech recognition.

YUNPENG JIN was born in 1998. He received the B.S. degree in software engineering from Nantong University, in 2021. He is currently pursuing the M.S. degree with Qinghai University, China. His research interests include computer vision and the identification of toxic weeds. He is engaged in researching semantic segmentation.

KAI LI was born in 1999. He received the B.S. degree in software engineering from Henan University, in 2021. He is currently pursuing the M.S. degree with Qinghai University, China. His research interests include computer vision and the identification of toxic weeds. He is engaged in researching instance segmentation and semantic segmentation.

HAIYANG LI was born in 1999. He received the B.S. degree in computer science and technology from Qinghai University, China, in 2020, where he is currently pursuing the M.S. degree. His research interests include computer vision and the identification of toxic weeds. He is engaged in researching and semantic segmentation.

XIANGJIE HUANG was born in 1998. He received the B.S. degree in civil engineering from Hunan University of Arts and Science, in 2020. He is currently pursuing the M.S. degree with Qinghai University, China. His research interests include computer vision and calculating grassland carrying capacity. He is dedicated to researching object detection.

YAOSHENG HAN was born in 1999. He received the B.S. degree in computer science and technology from the North University of China, in 2022. He is currently pursuing the M.S. degree with Qinghai University, China. His research interests include computer vision and calculating grassland carrying capacity. He is dedicated to researching object detection.