electronics

Article

E-HRNet: Enhanced Semantic Segmentation Using Squeeze and Excitation

Jin-Seong Kim [1], Sung-Wook Park [1], Jun-Yeong Kim [1], Jun Park [1], Jun-Ho Huh [2,3], Se-Hoon Jung [4,*] and Chun-Bo Sim [1,*]

[1] Interdisciplinary Program IT-Bio Convergence System, Sunchon National University, 255 Jungang-ro, Suncheon-city 57922, Jeollanam-do, Republic of Korea; 1235037@s.scnu.ac.kr (J.-S.K.); 411050@scnu.ac.kr (S.-W.P.); jy_kim@scnu.ac.kr (J.-Y.K.); j_park@scnu.ac.kr (J.P.)
[2] Department of Data Science, (National) Korea Maritime and Ocean University, Busan 49112, Gyeongsang-do, Republic of Korea; 72networks@kmou.ac.kr
[3] Interdisciplinary Major of Ocean Renewable Energy Engineering, (National) Korea Maritime and Ocean University, Busan 49112, Gyeongsang-do, Republic of Korea
[4] Department of Computer Engineering, Sunchon National University, 255 Jungang-ro, Suncheon-city 57922, Jeoolanam-do, Republic of Korea
* Correspondence: shjung@scnu.ac.kr (S.-H.J.); cbsim@scnu.ac.kr (C.-B.S.)

Abstract: In the field of computer vision, convolutional neural network (CNN)-based models have demonstrated high accuracy and good generalization performance. However, in semantic segmentation, CNN-based models have a problem—the spatial and global context information is lost owing to a decrease in resolution during feature extraction. High-resolution networks (HRNets) can resolve this problem by keeping high-resolution processing layers parallel. However, information loss still occurs. Therefore, in this study, we propose an HRNet combined with an attention module to address the issue of information loss. The attention module is strategically placed immediately after each convolution to alleviate information loss by emphasizing the information retained at each stage. To achieve this, we employed a squeeze-and-excitation (SE) block as the attention module, which can seamlessly integrate into any model and enhance the performance without imposing significant parameter increases. It emphasizes the spatial and global context information by compressing and recalibrating features through global average pooling (GAP). A performance comparison between the existing HRNet model and the proposed model using various datasets show that the mean class-wise intersection over union (mIoU) and mean pixel accuracy (MeanACC) improved with the proposed model, however, there was a small increase in the number of parameters. With cityscapes dataset, MeanACC decreased by 0.1% with the proposed model compared to the baseline model, but mIoU increased by 0.5%. With the LIP dataset, the MeanACC and mIoU increased by 0.3% and 0.4%, respectively. The mIoU also decreased by 0.1% with the PASCAL Context dataset, whereas the MeanACC increased by 0.7%. Overall, the proposed model showed improved performance compared to the existing model.

Keywords: deep learning; computer vision; CNN; attention

check for updates

Citation: Kim, J.-S.; Park, S.-W.; Kim, J.-Y.; Park, J.; Huh, J.-H.; Jung, S.-H.; Sim, C.-B. E-HRNet: Enhanced Semantic Segmentation Using Squeeze and Excitation. *Electronics* **2023**, *12*, 3619. https://doi.org/10.3390/electronics12173619

Academic Editor: Chiman Kwan

Received: 5 July 2023
Revised: 18 August 2023
Accepted: 25 August 2023
Published: 27 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Studies on convolutional neural networks (CNNs) in the field of computer vision have demonstrated the high accuracy and good generalization performance of CNNs for various tasks and open datasets. CNNs exhibit excellent generalization performances in several computer vision problems, including image classification, semantic segmentation, object detection, and human pose estimation. However, capturing complex relationships between channels or pixel positions in space is challenging because of insufficient feature extraction for global context and spatial information.

To solve this problem, one study combined a CNN and an attention module with a residual connection architecture by emphasizing global context information, leading to

improved performance [1]. In addition, CNN-based models undergo a downsampling process, in which a high-resolution image is transformed into a low-resolution image during the feature extraction process. Downsampling causes a loss of global context and spatial information. This problem is a major cause of performance degradation in semantic segmentation in which image features must be restored to the original image. A high-resolution network (HRNet) model was thus proposed to solve this resolution reduction problem [2]. Among the various models capable of performing semantic segmentation, HRNet was originally developed for human pose estimation and achieved superior performance. However, because the process of downsampling and upsampling human pose estimation is similar to semantic segmentation, it exhibits high performance in semantic segmentation.

HRNet is a parallelized model that maintains multi-scale resolution to efficiently learn global context and spatial information. The high- and low-resolution convolutional branches were kept parallel to extract and share features. This method can learn rich information and partially solve the problem of information loss by undergoing a convergence process in which branches share characteristics. Since then, the HRNet has undergone many improvements in terms of accuracy and speed in human pose estimation, classification, and object detection [3–5].

In this study, an attention module was added to every convolution block to improve the performance by reducing the loss of global context and spatial information during feature extraction. The method of adding an attention module enables more accurate semantic segmentation by more efficient information fusion between the high-resolution and low-resolution branches. The squeeze-and-excitation (SE) Block [6], published simultaneously as [1], was used as an attention module. The SE Block has been used in various fields, such as classification and semantic segmentation because it can be easily added to any model [7,8]. The SE Block can efficiently recalibrate the features by applying squeeze and excitation techniques. The SE Block was verified as effective in reducing errors while minimizing the increase in the number of parameters [6]. The technique of increasing accuracy while minimizing the increase in the number of parameters is easy in areas similar to autonomous driving, where lightweight and accuracy are crucial. To evaluate the performance of the proposed method, the Cityscapes [9], LIP [10], and PASCAL Context [11] datasets were used. These datasets have been widely used for semantic segmentation. The baseline model used was an HRNet model pretrained with ImageNet [12].

Experiments were then performed to compare the performances depending on the presence or absence of the attention module. The contribution of this study is that it proposes a method to improve performance while suppressing the increase in the number of parameters as much as possible by using an attention module. The global context information included in each channel was recalibrated to improve pixel segmentation and class classification performance. The effect of global context information on semantic segmentation was experimentally confirmed.

Section 2 describes the HRNet and the attention modules. Section 3 explains the HRNet based on the attention module. In Section 4, we compare the improvement in the accuracy of the proposed method with that of the existing method through experiments using various datasets. Finally, Section 5 presents the conclusions of this study.

2. Related Works

2.1. Semantic Segmentation

Upsampling is the task of restoring features extracted from semantic segmentation of the original image and classifying each pixel into a class. Spatial information of resolution is important for classifying pixel units. Therefore, the FCN [13], a model with improved performance, was first proposed by reducing the pixel location information loss and configuring all layers with convolution layers. Subsequently, SegNet [14] and UNet that [15] use an encoder–decoder structure were proposed. In addition, noting that the spatial information of different resolutions is important for performance improvement, Deeplabv3 [16] and PSPNet [17], using Atrous Convolution and ASPP, were proposed.

RefineNet [18], which combines feature maps of various resolutions using a refine block, showed improved performance because a high resolution provides rich spatial information. As another method using multi-resolution, research on transmitting and exchanging low-resolution information with a residual connection structure [19] has been published; in addition, several other studies have been published, such as combining multi-scale pyramid representations [20,21].

Existing CNN-based models have a pyramid structure in which the size of the convolution feature maps decreases as the depth increases [1,13]. However, HRNet maintains feature maps with a smaller size than the high-resolution branched branches while maintaining high-resolution feature maps in parallel. A new feature map is generated by merging the feature maps in branches with different resolutions. This method obtains richer information by exchanging information from different resolutions. Feature maps containing information at multiple resolutions allow high-quality upsampling, resulting in more accurate segmentation. The structure of HRNet is shown in Figure 1.

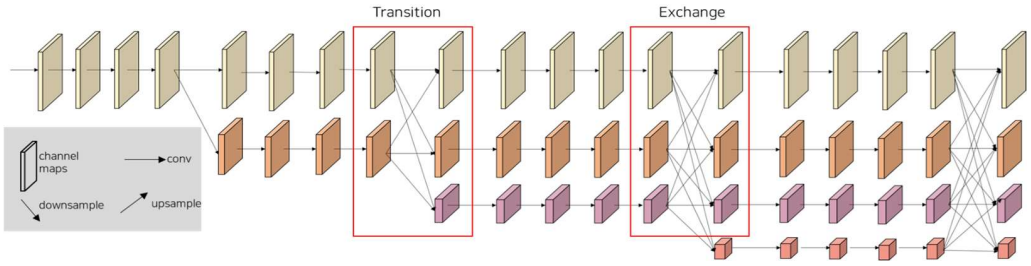

Figure 1. Example of HRNet architecture. There are four stages. The architecture consists of high-resolution convolutions with a transition unit and an exchange unit.

The HRNet consists of four stages. The first stage is a bottleneck structure with 64 channels like ResNet-50 [22]. The second, third, and fourth stages consist of transition and exchange units. The transition unit fuses feature maps of different branches to generate a new feature map. The exchange unit exchanges information on the feature maps of different branches. The overall structure of the HRNet is one in which the unit exchange is repeated four times after the unit transition and convolution. In HRNetV2-W18, W30, and W48, W is the number of channels with the highest-resolution convolution. The size of the convolution was 3×3, and the size of the first input image feature map was different for each dataset, as explained in detail in Section 3. When generating a new feature map using feature maps of different resolutions, downsampling or upsampling was performed to match the resolutions. For downsampling, a stride 2 convolution was performed when the resolution size was reduced by $1/2$. Stride 2 convolution was performed again when the resolution was reduced by $1/4$. When upsampling by $2\times$ or $4\times$, the maximum value was used, and upsampling was performed in one step without intermediate steps. The number of channels in a parallel branch doubled when the resolution was reduced by half. If the original resolution image size was 32 channels, $1/2$ had 64 channels, $1/4$ had 128 channels, and $1/8$ had 256 channels.

Despite the aforementioned efforts, HRNet continues to experience information loss during the feature extraction process, owing to the inherent characteristics of convolution-based models. These factors contribute to a decrease in resolution, which is a significant concern in semantic segmentation because they adversely affect the segmentation accuracy. To address this issue, this study proposes inserting an attention module immediately after each convolution to mitigate information loss and alleviate the resolution reduction problem.

2.2. Attention Module

The basic idea of attention in natural language processing is that the encoder refers to the entire input sentence once again at each timestamp at which the decoder predicts the output word. Rather than referencing the entire input sentence in equal proportions, we refer to the part of the input word related to the word to be predicted at that time. The

basic concept of the Attention technique is a dictionary data type consisting of key values applied to many fields of computer engineering; this is shown in (1):

$$Attention(Q, K, V) = AttentionValue \quad (1)$$

In Equation (1), Attention calculates the similarity between a given query and a key. The output similarity is then multiplied by each value mapped to a key. The sum of all values that reflected similarity was then obtained. Self-attention is an expanded form of attention [23]. The query, key, and value of existing attention are different values, whereas those of self-attention are the same. Self-attention recalibrates the channel by passing an input query and key through a 1×1 convolution. Subsequently, keys are transposed and multiplied to obtain the cosine similarity. The attention map is then outputted using softmax. Finally, a self-attention feature map is generated by multiplying the values that have undergone 1×1 convolution. Self-attention has expanded to various fields such as reinforcement learning, image captioning, and natural language processing [24,25]. It is also used to emphasize the relationship between context information and pixels [26]. Attention mechanisms have been used in many computer vision tasks to address the limitations of standard convolutions [27–30]. In some computer vision tasks, multi-head self-attention with a sufficient number of heads produced notable results in a study by Cordonnier et al. [31]. In addition, a standalone self-attention model in which all layers are composed of self-attention achieved excellent performance [32].

The attention module is mainly used in tasks where context information is important, such as visual question answering (VQA), image captioning, and scene character recognition [33,34]. However, when the concept of attention was expanded to self-attention, it began to be used in CNN. SENet uses an attention mechanism that captures the interactions between channels such that each channel can be assigned a different weight. This model can improve performance through different weightings per channel. A channel with a large weight is interpreted as an important feature, whereas a channel with a small weight is interpreted as containing less important information. Different weights were assigned to different channels of the feature map and multiplied. The module was used to assign different weights to each channel. The SE Block consists of two stages: squeeze and excitation. In the squeeze stage, global average pooling (GAP) is performed to make each channel of the image one-dimensional. In the recalibration stage, the squeezed vector passes through two fully connected layers: a rectified linear unit (ReLU) and a sigmoid. Finally, the flattened vector is multiplied by the image that has passed through a 1×1 convolution and the weight, which is squeezed information, to emphasize the important information. Figure 2 illustrates an SE Block [6]. Figure 3 shows the detailed architecture of the SE Block inserted into ResNet.

The hyperparameter in the SE Block is the reduction ratio, which reduces and increases the number of nodes in the fully connected layer and ReLU parts. As the reduction ratio decreased, the number of parameters increased. As the number of reduction ratio increased, the number of parameters decreased. That is, it is a hyperparameter related to changes in capacity and computational cost.

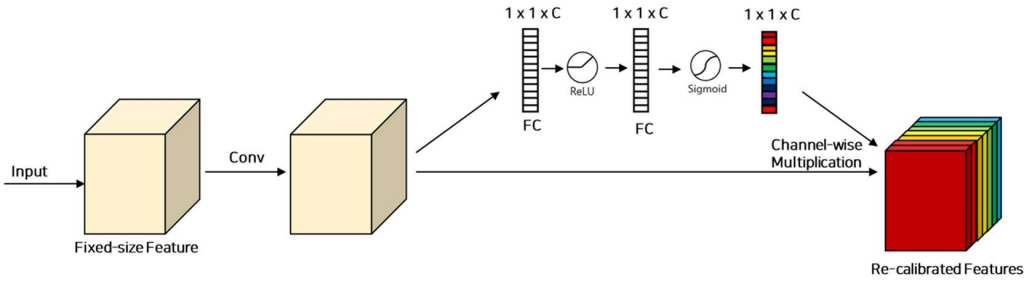

Figure 2. Details of the SE Block architecture.

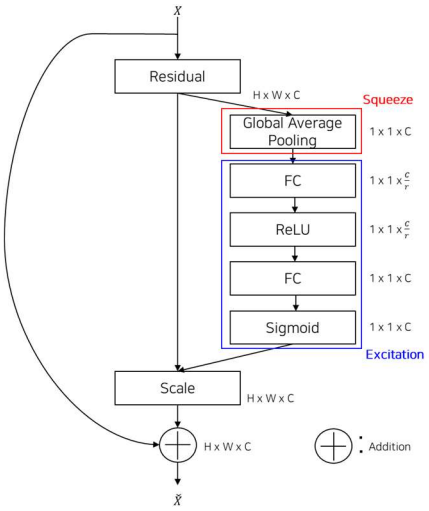

Figure 3. Details of the SE Block architecture added into the ResNet.

3. Proposed Method

Section 3 explains the structure of the proposed model and how it is combined with the attention module. The proposed method focuses on improving the upsampling performance in semantic segmentation by adding an attention module to the HRNet. When the attention module was added, an increase in the number of parameters was required. Thus, an SE Block with a low computational load was used. Figure 4 presents an overview of E-HRNet, where an SE Block, which is an attention module, is inserted at the end of each convolution block. In addition, all convolution blocks, except for the bottleneck block, have the same structure.

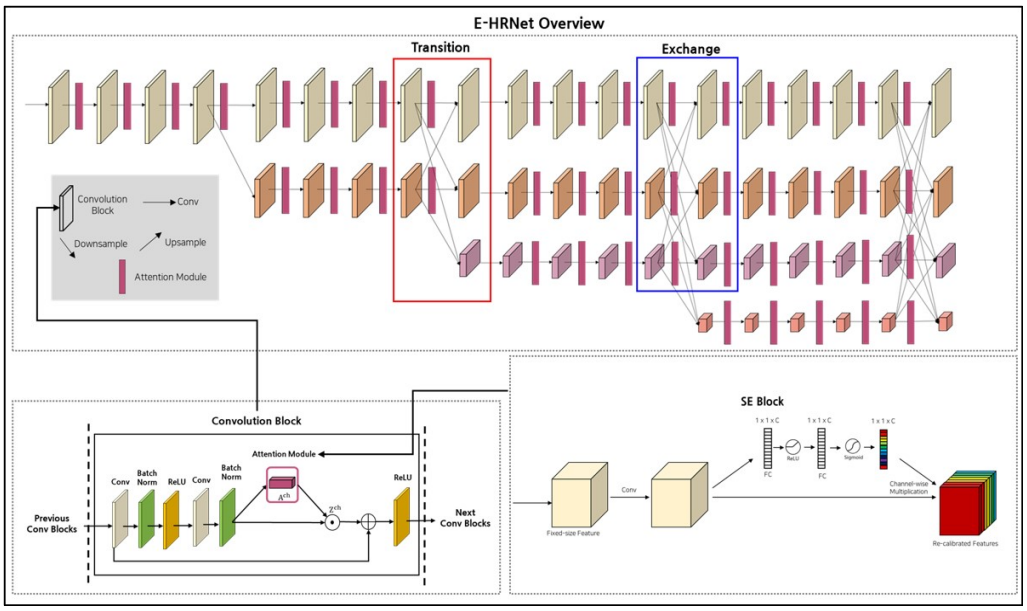

Figure 4. E-HRNet overview.

3.1. Details of HRNet Architecture

The detailed architecture of an existing HRNet is shown in Figure 1. The baseline model used was HRNetV2-W48. In HRNetV2-W48, the highest-resolution branch had 48 convolutional channels with resolutions of 1024 × 512 for Cityscapes, 473 × 473 for LIP, and 480 × 480 for PASCAL Context. Feature maps with 1/2, 1/4, and 1/8 resolutions were used only to exchange information at different resolutions. Therefore, attention modules can be easily added to all resolution branches per convolution block unit. There are four

stages. Unit transitions and exchanges were repeated to form the 2nd, 3rd and 4th stages. The unit transition and exchange consist of a multi-resolution group convolution and multi-resolution convolution, as shown in Figure 5a,b. Figure 5a shows a simple extension of the convolution with multiple resolutions. Multi-resolution group convolution divides an input channel into subsets of several channels and performs each convolution separately for different spatial resolutions.

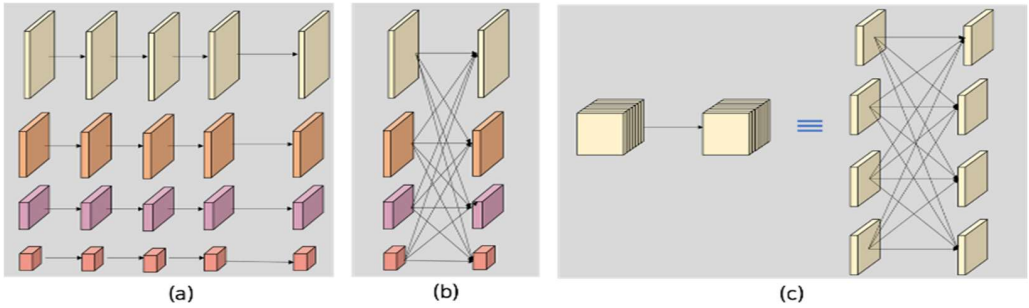

Figure 5. Multi-resolution block: (**a**) Multi-resolution group convolution and (**b**) multi-resolution convolution. (**c**) A normal convolution (left) is equivalent to a fully connected multi-branch convolution (right).

Figure 5b illustrates the multi-resolution convolution that exchanges and fuses features extracted from parallel branches with information from different resolutions. Multi-resolution convolution is similar to the multibranch full-connection method of a general convolution, as shown in Figure 5c. A normal convolution can be divided into several small convolutions. The input channels are divided into several subsets. The output channels are also divided into several subsets. The input and output subsets were connected in a fully connected manner. Each connection has a normal convolution. Each subset of the output channels was the sum of the convolution outputs for each subset of the input channels.

The difference from normal convolution is that in multi-resolution convolution, each subset of the channels has a different resolution. In addition, to reduce the resolution through downsampling, 2-stride 3×3 convolution was used to connect the input and output channels. Bilinear upsampling was performed while upsampling the downsampled feature map.

3.2. E-HRNet Architecture

A total of 71 ReLU layers were added by adding SE blocks to the existing HRNet model, consisting of 307 convolution layers, 306 batch normalizations, 269 ReLU layers, 4 bottleneck layers, 104 basic blocks, and 8 high-resolution modules. A total of 108 GAP layers were added for compression with one feature. Additionally, 206 fully connected layers, 108 sigmoid layers, and 108 SE Blocks were added. The existing number of parameters increased by 0.4 M from 65.8 M based on the Cityscapes dataset to 66.2 M, an increase of less than 1%. Giga floating point operations per second (GFLOPs) slightly increased to 0.0004.

Figure 6 illustrates the E-HRNet. The existing HRNet efficiently extracts features by fusing the features between parallel branches. However, information loss still occurred during downsampling. In the proposed model architecture, global context information within the object domain can be recalibrated by adding an attention module at the end of every convolution block to reduce information loss.

The SE Block is used in this context because of its ease of integration into any model and its ability to address the issue of information loss by recalibrating features with a minimal parameter increase. Specifically, the SE-Block-based attention module passes the input features through the GAP and squeezes each channel into one feature, that is, a scalar value. Subsequently, as shown in Figure 2, the importance of the feature squeezed through the fully connected layer and sigmoid is calculated as a probability value between 0 and 1 for each channel. The calculated importance is normalized as a weight and multiplied by an image that has undergone a 1×1 convolution to readjust the feature value.

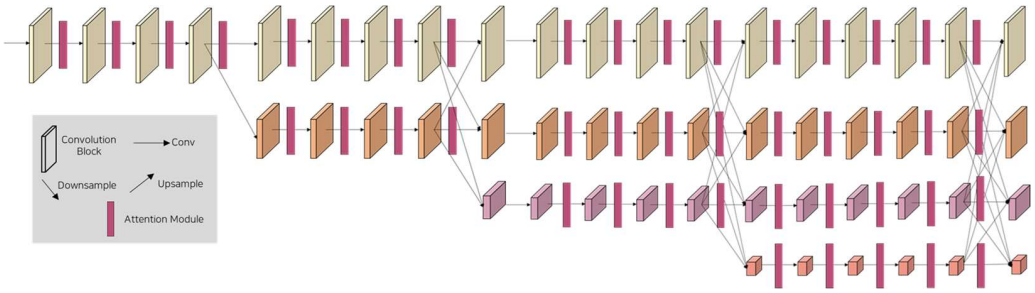

Figure 6. Details architecture of E-HRNet.

In this study, the SE block was selected to recalibrate the feature value of the global context information for each channel within a range where the number of parameters did not increase excessively. In addition, by adding an SE block to all the convolution processes, information can be extracted uniformly at both high and low resolutions.

3.3. Instantiation

To check the effect of the attention module on the segmentation accuracy, this study was implemented in a manner similar to that of HRNetV2. The network starts with a branch of a 2-stride, 3×3 convolution that reduces the feature-map resolution to 1/4. Stage 1 consists of 4 convolutional blocks, each of which comprise a 64-channel bottleneck. Subsequently, a 3×3 convolution is continued one-by-one to reduce the width of the feature map to C. C means 48 of HRNetV2-W48. Stages 2, 3 and 4 include 1, 4 and 3 multi-resolution blocks, respectively. The widths of the four resolution convolutions were double those of C, 2C, 4C and 8C. Each branch of the multi-resolution group convolution contains 4 convolution blocks. Each resolution contains two 3×3 convolutions. In Figure 7, the middle box enlarges the input size 4 times through bilinear upsampling of the feature map extracted from the four resolution branches.

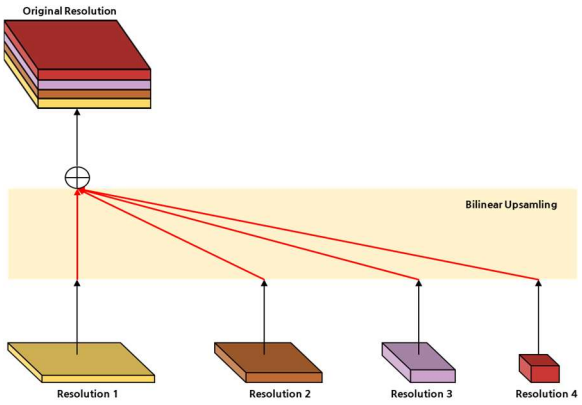

Figure 7. Concatenating the representation from all resolutions for semantic segmentation.

The outputs of all resolutions were then mixed using a 1×1 convolution to generate a 15C dimensional representation. Finally, a segmentation map with the original resolution is generated. Based on this architecture, a SE Block is added to every convolution block unit. Algorithm 1 shows the pseudocode of the E-HRNet. The code is written in Python. The deep learning library used was PyTorch. The SE Block was inserted at the end of the Basic Block that undergoes two convolutions. The SE Block learns the nonlinearity between channels through the fully connected layer and ReLU after being squeezed into a scalar value through adaptive average pooling. Finally, important information is emphasized through the sigmoid, and other information is zeroed out.

Pseudocode of E-HRNet (variables N, C, H, W denote sample number in a mini-batch, feature channels, image height, and image width, respectively) as Algorithm 1.

Algorithm 1: Attention Module Pseudocode, Torch-like

```
# input features with shape [N, C, H, W] = x
# N: batch size, C: channels, H: height, W: width
# SEBlock = Attention module
# reduction ratio = 16
def SEBlock (x)
    squeeze = torch.nn.AdaptiveAvgPool2d(x)
    excitation = torch.nn.Fully_connected(squeeze, out_channel/ratio)
    excitation = torch.nn.ReLU(excitation)
    excitation = torch.nn.Fully_connected(excitation/ratio, out_channel)
  excitation = torch.nn.Sigmoid(excitation)
  scale = x * excitation
  return scale
def Convolution Basicblock (x)
  residual = x
  out = torch.nn.Convolution(x)
  out = torch.nn.BatchNorm2d(out)
  out = torch.nn.ReLU(out)
  out = torch.nn.Convolution(out)
  out = torch.nn.BatchNorm2d(out)
  out = torch.nn.SEBlock(out)
  out += residual
  out = ReLU(out)
  return out
```

4. Experiments

Semantic segmentation is the task of assigning a label to each pixel. In this study, to verify the effect of the attention module on segmentation accuracy in semantic segmentation, the parameters, datasets, and training rules were set the same as those of the existing HRNetV2, except for the attention module. Cityscapes [9], a representative scene-parsing dataset, and LIP [10], a human-parsing dataset, were used. In addition, PASCAL Context [11], a general image dataset, is used. PASCAL Context extends the 2010 PASCAL-VOC. The HRNet-based models were pre-trained using ImageNet. Tables 1 and 2 list the hardware specifications and software versions used for development and testing.

Table 1. Hardware specifications.

Hardware	Specifications
CPU	Intel Core i7 7700k
Graphics Card	NVIDIA Geforce RTX 3090 24 GB
RAM	Samsung DDR4 32 GB
SSD	Samsung 850 Pro 512 GB

Table 2. Software version.

Software	Version
Operating System	Ubuntu 20.04.1 LTS
CUDA	11.2.67
cuDNN	8.1.0
Programing Language	Python 3.8.10
Pytorch	1.8.1

4.1. Cityscapes

The Cityscapes dataset consists of 5000 high-resolution and finely annotated scene images. These finely annotated images were divided into 2975 training, 500 validation, and

1525 testing images. There were 30 classes in total. In this study, 19 classes, excluding the empty and sparse categories, were used for the training and evaluation of efficient learning.

The batch size was set to six. The same training protocol as HRNetV2 [17,35] was used, except that a single GPU was used instead of multiple GPUs. An image with a resolution of 1024 × 2048 pixels was randomly cropped to 512 × 1024 pixels. Data were augmented using random scaling and random horizontal flip in the range of 0.5–2. The optimizer used stochastic gradient descent (SGD). The initial learning rate was 0.01, and the momentum was set to 0.9. The dampening was set to 0, and the weight decay was 0.0005. The nesterov was set to false, and the maximize was set to false. The foreach was set to none, and the differentiable was set to false. The learning rate schedule used a polylearning rate policy with a power of 0.9. The reduction ratio used for the SE Block was 16. The performance of the model was evaluated using a single-scale non-flipped dataset.

Table 3 presents a comparison of the number of parameters, GFLOPs, mIoU, and MeanACC of the HRNet and the proposed models with those of the Cityscapes validation set. The number of parameters increased by 0.4 M, and GFLOPs increased by 0.001 compared to HRNetV2-W48, which became the baseline model. MeanACC, the average accuracy of the pixel, decreased by 0.1%. However, the mIoU increased by 0.5% owing to the improved performance in segmenting the regions of objects corresponding to pixel classes.

Table 3. Results of HRNetV2-based semantic segmentation model with the Cityscapes validation set (single scale and no flipping). GFLOPs is calculated on RTX 3090 with input size of 1024 × 2048. The proposed method backbone is HRNetV2-W48.

Method	# Params [M]	GFLOPs [G]	mIoU [%]	MeanACC [%]
HRNetV2-W18	1.5	7.774	63.6	72.9
HRNetV2-W48	65.8	174.043	79.4	**87.1**
Proposed Method	66.2	174.044	**79.9**	87.0

Table 4 shows the mIoU comparison of existing models and the proposed model with the Cityscapes validation set. It achieved 4.2% higher performance than UNet++ [23], a relatively lightweight model. It also showed 1.2% and 0.1% higher performance than DeepLabv3 [16] and DeepLabv3+ [20] of similar weight, respectively.

Table 4. Results of semantic segmentation comparison with other models with Cityscapes dataset.

Method	Backbone	mIoU [%]
UNet++ [19]	ResNet-101	75.5
DeepLabv3 [16]	Dilated-ResNet-101	78.5
DeepLabv3+ [20]	Dilated-Xception-71	79.6
Proposed Method	HRNetV2-W48	**79.9**

Table 5 compares the mIoU, instance intersection over union (iIoU) classes, IoU categories, and iIoU categories between HRNet and the proposed model on the Cityscapes test set. While IoU evaluates how well a model segments an entire class, iIoU is a measure that determines how efficiently a model distinguishes individual objects within the same class by evaluating segmentation accuracy at the instance level. Utilizing both metrics simultaneously provides a comprehensive understanding of how well the model segments individual objects. The difference between class and category lies in the scope of consideration. For example, a class encompasses all of the individual classes, while a category groups similar classes into broader categories. 'Bus', 'Car', and 'Truck' are all grouped under the 'Vehicles' category. Overall, the proposed model demonstrated strong performance across all metrics, with a particularly noticeable improvement in iIoU. This suggests that the performance of the proposed model is more adept at segmenting individual objects than entire classes.

Table 5. Results of HRNetV2-based semantic segmentation model with the Cityscapes test set (single scale and no flipping). GFLOPs is calculated on RTX 3090 with input size of 1024 × 2048. The proposed method backbone is HRNetV2-W48.

Method	mIoU	iIoU Class	IoU Category	iIoU Category
HRNetV2-W18	63.9	38.0	85.7	69.3
HRNetV2-W48	77.2	55.0	91.1	79.1
Proposed Method	**77.5**	**55.5**	**91.2**	**79.9**

Table 6 shows the results of class-wise IoU on the Cityscapes test set. The proposed model demonstrated similar segmentation performance for large objects like 'sky' and 'buildings' compared to existing models, but excelled in segmenting relatively small and complex objects such as 'traffic lights', 'traffic signs', and 'fences'. The results of Tables 5 and 6 show that by emphasizing channel information, the characteristics of objects that belong to the same class or are small in size and easily confused can be mitigated.

Table 6. Class-wise results of HRNetV2-based semantic segmentation model with the Cityscapes test set.

Class	HRNetV2-W18	HRNetV2-W48	Proposed Method
Road	97.4	**98.6**	**98.6**
Sidewalk	78.4	**86.6**	**86.6**
Building	88.8	93.0	**93.1**
Wall	33.8	**56.7**	53.8
Fence	40.9	59.0	**60.0**
Pole	51.8	68.9	**69.5**
Traffic light	56.2	76.5	**77.0**
Traffic sign	64.1	**79.3**	79.2
Vegetation	91.8	**93.7**	**93.7**
Terrain	69.7	72.2	**72.9**
Sky	93.4	95.5	95.5
Person	76.5	86.8	**86.9**
Rider	51.7	70.7	**71.5**
Car	92.4	**95.8**	95.7
Truck	34.8	60.1	**63.0**
Bus	46.7	67.8	**67.2**
Train	36.0	62.1	**63.3**
Motorcycle	46.6	**68.9**	68.5
bicycle	63.1	75.4	**75.9**

Figure 8 shows the semantic segmentation prediction maps of the model trained on the Cityscapes dataset. HRNetV2-W18 exhibited relatively more misclassifications due to unclear boundaries between objects. On the other hand, HRNetV2-W48 demonstrated clearer boundaries between objects and fewer misclassifications compared to HRNetV2-W18. Our proposed model shares similarities with HRNetV2-W48; however, it displayed superior capabilities in accurately segmenting small and intricate objects that are easily overlooked. From these results, we can infer that the number of channels in convolution plays a significant role in segmentation performance. Additionally, we observed that information emphasis through attention modules has a meaningful impact on accurately segmenting intricate objects.

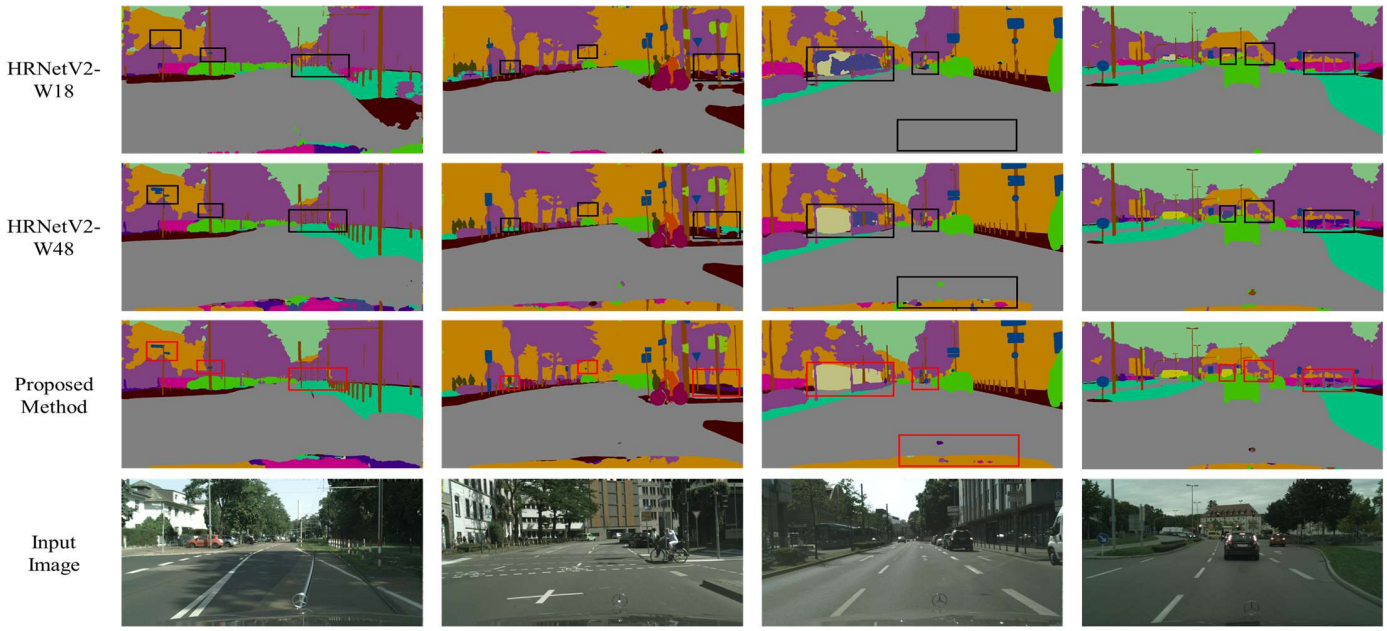

Figure 8. Comparison of proposed method and HRNetV2 for semantic segmentation prediction maps. The black boxes represent the baseline model, while the red boxes represent the proposed method.

4.2. LIP

The LIP dataset consists of 50,462 carefully annotated images of human body parts. The dataset was divided into 30,462 images for training and 10,000 images for validation. It consisted of 19 classes related to human parts and one background class.

The image was resized to 473 × 473 according to the training and test settings in [36]. The performance was evaluated as the average of the segmentation maps of the original and flipped images. The settings for the data augmentation and learning rate schedule and reduction ratio of SE Block were the same as those for Cityscapes. The training settings are the same as those in [26]. The optimizer used SGD. The initial learning rate was set to 0.01, and the momentum was set to 0.9. The dampening was set to 0, and the weight decay was set to 0.0005. The nesterov was set to false, and the maximize was set to false. The foreach was set to none, and the differentiable was set to false. The batch size was 8. The performance of the model was evaluated using a single-scale non-flipped dataset.

Table 7 shows a comparison of parameters, GFLOPs, mIoU, and MeanACC indicators of the existing HRNet model and the proposed model with the LIP validation set. The number of parameters increased by 0.4 M, and GFLOPs increased by 0.0004 compared to HRNetV2-W48, the baseline model. Both the object region and pixel class classification accuracy showed improvement, with MeanACC increased by 0.3% and mIoU increased by 0.4%.

Table 7. Results ofHRNetV2-based semantic segmentation model results with LIP validation set (single scale and no flipping). GFLOPs is calculated on the RTX 3090 and input size 473 × 473. The proposed method backbone is HRNetV2-W48.

Method	# Params [M]	GFLOPs [G]	mIoU [%]	MeanACC [%]
HRNetV2-W18	1.5	3.3798	13.4	19.1
HRNetV2-W48	65.8	75.9817	52.6	65.4
Proposed Method	66.2	75.9821	**53.0**	**65.7**

Table 8 shows the mIoU comparison of several models and the proposed model with the LIP validation set. The proposed model as a whole without additional data achieved the best performance.

Table 8. Results of semantic segmentation comparison with other models with LIP validation set.

Method	Backbone	Extra	mIoU [%]
Attention + SSL [10]	VGG-16	Pose	44.7
DeepLabv3+ [20]	Dilated-ResNet-101	-	44.8
MMAN [37]	Dilated-ResNet-101	-	46.8
SS-NAN [38]	ResNet-101	Pose	47.9
MuLA [39]	Hourglass	Pose	49.3
JPPNet [40]	Dilated-ResNet-101	Pose	51.3
Proposed Method	HRNetV2-W48	N	**53.0**

4.3. PASCAL Context

The PASCAL Context dataset consists of 4998 scene images for training and 5105 test images. This class consisted of 59 object classes and one background class.

The settings for the data augmentation and learning rate schedule and reduction ratio of SE Block were the same as those in Cityscapes. The optimizer used SGD. According to the training strategy in [41,42], the image size was resized to 480 × 480, and the initial learning rate was set to 0.004. The momentum was set to 0.9, and the dampening was set to 0. The weight decay was set to 0.001, the nesterov was set to false. The maximize was set to false, and the foreach was set to none. The differentiable was false. The batch size was 13. The test strategy was based on a previously described procedure [41,42]. The test image was resized to 480 × 480 pixels and input into the model. The output 480 × 480 segmentation map was resized to the original image size. The performance of the model was evaluated using a single-scale non-flipped dataset.

Table 9 shows a comparison of parameters, GFLOPs, mIoU, and MeanACC indicators of the HRNet model and the proposed model with the PASCAL Context test set. The number of parameters increased by 0.5 M, and GFLOPs increased by 0.0004 compared to HRNetV2-W48, the baseline model. The classification accuracy of the pixel class seemed to be improved, as the mIoU fell by 0.1%, whereas MeanACC increased by 0.7%.

Table 9. Results of HRNetV2-based semantic segmentation model results with PASCAL Context test set (single scale and no flipping) (60 classes, single scale, and no flipping). GFLOPs is calculated on RTX 3090 with input size of 480 × 480. The proposed method backbone is HRNetV2-W48.

Method	# Params [M]	GFLOPs [G]	mIoU [%]	MeanACC [%]
HRNetV2-W18	1.5	3.5484	21.1	29.8
HRNetV2-W48	65.8	76.8800	**45.3**	55.0
Proposed Method	66.3	76.8804	45.2	**55.7**

Table 10 shows the mIoU comparison of several models with the proposed model on the PASCAL Context test set. As in Table 9, 60 classes were evaluated, with the proposed model achieving the best performance.

Table 10. Results of semantic segmentation comparison with other models on PASCAL Context test set (60 classes).

Method	Backbone	mIoU [%]
FCN-8s [13]	VGG-16	35.1
BoxSup [43]	-	40.5
HO_CRF [44]	-	41.3
Piecewise [45]	VGG-16	43.3
Proposed Method	HRNetV2-W48	**45.2**

4.4. HRNet-Based Model Performance Comparison Results

In the mIoU comparison, the PASCAL Context dataset, which was designed for segmenting small objects, saw a decrease of 0.1%. On the other hand, the Cityscapes validation

set, intended for scene parsing, improved by 0.5%. In the experiments on the Cityscapes test set, the mIoU, iIoU class, IoU category, and iIoU category improved by 0.3%, 0.5%, 0.1%, and 0.8%, respectively. Additionally, the LIP dataset, designed for body parts parsing, experienced a 0.4% increase. The MeanACC comparison showed that the proposed model exhibited a decrease of 0.1% in Cityscapes, a scene understanding dataset, while showing an increase of 0.3% in LIP and 0.7% in the PASCAL Context dataset, compared with the existing HRNetV2-W48. Therefore, it can be seen that emphasizing global context information can influence the performance of segmenting boundaries between objects in scene understanding tasks; it also affects pixel classification accuracy more when segmenting small objects than relatively larger ones. Additional experiments to provide further evidence are included in Appendix A.

5. Conclusions

In this study, we propose an HRNet model that combines an attention module. The proposed method uses the SE Block as an attention module to reduce the loss of global context information. An attention module is introduced in each convolution block to mitigate the information loss, focusing on the information loss that occurs at every convolution. This approach emphasizes and preserves crucial information throughout a network, thereby effectively addressing the issue of information loss. The performance experiment compared the performances of the existing HRNet model and the proposed model using different learning strategies for each dataset. The number of parameters increased by 0.4 M in Cityscapes and LIP and by 0.5 M in PASCAL Context. The GFLOPs values increased by 0.001 in Cityscapes, and 0.0004 in LIP and Pascal Context. When using the Cityscapes dataset, the pixel class classification accuracy decreased slightly. However, the object-range segmentation performance improved. With the LIP dataset, all the performance metrics showed improvement. With PASCAL Context, the object region segmentation performance decreased slightly, whereas the pixel class classification performance improved. Compared to several other models, the best performance was achieved by the proposed model. Consequently, the attention module improved the performance without excessively increasing the complexity of the model. Furthermore, we observed that emphasizing global contextual information has a significant effect on performance.

In the future, it is expected that higher performance can be obtained by precisely adjusting the optimizer, learning policy, and hyperparameter values suitable for the proposed model. In future research, it will be necessary to develop an optimal attention module for the proposed model. Therefore, it is necessary to develop a new method to combine the extracted features for upsampling.

Author Contributions: Conceptualization, J.-S.K., S.-W.P., J.-Y.K., J.P., J.-H.H., S.-H.J. and C.-B.S.; data curation, J.-S.K. and J.-H.H.; formal analysis, J.-S.K., S.-W.P., J.-Y.K., J.P., S.-H.J. and C.-B.S.; funding acquisition, S.-H.J. and C.-B.S.; investigation, J.-S.K., S.-H.J. and C.-B.S.; methodology, S.-W.P., J.-Y.K., J.P. and J.-H.H.; project administration, S.-H.J. and C.-B.S.; resources, J.-S.K., S.-W.P., J.-Y.K., S.-H.J. and C.-B.S.; software, J.-S.K., S.-W.P., J.-Y.K., J.P., S.-H.J. and C.-B.S.; supervision, C.-B.S.; validation, J.-S.K., S.-W.P., J.-Y.K. and J.P.; visualization, J.-S.K., S.-W.P., J.-Y.K., J.P., J.-H.H. and S.-H.J.; writing—original draft, J.-S.K., S.-W.P., J.-Y.K., J.P., J.-H.H., S.-H.J. and C.-B.S.; writing—review and editing, J.-H.H., S.-H.J. and C.-B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Grand Information Technology Research Center support program (IITP-2023-2020-0-01489), supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP), and by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A3050843) and his research was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through Smart Farm Innovation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) and Rural Development Administration (RDA) and Ministry of Science and ICT (MSIT) (421028-3).

Data Availability Statement: Not applicable.

Acknowledgments: This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Grand Information Technology Research Center support program (IITP-2023-2020-0-01489), supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) and this research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A3050843) and by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through Smart Farm Innovation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA) and Rural Development Administration (RDA) and Ministry of Science and ICT (MSIT) (421028-3).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. More Segmentation

ADE20K

For several reasons, we conducted additional experiments with the ADE20K dataset [46]. First, ADE20K encompasses various scene categories and annotated objects, enabling the evaluation of models even in complex scenarios. Second, in experiments using the Cityscapes dataset, there was little difference in segmentation performance for relatively large objects such as 'buildings' and 'sky' compared to other models, and in experiments using the PASCAL Context dataset, the mIoU actually dropped further, necessitating more research. Lastly, the granularity of the ADE20K annotations is particularly suitable to underscore the strengths we claim in our proposed model.

The ADE20K dataset was used in ImageNet scene parsing challenge 2016. There are 150 classes and diverse scenes with 1038 image-level labels. The dataset was divided into 20,210 training, 4002 validation, and 3352 testing images. Since the test set does not provide labels, the model's performance is evaluated through validation. The batch size was set to nine. The same training protocol as HRNetV2 + OCR [47] was used, except that a single GPU was used instead of multiple GPUs. The image size was resized to 520 $\times$ 520. The settings for the data augmentation and learning rate schedule and reduction ratio of SE Block were the same as those in Cityscapes. The optimizer used SGD. The initial learning rate was 0.02, and the momentum was set to 0.9. The dampening was set to 0, and the weight decay was 0.0001. The nesterov was set to false, and the maximize was false. The foreach was set to none, and the differentiable was set to false.

Table A1 presents a comparison of the number of parameters, GFLOPs, mIoU, and MeanACC of the HRNet and the proposed models with those of the ADE20K validation set. The number of parameters increased by 0.4 M, and GFLOPs increased by 0.004 compared to HRNetV2-W48, which became the baseline model. MeanACC, the average accuracy of the pixel, decreased by 0.1%. However, the mIoU increased by 0.5% owing to the improved performance in segmenting the regions of objects corresponding to pixel classes. However, the mIoU was maintained, and the MeanACC improved by 0.4%. As a result, the fine object segmentation performance in the Cityscapes dataset was enhanced, and although the boundary segmentation performance between objects in PASCAL Context was slightly decreased, the accuracy of pixel classes improved. In other words, this indicates that the channel information emphasis feature of the proposed model is effective in mitigating characteristics that are easily confused with small objects.

Table A1. Results of HRNetV2-based semantic segmentation model results with ADE20K validation set (single scale and no flipping). GFLOPs is calculated on the RTX 3090 and input size 520 $\times$ 520. The proposed method backbone is HRNetV2-W48.

Method	# Params [M]	GFLOPs [G]	mIoU [%]	MeanACC [%]
HRNetV2-W18	1.5	4.5349	23.69	31.8
HRNetV2-W48	65.9	92.7324	**42.0**	53.5
Proposed Method	66.3	92.7328	**42.0**	**53.9**

References

1. Woo, S.; Park, J.; Lee, J.; Kweon, I. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
2. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
3. Li, R.; Huang, H.; Zheng, Y. Human Pose Estimation Based on Lite HRNet with Coordinate Attention. In Proceedings of the 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 15–17 April 2022; pp. 1166–1170.
4. Li, L.; Tia, T.; Li, H.; Wang, L. SE-HRNet: A Deep High-Resolution Network with Attention for Remote Sensing Scene Classification. In Proceedings of the IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 533–536.
5. Wang, F.; Piao, S.; Xie, J. CSE-HRNet: A context and semantic enhanced high-resolution network for semantic segmentation of aerial imagery. *IEEE Access* **2020**, *8*, 182475–182489. [CrossRef]
6. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
7. Gholamalinejad, H.; Khosravi, H. Vehicle Classification using a Real-Time Convolutional Structure based on DWT pooling layer and SE blocks. *Expert Syst. Appl.* **2021**, *183*, 115420. [CrossRef]
8. Zhang, B.; Qi, S.; Wu, Y.; Pan, X.; Yao, Y.; Qian, W.; Guan, Y. Multi-Scale Segmentation Squeeze-and-Excitation UNet with Conditional Random Field for Segmenting Lung Tumor from CT Images. *Comput. Methods Programs Biomed.* **2022**, *222*, 106946. [CrossRef] [PubMed]
9. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
10. Gong, K.; Liang, X.; Shen, X.; Lin, L. Look into person: Self-supervised structure-sensitive learning and A new benchmark for human parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 932–940.
11. Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; Yuille, A. The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 891–898.
12. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
13. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
14. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
15. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
16. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
17. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
18. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multipath refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5168–5177.
19. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
20. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 833–851.
21. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 432–448.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Vaswani, A.; Shazeer, N.; Parmar, M.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

24. Shen, X.; Yin, C.; Hou, X. Self-attention for deep reinforcement learning. In Proceedings of the 2019 4th International Conference on Mathematics and Artificial Intelligence (ICMAI), Chengdu, China, 12–15 April 2019; pp. 71–75.
25. Li, Z.; Li, Y.; Lu, H. Improve image captioning by self-attention. In *International Conference on Neural Information Processing*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 91–98.
26. Qi, J.; Wang, X.; Hu, Y.; Tang, X.; Liu, W. Pyramid Self-attention for Semantic Segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 480–492.
27. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. *arXiv* **2018**, arXiv:1803.02155.
28. Bello, I.; Zoph, B.; Le, Q.; Vaswani, A.; Shlens, J. Attention augmented convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3286–3295.
29. Andreoli, J.-M. Convolution, attention and structure embedding. In Proceedings of the Advances Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 8–14 December 2019.
30. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 1971–1980.
31. Cordonnier, J.-B.; Loukas, A.; Jaggi, M. On the relationship between self-attention and convolutional layers. In Proceedings of theEighth International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020; pp. 1–18.
32. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. In Proceedings of the Advances Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
33. Shao, X.; Xiang, Z.; Li, Y.; Zhang, M. Variational joint self-attention for image captioning. *IET Image Process.* **2022**, *16*, 2075–2086. [CrossRef]
34. Li, S.; Tang, M.; Guo, Q.; Lei, J.; Zhang, J. Deep neural network with attention model for scene text recognition. *IET Comput. Vis.* **2017**, *11*, 605–612. [CrossRef]
35. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
36. Ruan, T.; Liu, T.; Huang, Z.; Wei, Y.; Wei, S.; Zhao, Y. Devil in the details: Towards accurate single and multiple human parsing. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 4814–4821.
37. Luo, Y.; Zheng, Z.; Zheng, L.; Guan, T.; Yu, J.; Yang, Y. Macro-micro adversarial network for human parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
38. Zhao, J.; Li, J.; Nie, X.; Zhao, F.; Chen, Y.; Wang, Z.; Feng, J.; Yan, S. Self-supervised neural aggregation networks for human parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1595–1603.
39. Nie, X.; Feng, J.; Yan, S. Mutual learning to adapt for joint human parsing and pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 502–517.
40. Liang, X.; Gong, K.; Shen, X.; Lin, L. Look into person: Joint body parsing & pose estimation network and A new benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 871–885.
41. Zhang, H.; Dana, K.J.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
42. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2393–2402.
43. Dai, J.; He, K.; Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1635–1643.
44. Arnab, A.; Jayasumana, S.; Zheng, S.; Torr, P.H.S. Higher order conditional random fields in deep neural networks. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 524–540.
45. Lin, G.; Shen, C.; van den Hengel, A.; Reid, I.D. Efficient piecewise training of deep structured models for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3194–3203.
46. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene parsing through ADE20K dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 633–641.
47. Yuan, Y.; Chen, X.; Chen, X.; Wang, J. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv* **2019**, arXiv:1909.11065.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.