mathematics

MDPI

Article

Scene Recognition for Visually-Impaired People's Navigation Assistance Based on Vision Transformer with Dual Multiscale Attention

Yahia Said [1,2,3,*], Mohamed Atri [4], Marwan Ali Albahar [5], Ahmed Ben Atitallah [6] and Yazan Ahmad Alsariera [7]

[1] Remote Sensing Unit, College of Engineering, Northern Border University, Arar 91431, Saudi Arabia
[2] King Salman Center for Disability Research, Riyadh 11614, Saudi Arabia
[3] Laboratory of Electronics and Microelectronics (LR99ES30), University of Monastir, Monatir 5019, Tunisia
[4] College of Computer Sciences, King Khalid University, Abha 62529, Saudi Arabia
[5] School of Computer Science, Umm Al-Qura University, Mecca 24382, Saudi Arabia
[6] Department of Electrical Engineering, College of Engineering, Jouf University, Sakaka 72388, Saudi Arabia
[7] College of Science, Northern Border University, Arar 91431, Saudi Arabia
* Correspondence: yahia.said@nbu.edu.sa

Abstract: Notable progress was achieved by recent technologies. As the main goal of technology is to make daily life easier, we will investigate the development of an intelligent system for the assistance of impaired people in their navigation. For visually impaired people, navigating is a very complex task that requires assistance. To reduce the complexity of this task, it is preferred to provide information that allows the understanding of surrounding spaces. Particularly, recognizing indoor scenes such as a room, supermarket, or office provides a valuable guide to the visually impaired person to understand the surrounding environment. In this paper, we proposed an indoor scene recognition system based on recent deep learning techniques. Vision transformer (ViT) is a recent deep learning technique that has achieved high performance on image classification. So, it was deployed for indoor scene recognition. To achieve better performance and to reduce the computation complexity, we proposed dual multiscale attention to collect features at different scales for better representation. The main idea was to process small patches and large patches separately and a fusion technique was proposed to combine the features. The proposed fusion technique requires linear time for memory and computation compared to existing techniques that require quadratic time. To prove the efficiency of the proposed technique, extensive experiments were performed on a public dataset which is the MIT67 dataset. The achieved results demonstrated the superiority of the proposed technique compared to the state-of-the-art. Further, the proposed indoor scene recognition system is suitable for implementation on mobile devices with fewer parameters and FLOPs.

Keywords: visually impaired; navigation assistance; vision transformer; dual multiscale attention

MSC: 68T45; 68T07

check for updates

Citation: Said, Y.; Atri, M.; Albahar, M.A.; Ben Atitallah, A.; Alsariera, Y.A. Scene Recognition for Visually-Impaired People's Navigation Assistance Based on Vision Transformer with Dual Multiscale Attention. *Mathematics* **2023**, *11*, 1127. https://doi.org/10.3390/math11051127

Academic Editor: Junlin Hu

Received: 25 January 2023
Revised: 15 February 2023
Accepted: 21 February 2023
Published: 24 February 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

Visually impaired people are part of society and they struggle in performing regular daily activities that require navigating. Navigation is the main activity for humans which essentially requires vision to move easier and safer. Navigating in familiar spaces can be easy, however, in unfamiliar spaces, it is difficult to move from one point to another safely. According to the World Health Organization, the number of visually impaired people will increase to more than 2.2 billion [1]. This huge number of persons require personal assistance which cannot be provided by humans due to many reasons such as the limited number of assistants, the privacy of the visually impaired, and efficacity. Using a virtual assistant can be a reliable solution to this problem. Artificial intelligence [2] has been widely deployed in many fields over the past years, especially those related to the assistance of visually impaired people [3].

The navigation scene for a visually impaired person is very important for space understanding and adaptation. Indoor scene understanding helps the visually impaired person familiarize themselves with the space and to navigate easier. By knowing the type of indoor space such as a bathroom, bedroom, or supermarket, a visually impaired person does not need to ask for human help to know where he is and it is easier to reach his destination with more confidence while getting his privacy. Indoor scene recognition can be performed based on processing visual data. Artificial intelligence can be a great solution to solve the problem of indoor scene recognition.

Recent advances in image processing techniques proved that many difficult problems can be solved. Deep learning techniques [4] have been successfully deployed for solving computer vision tasks. The power of deep learning techniques for computer vision tasks comes from the use of the convolutional neural network that is inspired by the biological brain with a high learning capacity and great generalization power. However, indoor scenes represent a wide inter- and intraclass variation. In effect, images belonging to the same scene class have a large variability, and images for different scene classes represent a large similarity. Those challenges, in addition to others such as illumination, deformation, and orientation make the scene recognition task challenging and require powerful techniques to achieve the desired performance.

The recent transformers [5] have boosted the performance of many applications to a higher level. Transformers present high sequence-to-sequence modeling capacities designed originally for natural language processing (NLP) tasks. Its high performance in NLP attracted the attention of the computer vision community to evaluate its impact on image processing applications. Those applications have been widely dominated by convolutional neural networks (CNN) [6]. Previous applications such as object detection [7], pedestrian detection [8], traffic sign recognition [9], and traffic light detection [10] have been successfully solved using CNN models such as faster RCNN [11], you only look once (YOLO) [12], and single shot multibox detection (SSD) [13]. However, many efforts have been presented to show the power of transformers in vision tasks. Most recent works [14,15] focus on the combination of CNN and transformers to solve vision tasks. Subsequently, those works achieved promising results, though they require more computation compared to pure transformers. Vision transformers (ViT) were the first pure transformers that used image processing based on embedding sequences of image patches as input. It was the first convolution-free model used for image classification that achieved comparable results to the convolution models. However, training ViT requires a large-scale dataset to achieve high performance. Some works [16], proved that model regulation and data augmentation techniques can be a solution to train ViT with fewer data and achieve high performances. Based on those techniques, more works [17,18] have been proposed to enhance the efficiency of ViT for image classification tasks.

Following the same target of building a powerful ViT, we proposed a dual multiscale attention mechanism. In CNN models, it was proved that multiscale models can achieve better performances. Motivated by this attempt, deploying multiscale architecture to ViT was investigated to validate its efficiency. The main idea of dual multiscale attention is to combine image patches from different scales to extract reliable features. The proposed ViT process separates embedding from small and large image patches then multiple fusion was applied to combine features from different branches. For this purpose, we proposed four fusion methods to investigate their impact on the performances. The fusion was performed by creating an agent using a nonpatch token for each branch, then features are exchanged by a cross-attention mechanism. This method allows the generation of an attention map in linear time instead of quadratic time. To handle computation complexity, architectural adjustment was applied.

The proposed ViT was applied for indoor scene recognition. A public dataset was used for the training and evaluation of the proposed ViT. The Massachusetts Institute of

Technology (MIT) dataset [19] was used. It consists of 15,620 images from 67 categories. The dataset was designed for indoor scene recognition due to its challenging conditions.

The main contribution of this work is the following:

- We proposed an indoor scene recognition for the assistance of visually impaired people in their navigation;
- We proposed dual multiscale attention for ViT that extracts features at multiscales with a dual branch by processing large and small image patches;
- We proposed an effective fusion method to combine the extracted features through cross attention that reduces computations and memory occupation by processing data in linear time instead of quadratic time;
- Evaluating the proposed ViT on two different scene recognition datasets and high performances were achieved which demonstrates the efficiency of the proposed approach.

The remainder of this paper is structured as follows. Section two is reserved for investigating related works. In section three, we presented and detailed the proposed approach. Experimental results were provided and discussed in section four. In section five, we provide conclusions and future works.

2. Related Works

This work is related to two main research directions: ViT enhancement for image processing tasks and indoor scene recognition. Before the devolvement of transformers, attention modules have been widely deployed in CNN models to boost performance. In addition, the multiscale scheme was key to achieving more performance using the CNN models.

ViT enhancement—ViT was the first model based on transformers used for image recognition with high performance that outperformed CNN models. Since its first use, many techniques were proposed to achieve more performance.

A pyramid structure was proposed by Wang et al. [18] to extract features at different levels. The proposed structure was designed to overcome the challenges preventing the use of transformers for complex tasks such as object detection and instance segmentation. The proposed pyramid ViT introduced the use of a progressive shrinking pyramid to reduce computation complexity presented by ViT in addition to being trained on image-dense partitions. Furthermore, it combines the advantages of transformers and CNN models by learning multiscale and high-resolution feature maps.

As ViT requires a large-scale dataset for training, distillation for data was proposed in [20]. The main idea was to design a teacher-student strategy for transformers. This strategy is based on token distillation to make the student learn from the teacher through attention. Two kinds of distillation were proposed, namely soft distillation and hard-label distillation. Soft distillation aims to reduce the Kullback-Leibler divergence between the teacher output and student output. The hard-label distillation was designed to consider the prediction of the teacher as a true label for the student instead of using the annotation label for the student. This proposal has been proven to be better than traditional training for transformers.

Improved self attention [21] was proposed to learn abstract representation instead of all-to-all representation in regular self-attention. Regular transformers map N input to N output which requires quadratic time complexity and memory consumption. Centroid attention was proposed to map N inputs to M outputs while M < N. This proposal allows for the summarizing of the information in fewer outputs and smaller memory. The proposed centroid attention amortizes the gradient descent update of a clustering loss function which demonstrates a fundamental link between attention and clustering. The compression of the input into centroids allows the extraction of relevant features for the prediction and minimizes the computation complexity.

Jaegle et al. [22] proposed asymmetric attention in order to iteratively distill inputs into a tight latent bottleneck. The proposed attention allows the processing of very large inputs. The proposed perceiver was built based on transformers, though different from ViT. The perceiver combines a cross-attention module and a transformer module. The cross-attention module maps a combination of a latent array and byte array to a latent array while the transformer maps a latent array to a latent array. The size of the byte array is very large which can be decided based on the size of the input while the size of the latent array is very small and considered a hyperparameter to be searched.

A layer-wise token-to-token transformation was proposed in the T2T-ViT model [17] to encode the important local structure for each token instead of the naive tokenization proposed in ViT. It was noticed that the ViT model achieves low performance compared to CNN models in the case of training from scratch on small datasets. The main problem is related to two issues. First, naive tokenization was not able to extract local features such as edges and lines at neighboring pixels. Second, the design of the ViT backbone limits the richness of features under limited computation resources and limited training samples. To handle the mentioned issues, tokens were aggregated recursively to model neighborhood tokens which allowed for the extraction of tokens surrounding structure and reduced token length. Also, a deep-narrow backbone was designed to reduce computation overhead. The proposed enhancement reduced the number of parameters and FLOPs by half compared to the original ViT with an improvement of 3% in accuracy.

Indoor scene recognition—it is a very active research field due to its important applications such as robotic navigation and visually impaired navigation assistance. So, many works have been proposed to achieve good performance.

An indoor scene recognition was proposed in [23] for the assistance of visually impaired people in their navigation. The proposed approach was based on a CNN model with transfer learning techniques. The efficientNet model [24] was used to guarantee lightweight implementation on mobile devices. The efficientNet model proposed to scale network parameters to achieve a balance between accuracy and computation complexity. The proposed model was evaluated on two different datasets and good results were achieved. However, the achieved results were not enough to be implemented in real-world applications.

Miao et al. [25] proposed an indoor scene recognition system for the robot. The proposed approach relied on detecting objects, analyzing them, and then predicting the scene. In effect, a segmentation network was used to extract object features and learn the connection between the scene and a set of existing objects using a feature aggregation module. Next, the relation between the detected objects and the scene is estimated based on an object attention module and a global attention module. The proposed approach demonstrated its efficiency in estimating scenes based on object relation. However, the proposed approach was computationally extensive which prevents its implementation on embedded devices for robotic use.

A multimodal deep learning-based approach [26] was proposed for indoor scene recognition. The proposed approach aims to the fusion of many learning modalities such as speech and visual features to identify scenes in videos. Exploring data modalities such as audio, visual, and semantic features can be the key to the classification of indoor scenes in videos. The learning model was composed of the CNN model and two recurrent neural network (RNN) models based on the combination of a long short terms memory (LSTM) network and a convolutional LSTM network. The first RNN model consists of a joint fusion and the second one consists of a late fusion. To train the proposed approach, a dataset was collected from social media with a total of 3788 videos combined with a YouTube-8M dataset that contains 900 videos for 9 indoor scene classes. The achieved results were promising, however, the complexity and the high computation consumption prevent the use of the proposed method for real applications such as robotics.

3. Proposed Approach

In this work, we proposed a scene recognition for visually impaired people's navigation assistance based on a vision transformer with dual multiscale attention. The proposed method was developed based on ViT. So, we started by presenting the background of the ViT model. Then the proposed model is described in detail.

3.1. Vision Transformer

Vision Transformer (ViT) is a model used for image recognition tasks based on the transformer which is designed for NLP tasks based on sequence-to-sequence transformation. Its main concept is to split the input image into a sequence of patch tokens. The size of the patch is predefined and then each patch is linearly projected into tokens. The class of each image is added as an additional token to the sequence. The self attention in the transformer encoder does not consider position order and vision tasks require position information for precise prediction. So, ViT proposed a position embedding mechanism for each token considering the class token. Figure 1 presents an illustration of ViT with position embedding. After that, all tokens are fed into the transformer encoder and at the final stage, the class token is used for the classification. The transformer encoder is built based on stacked blocks composed of multiheaded self attention (MSA) and multilayer perceptron (MLP). The multilayer perceptron is composed of linear layers with one GELU nonlinear layer after the first linear layer and the hidden layers are expanded with a ratio, r. A residual connection is applied after each block and a layer normalization (LN) is applied before each block. Considering x_0 as an input of ViT, the output of the kth block (y_k) can be computed as Equation (1).

$$\begin{gathered} x_0 = \left[x_{cls}\middle|\middle|x_{patch}\right] + x_{pos} \\ y_k = x_{k-1} + MSA\ (LN(x_{k-1})) \\ x_k = x_k + MLP\ (LN(y_k)) \end{gathered} \tag{1}$$

where x_{cls} is the class token, x_{patch} is the patch token, and x_{pos} is the position embedding.

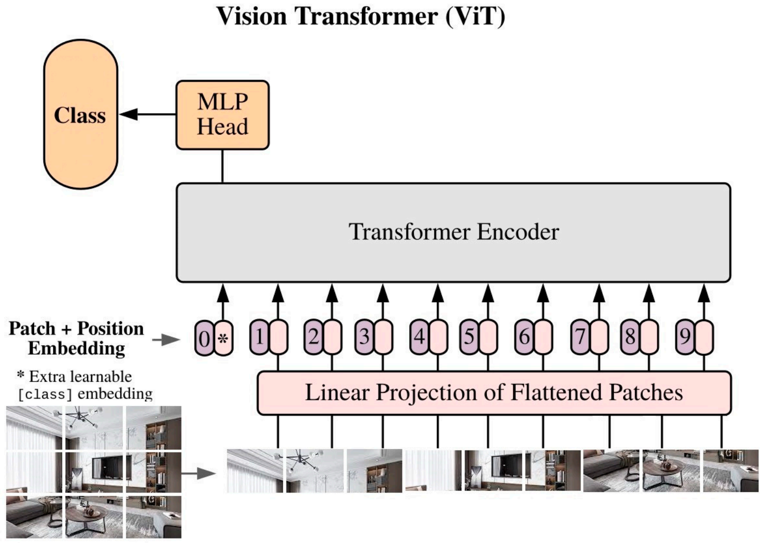
Vision Transformer (ViT)
Class
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure 1. ViT model with position embedding.

It is very important to mention the difference between ViT and CNN models considering the class token. CNN models generate the final embedding through features averaging over all spatial locations while the class is used to interact with the patch token for each transformer encoder to generate the final embedding in ViT. Considering this motivation, the class token was used as an agent to summarize patch tokens from all scales. Thus, we proposed dual multiscale attention based on the class token.

3.2. Proposed Dual Multiscale Attention ViT

The size of the patch directly affects the accuracy and the computation complexity. Smaller patch sizes result in better accuracy but require more FLOPs and memory. For example, using a patch size of 16 achieved higher accuracy compared to a patch size of 32 with a margin of 6%. However, using a patch size of 16 instead of 32 results in an increase of four times in computation complexity. Motivated by this fact, the proposed approach takes advantage of small-size patches while finding a good trade off between computation complexity and accuracy. First, we proposed a dual-branch scheme where different scales are used for each branch. Second, we proposed an effective fusion technique to fuse information from different branches. The proposed dual multiscale attention consists of two branches where the L-branch uses a large token size and more transformer encoders compared to the other branch and the S-branch uses a small token size and fewer transformer encoders. Wider embedding dimensions were associated with the L-branch and smaller ones for the S-branch. The two branches are fused L times at the next stage and the class token from the two branches is fused at the end for prediction. Figure 2 presents the proposed dual multiscale attention ViT. Furthermore, as in the original ViT, we deployed a learnable position embedding before the multiscale structure for each token in two branches.

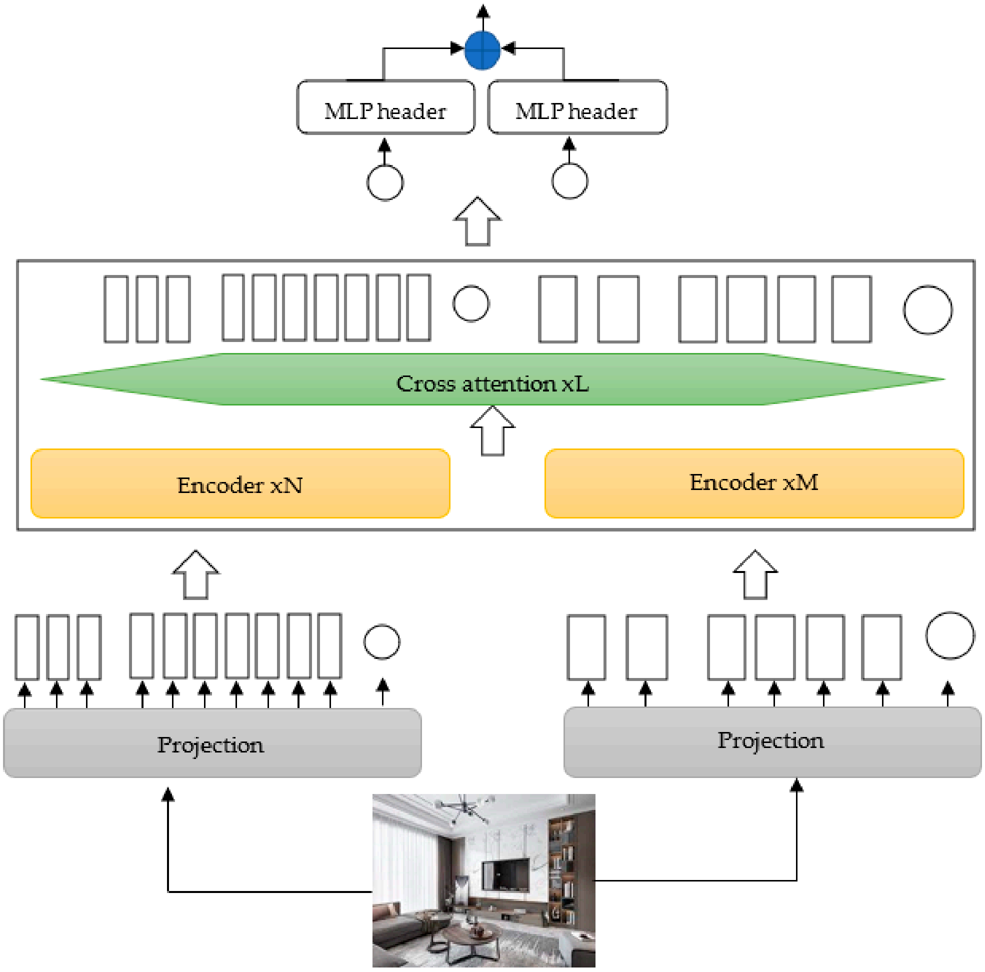

Figure 2. Proposed dual multiscale attention ViT.

To learn from multiscale features, it is important to use an effective fusion technique. As many techniques can be applied for feature fusion, we proposed four different fusion techniques. Figure 3 presents the proposed fusion techniques where there are three simple techniques and the proposed cross-attention fusion technique.

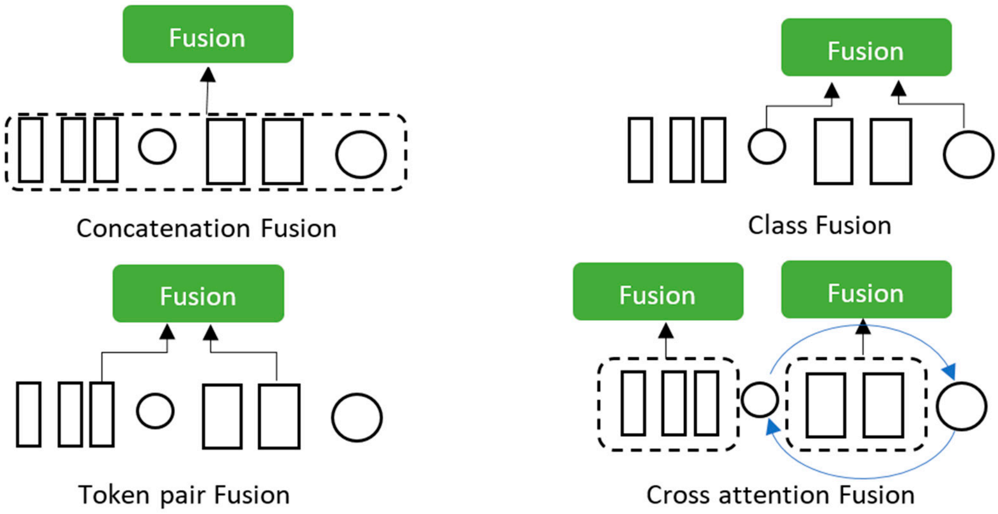

Figure 3. Proposed fusion techniques.

3.3. Proposed Fusion Techniques

Considering an input sequence x^i including patch and class tokens for both branches, the large branch and the small branch, x^i_{cls} is the class token, and x^i_{patch} is the patch token. Four different fusion techniques have been proposed for feature fusion from multiscale levels.

Concatenation fusion—a simple fusion technique based on the concatenation of all tokens from the large and small tokens. This technique does not consider the property of each branch. The extracted features are fused by a self-attention module. Thus, this fusion technique requires quadratic time. The output z^i of the concatenation fusion can be computed as Equation (2).

$$\begin{gathered} y = \left[f^l\left(x^l\right) \middle|\middle| \, f^s(x^s)\right] \\ o = y + MSA\ (LN(y)) \\ o = \left[o^l \middle|\middle| \, o^s\right] \\ z^i = g^i\left(o^i\right) \end{gathered} \tag{2}$$

where f^i is the projection function of the input patches and g^i is the back-projection function used to align features dimension.

Class fusion—the class token is deployed for prediction through the final embedding. This allows for considering it as abstract global features for each branch. So, this fusion technique is based on the sum of class tokens of both branches. This fusion is very effective and efficient since it requires one token from each branch to be performed. After fusing the class tokens, the information will be returned to patch tokens in the transformer encoder. The output z^i of the class fusion technique can be computed as Equation (3).

$$z^i = [g^i(\textstyle\sum_{j\in\{l,\ s\}} f^j(x^j_{cls}))||x^j_{patch}] \tag{3}$$

Token pair fusion—this fusion technique fuses tokens by pair from both branches. The spatial location was used for the fusion of tokens by combining them by pair. However, patch size and the number of patches are different in the two branches. So, an interpolation function was applied to align the spatial size before the fusion. Then, a pair of tokens, including class tokens, are fused. The output of token pair fusion can be computed as Equation (4).

$$z^i = [g^i(\sum_{j\in\{l,\ s\}} f^j(x^j_{cls}))||g^i\left(\sum_{j\in\{l,\ s\}} f^j\left(x^j_{patch}\right)\right)] \tag{4}$$

Cross-attention fusion—the main idea of this fusion technique is to use the class token of one branch in the other branch. So, the class token of the small branch will be used with the large branches and the class branch of the large branch will be used with the small branch. The class token was used as an agent for information exchange between branches. Passing the class token to the other branch will ensure including information at different scales. After fusion with the other branch, the class token will be projected back to its own branch in the next transformer encoder. At this end, it will transmit the learned information from one branch to the other. This fusion technique will enrich the features of the patch token in both branches. Figure 4 presents the cross-attention fusion. For a given branch, the cross-attention fusion starts by collecting patch tokens from the other branch and concatenating them with the class token. This procedure for the large branch can be computed as Equation (5).

$$x'^l = f^l\left(x^l_{cls}\right)\Big|\Big|x^l_{patch} \tag{5}$$

where f^l is the projection function for dimension alignment. After this step, the cross attention is performed between x'^l and x^l_{cls}. The only query is the class token since patch tokens' information are fused into the class token. The cross-attention fusion can be computed as Equation (6).

$$\begin{gathered} b = x'^l_{cls}w_b,\ k = x'^l w_k,\ v = x'^l w_v \\ A = softmax\left(\frac{bk^T}{\sqrt{\frac{C}{h}}}\right) \\ CA\left(x'^l\right) = A \times v \end{gathered} \tag{6}$$

where $w_b,\ w_k,\ w_v\ \in\ \mathbb{R}^{C\times(\frac{C}{h})}$ are hyperparameters to be learned. C is the dimension of the embedding and h is the number of heads. Since the only query in cross-attention fusion is the class token, the memory consumption and computation complexity are linear instead of quadratic in other fusion techniques. This fusion technique is more efficient in terms of computation. Furthermore, following the regular self attention, mutiheads were deployed in the cross attention. However, we eliminated the MLP for memory efficiency. Considering the large branch, the output of the cross-attention fusion can be computed as Equation (7).

$$\begin{gathered} y^l_{cls} = f^l\left(x^l_{cls}\right) + MCA(LN\left(\left[f^l\left(x^l_{cls}\right)||x^l_{patch}\right]\right)\Big) \\ z^l = [g^l\left(y^l_{cls}\right)\Big|\Big|x^l_{patch}] \end{gathered} \tag{7}$$

In the next section, we will demonstrate the superiority of cross-attention fusion compared to other fusion techniques empirically. Also, we present the achieved results and discuss the efficiency of the proposed dual multiscale attention for scene recognition.

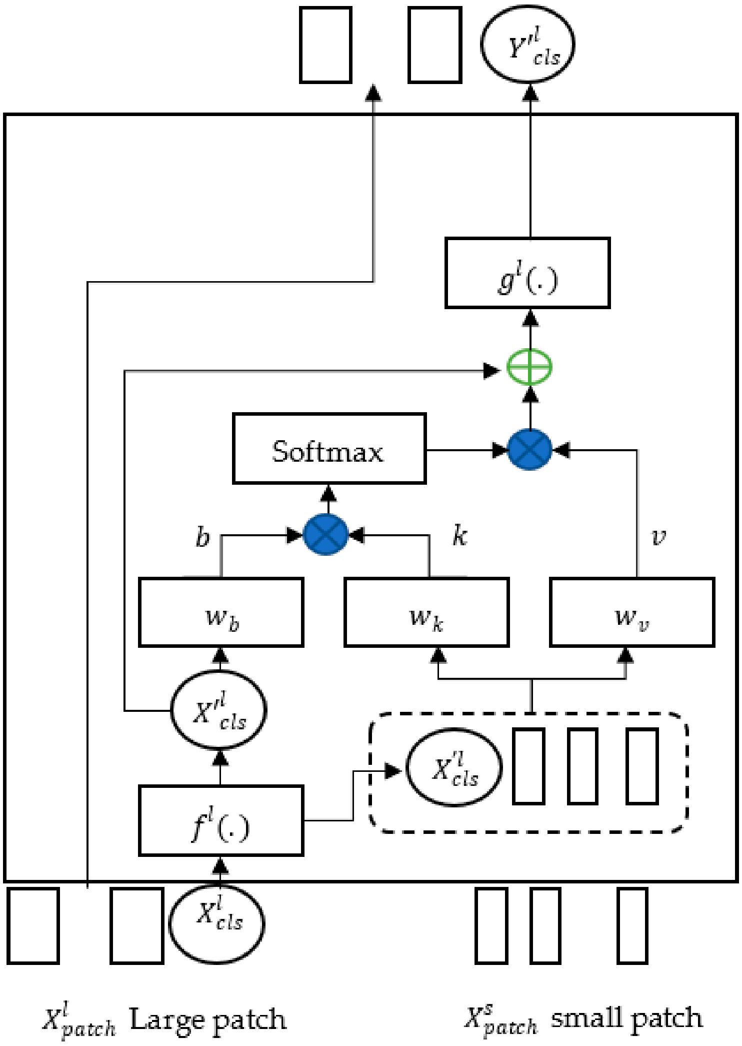

Figure 4. Cross-attention fusion.

4. Experiment and Results

In this section, we will present our experimental environments by detailing the available computation resources and the datasets used for the evaluation of the proposed approach. Next, we will discuss the achieved results after extensive experiments. Then, a comparative study will be presented to prove the performance of the proposed approach. We finish with an ablation study to evaluate the impact of each fusion technique and the effect of different configurations.

4.1. Experimental Environments

The proposed dual multiscale attention ViT was evaluated on a desktop powered with an intel i7 processor and Nvidia GTX 960 GPU. The performance of the proposed model was validated using the MIT67 dataset for indoor scene recognition [19]. The MIT67 dataset consists of 15,620 images from 67 categories. The dataset was designed for indoor scene recognition due to its challenging conditions. Samples from the dataset are presented in Figure 5. Since ViT requires a large-scale dataset for training, we started by training the proposed model on the imageNet21K dataset [27] and then use a pretrained model as initialization for training on the indoor scene dataset. Existing works proved that ViT-based models can be trained with a smaller dataset by applying a set of data augmentation techniques. The mixup [28], cutmix [29], rand augmentation [30], and random erasing [31] were applied as data augmentation techniques.

Figure 5. Samples from the MIT 67 dataset.

The proposed model was trained for 300 epochs and the batch size was fixed to eight. The learning rate was fixed at 0.004 and the weight decay was 0.05. Also, we applied a cosine linear-rate scheduler with a linear warm up. The input image resolution was resized to 224 × 224 and varied for comparison purposes. For adjusting the size of the position embedding, a bicubic interpolation was used.

The configuration of the proposed dual multiscale is presented in Table 1. Many configurations were evaluated to achieve the best performances. Thus, we varied the expansion ratio of the MLP, the dimension, and the depth of the embeddings. It was proved in the original ViT that generating the patch token using a CNN model can enhance the performance. Motivated by this, we proposed to generate the patch token using three convolution layers. For all evaluated models, the number of multiscale transformer encoders K was fixed to three, the number of transformer encoders in large branch N was fixed to one, and the number of cross-attention fusion module L was fixed to one.

Table 1. Configuration of the proposed dual multiscale attention (DMS) ViT.

Model	**Patch**	**Patch Size**		**Input Resolution**		**Heads**	**M**	**r**
		Large	**Small**	**Large**	**Small**			
DMS-ViT1	Linear	16	12	256	128	4	3	3
DMS-ViT2	Linear	16	12	384	192	6	5	3
DMS-ViT3	Linear	16	12	448	224	7	6	3
DMS-ViT1 *	3 convolutions	16	12	256	128	4	3	3
DMS-ViT2 *	3 convolutions	16	12	384	192	6	5	3
DMS-ViT3 *	3 convolutions	16	12	448	224	7	6	3

*: Patch generator based on 3 Convolution layers.

4.2. Results

The evaluation of the proposed models has proved their superiority. The accuracy per class of the DMS-ViT3 * is presented in Table 2. Interestingly, the proposed model shows a high accuracy for almost all classes except for some classes.

Table 2. Achieved accuracies for each class of the MIT67 dataset.

church inside	Elevator	Auditorium	buffet	classroom
98.3	97.5	97.4	98.6	99.5
greenhouse	bowling	cloister	concert hall	deli
96.4	97.4	98.8	97.3	97.9
dentaloffice	library	inside bus	closet	corridor
96.7	97.6	96.8	94.6	97.3
grocerystore	locker room	florist	studiomusic	hospitalroom
94.5	97.6	98.7	98.1	98.3
nursery	trainstation	bathroom	laundromat	stairscase
94.5	98.1	99.3	96.1	97.9
garage	gym	tv studio	videostore	gameroom
92.5	96.8	96.8	93.7	97.6
pantry	poolinside	inside subway	kitchen	winecellar
97.8	96.5	96.3	94.5	96.9
fastfood restaurant	restaurant kitchen	clothingstore	operating room	Computerroom
99.8	97.7	98.9	95.3	97.6
bookstore	waitingroom	dining room	bakery	livingroom
95.8	95.3	97.6	98.4	96.3
movietheater	bedroom	toystore	Casino	airport inside
96.7	97.7	98.4	98.2	97.3
artstudio	lobby	hairsalon	subway	warehouse
96.3	98.8	96.1	97.3	97.4
meeting room	children room	shoeshop	kindergarden	restaurant
97.3	98.3	95.4	97.5	96.8
museum	Bar	jewelleryshop	laboratorywet	mall
98.3	97.7	95.7	97.5	97.6
office	prison cell			
95.1	97.2			

The DeiT is a very powerful transformer-based model, so we consider it a comparison baseline model. The model was reproduced using our experimental environment. Table 3 presents the achieved results on the MIT67 dataset and a comparison against the DeiT models. The presented results demonstrate that the proposed DMS-ViT outperforms the DeiT model by a large margin in terms of accuracy while being computationally efficient.

Table 3. Comparison against the DeiT model on the MIT67 dataset.

Model	Top-1 Accuracy (%)	FLOPs (G)	Speed (FPS)	Parameters (M)
DeiT-Ti	89.2	1.3	45	5.7
DMS-ViT1	92.4	1.8	36	8.6
DMS-ViT1 *	93.1	2	34	8.8
DeiT-S	90.8	4.6	26	22.1
DMS-ViT2	93.5	5.8	23	27.4
DMS-ViT2 *	94.3	6.1	21	28.2
DeiT-B	94.6	17.6	15	68.6
DMS-ViT3	96.1	9.1	12	43.3
DMS-ViT3 *	96.8	9.5	11	44.2

*: Patch generator based on 3 Convolution layers.

Based on the achieved results, the DMS-ViT outperforms the DeiT model while using the same configuration. This proved the effectiveness of the proposed cross-attention fusion and the dual branch input for learning multiscale features from the input image. For large models, the DMS-ViT outperforms the baseline by 1.6% when using a linear projection and by 2.2% when using a convolutional projection of patch tokens while getting fewer parameters and GFLOPs. Surprisingly, generating patch tokens using convolution layers boosted the performance by 0.7% for DMS-ViT3 and by 1.2 for DMS-ViT1. It was noticed that the effectiveness of the convolution generator becomes weaker when the number of transformers encoder becomes higher. Despite this weakness, the model still gets an improvement of 0.7% in accuracy.

In addition to high accuracy, computation effectiveness must be considered since we targeted embedded implementation. The DMS-ViT1 and DMS-ViT2 have more GFLOPs and parameters compared to the baseline, however, the DMS-ViT3 has fewer GFLOPs and parameters compared to the DeiT model. Considering the model with the highest accuracy, the proposed DMS-ViT3 is more efficient.

For more convenience and to demonstrate the efficiency of the proposed model, we presented a comparative study against the state-of-the-art works for indoor scene recognition. Table 4 presents a comparison between the proposed model and state-of-the-art models for indoor scene recognition on the MIT67 dataset.

Table 4. Comparison against state-of-the-art models for scene recognition on the MIT67 dataset.

Model	Accuracy (%)	Speed (FPS)	FLOPs (G)	Parameters (M)
efficientNet-B0 [23]	92.35	41	0.39	5.3
DMS-ViT1	92.4	36	1.6	6.9
DMS-ViT1 *	93.1	34	1.8	8.6
efficientNet-B5 [23]	93.57	24	9.9	30
DMS-ViT2	93.5	23	2.3	12.1
DMS-ViT2 *	94.3	21	2.7	16.7
efficientNet-B7 [23]	95.6	13	9.9	66
DMS-ViT3	96.1	12	37	22.3
DMS-ViT3 *	96.8	11	5.6	26.5

*: Patch generator based on 3 Convolution layers.

To the best of our knowledge, we are the first to investigate the use of transformer-based models for indoor scene recognition. Most computer vision applications, including scene recognition, were dominated by CNN-based models. In this study, we presented a comparison against a state-of-the-art model which was efficientNet. This model was designed for computation efficiency targeting its implementation on mobile devices. As shown in Table 3, our models outperformed the efficientNet models in terms of accuracy while getting lower inference speeds. Despite achieving lower inference speed, the proposed models are still suitable for real-time application with the advantage of higher accuracy. The achieved results proved the efficiency of the proposed model for indoor scene recognition. It is important to note that the DMS-ViT1 * has achieved the best trade-off between accuracy and processing speed. Also, DMS-ViT3 * and DMS-ViT2 * present a good trade off compared to efficientNet, which is slightly faster but less accurate. Considering the target application, the proposed DMS-ViT1 * is the best candidate with reasonable accuracy and high processing speed while being suitable for implementation on embedded devices.

4.3. Ablation Study

An ablation study was performed to investigate the impact of different contributions. First, the impact of the fusion technique was analyzed. Second, the impact of the model configuration and the choice of hyperparameters. Finally, we investigated the impact of the proposed dual-branch scheme.

Impact of fusion technique—it was proven that the fusion technique had a wide impact on performance and computation. As mentioned earlier, we proposed four fusion techniques and evaluated each one separately. The fusion techniques were evaluated on the imageNet21K dataset for better analyses. Table 5 presents a comparison between different fusion techniques.

Table 5. Comparison between different fusion techniques.

Fusion	Top-1 Accuracy (%)	FLOPs (G)	Parameters (M)
None	80.2	5.3	23.7
Concatenation	80	7.6	27.5
Class fusion	80.4	5.4	24.3
Token pair	80.5	5.5	24.4
Cross attention	81.2	5.6	26.5

Among the fusion techniques, the cross-attention fusion is the best with the highest accuracy balanced with GFLOPs and parameters. Surprisingly, the concatenation fusion achieved the worst performance even compared to a model with no fusion in addition to a huge increase in GFLOPs and parameters.

Impact of patch size for dual branch—to investigate the effect of patch size in the large and small branches, we performed many experiments with different sizes such as 8 for the small branch, 16 for the large branch, 12 for the small branch, and 16 for the large branch. The achieved results presented in Table 5 prove that the (12, 16) is better than the (8, 16) in terms of accuracy while having fewer GFLOPs and parameters. Normally, the (8, 16) should achieve better performance since it provides fine-grained features by the small branch, however, it was the worst due to the large granularity difference between the branches. The (12, 16) achieved better results due to the smooth learning of features from both branches.

Impact of channel width and depth in the small branch—the main target of this work was to build a transformer-based model with a lightweight size and low computation complexity. So, we investigated the impact of the width and depth of the small branch on the computation complexity. Extensive experiments proved that using a lightweight small branch is more suitable. As shown in Table 5, increasing the depth and width of the small branch increased the GFLOPs and parameters without any accuracy improvement. It was proven that the large branch is charged for features extraction and the small branch provides additional features.

Impact of the number of fusion modules and the number of multiscale transformer encoders—to speed up the fusion frequency, there were two methods that increased the number of fusion modules (L) or increased the number of transformer encoders (K). The latter method required reducing M to maintain the depth of the model. The reported results in Table 6 proved that increasing the fusion frequency did not provide any enhancement in accuracy while increasing the GFLOPs and parameters. So, using more fusion modules or more transformer encoders had a negative impact since it increased computation complexity without any accuracy improvements.

Table 6. Ablation study with different model configurations of the imageNet21k dataset.

Model	Patch Size		Resolution		K	N	M	L	Top-1 Accuracy (%)	FLOPs (G)	Parameters (M)
1	12	16	192	384	3	1	4	1	81	5.6	26.7
2	8	16	192	384	3	1	4	1	80.8	6.7	26.7
3	12	16	384	384	3	1	4	1	80.1	7.7	31.4
4	12	16	192	384	3	2	4	1	80.7	6.3	28.2
5	12	16	192	384	3	1	2	2	81	5.6	28.9
6	12	16	192	384	6	1	4	1	80.9	6.6	31.1

Impact of the class token exchange—we conducted many experiments to show the importance of the class token. The proposed model was evaluated without the class token exchange in the cross-attention fusion then it was evaluated with a class token exchange. The first experiment without class token exchange achieved a top-one accuracy of 80% on the imageNet21K dataset while in the second experiment, the model achieved an accuracy of 81%. This experiment showed that the class token exchange has a positive impact on performance due to passing summarized information from one branch to another.

5. Conclusions

Navigation assistance for visually impaired people is very important to allow them to perform their daily activities freely and safely. Indoor scene recognition is very important for scene understanding which facilitates navigating in a familiar space. In this work, we proposed a scene recognition model based on transformers. A dual multiscale attention ViT was proposed to enhance the accuracy of existing ViT for scene recognition. The proposed model consists of extracting features from a large branch and a small branch. To combine the extracted features effectively, a cross-attention fusion technique was proposed. Extensive experimentation proved the efficiency of the proposed model for indoor scene recognition compared to existing transformer-based models and state-of-the-art CNN models for indoor scene recognition on the MIT67 dataset. The proposed DMS-ViT1 * has achieved the best trade off between accuracy and processing speed while presenting the possibility of implementation on mobile devices. Since this work investigated the use of dual branch and multiscale ViT for indoor scene recognition, future work will develop more complex applications such as object detection and instance segmentation based on the proposed model.

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

Funding: This research was funded by King Salman Center for Disability Research, grant number KSRG-2022-023.

Data Availability Statement: Data will be made available on request.

Acknowledgments: The authors extend their appreciation to the King Salman Center for Disability Research for funding this work through Research Group no KSRG-2022-023.

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

References

1. Blindness and Vision Impairment. Available online: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed on 25 July 2022).
2. Strong, A.I. Applications of artificial intelligence & associated technologies. *Science* **2016**, *5*, 64–67.
3. Afif, M.; Ayachi, R.; Pissaloux, E.; Said, Y.; Atri, M. Indoor objects detection and recognition for an ICT mobility assistance of visually impaired people. *Multimed. Tools Appl.* **2020**, *79*, 31645–31662. [CrossRef]
4. Goodfellow, I.; Yoshua, B.; Aaron, C. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
6. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
7. Ayachi, R.; Said, Y.; Atri, M. A Convolutional Neural Network to Perform Object Detection and Identification in Visual Large-Scale Data. *Big Data* **2021**, *9*, 41–52. [CrossRef] [PubMed]
8. Ayachi, R.; Said, Y.; Ben Abdelaali, A. Pedestrian detection based on light-weighted separable convolution for advanced driver assistance systems. *Neural Process. Lett.* **2020**, *52*, 2655–2668. [CrossRef]
9. Ayachi, R.; Afif, M.; Said, Y.; Ben Abdelali, A. Traffic Sign Recognition Based On Scaled Convolu-tional Neural Network for Advanced Driver Assistance System. In Proceedings of the 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), Genova, Italy, 9–11 December 2020; pp. 149–154.
10. Ayachi, R.; Afif, M.; Said, Y.; Ben Abdelali, A. An Embedded Implementation of a Traffic Light Detection System for Advanced Driver Assistance Systems. In *Industrial Transformation*; CRC Press: Boca Raton, FL, USA, 2022; pp. 237–250.
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9. [CrossRef] [PubMed]
12. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.
14. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–13.
15. Srinivas, A.; Lin, T.-Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16519–16529.
16. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
17. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F.E.H.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 558–567.
18. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578. [CrossRef]
19. Quattoni, A.; Torralba, A. Recognizing indoor scenes. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 413–420.
20. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; Volume 139, pp. 10347–10357.
21. Wu, L.; Liu, X.; Liu, Q. Centroid transformers: Learning to abstract with attention. *arXiv* **2021**, arXiv:2102.08606.
22. Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; Carreira, J. Perceiver: General perception with iterative attention. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 4651–4664.
23. Afif, M.; Ayachi, R.; Said, Y.; Atri, M. Deep Learning Based Application for Indoor Scene Recognition. *Neural Process. Lett.* **2020**, *51*, 2827–2837. [CrossRef]
24. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 6105–6114.

25. Miao, B.; Zhou, L.; Mian, A.S.; Lam, T.L.; Xu, Y. Object-to-scene: Learning to transfer object knowledge to indoor scene recognition. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech, 27 September–1 October 2021; pp. 2069–2075.
26. Glavan, A.; Talavera, E. InstaIndoor and multi-modal deep learning for indoor scene recognition. *Neural Comput. Appl.* **2022**, *34*, 6861–6877. [CrossRef]
27. Ridnik, T.; Ben-Baruch, E.; Noy, A.; Zelnik-Manor, L. Imagenet-21k pretraining for the masses. *arXiv* **2021**, arXiv:2104.10972.
28. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
29. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032.
30. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703.
31. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008.

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.