j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

Implementation Of The Swin Transformer and Its Application In Image Classification

[1]Rasha. A. Dihin, [2]Ebtesam N. Al Shemmary, [3]Waleed Ameen Al Jawher

[1]Department of Computer Science, University of Kufa, Najaf, Iraq,
. [2]IT Research and Development Center, University of Kufa, Najaf, Iraq
[3]Collge of Engenering, Uruk University, Baghdad, Iraq
rashaa.aljabry@uokufa.edu.iq.

Abstract There are big differences between the field of view of the calculator and the field of natural languages, for example, in the field of vision, the difference is in the size of the object as well as in the accuracy of the pixels in the image, and this contradicts the words in the text, and this makes the adaptation of the transformers to see somewhat difficult.Very recently a vision transformer named Swin Transformer was introduced by the Microsoft research team in Asia to achieve state-of-the-art results for machine translation. The computational complexity is linear and proportional to the size of the input image, because the processing of subjective attention is within each local window separately, and thus results in processor maps that are hierarchical and in deeper layers, and thus serve as the backbone of the calculator's vision in image classification and dense recognition applications. This work focuses on applying the Swin transformer to a demonstrated mathematical example with step-by-step analysis. Additionally, extensive experimental results were carried out on several standardized databases from CIFAR-10, CIFAR-100, and MNIST. Their results showed that the Swin Transformer can achieve flexible memory savings. Test accuracy for CIFAR-10 gave a 71.54% score, while for the CIFAR-100 dataset the accuracy was 46.1%. Similarly, when the Swin transformer was applied to the MNIST dataset, the accuracy increased in comparison with other vision transformer results.

Crossref 10.36371/port.2023.4.2

Keywords: *Image classification, Object detection, Swin transformer ST, Vision transformer ViT.*

1. INTRODUCTION

Vision transformers (ViTs) were first proposed for the machine translation task in the Natural Language Processing NLP domain. The transformer-based methods have accomplished innovative performance in a variety of tasks [1]. The ViT's disadvantage is that it necessitates pre,-training on its larg,e ,dataset, [2-4]. Transformers have been widely used in numerous vision problems, especially for visual recognition and detection [5].

The pioneering work of Swin Transformer [6] has two outstanding and effective contributions that distinguish it from other transformers:

a) The feature hierarchical scheme offers outstanding performance in terms of computational complexity. It is linear, and for this feature (ST) is the backbone of (CV) tasks.
b) The second advantage of ST is that it proposes a design equipped with variable-size windows between layers for successive attention, and this improves the power of the modeling [7].

The model proposed by Gu. Yeonghyeon *et al* [8] STHarDNet, which combines the blocks of Swin adapters with U-Net, is light in weight because it contains. The proposed model was used in the segmentation of MRI images taken to diagnose stroke, and it gave superior results. The Swin was used in the first communication layer in the encryption stage to preserve the features of the Swin hierarchical switch STHarDNet is similar to the CNN model in that it takes into account the constraints and completes the task, therefore it is considered the best.

Liao. Zhihao .el.at [9] The Swin-PANet model is proposed as a window-dependent self,-attention mechanis.m using ST in the intermediate supervision network, which has been called the Prior Attention Network. In this case, there was an increase in the improvement of the details in the limits, and the proposed Swin- PANet was used to diagnose skin cancer, and it gave efficient results in improving the accuracy of segmentation and outperformed the modern models, but despite that, it still suffers from a set of limitations, for example, the ability to transfer learning.

L. Jingyun *et al* [10] proposed a powerful basic image recovery model based on the Swin Transformer named SwinIR. The proposed model consists of three parts, firstly extracting shallow features, secondly extracting deep features, and finally creating HR modules for re-basis, SwinIR achieved superior image recovery performance in these three tasks and six different settings. namely: “classic image SR”, grayscale image denoising, RGB image

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

denoising, real-world image SR, light,weight image, and JPEG.

H. Siyuan *et al* [11] proposed a new method for classifying remote sensing images called (TSTNet), which is a two-stream network using Swin transformer, and which was used for each stream and considered the basis for it.

Thus, when comparing the performance of the proposed network with modern models, it was found that the network gives good performance by classifying on a difficult and available data set.

A. Hatamizadeh *et al* [12] reformulated the problem semantic segmentation of 3D "brain tumor" as a Seq-to-Seq prediction issue by developing the Swin UNEt TRansformers (Swin UNETR) model. In this model, the input data of multi-modal was converted onto a 1D embedding sequence and utilized as the input to a hierarchical ST representing the encoder. In the validation phase, this model was listed among the top-performing techniques, and in the testing phase, it displayed competitive performance.

This paper will cast as a specific case of extending the “self-attentio,n mechanism” of the Swin Transformer by demonstrating its performance on a given matrix. It will analyze the intermediate results and explains the arbitrary relations between any two elements of the input during the implementations. In addition, it will explore the work on modeling labeled and directed graphs. On the three tasks, the Swin Transformer performance architecture was compared to the past “state-of-the-art”. Several experiments were conducted on “CIFAR-,10”, “CIFAR-,100”, and “MNIST” for image classification. The paper was organized into five sections: the second section covers the implantation of the Swin Transformer Block. ,Section 3 elucidates the application of the Swin transform, s ection’ 4 states the results, and the final section demonstrates several conclusions.

2. BASIC CONCEPT OF A TRANSFORMER

The main structure of any Transformer [6-8] uses an encoder-decoder structure where we note the use of layers in both encoder and decoder as shown in Figure 1

The encryption layer also contains two important sub-layers: first, the self-attention, and second, the position-wise feed-forward layer. On the other hand, the decoding layer contains three sub-layers:1) self-attention,layer, 2) ‘encoder’-decoder attention, and 3) a ‘positio,n-wise feed’-,forward layer. The transformers in general use a special type of connection around each of the sub-layers called residual connections. In addition, it uses an essential process known as layer normalization. One of the important stages in the structure of transformers is the masking in the decoder. This masking which uses self-attention prevents a given output position from incorporating information about future output positions during training. Both the input components of the encoder and the decoder are locally encoded based on sinusoids at different frequencies and all this is before entering the first layer and therefore useful in making the model generalizable during training, and also helps the model to learn by relative position and when comparing this property with representations Relative position is paradoxical because the relative position of its representations is fixed and this, of course, helps spread the location information to other higher layers.

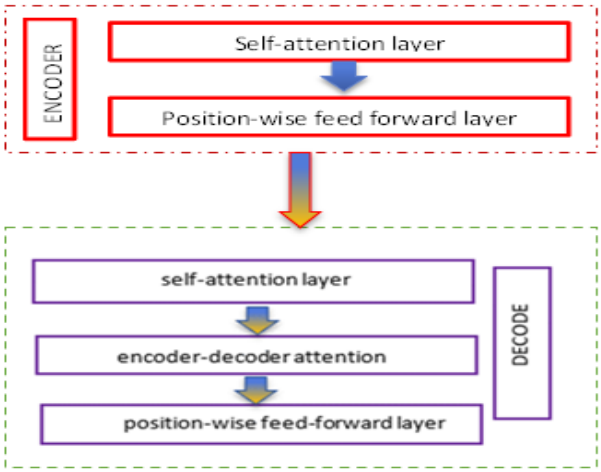

***Figure 1**: Transformer main blocks*

3. THE SHIFTED WINDOW TRANSFORMER (ST)

Shifted Window Transformer or Swin transformer (ST) constructs hierarchical feature maps of the image by combining their patches into deeper layers. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows as well as allowing the cross-window connection. The cost is computationally linear because the processing of self-attention within each local window is proportional to the size of the image entered. Thus, it generates feature maps with a single low resolution in comparison with earlier vision

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

transformers. Therefore, it can serve as a foundation for image and texture categorization, and speech processing algorithms. There are many Challenges in Vision Applications using the Shifted Window Transformer. Transformer-based models are all set tokens in size and may therefore be inappropriate for some CV applications

The resolution of the pixels in the images is higher than the words in the text segments, and this is another variation between CV and NLP. The last difference is that some CV tasks may require detailed predictions at the pixel level. For example, semantic segmentation is not possible in high-resolution images because it causes quadratic computational complexity. As a solution, the Microsoft research team in Asia presents the Swin transformer as a general-purpose Transformer foundation that creates deep network maps. The hierarchical diagram of the Swin Transformer is given in Figure 2.

Figure 2 shows how ST builds a hierarchical representation by starting with tinypatches and then merging these neighboring patches to get deeper layers.

Where by using these feature maps that are hierarchical, the computational complexity is linear so that it divides the image into non-overlapping windows, as well as the number of corrections is constant, unlike other transformers, which are distinguished maps with one precision and therefore the computational complexity is quadratic

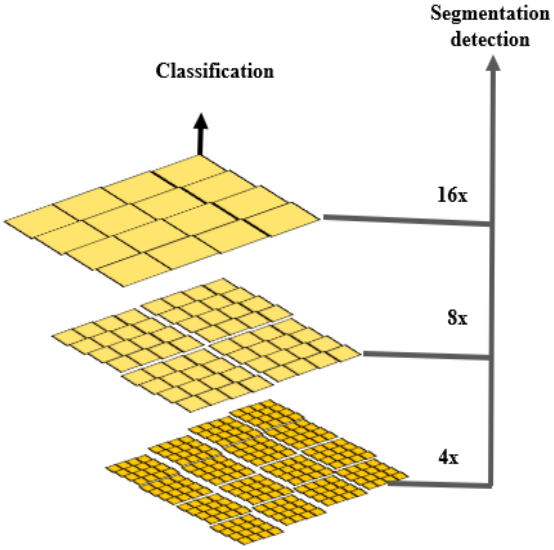

Figure 2: *The Hierarchical of the Swin Transformer*

4. STRUCTURE DESIGN

Moving the window between successive ST self-attention levels switches to final layer windows. This key feature of the ST greatly enhances modeling capabilities and because all query patches that are inside the window share the same key set, this simplifies access to device memory [6]. Figure 3 shows the architecture of the ST in its simplest form. The input RGB image is split into several nonn-overlapping, patches using the patch splitter used in ViT, where each patch is treated as a "token", where individual pixels are set to be a sequence of “RGB” values.

The input image is routed via the patch partition layer and split into patchess of size 4x4 to create patch tokens with the shape (W/4, H/4, and 4x4 channel). In stage 1, the resulting patch tokens are subjected to linear embedding. They are then fed into two linked ST blocks to make tokens of (W/4, H/4, C), where C may be any dimension. Patch merging and Swin transformer blocks are used in stages 2, 3, and 4, respectively. In stages 2, 3, and 4, the shape of the tokens is (W1/81, H1/81, 2C),(W1/161, H1/161, 4C1), and (W1/321, H1/321, 8C1) respectively.

Swin Transformer can be created using transformed windows. The normal multi-headed self-attention unit has been replaced. The rest of the layers remain the same. This layer is followed by a non-linear unit of Gaussian error. Next, Layer normalization is applied before each multi-head self. -attention module. Finally, a residual connection is applied after each of the above modules mentioned before.

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

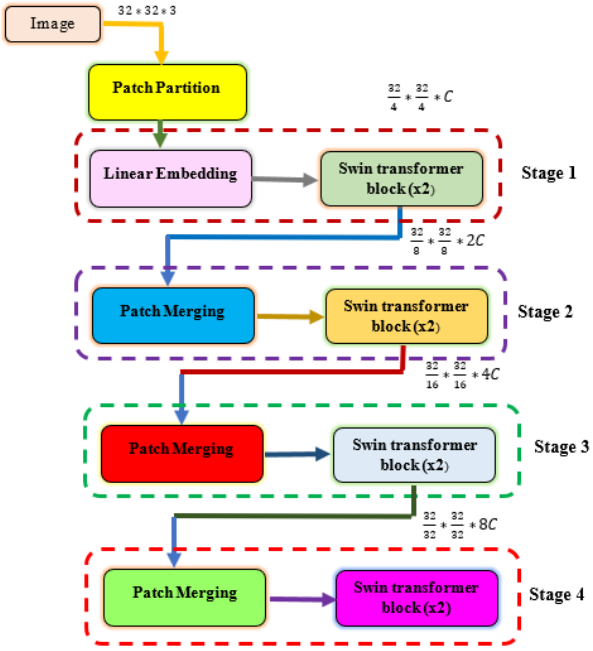

Figure 3: *The architecture of the Swin transformer*

The following steps demonstrate how these module computations of the Swin transformer are implemented considering a given matrix:

Step 1: Give a matrix R of 8x8 pixels as shown in Figure 4.

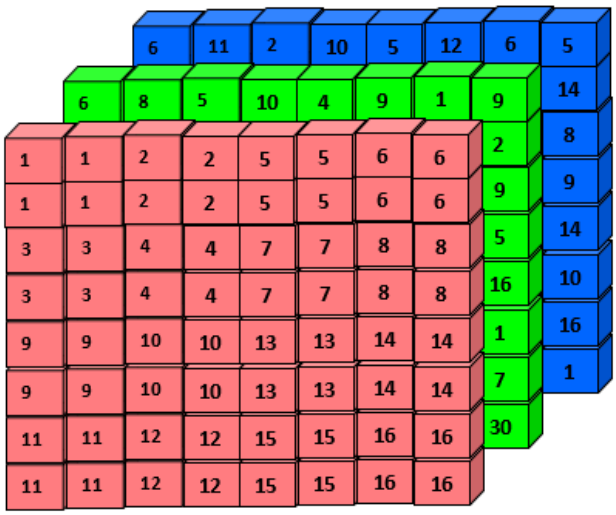

Figure 4: *The matrix R.*

Step 2: Self-attention in non-overlapped blocks.

Self-attention In sub-layers use multiple attention heads where the results in each heads are sequential and a linear transformation will be applied to the parameters in them. The first Swin Transformer (Swin-T) block module uses a regular window partitioning designing. The next block adopts a shifted window partitioning strategy, while the first regular block partitioning begins from the top-left pixel [3], [11]. Hence the 8x8 given feature map is evenly divided into 2x2 windows with 4x4 items each, as illustrated in Figure 3. Each patch is called a "token" with a size of 4x4x3=48 pixels, where 3 is for the RGB channel and 4 is the height and width of the square patch, as shown in Figure 5. In patch partition, if M=4, if matrix 8*8, the patch=64. After patch partition, the size of the image will become:

Size of image = 2 x 2 x 48, $\frac{H}{4} \times \frac{W}{4} \times C = \frac{8}{4} * \frac{8}{4} * 48 = 192$

Then, the number of channels will be converted from 48 to C through a full link layer, also known as the Linear Embedding layer. After finishing the patch, it will go through Linear Embedding [13].

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

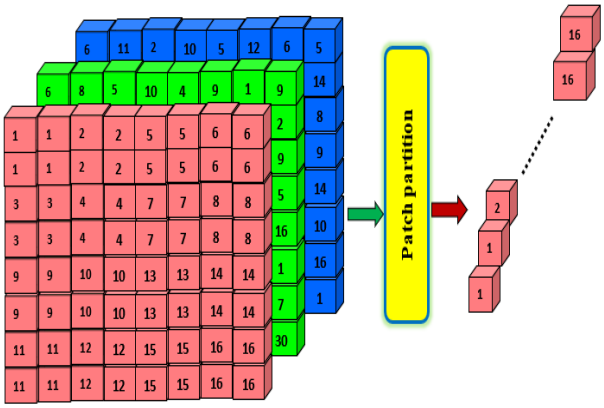

Figure 5: *Patch partition diagram.*

Step 3: The matrix R in this stage 1 is shown in Figure 6, where the number of symbols maintained in the transformer blocks is $(\frac{H1}{41} \times \frac{W1}{41})$. Figure 6 shows how each 4x4 pixel (3 channels) is flattened into 1x1 patches (48 channels). The matrix R with 8x8 pixels (3 channels) is processed in this way, then (2, 2) patches, 48 channels, and a feature map with the size of (2, 2, 48) will be received, (see the second stereogram with the green part in Figure 6). Input features (W / 8 × H / 8 × 8C1) apply a linear layer to it, so to increase the dimensions of this feature to 2 × the original dimension to ((21× the original dimension1)) (W/8×H/8×8C). To expand the accuracy of the features of entering to (W,/8×H1/8×8C) and reduce the dimension to a quarter of the input dimension (W1/8×H1/8×2C → W/4×H/4×C), the scientific order is rearranged.Patch partition and Linear Embedding are directly combined into one, a convolution core with a size of 4x4, and a stripe of 4 that is used directly to convert the number of channels from 3 to C.

Step 4: The matrix R in stage 2 is shown in Figure 7 Number of symbols reduced by a multiple of 2x21 = 4 (2x precision downsampling).The output from it is set to 2C. feature transformation used with Swin transformer blocks with the resolution remaining at: $\frac{H}{8} * \frac{W}{8} * 2C$, for matrix R $\frac{8}{8} * \frac{8}{8} * 2 * 48 = 1 * 1 * 96 = 96$.

Assuming the input patch merging is an 8x8 single-channel feature map (feature map), Patch Merging will divide each 4x4 adjacent pixel into a patch, and then divide the same position in each patch (the same Color) pixels into 4 feature maps (putting them together). The depth of the feature map is changed from C to 2C. This basic example shows that after going through the patch merging layer, the height and width of the feature map are halved, but the depth is doubled. To reduce the resolution, adjust the number of channels used, and complete the hierarchical design of the ST, merge correction was used [14]. In this case, each downsample is two, and elements are chosen at every other point in the row and column directions before being spliced together to expand, as seen in Figure 8. Next, apply the Swin transformer block to the result.

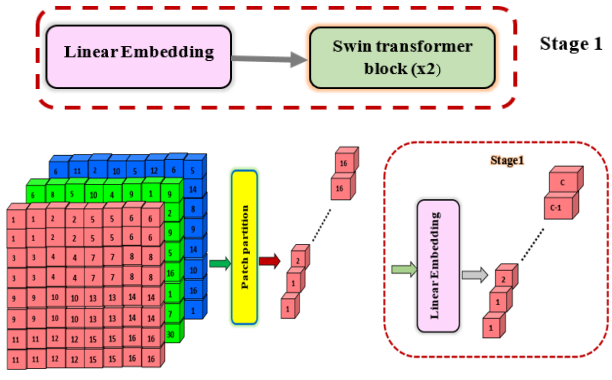

Figure 6: *Stage 1 representation.*

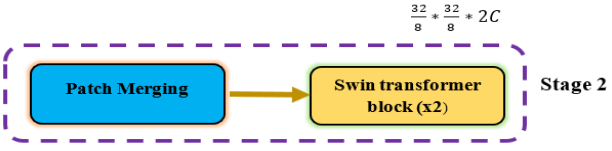

Figure 7: *Stage 2 representation*

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

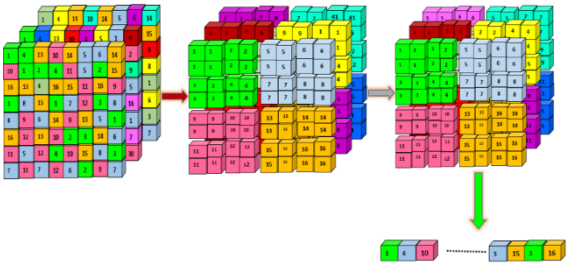

***Figure 8**: Patch merging process.*

Step 5: In the Sstage3, patch merge is applied which aggregates 2 x 2 adjacent patches, then a linear layer with 4C-D is applied and by this process the number of tokens is reduced by multiples of (6 x precision downsampling) kept at $(\frac{H}{16} * \frac{W}{16} * 4C)$ as shown in Figure 9 below, for matrix R $\frac{8}{16} * \frac{8}{16} * 4 * 48 = 4.$

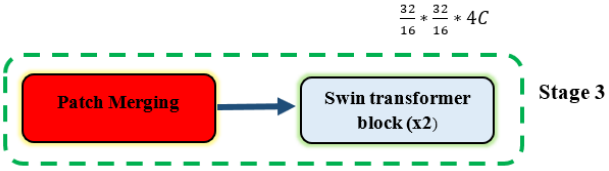

***Figure 9:** Stage 3 representation*

Step 6: In stage 4, The patch merge layer aggregates the features of each set of patches that are contiguous (2x2) and then applies a linear layer (8C-D) to the contiguous features. This decreases the number of "tokens" by a factor of (2x resolution downsampling) remaining at $\frac{H}{32} * \frac{W}{32} * 8C$ as shown in Figure 10, for matrix R $\frac{8}{32} * \frac{8}{32} * 8 * 48 = 24.$

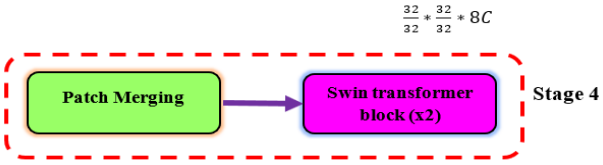

***Figure 10:** Stage 4 representation*

The stages (1, 2, 3, and 4) work together to provide a hierarchical representation on the same resolution as the feature map.

Step 7: Generation matrix for 4 blocks (R1, R2, R3, R4) as shown in Figure 11, each block with a size of 4 × 4 elements. Where block R1 take the first element in the block partitioning **[1, 2, 2, 2; 1, 1, 2, 2; 3, 3, 4, 4; 3, 3, 4, 4]**, while block R2 will take the second block from the block partitioning **[5, 5, 6, 6; 5, 5, 6, 6; 7, 7, 8, 8; 7, 7, 8, 8]**, block R3 take the third block from the block partitioning **[9, 9, 10, 10; 9, 9, 10, 10; 11, 11, 12, 12; 11, 11, 12, 12]** , and lastly, block R4 takes a block from block partitioning **[13, 13, 14, 14; 13, 13, 14, 14; 15, 15, 16, 16; 15, 15, 16, 16].**

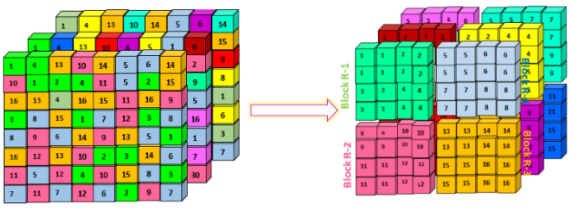

***Figure 11**: Self -attention within each block.*

Step 8: Compute the global bull attention mechanism (MSA) as in Equation 1. Its computational complexity is the same as (hw) square correlation [15].

$\Omega\ (M\ S\ A\ )=4\ h\ w\ C2+2\ (\ h\ w\ )^2C$ (1)

Where;

- h denotes the feature map height1.
- W1 is the 1feature map width1.
- C1 is the 1feature map depth1.
- M denotes the size of each window (Windows).

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

To implement on matrix R in Figure 2 with (H=8, W=8, C=48, M=4):

$$\Omega(MSA) = 4hwC2 + 2(hw)^2C \quad (1)$$

$$= 4x8x8x(48)^2 + 2(8x8)^2x48$$

$$= \mathbf{983040} \text{ (quadratic complexity)}$$

The standard transformer for vision conducts global self-attention, which has quadratic complexity concerning the number of tokens leading to intensive computational cost [16].

Step 9: Compute window1-based multi1-head1 attention mechanism (W1-MSA) as in Equation 2, because the computational complexity is huge in (*M S A)* [15]:

$$\Omega(W{-}MSA) = 4hwC2 + 2M^2hwC \quad (2)$$

To implementation on matrix A in Figure 1 with (H=8, W=8, C=48, M=4).

$$= 4x8x8x(48)^2 + 2x4^2x8x8x48$$

$$= \mathbf{688128} \text{ (linear complexity)}$$

Step 10: ShiftedR window 1partitioning in successive blocksS, see Figure 12. ST is constructed by substituting the conventional (MSAA) module in a Transformer block with a window-shifted moduleL. It adopts the configuration of windows in the next module, in which each of the previous blocks is divided, resulting in new windows in the next block, The self-attention computation in the new blocks crosses the boundaries of the previous blocks, providing connections across blocks between neighboring non-overlapping windows. Two consecutive Swin Transformer block alternates are used in both W-MSA and SW-MSA.

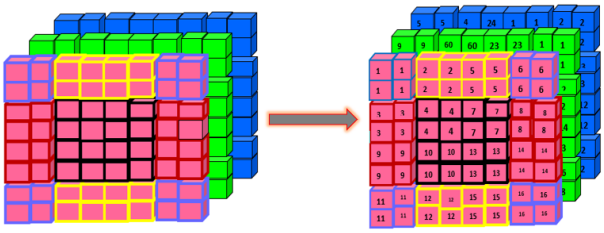

Figure 12: *Block partition.*

Where shifted block partitioning generates more windows, the windows increase by **(x 2.25).** The number of a window in the Self-attention of non-overlapped blocks is **(2×2)=4** windows while in this case is **(3×3)=4×2.25=9** windows with some of the blocks will be smaller.

Step 11: The effective batch calculation for shifted configuration. Shifted window dividing generates more windows, and some of the windows will be smaller. An efficient batch computing strategy based on cyclic shifting towards the top-left side offers a better workaround. Figure 13 shows the computing method of cyclic shifting towards the top-left side. Figure 14 shows the implementation of cyclic shifting on the matrix R**.** Where apply cyclic shifting by (-2, -2) to the matrix R to the outcome as illustrated in Figure 14.

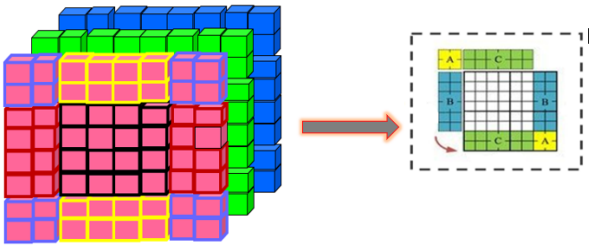

Figure 13*: Cyclic shifting computation.*

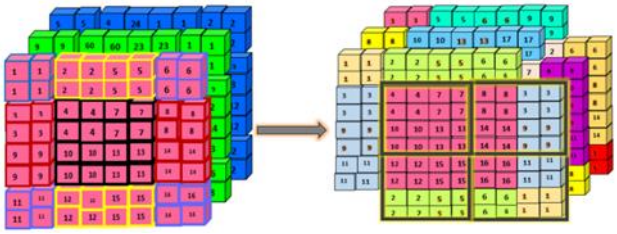

Figure 14: *The implementation of cyclic shifting on matrix R*

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

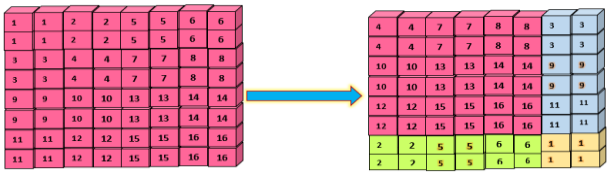

Figure 15: *The implementation of cyclic shifting by (-2, -2) to the matrix R.*

Step 12: Masked-MSA. The batched block may be formed of many sub-blocks that are not contiguous to the feature maps after the cyclic shift; hence, Within each sub-block to determine to limit "self-attention" a technique called Masking is used. The cyclic shift maintains the same numbers of the batched blocks as standard block splitting while still being computationally efficient, as seen in Figure 16. Figure 17 illustrates how to find masked-MSA from the matrix R.

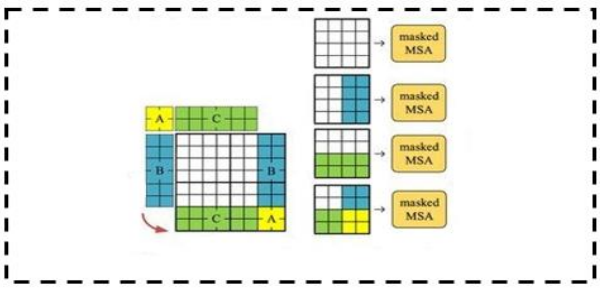

Figure 16: *Masked-MSA mechanism.*

Step 13: Compute Shifted window partitioning in successive blocks.

ST was built by adopting SW-MSA, where this unit relies on transformed windows instead of (MSA) and the rest of the layers were kept as they are, and this layer is followed by two layers of (MLP) and "GELU" where this layer is non-linear. Both the MSA layer and the MLP layer were preceded by the LN layer, and because this mechanism lacks connections across windows, this means that its modeling ability is limited as in Figure 18. Equations (3-6) state the mathematical expression of W-MSA and SW-MSA as follows [15]:

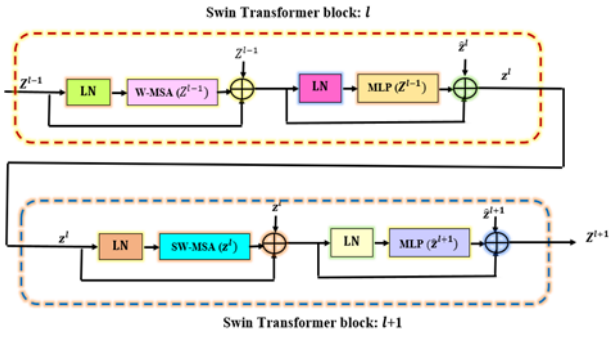

Figure 18: *An example of the Swin transformer connection.*

$$\hat{z}^{l} = W - MSA\big(LN(Z^{l-1})\big)1 + Z1^{l-1} \quad (3)$$

$$z^{l} = MLP\big(LN(Z^{l-1})\big) + \hat{z}^{l\text{|}} \quad (4)$$

$$\hat{z}^{l+1} = SW - MSA\big(LN(z^{l})\big) + z^{l} \quad (5)$$

$$Z^{l+1} = MLP\big(LN(\hat{z}^{l+1})\big) + \hat{z}^{l+1} \quad (6)$$

Where:

- $\hat{z}^{l}$: the item is in the current block.
- Z^{l-1}: the item in the previous block.
- L N: layer-norm.
- M L P: multi-layer perceptron.
- W-MSA: window self-attention.
- SW-MSA: shift window self-attention

1. To find $\hat{z}^{l}$ apply Equation (3) as shown in Figure 19.

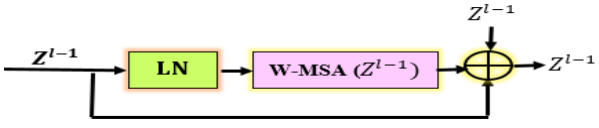

Figure 19: *The computation of the item in the current block.*

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

First, need to find the LN for (Z^{l-1}) where layer normalization (LN) is a key part of the transformer for stable training and faster convergence. Equation 7 applies the LN over each sample z∈ R_d, as shown in Figure 20 as follows [17],[18]:

$$LN(z) = \frac{z-\mu}{\delta} \circ \gamma + \beta \qquad (7)$$

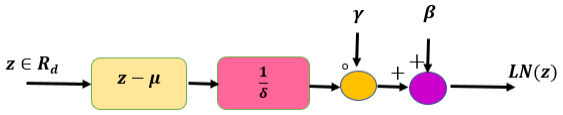

Figure 20: *The LN(z) Computation.*

Where:

- μ' ∈ R, δ ∈ R: is the feature's mean and standard deviation, respectively.
- (◦): is the "element-wise dot",
- γ' ∈ R_d, β ∈ R_d: are affine transform parameters that can be learned, where γ refers to gain (scale factor) and β refers to biases (offset) the value initialize β equals 0 value initialize γ equals 1 and if the affine transform = true.
- Now, to compute μ in LN apply the Equation 8 as follows [18]:

$$\mu = \frac{1}{m}\sum_{j=1}^{m} z_{i,j} \qquad (8)$$

- To compute δ apply Equation 9 as follows [19]:

$$\sigma = \frac{1}{m}\sqrt{\left(\sum_{j=1}^{m} z_{I,j} - \mu_{I,j}\right)^2} \qquad (9)$$

After computing the (LN) for Z^{l-1} apply the Window -Multi-Head Self Attention to compute (LN) need to detect the number of channels, mini Batch Size, and the sequence Length ($W - MSA$) as shown in step 2 to find the item in the current block (z ^ l). To implement matrix R take the block R1= [1 1 2 2; 1 1 2 2; 3 3 4 4; 3 3 4 4], the number of channels =3, min -Batch Size =4 and sequence Length=1 then the $\mu = 2.500$ and $\sigma = 1.1547$, γ =1, β= 0. Now, compute the block R1 by applying Equation 7. The result =[-1, -1, -1,-1, -1, -1, -1, -1, 1, 1, 1, 1, 1, 1, 1, 1].The same thing applies to block R2, block R3 and block R4, after that the result passed over the W-MSA as shown in step 3 where the result from W-MSA is the summation with the item in the previous block (Z^{l-1}).

2. Find z^l as shown in Figure 21. To compute z^l need to find LN for the item in the previous block (Z^{l-1}) , to implement matrix R take the block R1= [1 1 2 2; 1 1 2 2; 3 3 4 4; 3 3 4 4], the result LN(Z^{l-1})= [-1, -1, -1, -1, -1, -1, -1, -1, 1, 1, 1, 1, 1, 1, 1, 1], then, as shown in Equation 10, MPL is used between self-,attention layers for feature transformation and non-linearity [19]:

$$\mathrm{MLP(X)} = \mathrm{FC}\left(\sigma\left(\mathrm{FC(X)}\right)\right) \qquad (10)$$

$$\mathrm{FC(X)} = \mathrm{XW} + \mathrm{b} \qquad (11)$$

Where:

- W: the weight.
- b: the fully-connected layer's bias term.
- σ (·): an activation function like GELU.

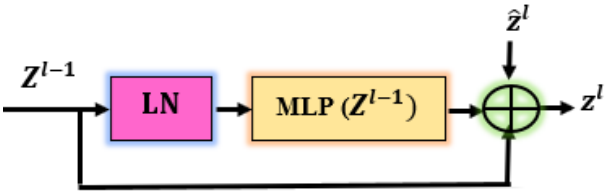

Figure 21*: Finding node in the hidden layer.*

To apply Equation 11 needs to find the first FC. When using FC, a weight matrix must be calculated before adding a bias vector. Then, as indicated in Equation 12, use the activation function GELU (gaussian error linear unit) as an activation function on the outcome of fully connected (FC) as follows [18],[20]:

$$\emptyset(x1) = 0.51x(1 + \tanh\left[\sqrt{2/\pi}\ (x_{\text{ي}} + 0.7044715x^3)\right] \qquad (12)$$

After that apply fully connect (FC) again on the result from MLP summation with $\hat{z}^l$.

3. Compute $\hat{z}^{l+1}$ by applying Equation 5 as shown in Figure 22. Finding the layer normalization (LN) first, and then applying Shift Window-Multi-head Self Attention SW-MSA to the result, as illustrated in steps (6-8).

4. Compute Z^{l+1} by applying Equation 6 as shown in Figure 23. First, find the layer normalization (LN), and then apply the Mulli Multi-Layer Perceptron (MLP) to the LN result.

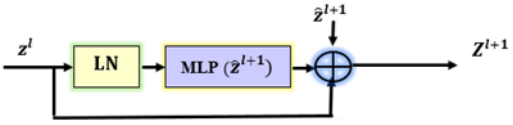

Figure 23: *Computation of* Z^{l+1}

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

Step 14: Reverse Cyclic shift. Figure 24 depicts the implementation of the reverse cyclic shift to retrieve the original matrix. Figure 25 shows the implementation of the reverse cyclic shift on the matrix resulting from the cyclic shift in step 8.

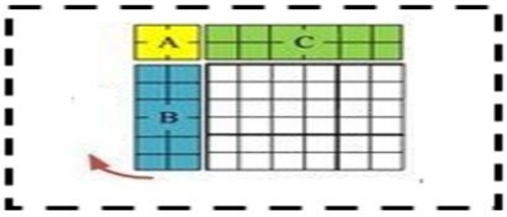

***Figure 24**: Reverse Cyclic shift implementation.*

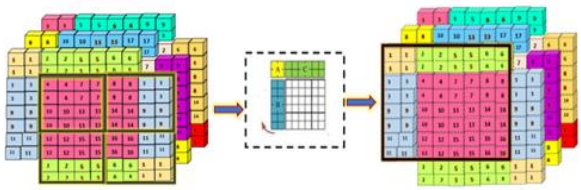

***Figure 25:** Implementation of the reverse cyclic shift on the matrix from step 8.*

5. DATASET APPLICATION OF SWIN TRANSFORMER

Vision transducers have wide applications including image recognition as well as image classification, object detection and segmentation. For the application of the Swin transducer, in this paper it is applied to two datasets.as follows:

1. CIFAR-10 dataset, The project was funded by the Canadian Institute for Advanced Research, a collection of 600 images from each of the 100 classes was gathered to be used in this work. This is referred to as the CIFARt-100b dataset. This dataset was collected using the same methods as CIFARn-10o. CIFAR-100 classesvv are mutually exclusive of CIFAR-10 classes, CIFAR-10 and CIFAR-100 are subsets of the 808 million annotated tiny image datasets [21]. CIFAR-10 and CIFAR-100n bdatasets consist of 502,000 training and 10j,000 testy images of 321×327 resolution with a total number of classes 10v and 100u, respectively [22],[21].

2. The MNIST dataset, the Modified National Institute of Standards and Technology database introduced by LeCun et al. in 1998, is a decent database for individuals consisting of 10-class “handwritten digits”. The MNIST dataset has a training set of 60,000l instances and a testv set of 10,000c examples, the image with a resolution of 28x28 and a total of x10 classes. MNIST's popularity stems from its small size [23].

6. RESULTS

The ST's shifted windowing technique improves performance by connecting via windows and by restricting the account for self-attention to local windows that are not overlapping. Many parameters must be selected before conducting the Swin transform on a dataset, such as a patch size, the number of heads, the embedding dimension, the number of multi-layer perceptrons, the size of the window, and the size of the shift. Where the parameter values in this work are ((2x2), 8, 64, 256, 2, and 1), respectively. In this paper, the Swin Transformer achieves strong performance on the CIFAR -10 and CIFAR-100 datasets. Table1 and (Figure 26 Appendix) shows the performance of the Swin transform model on the CIFAR-10 datasets. Table2 and (Figure 27 Appendix) shows the performance of the Swin transform model on the CIFAR-100 datasets. Table 3 and (Figure 28 Appendix) show the performance of the Swin transform on the MNIST dataset.

***Table 1.** The classification accuracy and loss on the CIFAR-10 dataset.*

Epoch	Train - Accuracy	Tarin - loss	Val - Accuracys	Val-loss
1	0.3325	1.9498	0.4152	1.7837
10	0.6328	1.6481	0.6254	1.3497
20	0.6852	1.2335	0.6692	1.2787
30	0.7141	1.1782	0.6946	1.2235
40	0.7288	1.1525	0.7080	1.2062
50	0.7438	1.1222	0.7272	1.1581
60	0.7561	1.0971	0.7272	1.1745
70	0.7628	1.0820	0.7112	1.1832
80	0.7718	1.0660	0.7262	1.1653

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

Table 2. *Classification accuracy and loss on CIFAR-100 dataset.*

Epoch	Train - Accuracy	Train -loss	Val - Accuracy	Val -loss
1	0.0798	4.1651	0.1288	3.9402
10	0.3442	2.9926	0.3338	3.0327
20	0.4055	2.7698	0.3896	2.8351
30	0.4380	2.6544	0.4078	2.7628
40	0.4616	2.5686	0.4184	2.7438
50	0.4788	2.5159	0.4318	2.7099
60	0.4960	2.4707	0.4362	2.6810
70	0.5048	2.4331	0.4356	2.6860
80	0.5122	2.4086	0.4452	2.6386

Through Table 1 and Table 2, It can be noted that the Swin transform performed well when applied to CIFAR-10 and CIFAR-100 datasets, and the accuracy increases in both training and validation as the number of epochs increases, while the loss is lower in both training and validation as the number of epochs increases. According to the results, the Swin transform produces better outcomes with CIFAR-10 than with CIFAR-100. Whereas the test accuracy for CIFAR-10 was 71.54%, it was 46.1% for the CIFAR-100 dataset. Table 3 indicates that when the Swin transform is applied to the MNIST dataset, the accuracy increases in both training and validation as the number of epochs increases.

Table 3. *Classification accuracy and loss on MNIST dataset.*

Epoch	Train - Accuracy	Train - loss	Val - Accuracy	Val - loss
1	0.0834	4.1474	0.1196	3.9148
10	0.3423	2.9891	0.3288	3.0287
20	0.4024	2.7730	0.3692	2.8723
30	0.4389	2.6501	0.3922	2.7989
40	0.4622	2.5711	0.4214	2.7222
50	0.4791	2.5199	0.4204	2.7206
60	0.4950	2.4699	0.4358	2.6942
70	0.5034	2.4313	0.4334	2.6717
80	0.5172	2.3921	0.4318	2.6806

6. CONCLUSION

The Swin Transform is (VT), where its computational complexity is proportional to the size of the input image, because its representation is a hierarchical representation.Furthermore, ST beats prior best techniques in COCOR objectQ identification and “ADE20K” semantic segmentation. The properties of ST make it suited for a wide range of vision applications, including dense prediction tasks and image classification, such as “object recognition” and “semantic segmentation”. According to the researchers, ST's excellent performance on many vision challenges will support the integration of vision and language signal modeling. The researchers proved that shifting Window-based self-attention diversion is effective and appropriate for visual problems and is an essential component of ST. This work used the Swin transform on a mathematical example and detailed all of the transformation processes. It was also used to classify images into many classes using three different datasets (CIFAR-10, CIFAR-100, and MNIST). The results demonstrated that the Swin transform performs well in classifying images with linear computational complexity as compared to the ViT transformer's quadratic computational complexity. By the incorporation of Swin transformer with transformation techniques, the process of selecting, combining, generating or adapting several features to efficiently solve accuracy and computation time problems. One of the motivations for studying Swin transformer is to build systems which can handle classes of problems rather than solving just one problem [23-34].

REFERENCES

[1] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, “UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-wise Perspective with Transformer,” *arXiv Prepr. arXiv2105.05537*, 2021, [Online]. Available: http://arxiv.org/abs/2109.04335.

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

[2] W. Wang, E. Xie, X. Li, and D.-P. Fan, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions arXiv:2102.12122v2," *arXiv:2102.12122v2 [cs.CV]*, 2021.

[3] J. Yang, C. Li, P. Zhang, X. Dai, and B. Xiao, "Focal Attention for Long-Range Interactions in Vision Transformers," *NeurIPS (Spotlight)*. pp. 1–21, 2021.

[4] D. Lu, J. Wang, Z. Zeng, B. Chen, S. Wu, and S.-T. Xia, "SwinFGHash: Fine-grained Image Retrieval via Transformer-based Hashing Network," *Bmvc*. 2021.

[5] H. Song, D. Sun, S. Chun, and V. Jampani, "An Extendable, Efficient and Effective Transformer-based Object Detector," *arXiv:2204.07962v1*, 2022.

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, and Y. Wei, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows arXiv:2103.14030v2," *arXiv:2103.14030v2*, 2021.

[7] L. Wang, R. Li, C. Duan, C. Zhang, and X. Meng, "A Novel Transformer based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images," *Geosci. Remote Sens. Lett.*, 2021.

[8] Y. Gu, Z. Piao, and S. J. Yoo, "STHarDNet: Swin Transformer with HarDNet for MRI Segmentation," *Appl. Sci.*, 2022.

[9] Z. Liao, N. Fan, and K. Xu, "Swin Transformer Assisted Prior Attention Network for Medical Image Segmentation," *Appl. Sci.*, 2022.

[10] J. Liang, J. Cao, G. Sun, and K. Zhang, "SwinIR: Image Restoration Using Swin Transformer," *arXiv:2108.10257v1*, 2021.

[11] S. Hao, B. Wu, K. Zhao, and Y. Ye, "Two-Stream Swin Transformer with Differentiable Sobel Operator for Remote Sensing Image Classification," *Remote Sens.*, 2022.

[12] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images." 2022, [Online]. Available: http://arxiv.org/abs/2201.01266.

[13] H. Wu, B. Xiao, N. Codella, and M. Liu, "CvT: Introducing Convolutions to Vision Transformers Haiping," *IEEE*, 2021.

[14] L. Yan, J. Huang, H. Xie, P. Wei, and Z. Gao, "Efficient Depth Fusion Transformer for Aerial Image Semantic Segmentation," *Remote Sens.*, 2022.

[15] Z. Liu *et al.*, "Video Swin Transformer." 2021, [Online]. Available: http://arxiv.org/abs/2106.13230.

[16] W. Wang, L. Yao, L. Chen, and B. Lin, "CROSSFORMER: A VERSATILE VISION TRANSFORMER HINGING ON CROSS-SCALE ATTENTION Wenxiao," *arXiv:2108.00154v2*, 2021.

[17] J. Xu, X. Sun, Z. Zhang, and G. Zhao, "Understanding and Improving Layer Normalization Jingjing," *Conf. Neural Inf. Process. Syst.*, 2019.

[18] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *arXiv:1607.06450v1*. 2016.

[19] I. Tolstikhin *et al.*, "MLP-Mixer: An all-MLP Architecture for Vision." 2021, [Online]. Available: http://arxiv.org/abs/2105.01601.

[20] J. Guo, K. Han, H. Wu, and C. Xu, "CMT: Convolutional Neural Networks Meet Vision Transformers Jianyuan," *arXiv:2107.06263v2*, 2021.

[21] J. Ahn, J. Hong, J. Ju, and H. Jung, "Rethinking Query, Key, and Value Embedding in Vision Transformer under Tiny Model Constraints," *arXiv:2111.10017v1*, 2021, [Online]. Available: http://arxiv.org/abs/2111.10017.

[22] E. In, "Lmsa: Low-Relation Mutil-Head Self- Attention Mechanism in Visual Transformer," pp. 1–11, 2022.

[23] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms," *arXiv:1708.07747v2*, 2017, [Online]. Available: http://arxiv.org/abs/1708.07747.

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

[24] Mr Hamid M Hasan, Waleed A. Mahmoud Al- Jawher, Majid A Alwan "3-d face recognition using improved 3d mixed transform" Journal International Journal of Biometrics and Bioinformatics (IJBB), Volume 6, Issue 1, Pages 278-290, 2012.

[25] Waleed A. Mahmoud, MS Abdulwahab, HN Al-Taai "The Determination of 3D Multiwavelet Transform" IJCCCE, Volume 2, Issue 4, pages 28-46 2005.

[26] Waleed Ameen Mahmoud "A Smart Single Matrix Realization of Fast Walidlet Transform" Journal International Journal of Research and Reviews in Computer Science, Volume 2, Issue, 1, Pages 144-151, 2011.

[27] Abbas Hasan Kattoush, Waleed Ameen Mahmoud Al-Jawher, Osama Q Al-Thahab "A radon-multiwavelet based OFDM system design and simulation under different channel conditions" Journal of Wireless personal communications, Volume 71, Pages 857-871, 2013.

[28] . Hadeel Al-Taai Walid Mahmoud, Mutaz Abdulwahab "New fast method for computing multiwavelet coefficients from 1D up to 3D" Proc. 1st Int. Conference on Digital Comm. & Comp. App., Jordan, Pages 412-2

[29] Abbas H Kattoush, Waleed A Mahmoud, Ali Shaheen, Ahed Ghodayyah "The performance of proposed one dimensional serial Radon based OFDM system under different channel conditions" The International Journal of Computers, Systems and Signals, Volume 9, Issue 2, Pages 412-422, 2008.

[30] Walid A Mahmoud, Majed E Alneby, Wael H Zayer "2D-multiwavelet transform 2D-two activation function wavelet network-based face recognition" J. Appl. Sci. Res, vol. 6, issue 8, 1019-1028, 2010.

[31] Waleed A Mahmoud, MR Shaker "3D Ear Print Authentication using 3D Radon Transform" proceeding of 2nd International Conference on Information & Communication Technologies, Pages 1052-1056, 2006.

[32] Waleed A Mahmoud, Afrah Loay Mohammed Rasheed "3D Image Denoising by Using 3D Multiwavelet" AL-Mustansiriya J. Sci, vol 21, issue 7, pp. 108-136, 2010.

[33] AHM Al-Heladi, W. A. Mahmoud, HA Hali, AF Fadhel "Multispectral Image Fusion using Walidlet Transform" Advances in Modelling and Analysis B, vol 52, issue 1-2, pp. 1-20, 2009.

[34] W. A. Mahmoud & I.K. Ibraheem "Image Denoising Using Stationary Wavelet Transform" Signals, Inf. Patt. Proc. & Class., vol 46, issue 4, pp. 1-18, 2003.

Appendix

The performance of the Swin transform model on (CIFAR-10, CIFAR-100, and MNIST) datasets.

Figures 26, 27, and 28 show the training and validation classification accuracy, and loss over 80 epochs for CIFAR-10, CIFAR-100, and MNIST datasets, respectively.

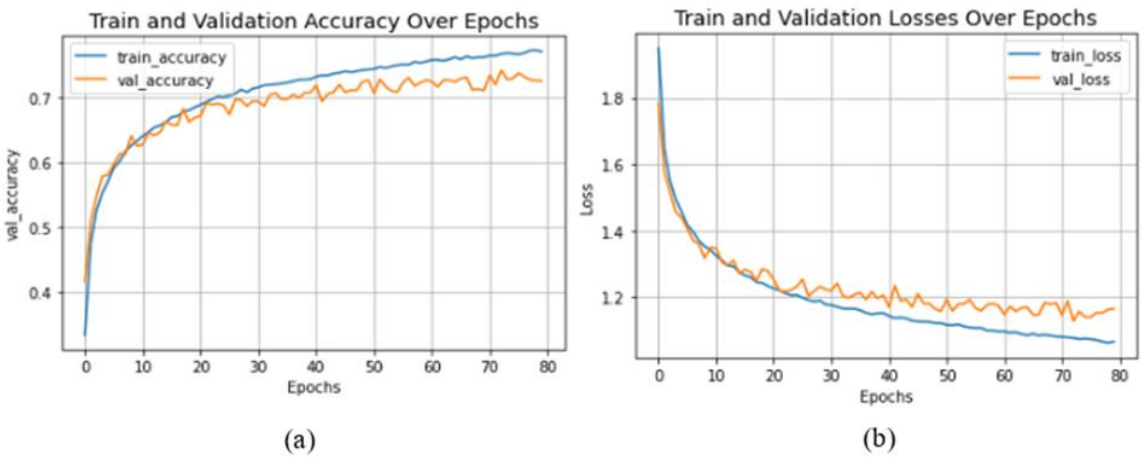

Figure 26: *Training and validation over epoch for CIFAR-10 dataset, (a) accuracy, (b) loss, (epochs=80).*

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

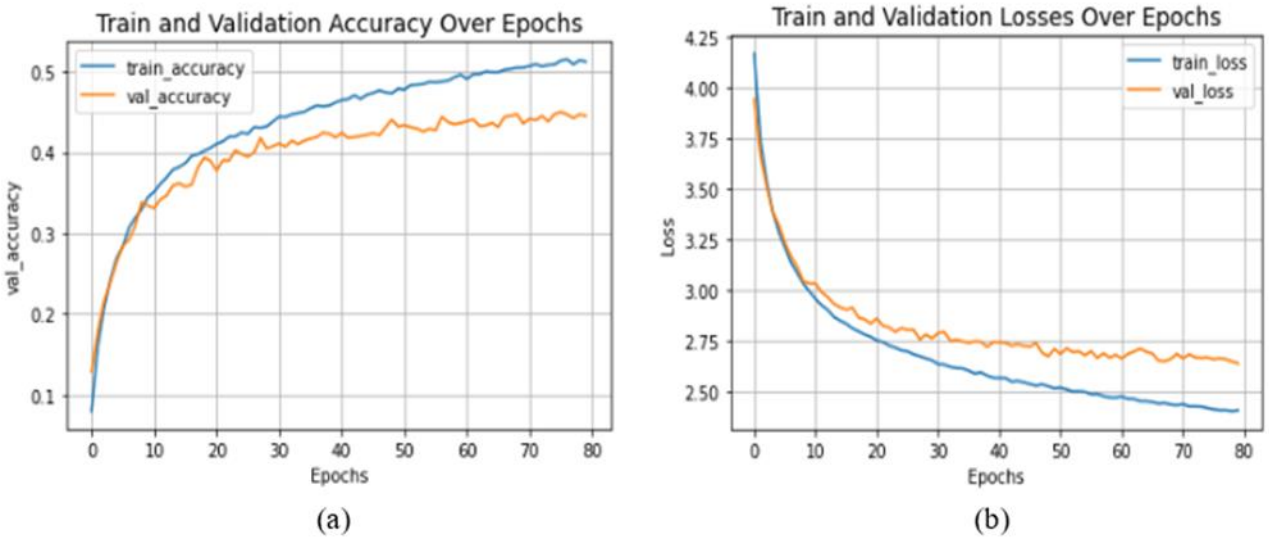

Figure 27: *Training and validation over epoch for CIFAR-100 dataset, (a) accuracy, (b) loss, (epochs=80).*

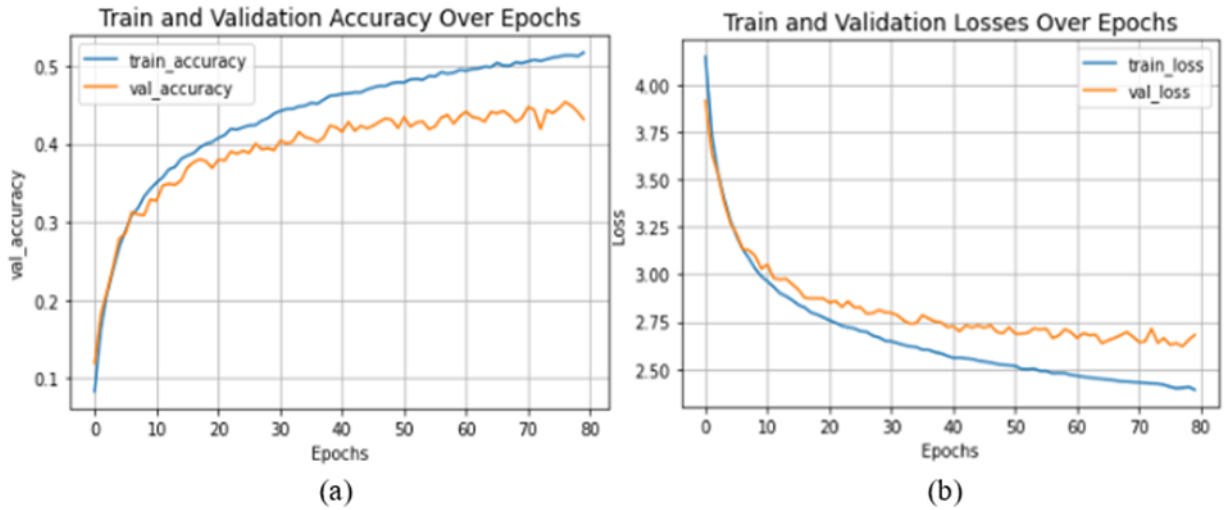

Figure 28: *Training and validation over epoch for MNIST dataset, (a) accuracy, (b) loss, (epochs=80).*

Rasha. A. Dihin, Ebtesam N. Al Shemmary, Waleed Ameen Al Jawher, 2023. Implementation Of The Swin Transformer and Its Application In Image Classification. *Journal port Science Research*, 6(4), pp. 318-331. https://doi.org/10.36371/port.2023.4.2