Salient Mask-Guided Vision Transformer for Fine-Grained Classification

Dmitry Demidov, Muhammad Hamza Sharif, Aliakbar Abdurahimov, Hisham Cholakkal and Fahad Shahbaz Khan

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, U.A.E.

Keywords: Vision Transformer, Self-Attention Mechanism, Fine-Grained Image Classification, Neural Networks.

Abstract: Fine-grained visual classification (FGVC) is a challenging computer vision problem, where the task is to automatically recognise objects from subordinate categories. One of its main difficulties is capturing the most discriminative inter-class variances among visually similar classes. Recently, methods with Vision Transformer (ViT) have demonstrated noticeable achievements in FGVC, generally by employing the self-attention mechanism with additional resource-consuming techniques to distinguish potentially discriminative regions while disregarding the rest. However, such approaches may struggle to effectively focus on truly discriminative regions due to only relying on the inherent self-attention mechanism, resulting in the classification token likely aggregating global information from less-important background patches. Moreover, due to the immense lack of the datapoints, classifiers may fail to find the most helpful inter-class distinguishing features, since other unrelated but distinctive background regions may be falsely recognised as being valuable. To this end, we introduce a simple yet effective Salient Mask-Guided Vision Transformer (SM-ViT), where the discriminability of the standard ViT's attention maps is boosted through salient masking of potentially discriminative foreground regions. Extensive experiments demonstrate that with the standard training procedure our SM-ViT achieves state-of-the-art performance on popular FGVC benchmarks among existing ViT-based approaches while requiring fewer resources and lower input image resolution.

1 INTRODUCTION

Fine-grained visual classification (FGVC) is a challenging computer vision task that aims to detect multiple sub-classes of a meta-category (e. g., car or airplane models (Krause et al., 2013; Maji et al., 2013), animal or plant categories (Horn et al., 2018; Nilsback and Zisserman, 2008), etc.). This type of image classification, compared to the traditional one (Deng et al., 2009; Zhou et al., 2014), involves larger inter-class similarity and a lack of data per class (Ding et al., 2019). This complex yet essential problem has a wide range of research and industrial applications including, but not limited to, autonomous driving, visual inspection, object search, and identification.

One of the important keys to solving the FGVC problem is to detect more distinguishable regions in an image (Wang et al., 2018; Luo et al., 2019), and the computer vision community has produced various solutions attempting to do so for the past years (Khan et al., 2011; Khan et al., 2015; Zheng et al., 2019a). The most recent achievements, based on deep neural networks (He et al., 2016), mainly represent localisation-based and attention-based methods (following (Wang et al., 2021b; He et al., 2022)). The localisation-related approaches aim to learn discriminative and interpretable features from specific regions of input images. Such approaches initially utilised properly annotated parts of an object in each image (Berg and Belhumeur, 2013; Xie et al., 2013;

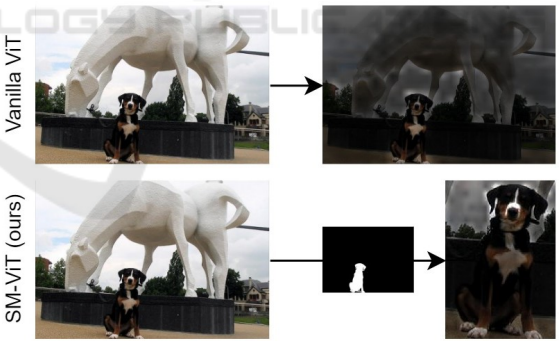

Figure 1: Visualised attention performance comparison of vanilla ViT (first row) and our SM-ViT (second row). For ViT we demonstrate the averaged attention map for the final class token. While for SM-ViT we first show the extracted saliency mask from the salient object detection module, and then the final class token's averaged attention map augmented according to this mask.

Demidov, D., Sharif, M., Abdurahimov, A., Cholakkal, H. and Khan, F.
Salient Mask-Guided Vision Transformer for Fine-Grained Classification.
DOI: 10.5220/0011611100003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 27-38
ISBN: 978-989-758-634-7; ISSN: 2184-4321

Huang et al., 2016), however, impractically laborious and costly densely annotated datasets along with the slow inference of the final model initiated more advanced techniques. Recent works on localisation-based methods (Ding et al., 2019; Ge et al., 2019), adopt a region proposal network able to predict regions potentially containing the discriminative features. This information is further passed to a backbone in order to extract features from these regions (Chen et al., 2009). A drawback of such methods is that they are tend to consider the predicted regions as independent patches (He et al., 2022), what may result in inefficiently large bounding boxes simply containing more foreground information than the other potentially more discriminative but smaller proposals. Another issue is that such extra modules often require solving an individual optimisation problem.

Recently, attention-based methods (Xiao et al., 2015; Zheng et al., 2021) leveraging vision transformer (ViT) architecture (Dosovitskiy et al., 2021) have achieved noticeable results on image classification problems. ViT considers an input image as a sequence of its patches, what allows the model to aggregate important information from the whole image at a time. A self-attention mechanism further attempts to detect the most discriminative patches, which help to automatically find the important regions in an image. This way of processing makes the model able to capture long-range dependencies beneficial for classification (Chen et al., 2021). Such an ability of the attention-based methods to efficiently learn distinctive features can also be helpful for the FGVC problem as well. Nevertheless, a recent study (He et al., 2022) investigating the performance of vanilla ViT on FGVC indicates that the class token, deciding on the final class probabilities, may pay more attention to global patches and concentrate less on local ones, which can hamper the performance in fine-grained classification. Recent ViT-based approaches for FGVC (Wang et al., 2021b; He et al., 2022) typically attempt to solve this issue by introducing an extra module, which is responsible for better segregation of class token attention by implicit distinguishing of potentially discriminative regions while disregarding the rest. However, these methods may struggle to effectively focus on more discriminative regions due to only relying on the self-attention mechanism, resulting in the classification token likely aggregating global information from less valuable background regions. Moreover, despite the accuracy improvement, such methods mainly introduce significantly more computations or trainable parameters.

Contributions. *(1)* In this work, we introduce a simple yet effective approach to improve the performance of the standard Vision Transformer architecture at FGVC. Our method, named Salient Mask-Guided Vision Transformer (SM-ViT), utilises a salient object detection module comprising an off-the-shelf saliency detector to produce a salient mask likely focusing on the potentially discriminative foreground object regions in an image. The saliency mask is then utilised within our ViT-like Salient Mask-Guided Encoder (SMGE) to boost the discriminability of the standard self-attention mechanism, thereby focusing on more distinguishable tokens. *(2)* We argue that, in the case of fine-grained classification, the most important features are in the foreground and come from the main (salient) object in an image. However, unlike some of the previous SOTA ViT-based works, we do not completely disregard the less recognisable image parts but rather guide the attention scores towards the more beneficial salient patches. *(3)* Moreover, we address the well-known problem of the immense lack of the datapoints in FGVC datasets, when classifiers often fail to find truly helpful inter-class distinguishing features, since unrelated but distinctive background regions may be falsely recognised as being valuable within the little available information provided by a training data set. Therefore, by encouraging the self-attention mechanism to pay its ”attention” to the salient regions, we simply enforce it to concentrate its performance within the main object and, therefore, to find the truly distinguishing cross-class patches. *(4)* To the best of our knowledge, we are the first to explore the effective utilisation of saliency masks in order to extract more distinguishable information within the ViT encoder layers by boosting the discriminability of self-attention features for the FGVC task. *(5)* We experimentally demonstrate that the proposed SM-ViT effectively reduces the influence of unnecessary background information while also focusing on more discriminative object regions (see Fig. 1). Our comprehensive analysis of extensive experiments on three popular fine-grained recognition datasets (Stanford Dogs, CUB, and NABirds) demonstrates that with the standard training procedure the proposed SM-ViT achieves state-of-the-art performance on FGVC benchmarks, compared to existing ViT-based approaches published in literature. *(6)* Another important advantage of our solution is its integrability, since it can be fine-tuned on top of a ViT-based backbone or can be integrated into a Transformer-like architecture that leverages the standard self-attention mechanism. The code and models are shared at: *https://github.com/demidovd98/sm-vit*.

2 RELATED WORK

2.1 Fine-Grained Visual Classification

Besides the plain feature-encoding CNN-based solutions (Yu et al., 2018; Zheng et al., 2019b; Gao et al., 2020) simply extracting high-order image features for recognition, current specific solutions for the FGVC problem are mainly related to two following groups based on a method used: localisation-based and attention-based approaches. The former aim to explicitly detect discriminative regions and perform classification on top of them, and the latter aim to predict the relationships among image regions and classify the object by this information.

Early localisation-based methods (Berg and Belhumeur, 2013; Huang et al., 2016), initially proposed for densely annotated datasets with bounding boxes for important regions, first locate the foreground object and its parts and then classify the image based on this information. Despite the relatively better performance, such solutions highly rely on the manual dense annotations including one or multiple bounding boxes per image, what makes them practically inapplicable to the real world scenarios. As a solution for this problem, in (Ge et al., 2019) the authors leverage weakly-supervised object detection and instance segmentation techniques to first predict multiple coarse regions and further choose the most distinguishable ones. In later works, (He and Peng, 2017) suggested using additional spatial constraints to improve the quality of the chosen parts, (Wang et al., 2020) presented an approach to utilise potential correlations among parts in order to select the best ones, and (Yang et al., 2021) presented a method able to first create a database of region features and then correct the class prediction by re-ranking the detected global and local information. Despite the better performance, such approaches usually require a separate, properly constructed, detection branch, which complicates the overall architecture and noticeably increases training and inference time. In addition, complete cropping of less important regions does not always increase the model's accuracy.

As an alternative, attention-based methods are able to perform both classification and localisation steps simultaneously and with no additional data, by predicting the discriminative parts inside the self-attention mechanism. For example, in (Zhao et al., 2017) authors propose leveraging visual attention to extract different attention maps and find the important information in them. A multi-level attention technique is presented in (Xiao et al., 2015), where the final model is capable of filtering out the common among classes regions. Later works (Yu et al., 2018) and (Zheng et al., 2021) demonstrate modified architectures with the integration of a cross-layer bilinear pooling mechanism and a progressive attention technique, respectively, where both aim to progressively improve the region prediction performance. Several other recent solutions for FGVC are (Yu et al., 2021; Zhao et al., 2022), which demonstrate different improvements for distinctive regions localisation, and (Behera et al., 2021; Zhu et al., 2022; Do et al., 2022; Diao et al., 2022; Sun et al., 2022), which propose complex techniques for mainly marginal performance increasing. Although these methods actually show some improvement, they mostly come with a few noticeable drawbacks: a significant computational cost or a more complex architecture, resulting in a lack of interpretability and dataset-specific solutions.

2.2 Vision Transformer

Initially discovered to process sequences of text in natural language processing (NLP) (Vaswani et al., 2017), the Transformer architecture with its self-attention mechanism has shown a great success in that field (Devlin et al., 2019; Dai et al., 2019; Tsai et al., 2019) and was later extended by researchers to computer vision (CV) tasks. After the proposal of the Vision Transformer architecture (Dosovitskiy et al., 2021), which demonstrated the SOTA performance on multiple problems, the community has been gradually exploring ViT's abilities by using it as a backbone for popular CV problems, such as image classification (Lee et al., 2021), object detection (Carion et al., 2020; Zhu et al., 2021), segmentation (Xie et al., 2021a; Wang et al., 2021a; Zheng et al., 2021) and others (Girdhar et al., 2019; Sun et al., 2020; He et al., 2021). Simple integration of the ViT architecture into other backbones and techniques has gained promising achievements and still remains SOTA for various problems. However, only a few studies investigate the properties of Vision Transformer on the FGVC problem, where the vanilla ViT model shows worse performance than its CNN counterparts.

One of the pioneers leveraging ViT on FGVC tasks is TransFG framework (He et al., 2022), which is the first to propose a solution to automatically select the distinguishable image patches and later use them for the final classification step. However, in order to achieve better results, this method uses overlapping patches, what requires significantly more resources, compared to vanilla ViT. Further proposed FFVT (Wang et al., 2021b) uses its special MAWS module for feature fusion, what makes it able to aggregate more local information from the ViT encoder

layers, what, as the authors stated, improves the original ViT feature representation capability. However, its overall idea includes selecting the patches with the most attention scores and then disregarding the other ones. This concept, based on the imperfect self-attention mechanism, may increase the negative for FGVC effect of background patches.

2.3 Mask-Guided Attention

Similar to our work, recently proposed mask-guided attention techniques, mostly based on primitive saliency models, have also demonstrated promising results in detection and re-identification tasks (Xie et al., 2021b). However, only few studies have attempted to explore its ability to be helpful for fine-grained visual classification. For example, in (Song et al., 2018) the authors suggest adding mask information as an additional input channel for a CNN in order to separately learn features from the original image and both foreground and background masked copies of it. Another work, (Wang et al., 2021c), investigates the capacity of such models when trained on more difficult patchy datasets, where the authors offer to use the predicted mask to guide feature learning in the middle-level backbone layers. In addition to that, some other early writings also discuss the ways of leveraging the saliency information to guide the learning process by simply using the mask as an extra input (Mechrez et al., 2018; Hagiwara et al., 2011). Another close to our idea recent approach is (Jiang et al., 2022), where the authors suggested using convolution kernels with different sizes for input patches. These feature activations are embedded into a ViT-like encoder in order to increase its locality and translation equivariance qualities for binary medical image segmentation problems. However, the authors do not directly use the actual saliency information, and rather assume that different activation maps are coming from different types of the same single positive class. Recently, several works (Gatys et al., 2017; Mejjati et al., 2020) suggested integration of saliency prediction as an additional component in a loss function in order to improve the training procedure. This is achieved by feeding both the image and the predicted target saliency map, so that the model is trained to produce outputs similar to the map. Another idea of leveraging saliency information in attention-based models is described in (Yu et al., 2022). The authors suggested using attention maps from each encoder layer to generate a saliency binary mask, which is then used for model pruning. All these methods indeed demonstrate that the fusion of saliency information with the main architecture's data flow can be beneficial and efficient, however they do not directly utilise the saliency and rather use it as collateral data, what results in a less noticeable improvement.

3 METHOD

3.1 Vision Transformer Framework

Following the original Transformer (Vaswani et al., 2017) made for NLP tasks, the standard Vision Transformer architecture (Dosovitskiy et al., 2021) also expects an input to be a 1D sequence of tokens. Therefore, in order to adapt it to computer vision problems, 2D input images need to be first cropped into smaller 2D patches and then flattened into 1D vectors, so that the input dimension changes are following: $(H,W,C) \rightarrow (N_p,P,P,C) \rightarrow (N_p,(P^2 * C))$, where H,W,C are initial image sizes, P is a predefined patch size, and $N_p = H * W/P^2$ is the number of such patches. Next, using a trainable linear projection, transformed patches are mapped into a latent embedding space of dimension D, which is the vector size of all tokens throughout all ViT layers. Learnable 1D position embeddings are further added to the patches in order to preserve the information about spatial relations among them. Therefore, the resulting token embedding procedure is as follows:

$$\mathbf{z}_0 = [x_{cls}; x_p^1\mathbf{E}; x_p^2\mathbf{E}; ...; x_p^{N_p}\mathbf{E}] + \mathbf{E}_{pos}, \qquad (1)$$

where $\mathbf{E} \in \mathbb{R}^{(P^2*C)\times D}$ is the trainable linear projection and $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$ is the position embedding.

As the last input preparation step, an extra learnable class token is pre-pended to the sequence, so that it can interact with the image patches, similar to (Devlin et al., 2019). This token, fed to the encoder with the sequence of embedded patches, is supposed to aggregate the information from the image tokens in order to summarise the image representation and convey it to a classification head.

In more detail, the vanilla ViT encoder component, same as in (Vaswani et al., 2017), consists of several repeating encoding layers utilising the multi-head self-attention (MSA) mechanism, MLP blocks, and both layer normalisation (LN) and residual connection techniques (Baevski and Auli, 2019; Wang et al., 2021a). More specifically, MSA used in vanilla ViT is an extension of the ordinary self-attention mechanism, represented by this equation:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (2)$$

where d_k is a scaling factor equal to the dimension number of Q,K,V, which are queries, keys, and values respectively, derived from the input patches.

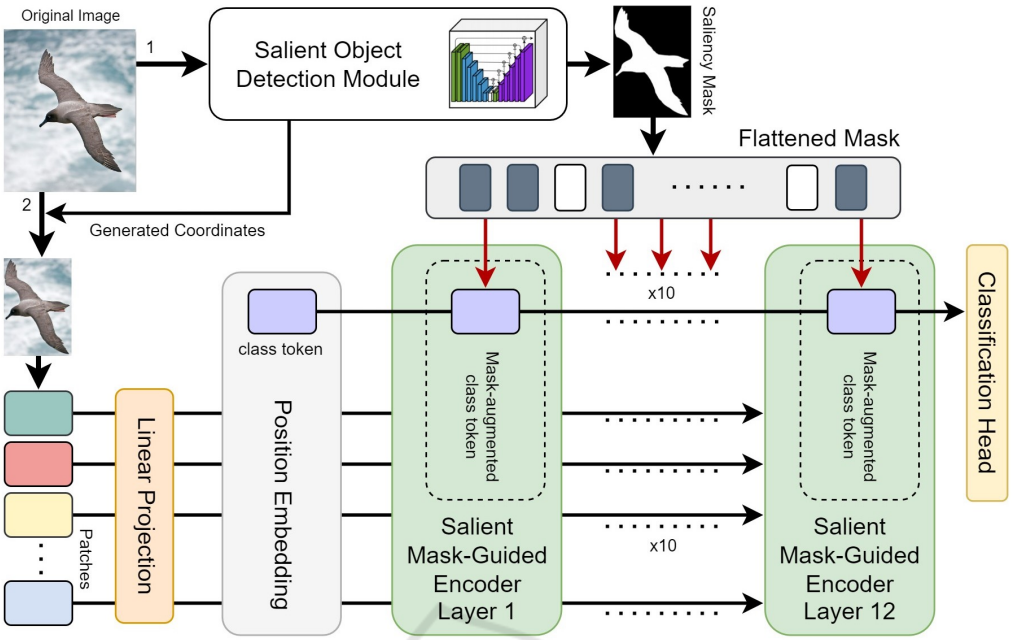

Figure 2: The overall architecture of our proposed SM-ViT. An image is first fed into the salient object detection module to extract its saliency mask and automatically generate a bounding box, which are then used to prepare a binary mask and to crop the image respectively. Further, the cropped image is fed into the ViT-like architecture, where it is first split into patches, projected into the embedding space, the positional embedding is added to the patches, and a class token is prepended. Next, the resulted sequence of tokens is passed through each layer of our Salient Mask-Guided Encoder (SMGE), where inside the multi-head self-attention mechanism the flatten binary mask is used to augment attention scores of the class token accordingly. Lastly, the class token values from the last SMGE layer are passed to a classification head to perform categorisation.

Eventually, the classification head, implemented as a multi-layer perceptron (MLP) with a hidden layer, is attached to the class token $\mathbf{z}_L^0$ in the last encoder layer L and is responsible for the final category prediction, based on the aggregated information.

3.2 Salient Mask-Guided ViT

Overall Architecture. The ViT architecture, initially designed for less fine-grained problems, is supposed to capture both global and local information, what makes it spend a noticeable part of its attention performance on the background patches (Dosovitskiy et al., 2021). This property makes vanilla ViT perform worse on FGVC tasks, since they usually require finding the most distinguishable patches, which are mostly the foreground ones. In order to resolve this issue, we propose Salient Mask-Guided Vision Transformer (SM-ViT), which is able to embed information coming from a saliency detector into the self-attention mechanism. The overall architecture of our SM-ViT is illustrated in Fig. 2.

Salient Object Detection Module. At the initial step, we utilise a salient object detection (SOD) module for saliency extraction. Our proposed method employs a popular deep saliency model, U2-Net (Qin et al., 2020), pre-trained on a mid-scale dataset for salient object detection (Wang et al., 2017). We chose this particular solution since its nested U-shaped architecture predicts saliency based on rich multi-scale features at relatively low computation and memory costs. First, an input image is passed through the SOD module set up in a test mode, which further generates the final non-binary saliency probability map. In the next phase, the model output is normalised to be within the values $[0...1]$ and then converted into a binary mask by applying a threshold d_α on each mask's pixel a_i according to the following equation:

$$a_i = \begin{cases} 1, & \text{if } a_i \geq d_\alpha \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where d_α is a pixel's intensity threshold. According to the authors' recommendations (Qin et al., 2020), we set it to 0.8 for all our experiments. Finally, the resulting binary mask and a bounding box for the found salient object(s) (in the form of the minimum and maximum 2D coordinates of the positively thresholded pixels) are extracted and saved.

An important note is that our solution also takes into account the cases when a mask is not found or is corrupted, and, if so, the initial probability map is first refined again with Eq. (3) using a threshold $d_\alpha = 0.2$,

which allows more pixels to be considered positive. If the mask is not restored even after refining, its values are automatically set as positive for the central 80 % of the image pixels.

The extracted binary mask and bounding box are further passed into our SMGE.

Salient Mask-Guided Encoder. The core module of SM-ViT is our novel Salient Mask-Guided Encoder (SMGE), which is a ViT-like encoder modified to be able to receive and process saliency information. Its main purpose is to increase the class token attention scores for the image tokens containing foreground regions. Initially, an image, cropped according to the extracted in SOD module bounding box, in a form of patches is projected into linear embeddings, and a position embedding is added to it. Next, instead of the standard ViT's encoder, our SMGE takes its place functioning as an improved self-attention mechanism. In order to understand the intuition behind our idea, we need to point out that the way of attention obtaining in the vanilla ViT encoder (refer to Eq 2) makes the background and foreground patches equally important and does not discriminate valuable for FGVC problems salient regions of the main object(s) in an image. Taking this issue into account, our solution is to increase attention scores for the patches that include a part of the main (salient) object in them. However, due to the nature of the self-attention mechanism and the non-linearity used in it, one can not simply increase the final attention values themselves, since it will break the major assumptions of the algorithm (Dosovitskiy et al., 2021). Therefore, in order to solve this problem, we apply changes to attention scores calculated right before the softargmax function (also known as softmax), according to the saliency mask provided by the salient object detection module. For this purpose, the binary mask is first flattened into a 1D vector and a value for the class token is prepended to it, so that, similar to Eq 1, the size of the resulting mask matches the number of tokens ($N_p + 1$):

$$\mathbf{m} = [m_{cls}; m_p^1; m_p^2; ...; m_p^{N_p}], \tag{4}$$

where m_{cls} is always positive since the attention of the class token to itself is considered favourable (Dosovitskiy et al., 2021). Further, a conventional attention scores matrix X_{scor} is calculated in each head:

$$X_{scor} = \frac{QK^T}{\sqrt{d_k}}V \tag{5}$$

Next, the maximum value $\mathbf{x}_{max}$ among the attention scores of the class token to each patch is found for each head. These values are further used to modify the attention scores of the class token by increasing the unmasked by $\mathbf{m}$ ones with a portion of the largest found value $\mathbf{x}_{max}$, what is calculated for every head:

$$\mathbf{x}_{scor_{cls}} = \begin{cases} x^i_{scor_{cls}} + (x_{max} * d_\theta) & \text{if } m_i = 1 \\ x^i_{scor_{cls}} & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathbf{x}_{scor_{cls}}$ is a row in the matrix of attention scores X_{scor} belonging to the class token, d_θ is a coefficient controlling the portion of the maximum value to be added, and $i \in [1, 2, ...N_p, N_p + 1]$. We provide an ablation study on the choice of coefficient d_θ in Section 4.3. Finally, following Eq 2, the rows of the resulted attention scores matrix X'_{scor}, including the modified values in its $\mathbf{x}_{scor_{cls}}$ row, are converted into probability distributions using a non-linear function:

$$Y = softmax(X_{scor}) \tag{7}$$

Eventually, similar to the multi-layer vanilla ViT encoder, the presented algorithm is further repeated at each SMGE's layer until the classification head, where the standard final categorisation is done based on the class token aggregating the information from the "highlighted" regions throughout SMGE. To summarise, our simple yet efficient salient mask-guided encoder changes the vanilla ViT encoder by modifying its standard attention mechanism's algorithm (in Eq 2) with Eq 4-7. Therefore, relatively to the vanilla ViT encoder, our SMGE only adds pure mathematical steps, does not require extra training parameters, and is not resource costly.

4 EXPERIMENTS AND RESULTS

In this section, we describe in detail the setup used for our experiments, compare the obtained results with current state-of-the-art achievements, and provide an ablation study containing quantitative and qualitative analysis. We demonstrate and explore the ability of our SM-ViT to utilise saliency information in order to improve its performance on FGVC problems.

4.1 Experiments Setup

Table 1: The details of three fine-grained visual classification datasets used for the experiments.

Dataset	Categories	Classes	Images
Stanford Dogs	Dogs	120	20,580
CUB-200-2011	Birds	200	11,788
NABirds	Birds	555	48,562

Datasets. We explore the properties of our SM-ViT on three different popular benchmarks for FGVC: Stanford Dogs (Khosla et al., 2011), CUB-200-2011

Table 2: Accuracy comparison on three FGVC datasets for our SM-ViT and other SOTA ViT-based methods. The considered methods use the ViT-B/16 model pre-trained on the ImageNet-21K dataset. All of them are then fine-tuned with the standard ViT training procedure with no overlapping patches. The best accuracies are highlighted in bold.

Method	Backbone	Stanford Dogs	CUB-200-2011	NABirds
ViT (Dosovitskiy et al., 2021)	ViT-B/16	91.4	90.6	89.6
TPSKG (Liu et al., 2022)	ViT-B/16	91.8	91.0	89.9
DCAL (Zhu et al., 2022)	ViT-B/16	-	91.4	-
TransFG (He et al., 2022)	ViT-B/16	91.9	91.5	90.3
SIM-Trans (Sun et al., 2022)	ViT-B/16	-	91.5	-
AFTrans (Zhang et al., 2022)	ViT-B/16	91.6	91.5	-
FFVT (Wang et al., 2021b)	ViT-B/16	91.5	**91.6**	90.1
SM-ViT (Ours)	ViT-B/16	**92.3**	**91.6**	**90.5**

(Welinder et al., 2010), and NABirds (Van Horn et al., 2015) (for more details, see Table 1). From the chosen datasets, Stanford Dogs and CUB-200-2011 are considered medium-sized FGVC benchmarks, and NABirds is a large-sized one. We also emphasise that, despite its size, the Stanford Dogs dataset includes images with multiple objects, artificial objects, and people. It makes the task harder for the saliency extraction module due to the pre-training objective's shift towards humans.

Baselines and Implementation Details. For all our experiments, the backbone for the classification part is Vision Transformer, more specifically, a ViT-B/16 model pre-trained on the ImageNet-21K (Deng et al., 2009) dataset with 224x224 images and with no overlapping patches. For the saliency detection module, a U2-Net model pre-trained on the DUTS-TR (Wang et al., 2017) dataset, is used with constant weights.

Following common data augmentation techniques, unless stated otherwise, the image processing procedure is as follows. For the saliency module, as recommended by the authors (Qin et al., 2020), the input images are resized to 320x320 and no other augmentations are applied. For our SM-ViT, the images are resized to 400x400 for the Stanford Dogs and CUB-200-2011 datasets and to 448x448 for NABirds (due to its higher-resolution images), without cropping for all datasets. Next, only for the training process, random horisontal flipping and colour jittering techniques are applied. The threshold d_θ in Eq. (6) is set to 0.25 for CUB and NABirds and to 0.3 for Stanford Dogs (see Section 4.3 for ablation details). All our models are trained with the standard SGD optimiser with a momentum set to 0.9 and with a learning rate of 0.03 for CUB and NABirds, and 0.003 for Stanford Dogs, all with cosine annealing for the optimiser scheduler. The batch size is set to 32 for all datasets. Pre-trained with 224x224 images ViT-B/16 weights are load from the official ViT (Dosovitskiy et al., 2021) resources. For a fair comparison, we also reimplement some of the methods with the above-mentioned preset while also following their default settings.

All our experiments are conducted on a single NVIDIA RTX 6000 GPU using the PyTorch deep learning framework and the APEX utility.

4.2 Comparison with State-of-the-Art

In this subsection, we compare the performance of our SM-ViT with other ViT-based SOTA methods on three FGVC datasets. Before discussing the results, we need to emphasise that our initial goal is to provide an improved Vision Transformer architecture which is able to perform on FGVC problems better than the original ViT and also can easily replace it in other works where it is used as a backbone. We are designing an approach to improve the baseline without additional training parameters and significant architecture changes, rather than simply providing the best-performing but hardly applicable solution "by all means". Keeping this in mind, in Table 2 we compare our SM-ViT with other officially published SOTA ViT-based approaches, which only use the ViT-B/16 backbone with no significant changes and do not require a lot of extra computations or training parameters compared to vanilla ViT. There also exist other methods which mainly use significantly more complex solutions, either requiring noticeably more training time and resources, or using more sophisticated and less popular backbones.

The results on Stanford Dogs demonstrate that besides a significant improvement of 0.9 % over vanilla ViT, our method is also superior among other approaches utilising the unchanged ViT-B/16 backbone showing a margin of 0.4 % to the second best counterpart.

It is important to mention that images in this dataset often include multiple extraneous objects (e.g. other animals, multiple categories, artificial objects) besides the main category, which may negatively affect the performance of the salient object detection

Figure 3: Visualisation of vanilla ViT and our SM-ViT results on different datasets. The first row shows original images, while the second and third rows demonstrate averaged by all heads attention maps generated by the class token at the final encoder layer of vanilla ViT and SM-ViT, respectively. Brightness intensity represents the total amount of attention, where the more attention the class token pays to a region, the brighter it is, and the other way around.

module. Nevertheless, our SM-ViT still manages to demonstrate the best result. On CUB our solution outperforms vanilla ViT by noticeable 1.0 % and also shares the Top-1 performance with another ViT-based solution, FFVT, which shows noticeably lower performance on the other datasets.

For NABirds, our method improves vanilla ViT performance by up to 90.5 %, showing a margin of 0.9 % over the predecessor and of 0.2 % over the closest counterpart.

We also emphasise that, our approach provides almost equally large performance increase on each considered dataset, what certainly demonstrates its ability to adapt to a problem and generalise better. We point out that, unlike some of the considered dataset-specific methods, our goal was not the proposal of an over-optimised on a singular dataset solution, but rather a potentially widely-applicable automatic approach with fewer heuristic parameters to adjust.

4.3 Ablation Studies

4.3.1 Effect of D_θ Coefficient

With the goal to investigate the influence of the heuristic part of our solution, we provide an ablation study on the effect of different values for hyper-parameter d_θ in Eq 6. The results of the experiments can be found in Table 3. Therefore, following the ablation results, the best d_θ value for CUB and NABirds is 0.25, and for Stanford Dogs is 0.3. One can also observe that performance is better with this coefficient value within the $0.2 - 0.3$ range, so we can suggest that for other FGVC datasets this range can be a good starting point. We assume that this is the case due to the nature of the self-attention mechanism. Making the attention scores too large compared to other patches makes it too discriminative compared to the less distinguishable and background regions, since the final difference among output values grows exponentially. In addition, one know that the attention mechanism is not perfect and can not always identify the most important regions, let alone the fact that the performance of saliency detectors is also imperfect so they may produce messy predictions. Moreover, in some cases some information about background can be especially helpful so it may be unnecessary to completely crop out the background patches.

Table 3: The effect of different values for hyper-parameter d_θ in Eq 6. Training procedure and the rest hyper-parameters remain unchanged. The best performance is highlighted in bold.

Value of d_θ	Stanford Dogs	CUB	NABirds
0.1	92.1	91.4	90.3
0.25	92.2	**91.6**	**90.5**
0.3	**92.3**	91.5	90.4
0.5	92.1	91.3	90.2
1.0	92.0	91.1	90.1

All these points make the choice of d_θ a trade-off between the performance of the utilised saliency extractor and attention-based backbone.

Table 4: The effect of our SMGE method, applied at different stages in our SM-ViT. The indicated performance is for the CUB-200-2011 dataset with the standard ViT training and validation procedure. For inference experiments, the results represent the average time per image in a batch size of 16.

Method	Img. Resolution	SMGE in training	SMGE in inference	Accuracy	Inference time, relative increase
ViT	448x448	×	×	90.6	x1.0
ViT + SMGE	448x448	×	✓	90.8	x2.2
ViT + SMGE	400x400	✓	×	91.1	x0.6
SM-ViT (Ours)	400x400	✓	✓	91.6	x1.4

4.3.2 Effect of SMGE

In Table 4 we provide the results of our SM-ViT with our SMGE module at different stages to prove that although mask cropping with the saliency mask automatically generated by the SOD module indeed improves performance, the main effect is achieved mostly by our SM-ViT method altogether. We can see that even with the SMGE module applied during training only, its ability to effectively embed saliency information into the self-attention mechanism allows the model to still have better accuracy than the original ViT has. For better understanding of this idea, the visualised outputs of two different SMGE setups are provided in Figure 4 for comparison. We can observe that SM-ViT, trained with SMGE and then utilised without SMGE during inference, still produces good-quality attention maps even without the helpful cropping and explicit attention scores augmentation techniques. It can be seen that in both cases the attention maps generated for the same images by SM-ViT are more pronounced and cover more potentially discriminative regions, compared to vanilla ViT.

Figure 4: Visual comparison of class token averaged attention maps at the last encoder layer of our SM-ViT, first fine-tuned with SMGE, and then used with disabled (first row) and enabled (second row) SMGE during inference.

It is also worth mentioning that our solution not only outperforms (or is on par with) other SOTA ViT-based methods but also does it with lower resolution inputs. In addition, according to the ablation results in Table 4, our SM-ViT is faster and still more accurate than vanilla ViT when used with SMGE disabled for inference. It becomes possible due to the fact that our solution utilises saliency masks automatically generated by the SOD module. These binary masks are used to crop the main salient object in an image before it is processed by SMGE. This noticeably reduces the original input images and, therefore, allows the model to perform well enough with lower resolutions. More specifically, our SM-ViT achieves SOTA results with images of 400x400 resolution, compared to its counterparts that require at least 448x448 inputs to achieve a lower or comparable performance. In particular, while achieving better than vanilla ViT results, our SM-ViT requires 15 % fewer computations for trainable parameters when using SMGE and 40% less inference time when it is disabled for inference, which are significant benefits for a resource-intensive ViT-based module.

Table 5: The effect of different components of our SMGE on the overall performance. The indicated accuracies are for the CUB-200-2011 dataset.

Method	Saliency Cropping	Guided Attention	Acc.
ViT	×	×	90.6
ViT + SMGE	✓	×	90.9
SM-ViT (Ours)	✓	✓	91.6

In order to understand the influence of our applied techniques inside SMGE, we provide a performance comparison with different setups in Table 5. Based on the results, we emphasise that although mask cropping actually helps to improve the accuracy, it is not the main reason for the performance increase.

4.3.3 Qualitative Analysis

To better understand the significance of our SM-ViT, we demonstrate its real outputs and compare them with vanilla ViT ones using images from different datasets. The comparison is shown in Fig 3, where the first row contains input images, while the second and

third rows demonstrate class token attention maps of vanilla ViT and our SM-ViT, respectively. The attention maps are obtained by averaging the weights of all heads at the last encoder layer, and brightness represents the amount of attention, where the brighter the regions are, the more attention the class token pays to it, and the other way around. From the results, it is noticeable that SM-ViT provides more focused on the salient object attention maps, which cover more diverse and potentially discriminative parts of the main objects (e. g., colourful feathering, beak, wings shapes, and distinguishable colour patterns). Such an ability also makes the category predictions more robust to natural augmentations and also allows the class token to detect more visually distinctive parts at the same time. From our extensive qualitative analysis of visualised attention maps obtained by random images from the datasets, we noticed that SM-ViT mostly predicts either better or similar to vanilla ViT attention maps, and only rarely it highlights visually less discriminative regions.

5 CONCLUSION

In this work, we propose a novel SM-ViT method able to improve the performance of vanilla Vision Transformer on FGVC tasks by guiding the attention maps towards potentially more important foreground objects and, therefore, reducing its "spreading" to less distinguishable regions. Our core, simple yet efficient salient mask-guided encoder boosts attention efficiency by simply utilising saliency information, does not require additional training parameters, and is relatively not resource costly. Experimental results demonstrate that with a comparable amount of resources, our SM-ViT is able to produce better than (or similar to) SOTA results among other ViT-based approaches while remaining noticeably efficient. Based on the promising results, we expect our solution to improve performance on other FGVC datasets containing classes naturally similar to the ones used for saliency module pre-training. Moreover, we emphasise that our proposed SMGE can be further extended to other popular ViT-like backbones with the conventional self-attention mechanism (e.g., DeiT (Touvron et al., 2021), Swin-T (Liu et al., 2021)). In addition, other, more powerful salient object detectors producing standard saliency maps, can be used. Therefore, we believe that SM-ViT has great potential to further boost the performance of various FGVC setups and could be a good starting point for future work.

REFERENCES

Baevski, A. and Auli, M. (2019). Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*.

Behera, A., Wharton, Z., Hewage, P., and Bera, A. (2021). Context-aware attentional pooling (cap) for fine-grained visual classification. *arXiv pre-print arXiv:2101.06635*.

Berg, T. and Belhumeur, P. N. (2013). Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 955–962.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings*, page 213–229.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv pre-print arXiv:2102.04306*.

Chen, W., Liu, T.-y., Lan, Y., Ma, Z.-m., and Li, H. (2009). Ranking measures and loss functions in learning to rank. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics*, pages 4171–4186.

Diao, Q., Jiang, Y., Wen, B., Sun, J., and Yuan, Z. (2022). Metaformer: A unified meta framework for fine-grained recognition. *arXiv pre-print arXiv:2203.02751*.

Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., and Jiao, J. (2019). Selective sparse sampling for fine-grained image recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6598–6607.

Do, T., Tran, H., Tjiputra, E., Tran, Q. D., and Nguyen, A. (2022). Fine-grained visual classification using self assessment classifier. *arXiv pre-print arXiv:2205.10529*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Gao, Y., Han, X., Wang, X., Huang, W., and Scott, M. (2020). Channel interaction networks for fine-grained image categorization. *AAAI Conference on Artificial Intelligence*, 34(07):10818–10825.

Gatys, L. A., Kümmerer, M., Wallis, T. S. A., and Bethge, M. (2017). Guiding human gaze with convolutional neural networks. *arXiv pre-print arXiv:1712.06492*.

Ge, W., Lin, X., and Yu, Y. (2019). Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3029–3038.

Girdhar, R., Carreira, J. J., Doersch, C., and Zisserman, A. (2019). Video action transformer network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 244–253.

Hagiwara, A., Sugimoto, A., and Kawamoto, K. (2011). Saliency-based image editing for guiding visual attention. *PETMEI'11 - Proceedings of the 1st International Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*.

He, J., Chen, J.-N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., and Wang, C. (2022). Transfg: A transformer architecture for fine-grained recognition. *AAAI Conference on Artificial Intelligence*, 36(1):852–860.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

He, S., Luo, H., Wang, P., Wang, F., Li, H., and Jiang, W. (2021). Transreid: Transformer-based object re-identification. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 14993–15002.

He, X. and Peng, Y. (2017). Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Horn, G. V., Aodha, O. M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2018). The inaturalist species classification and detection dataset. In *2018 Conference on Computer Vision and Pattern Recognition*, pages 8769–8778.

Huang, S., Xu, Z., Tao, D., and Zhang, Y. (2016). Part-stacked cnn for fine-grained visual categorization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1173–1182.

Jiang, Y., Xu, K., Wang, X., Li, Y., Cui, H., Tao, Y., and Lin, H. (2022). Satformer: Saliency-guided abnormality-aware transformer for retinal disease classification in fundus image. In *The 33st International Joint Conference on Artificial Intelligence*, pages 987–994.

Khan, F., Weijer, J., Bagdanov, A., and Vanrell, M. (2011). Portmanteau vocabularies for multi-cue image representation. In *Advances in Neural Information Processing Systems*, volume 24.

Khan, F. S., Anwer, R. M., van de Weijer, J., Felsberg, M., and Laaksonen, J. (2015). Compact color–texture description for texture classification. *Pattern Recognition Letters*, 51:16–22.

Khosla, A., Jayadevaprakash, N., Yao, B., and Fei-Fei, L. (2011). Novel dataset for fine-grained image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561.

Lee, S. H., Lee, S., and Song, B. C. (2021). Vision transformer for small-size datasets. *arXiv pre-print arXiv:2112.13492*.

Liu, X., Wang, L., and Han, X. (2022). Transformer with peak suppression and knowledge guidance for fine-grained image recognition. *Neurocomputing*, 492.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002.

Luo, W., Yang, X., Mo, X., Lu, Y., Davis, L. S., Li, J., Yang, J., and Lim, S.-N. (2019). Cross-x learning for fine-grained visual categorization. *2019 International Conference on Computer Vision*, pages 8241–8250.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. (2013). Fine-grained visual classification of aircraft. *arXiv pre-print arXiv:2102.04306*.

Mechrez, R., Shechtman, E., and Zelnik-Manor, L. (2018). Saliency driven image manipulation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1368–1376.

Mejjati, Y. A., Gomez, C. F., Kim, K. I., Shechtman, E., and Bylinskii, Z. (2020). Look here! a parametric learning based approach to redirect visual attention. In *ECCV 2020: 16th European Conference*, page 343–361.

Nilsback, M.-E. and Zisserman, A. (2008). Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M. (2020). U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404.

Song, C., Huang, Y., Ouyang, W., and Wang, L. (2018). Mask-guided contrastive attention model for person re-identification. In *2018 Conference on Computer Vision and Pattern Recognition*, pages 1179–1188.

Sun, H., He, X., and Peng, Y. (2022). Sim-trans: Structure information modeling transformer for fine-grained visual categorization. *arXiv pre-print arXiv:2208.14607*.

Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C., and Luo, P. (2020). Transtrack: Multiple object tracking with transformer. *arXiv pre-print arXiv:2012.15460*.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021). Training data-efficient image transformers &; distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 10347–10357.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *The 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., and Belongie, S. (2015). Building a bird recognition app and large scale dataset with citizen scientists. In *Computer Vision and Pattern Recognition (CVPR)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Wang, H., Zhu, Y., Adam, H., Yuille, A., and Chen, L. (2021a). Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5459–5470.

Wang, J., Yu, X., and Gao, Y. (2021b). Feature fusion vision transformer for fine-grained visual categorization. In *2021 British Machine Vision Conference (BMVC)*.

Wang, J., Yu, X., and Gao, Y. (2021c). Mask guided attention for fine-grained patchy image classification. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., and Ruan, X. (2017). Learning to detect salient objects with image-level supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, Y., Morariu, V. I., and Davis, L. S. (2018). Learning a discriminative filter bank within a cnn for fine-grained recognition. In *2018 Conference on Computer Vision and Pattern Recognition*, pages 4148–4157.

Wang, Z., Wang, S., Yang, S., Li, H., Li, J., and Li, Z. (2020). Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning. In *2020 Conference on Computer Vision and Pattern Recognition*, pages 9746–9755.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. (2010). Caltech-ucsd birds 200. Technical Report CNS-TR-201, Caltech.

Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., and Zhang, Z. (2015). The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850.

Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., and Luo, P. (2021a). Segmenting transparent object in the wild with transformer. *arXiv pre-print arXiv:2101.0846*.

Xie, J., Pang, Y., Khan, M. H., Anwer, R. M., Khan, F. S., and Shao, L. (2021b). Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection. *IEEE Transactions on Image Processing*, 30:3872–3884.

Xie, L., Tian, Q., Hong, R., Yan, S., and Zhang, B. (2013). Hierarchical part matching for fine-grained visual categorization. In *2013 IEEE International Conference on Computer Vision*, pages 1641–1648.

Yang, S., Liu, S., Yang, C., and Wang, C. (2021). Re-rank coarse classification with local region enhanced features for fine-grained image recognition. *arXiv preprint arXiv:2102.09875*.

Yu, C., Zhao, X., Zheng, Q., Zhang, P., and You, X. (2018). Hierarchical bilinear pooling for fine-grained visual recognition. In *2018 European Conference on Computer Vision (ECCV)*, pages 595–610.

Yu, F., Huang, K., Wang, M., Cheng, Y., Chu, W., and Cui, L. (2022). Width & depth pruning for vision transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3143–3151.

Yu, X., Zhao, Y., Gao, Y., Yuan, X., and Xiong, S. (2021). Benchmark platform for ultra-fine-grained visual categorization beyond human performance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10285–10295.

Zhang, Y., Cao, J., Zhang, L., Liu, X., Wang, Z., Ling, F., and Chen, W. (2022). A free lunch from vit: adaptive attention multi-scale fusion transformer for fine-grained visual recognition. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3234–3238.

Zhao, B., Wu, X., Feng, J., Peng, Q., and Yan, S. (2017). Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*, 19(6):1245–1256.

Zhao, Y., Yu, X., Gao, Y., and Shen, C. (2022). Learning discriminative region representation for person retrieval. *Pattern Recognition*, 121:108229.

Zheng, H., Fu, J., Zha, Z., and Luo, J. (2019a). Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5007–5016.

Zheng, H., Fu, J., Zha, Z.-J., and Luo, J. (2019b). Learning deep bilinear transformation for fine-grained image representation. In *Advances in Neural Information Processing Systems*, volume 32.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., and Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6877–6886.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, page 487–495.

Zhu, H., Ke, W., Li, D., Liu, J., Tian, L., and Shan, Y. (2022). Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4682–4692.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021). Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*.