ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder

Sangwon Kim [1] Jaeyeal Nam [1] Byoung Chul Ko [1]

Abstract

Vision transformers (ViTs), which have demonstrated a state-of-the-art performance in image classification, can also visualize global interpretations through attention-based contributions. However, the complexity of the model makes it difficult to interpret the decision-making process, and the ambiguity of the attention maps can cause incorrect correlations between image patches. In this study, we propose a new ViT neural tree decoder (ViT-NeT). A ViT acts as a backbone, and to solve its limitations, the output contextual image patches are applied to the proposed NeT. The NeT aims to accurately classify fine-grained objects with similar inter-class correlations and different intra-class correlations. In addition, it describes the decision-making process through a tree structure and prototype and enables a visual interpretation of the results. The proposed ViT-NeT is designed to not only improve the classification performance but also provide a human-friendly interpretation, which is effective in resolving the trade-off between performance and interpretability. We compared the performance of ViT-NeT with other state-of-art methods using widely used fine-grained visual categorization benchmark datasets and experimentally proved that the proposed method is superior in terms of the classification performance and interpretability. The code and models are publicly available at https://github.com/jumpsnack/ViT-NeT.

1. Introduction

Transformer structures, which have recently led to successful results in natural language processing (NLP), have also been applied to computer vision. In the same manner in which NLP divides a sentence into several words and allows the transformer to learn the association between each word, a vision transformer (ViT) (Dosovitskiy et al., 2020) learns the association between image patches. Consequently, it has been proven that the ViT structure performs better in the field of vision than a convolutional neural network (CNN) structure. However, because the ViT and CNN approaches are based on a black-box learning mechanism, there are limitations in terms of the interpretability of the model. Thus, demand is increasing for information on the model interpretability, an explanation of the results, and the model performance.

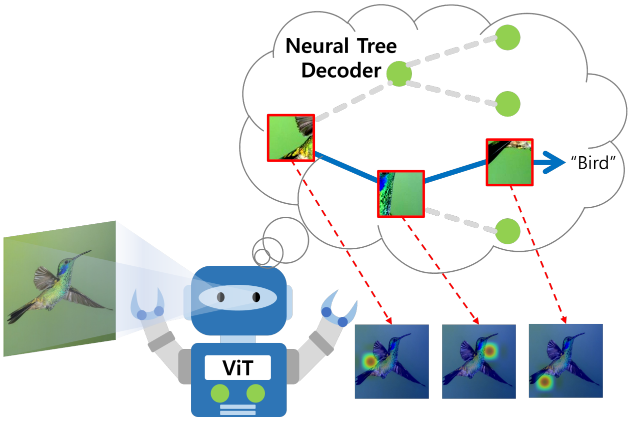

Figure 1. A ViT-NeT is a novel way of using a neural tree decoder to grant intrinsic interpretability to the ViT. In particular, the proposed NeT supports the interpretation of the sequential decision-making procedure for a given image. At each tree node, attention is concentrated on the glitter rump, green throat, and tail covert, and the image is classified as a "Bird" through a specific node path. ViT-NeT is compatible with any ViT methods and datasets for fine-grained visual classification.

In the field of computer vision, interpretability can be broadly divided into two approaches, i.e., those that provide attention (e.g., heat map and saliency) to the prediction results and those using the interpretability (transparency) of the model itself. First, attention based explainable models (Samek et al., 2017; Selvaraju et al., 2017; Chattopadhay et al., 2018; Itaya et al., 2021) provide saliency to the image pixels to explain why a predictive model classifies objects into specific classes. However, these methods have limita-

[1] Department of Computer Engineering, Keimyung University, Daegu, South Korea. Correspondence to: Byoung Chul Ko <niceko@kmu.ac.kr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

tions in that a person must infer the description of the model from the attention map again and cannot know the decision-making process for why the model emphasizes such pixels. Second, interpretable models (Chen & Guestrin, 2016; Ke et al., 2017; Dorogush et al., 2018; Lundberg et al., 2020; Kim et al., 2021) are structures in which the model itself is transparent, or the decision-making step is understandable. The interpretability of a model is closely related to its complexity and performance. From the perspective of complexity, the decision tree originally designed as an interpretable structure has an extremely simple make up, and therefore has an advantage in terms of the interpretability compared to the deep neural network (DNN). To overcome the low interpretability of DNNs while maintaining a high performance on fine-grained images, some algorithms (Chen et al., 2019a; Ji et al., 2020; Wan et al., 2021; Nauta et al., 2021) have attempted to combine DNNs with decision trees. Such algorithms provide an attention map to the user for reasons of classification as well as explanations of the classification process. However, these interpretable models still have limitations in explaining the decision-making processes, and there is a trade-off between interpretability and performance. For example, ProtoTree (Nauta et al., 2021) achieved the highest fine-grained image classification performance using an ensemble of a DNN and a decision tree; however, the interpretability was eventually lowered again owing to the learning complexity of the ensemble model.

In this study, we propose an interpretable vision transformer neural tree (ViT-NeT) that supports excellent fine-grained visual categorization (FGVC) and provides model interpretability with a simultaneous visual explanation. ViT-NeT is composed of a combination of a ViT encoder and soft neural tree decoder (NeT). Although the role of ViT is to achieve a high level classification performance using high-quality local and global features as well as the attention information, a neural tree acts as a discriminant decoder that interprets the decision-making process of the ViT and routes the images hierarchically. Therefore, ViT-NeT can largely solve the interpretability and performance trade-off that occurs in the existing ensemble models. Figure 1 shows the approximate operating structure of ViT-NeT. Figure illustrates which image patch receives attention for each node, how the image patch is branched for each node, and finally, the fine-grained classification.

This study is the first to combine ViT with neural decision trees to achieve a high interpretability and classification performance for fine-grained images. The contributions of the ViT-NeT proposed in this study are as follows:

- Unlike existing interpretable models, a high performance on fine-grained image classification is achieved competitively with state-of-the-art (SOTA) CNNs and ViTs, and an analysis is possible without a trade-off between performance and interpretability.

- The attention weight of a vanilla ViT supports only image interpretability, such as an attention map, whereas the proposed method makes the decision model itself transparent, enabling the decision-making process and image explanation at the same time.

- Molecule prototyping allows a simultaneous interpretation of the positive and negative components of data at the nodes of the neural tree.

- We present qualitative and quantitative evidence for the excellent interpretability of the decision-making of the ViT-NeT model for various test data.

2. Related Works

FGVC with ViTs Until recently, visual classification was the most fundamental field of deep-learning research; however, the performance of SOTA models has already reached its limit. Furthermore, fine-grained categorization, such as distinguishing visually similar animals and plants, is a challenging problem that goes beyond the limits of both human and machine performance. This limitation arises because most images have both a high inter-class correlation and a low intra-class correlation. Similar to other vision tasks, the FGVC task is driven by a CNN, which has achieved the highest benchmark performance. As various ViT-based methods (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021) have recently outperformed CNNs and demonstrated a SOTA performance, ViT has begun to be applied to FGVC (He et al., 2021; Zhang et al., 2021). For example, TransFG (He et al., 2021) exploited attention-based patch distillation and contrastive loss to extract robust local image patches and discriminative regions. AFTrans (Zhang et al., 2021) proposed a selective attention collection module that operates using a Siamese architecture that shares the weight parameters. However, because the attention module of the transformer is applied based on the implicit assumption that it would be extracted from a local highly discriminative part, there is a limitation in that the reliability of an attention map generated during the training process will be ambiguous.

Visual explanation with ViTs With the improving performances of CNNs, research on visual explanations is being actively conducted to secure the reliability necessary for its practical application. Thus far, visual explanation contributions in computer vision (Binder et al., 2016; Ribeiro et al., 2016; Selvaraju et al., 2017; Shrikumar et al., 2017; Lundberg & Lee, 2017; Chattopadhay et al., 2018; Itaya et al., 2021) have mostly focused on CNNs. Although ViTs are emerging as a new learning paradigm, few practical studies have been conducted on visual explanations. A common

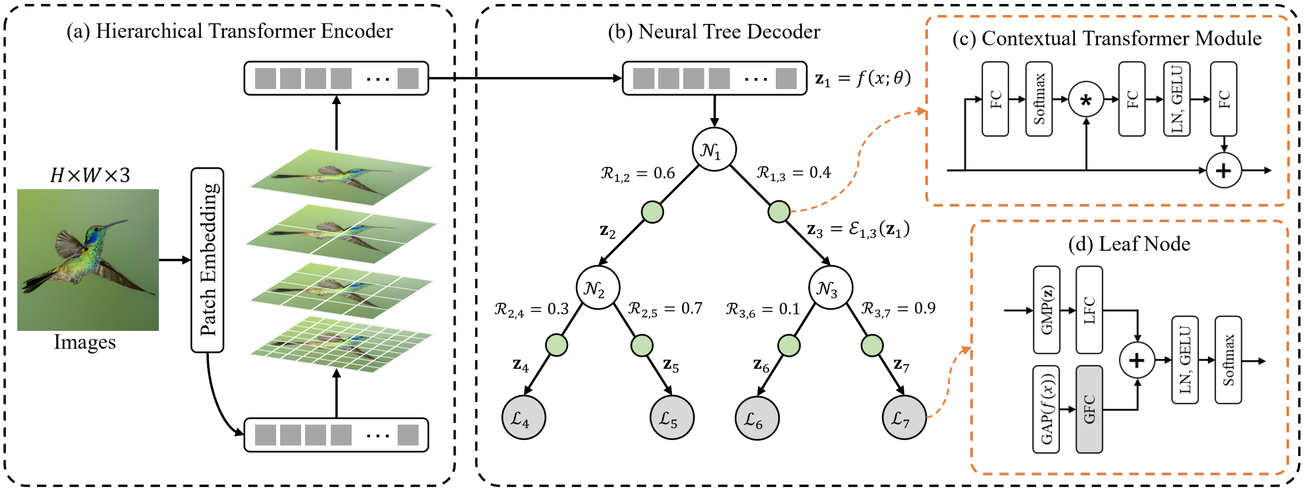

Figure 2. Overview of our ViT-NeT model formed using (a) the hierarchical transformer encoder, (b) neural tree decoder, (c) contextual transformer module, and (d) leaf node.

way to explain the output of a ViT is to visualize the attention maps. Because the attention maps created through self-attention are applied using only the query and key, there is a disadvantage in that the overall operation of the model is not considered. The major issue with the attention-based ViT methods is that the attention of the previous layer is combined nonlinearly with the attention of the next layer. To solve this problem, an attention rollout (Abnar & Zuidema, 2020) combining attention maps and assuming linearity has been proposed; however, it is difficult to avoid ambiguity in the attention maps, which may lead to incorrect correlations between image patches. Although CNNs and ViT-based visual explanations aim to emphasize a meaningful area, it is impossible to explain the decision-making process; in addition, they have limitations in that they support only a global interpretation. Therefore, for true visual explanations, we require a new model that is capable of image interpretability and at the same time can simultaneously explain the decision-making process.

Model interpretability with neural trees A decision tree is a powerful way to solve specific tasks at a single feature level. Although a decision tree architecture has intrinsic interpretability, it has a fatal flaw making it unsuitable for computer vision tasks. To overcome this hurdle, many studies (Hinton et al., 2015; Kontschieder et al., 2015; Balestriero, 2017; Tanno et al., 2019; Ji et al., 2020; Alaniz et al., 2021; Wan et al., 2021; Nauta et al., 2021) have introduced differentiable neural trees compatible with various vision tasks. ANTs (Tanno et al., 2019) exploits an adaptively configurable neural tree mechanism with trainable tree components, that is, nodes, edges, and leaves. However, it has a limited interpretation and is unsuitable for large-scale datasets. Although NBDT (Wan et al., 2021) applies a sequentially interpretable neural tree, it must use parameters induced from trained CNNs and requires WordNet (Miller, 1995) to define the interpretable tree architecture. With ProtoTree (Nauta et al., 2021), a new method is introduced for building an interpretable tree that can visualize decision-making with prototypes. Nonetheless, it has certain disadvantages in that the best performance only occurs when it is a formed ensemble, the ensemble of the tree eventually reduces the interpretability, and the training scheme is complex.

3. Approach

Although attention weights from a stack of ViT layers provide interpretable clues, existing interpretable models based on ViT (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021) only provide post-hoc explanations. Although intrinsic explanations are ideal cases for interpretable models, it is generally the case that the more interpretable the model structure is, the less accuracy it achieves. To achieve a better balance between the interpretability and accuracy of the model, we propose a novel, structurally interpretable, and visually explainable ViT-NeT model by combining a hierarchical transformer task with a neural tree decoder.

An overview of the proposed ViT-NeT is shown in Figure 2. First, we use a hierarchical vision-based transformer (Liu et al., 2021) among the ViT methods as a backbone for a feature representation. This method can consider the variations in small and large objects in an image simultaneously using a shifted window. Next, by changing the local level feature prototyping (Chen et al., 2018; Nauta et al., 2021) into a more human-friendly approach, the system was designed to improve the fine-grained classification performance while easily interpreting the decision-making process and enabling a visual explanation.

Algorithm 1 Training a ViT-NeT

```
Input: Training set $\boldsymbol{\mathcal{T}}$, number of epochs $E$, hierarchical transformer encoder $f$, depth of tree $d$, encoder parameter $\theta$, decoder parameter $\omega$

Initialize $\mathbb{T}$ = init tree($d; \omega$)
Initialize leaf node set $\mathbb{L} = \emptyset$
Get leaf node $ids$ from $\mathbb{T}$ and assign it into $\mathbb{L}$

function traverse($i, \mathbf{z}_i$)
  if $i \in \mathbb{L}$ then
    return $\mathcal{L}(\mathbf{z}_i)$
  end if
  $\mathcal{R}_{i,2\times i}(\mathbf{z}_i) = [\mathcal{N}(\mathbf{z}_i)]_0^1$
  $\mathcal{R}_{i,2\times i+1}(\mathbf{z}_i) = 1 - [\mathcal{N}(\mathbf{z}_i)]_0^1$
  $l_{distrib} = \text{traverse}(2 \times i, \mathcal{E}_{i,2\times i}(\mathbf{z}_i))$
  $r_{distrib} = \text{traverse}(2 \times i + 1, \mathcal{E}_{i,2\times i+1}(\mathbf{z}_i))$
  return $\mathcal{R}_{i,2\times i}(\mathbf{z}_i) \times l_{distrib} + \mathcal{R}_{i,2\times i+1}(\mathbf{z}_i) \times r_{distrib}$
end function

for $e \in \{1, ..., E\}$ do
  randomly split $\boldsymbol{\mathcal{T}}$ into $B$ mini-batches
  for $(x_b, y_b) \in \{\boldsymbol{\mathcal{T}}_1, ..., \boldsymbol{\mathcal{T}}_b, ..., \boldsymbol{\mathcal{T}}_B\}$ do
    $\mathbf{z}_1^b = f(x_b; \theta)$
    $\hat{y}_b = \text{traverse}(1, \mathbf{z}_1^b)$
    compute loss($\hat{y}_b, y_b$)
    update parameter $\theta$ and $\omega$ with AdamW
  end for
end for
```

Hierarchical transformer encoder ViT, which applies a transformer to image classification, has recently been proposed, and various attempts have been made to fill the gap between ViT and a CNN. Among them, the Swin Transformer (Liu et al., 2021) uses a shifted window mechanism to induce a powerful expression for dealing with the various problems of early ViT, such as inductive bias and computational complexity. The single-window approach used in conventional transformers exhibits some weaknesses in various perturbed images, which, unlike a CNN, must maintain an inductive bias. In the same way as a traditional CNN ensures robustness in many different cases, the Swin transformer inherits hierarchical feature encoding, similar to a feature pyramid, from the CNN framework. A Swin transformer uses a shift window to change the window size and process multiple patches with self-attention. Because the size of the patch included in the window varies, an attention map was created to easily detect small and large objects. The self-attention of a vanilla transformer applies a quadratic number of computations on the input image, whereas the Swin transformer conducts a linear computation on the image. This has the advantage of not only increasing the size of the model with a relatively small number of computations but also increasing the inference speed. We used the Swin transformer as a backbone model to extract the attention map features for fine-grained classification and a visual explanation of the images.

Interpretable neural tree decoder Existing interpretable methods are simple and effective but have limitations in that they provide only one option between explainability and performance. To achieve a better trade-off between accuracy and interpretability, we propose a NeT that uses a perfect binary tree and a differentiable routing mechanism. The proposed NeT consists of sets of nodes $\mathcal{N}(\cdot)$, leaves $\mathcal{L}(\cdot)$, and edges $\mathcal{E}_{i,j}(\cdot)$ between a parent node i and a child node j. Because the model uses a perfect binary tree, each internal node has two child nodes: $\mathcal{N}_{2\times i}$ and $\mathcal{N}_{2\times i+1}$. Given the encoder output $\mathbf{z}_1 = f(x; \theta)$, the proposed NeT predicts the final class labels $\hat{y}$ in a soft-decision manner. Algorithm 1 introduces the training procedure through a recursive tree traversing ViT-NeT.

The proposed NeT of the ViT-NeT model is aimed specifically at classifying objects that have similar inter-class correlations and dissimilar intra-class correlations. Although a global classification can be applied based on relatively simple global cues, fine-grained classification requires discriminant regions of subdivided categories that are highly localized. In this study, to capture a small discriminant feature, prototypes were applied to condensed image patches.

Branch routing The NeT uses prototypes to find discriminative regions among the image patches. When the discriminative feature is sampled, the routing direction is determined such that the samples are sent to the child in the differentiable routing module. Specifically, each internal node corresponds to a trainable prototype $\mathcal{P}_i \in \mathbb{R}^{H_1 \times W_1 \times D}$. Each prototype is a trainable tensor that is used to measure the squared L^2 routing score $\mathcal{N}_i \in [0, 1]$ against the reshaped image patches $\mathbf{z} \in \mathbb{R}^{H \times W \times D}$:

$$\mathcal{N}(\mathbf{z}_i) = \max_{\tilde{\mathbf{z}} \in \text{patches}(\mathbf{z}_i)} \log((\|\tilde{\mathbf{z}} - \mathcal{P}_i\|_2^2 + 1)/(\|\tilde{\mathbf{z}} - \mathcal{P}_i\|_2^2 + \epsilon)), \tag{1}$$

The routing score $\mathcal{N}(\mathbf{z}_i)$ of the i-th internal node gives the similarity between the nearest condensed image patch $\tilde{\mathbf{z}} \in \mathbb{R}^{1 \times D}$ and prototype $\mathcal{P}_i$. To define the similarity between a prototype and the image patches in each node, inspired by (Chen et al., 2018), we use a convolutional operation to measure the distance, where each prototype operates as a convolutional kernel over $\mathbf{z}$ and derives the distance between the prototype and its current $\mathbf{z}$. To induce similarity from the distance, we use a logarithmic similarity measure in Equation (1), which has a high similarity when the distance is closer to its receptive patch. Following Equation (1), we define the routing scores for each child node as

$\mathcal{R}_{i,2\times i}(\mathbf{z}_i) = [\mathcal{N}(\mathbf{z}_i)]_0^1$ and $\mathcal{R}_{i,2\times i+1}(\mathbf{z}_i) = 1 - [\mathcal{N}(\mathbf{z}_i)]_0^1$.

Contextual transformer enhancing A contextual transformer module is used to enforce the model to capture discriminative patches. Depending on the fact that the empirical receptive field is much smaller than the theoretical receptive field, the discriminative representation should be formed by a larger receptive field. In this regard, we focused on enhancing the global features into tree edges across all discretized image patches. As shown in Figure 2 (c), the contextual transformer module (CTM) provides a better description of the object by aggregating the global context to the patches of each position. Specifically, given that image patches are squeezed along the side channel dimensions to 1-dim, softmax is applied on the patch dimensions. This condensed context matrix is multiplied by the given image patches to assign global context information. Subsequently, the matrix is projected, followed by a couple of fully connected layers with GELU and layer normalization. Finally, the enhanced contextual image patches $\mathbf{z}_j$ using Equation (2) are fed into the corresponding child nodes/leaves.

$$\mathbf{z}_j = \mathcal{E}_{i,j}(\mathbf{z}_i) = \mathrm{CTM}(\mathbf{z}_i), \tag{2}$$

Label prediction Each leaf in the tree decoder corresponds to the leaf prediction module $\mathcal{L}(\cdot)$ for predicting the class probability over the K classes that need to be learned. Let $\rho(\mathbf{z}_1)$ be the accumulated routing score of the encoder output $\mathbf{z}_1$ passing from the root node to the l-th leaf node in a set of edges of a specific path p_l. The accumulated routing score, $\rho(\mathbf{z}_1)$, is calculated using Equation (3).

$$\rho(\mathbf{z}_i) = \prod_{(i,j)\in p_l} \mathcal{R}_{i,j}(\mathcal{E}_{i,j}(\mathbf{z}_i)), \tag{3}$$

Each leaf prediction module is formed through global average pooling (GAP), global max pooling (GMP), a global fully connected layer (GFC) with shared weights, a local fully connected layer (LFC), and layer normalization (LN), as shown in Figure 2 (d). The formula for the leaf prediction module is as follows:

$$\mathcal{L}(\mathbf{z}_l, x) = \mathrm{LN}(\mathrm{LFC}(\mathrm{GMP}(\mathbf{z}_l)) + \mathrm{GFC}(\mathrm{GAP}(f(x;\theta)))), \tag{4}$$

In Equation (4), whereas the leaf node learns the local view with a given enhanced context patch and LFC, the GFC fuels the leaf to learn its representation from a global contextual view.

The final prediction $\hat{y}$ is computed as the summation of all leaf predictions $\mathcal{L}$ multiplied by the accumulated routing scores ρ_l, such that

$$\hat{y} = \sum_{l\in\mathbb{L}} \sigma(\mathcal{L}(\mathbf{z}_l, x)) \cdot \rho_l(\mathbf{z}_1), \tag{5}$$

The final prediction $\hat{y}$ is optimized using a negative logarithmic likelihood loss with the ground truth label y.

4. Experiments

This section introduces the detailed setup, including the datasets and training hyper-parameters. A quantitative analysis was conducted, followed by ablation studies and qualitative analyses.

4.1. Experiment Setup

Datasets We evaluated our ViT-NeT on three FGVC datasets: CUB-200-2011 (Wah et al., 2011), Stanford Cars (Krause et al., 2013), and Stanford Dogs (Khosla et al., 2011), and compared our model with previous SOTA models in terms of accuracy and interpretability.

Implementation details ViT-NeT was implemented in PyTorch. Our hierarchical transformer encoder f contains Swin transformer layers (Liu et al., 2021) pre-trained using ImageNet-22K. First, we resized the input images to a pixel resolution of 448×448 through random cropping for training and center cropping for testing. Following the procedure of (Liu et al., 2021), we split the image into small size 4 patches and set the group window size to 14. An AdamW optimizer was employed with a momentum of 0.9. The learning rate was initialized as 2e-5 for CUB-200-2011, 2e-4 for Stanford Dogs, and 2e-3 for Stanford Cars. The batch size was set to 16. Training and testing were conducted using four NVIDIA Tesla V100 32GB GPUs with APEX.

4.2. Quantitative Analysis

Evaluation on the CUB-200-2011 Dataset The CUB-200-2011 (Wah et al., 2011) dataset consists of 11,788 images in 200 classes with 5,994 training images and 5,794 testing images. Table 1 shows the comparison results with various SOTA methods for both the CNN and ViT regimes. Compared to the best result of AFTrans (Zhang et al., 2021), the proposed ViT-NeTs show better results and outweighs all CNN-based methods on the CUB-200-2011 dataset. When the backbone network of ViT-NeT is replaced with base data-efficient image transformers (DeiT-B)(Touvron et al., 2021), the performance is degraded by 1.5% in comparison to the base Swin transformer (SwinT-B). When compared to the base transformer models DeiT and SwinT, the proposed neural tree decoder fuels an overall performance of approximately 2.5% and 3.2%, respectively. From the results, we can confirm that for the CUB-200-2011 dataset,

Table 1. Top-1 Accuracy comparison on CUB-200-2011.

Method	Backbone	Top-1 (%)
ProtoTree† (Nauta et al., 2021)	ResNet-50	82.2
STN (Jaderberg et al., 2015)	Inception	84.1
ResNet-50 (He et al., 2016)	ResNet-50	84.5
MA-CNN (Zheng et al., 2017)	VGG-19	86.5
DCL (Chen et al., 2019b)	VGG-16	86.9
TASN (Zheng et al., 2019b)	VGG-19	87.1
DFL-CNN (Wang et al., 2018)	ResNet-50	87.4
NTS-Net (Yang et al., 2018)	ResNet-101	87.9
DCL (Chen et al., 2019b)	ResNet-50	87.8
TASN (Zheng et al., 2019b)	ResNet-50	87.9
DBTNet (Zheng et al., 2019a)	ResNet-101	88.1
FDL (Liu et al., 2020)	DenseNet-161	89.1
PMG (Du et al., 2020)	ResNet-50	89.6
API-Net (Zhuang et al., 2020)	DenseNet-161	90.0
StackedLSTM (Ge et al., 2019)	GoogleNet	90.4
DeiT (Touvron et al., 2021)	DeiT-B	87.6
SwinT (Liu et al., 2021)	SwinT-B	88.4
TransFG (He et al., 2021)	ViT-B/16	90.9
AFTrans (Zhang et al., 2021)	ViT-B/16	91.5
ViT-NeT	DeiT-B	90.1
ViT-NeT	SwinT-B	**91.6**

† 224 image input

when the proposed NeT is combined with SwinT, it shows the best fine-grained classification performance compared to the CNN-based methods as well as base transformers.

Evaluation on the Stanford Dogs Dataset The Stanford Dogs dataset (Khosla et al., 2011) is a fine-grained dataset of 100 different breeds formed by 20,580 images, which is formed by 12,000 images for training and 8,580 images for testing. The evaluation results are listed in Table 2. Our ViT-NeT outperformed most of the comparison methods, particularly on the CNN backbone. Moreover, the performance of our model is on par with the attention-based variants (He et al., 2021; Zhang et al., 2021) and exceeds the accuracy of the black-box or base ViT-based methods. Base DeiT and SwinT provide a lower performance than when combined with our NeT decoder, DeiT-B+NeT and SwinT-B+NeT, i.e., 91.5% versus 93.6%, and 88.0% versus 90.3%, respectively. When our ViT-NeT model uses SwinT-B as a backbone, the top-1 accuracy is slightly diminished by 3.3% compared to the DeiT-B backbone. This is due to the problem that base SwinT's classification performance is somewhat inferior in images with a complex background. However, when combined with DeiT, it significantly outperforms the base DeiT as well as SOTA AFTrans. As a result, the proposed NeT proved that the fine-grained classification performance can be improved by using the delivered condensed image patches with a neural tree and a differentiable routing mechanism regardless of the type of ViT.

Table 2. Top-1 Accuracy comparison on Stanford Dogs.

Method	Backbone	Top-1 (%)
MaxEnt (Dubey et al., 2018)	DenseNet-161	83.6
FDL (Liu et al., 2020)	DenseNet-161	84.9
DFL-CNN (Wang et al., 2018)	ResNet-50	84.9
RA-CNN (Fu et al., 2017)	VGG-19	87.3
Cross-X (Luo et al., 2019)	ResNet-50	88.9
SEF (Luo et al., 2020)	ResNet-50	88.8
API-Net (Zhuang et al., 2020)	ResNet-101	90.3
DeiT (Touvron et al., 2021)	DeiT-B	91.5
SwinT (Liu et al., 2021)	SwinT-B	88.0
TransFG (He et al., 2021)	ViT-B/16	90.4
AFTrans (Zhang et al., 2021)	ViT-B/16	91.6
ViT-NeT	DeiT-B	**93.6**
ViT-NeT	SwinT-B	90.3

Table 3. Top-1 Accuracy comparison on Stanford Cars.

Method	Backbone	Top-1 (%)
ProtoTree† (Nauta et al., 2021)	ResNet-50	86.6
RA-CNN (Fu et al., 2017)	VGG-19	92.5
MaxEnt (Dubey et al., 2018)	DenseNet-161	93.0
DFL-CNN (Wang et al., 2018)	ResNet-50	93.1
SEF (Luo et al., 2020)	ResNet-50	94.0
FDL (Liu et al., 2020)	DenseNet-161	94.2
Cross-X (Luo et al., 2019)	ResNet-50	94.6
MMAL (Balikas et al., 2017)	ResNet-50	95.0
PMG (Du et al., 2020)	ResNet-50	95.1
API-Net (Zhuang et al., 2020)	DenseNet-161	**95.3**
DeiT (Touvron et al., 2021)	DeiT-B	92.4
SwinT (Liu et al., 2021)	SwinT-B	94.5
TransFG (He et al., 2021)	ViT-B/16	94.1
AFTrans (Zhang et al., 2021)	ViT-B/16	95.0
ViT-NeT	DeiT-B	94.7
ViT-NeT	SwinT-B	95.0

† 224 image input

Evaluation on the Stanford Cars Dataset The Stanford Cars dataset (Krause et al., 2013) contains 16,185 images in 196 classes, including 8,144 images for training and 8,041 images for testing. As shown in Table 3, the proposed ViT-NeT based on SwinT-B showed a top-1 accuracy of 95.0%. This accuracy is equivalent to that of AFTrans (Zhang et al., 2021), which achieves a SOTA performance among the ViT-based methods. The CNN-based SOTA method API-Net (Zhuang et al., 2020) employs a complex DenseNet-161 and an anti-perturbation inference design with a generative prediction strategy to exploit the discriminative features, achieving a top-1 accuracy of 95.3%. However, this method has a disadvantage in that it takes a long time to learn and test because of the complex network. By combining a NeT, ViT-NeT improves the performance of the base encoders, DeiT and SwinT. Combining proposed NeT model with DeiT-B and SwinT-B, we found that the performance improved by 2.3% and 0.5% compared to base DeiT and SwinT, respectively.

Table 4. Ablation study between pooling methods on CUB-200-2011 dataset.

Pooling	Top-1 (%)
GAP	91.4
GMP	91.6

Table 5. Ablation study on the contextual transformer module on CUB-200-2011 dataset.

Backbone	Use CTM	Top-1 (%)
DeiT-B	✗	89.3
	✓	90.1
SwinT-B	✗	90.7
	✓	91.6

In experiments on three datasets, we commonly found that performance could be significantly improved by combining NeT with base ViTs. When the Swin transformer was used as the backbone of NeT, the overall performance was high, but the Stanford Dogs dataset showed slightly lower performance. Therefore, it is necessary to improve the performance of NeT's CMT and leaf prediction module so that it can consistently achieve high performance regardless of the backbone and dataset type.

4.3. Ablation Study

Effectiveness of tree depth We graphed the performance changes using the tree depth, which is a key factor defining a tree architecture. To validate the effectiveness of the tree height, we defined five variants with different depths {3, 4, 5, 6, 7} of trees on the CUB-200-2011, Stanford Dogs and Stanford Cars datasets. As shown in Figure 3, the proposed ViT-NeT achieves the best performance when the depth of the tree is set to 4 in CUB-200-2011 dataset. In the case of a depth of 3, the model has a limited capacity, which is insufficient for representing subtle differences in the dataset. Conversely, when the depth of the tree is set to over 5, an excessive number of trainable parameters for the dataset lead the model to an overfitting, as shown in Figure 4. This issue can also be validated through a sequential analysis of the decision-making process, and it can be confirmed that the activations are derived from the overlapping characteristics of the given image.

In the case of the Stanford Cars and Stanford Dogs datasets, a significant performance was shown when the tree depth was 5 or more. In the case of the Stanford Cars dataset, there was no significant difference in the performance depending on the tree depth, and the best performance occurred at a depth of 6. Similarly, in the Stanford Dogs dataset, the best performance was achieved at a depth of 6, although the performance difference was larger with the tree depth. From the experiment results, we can see that the tree depth of NeT should differ according to the complexity of the dataset, and that increasing the depth does not improve the performance but rather leads to an overfitting.

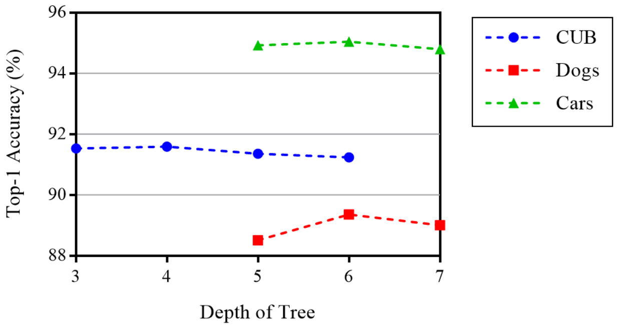

Figure 3. Top-1 Accuracy of a ViT-NeT with effect of the depth of the neural tree decoder on CUB-200-2011, Stanford Dogs and Stanford Cars.

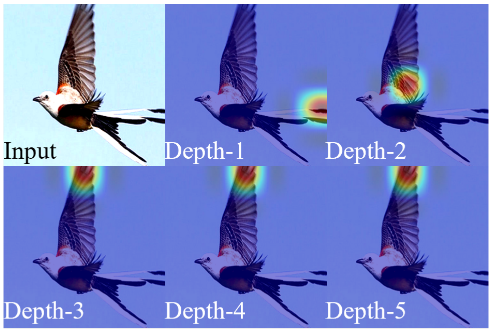

Figure 4. Prototype responses of a specific decision path in the NeT with the depth of 5. As the tree depth increases, model overfitting occurs, resulting in identical prototype responses at the tip of the bird's wing.

Effectiveness of leaf nodes Leaf nodes play an important role in projecting routed patches into subcategorical spaces. In Section 3, we suggest that each leaf node be configured with the local projection and global guidance projection. To allow the leaves to learn robust condensed patches, we considered two options: GAP and GMP. From the comparison results presented in Table 4, we found that using GMP leads to the top-1 accuracy (0.2%), beating out the GAP, on the CUB-200-2011 dataset. We identified that the GMP operation can gather more contextual information by focusing on the maximum response of the patches rather than encouraging a leaf to focus on the average responses between patches.

Effectiveness of contextual transformer module We designed a neural tree decoder using a CTM at all tree edges. We believe that the CTM enhances the global context information between discretized patches. To evaluate the

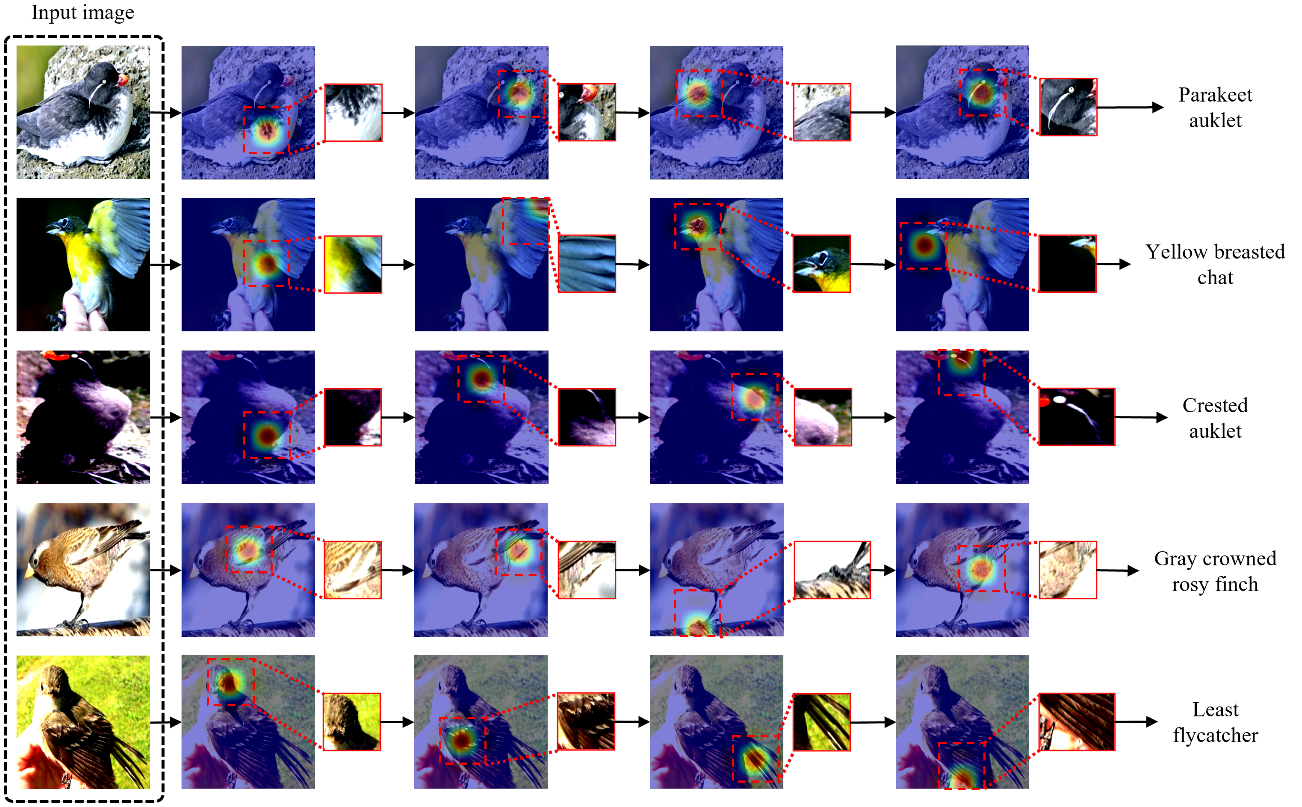

Figure 5. Visualized local interpretations showing sequential decision-making on randomly sampled images. The proposed NeT found tails, beaks, wings, feathers, claws, and eyes in the given images.

effectiveness of the CTM, we conducted a performance comparison on two backbone networks with two variants, i.e., *w*/CTM and *wo*/CTM. As shown in Table 5, when using DeiT as the backbone with the CTM of ViT-NeT, the top-1 accuracy is increased from 89.3% to 90.1%. Similarly, when SwinT is used as the backbone with the CTM of ViT-NeT, the top-1 accuracy increases from 90.7% to 91.6%. We confirm that the CTM can fuse various types of global context information within the condensed image patches for achieving a better performance.

4.4. Qualitative Analysis

We present an intrinsically interpretable NeT method. To validate the interpretability of the decision-making, we provide qualitative evaluations that show visualized sequential decision-making against randomly sampled images. Figure 5 shows the validation results of the decision path for a given image. To visualize a practical decision-making procedure, we redefine the decision routing in a hard decision manner, which greedily feeds condensed image patches to a specific child node with a larger routing score. As shown in this figure, The proposed NeT finds local discriminative regions in a human-friendly manner. Similar to (Chen et al., 2018; Nauta et al., 2021), some results attempt to interpret the background. Owing to the nature of the dataset, it is difficult to state that it is completely incorrect because the species of bird may be divided according to the habitat. For example, because auklets live in the sea, there may be a lot of water or stones in the background, whereas flycatchers mainly feed on insects, and thus they may be surrounded by trees or grass. We found that the results of the prototype trained at the root node from the proposed model are somewhat different from the results of the leaf. In the root node, there are cases in which the prototype mainly focuses on the color or texture, which may divide a large feature space into sub-decision regions owing to the nature of the binary tree.

5. Conclusion

A CNN and ViT, which are leading SOTA approaches in image classification, both have a flaw in that they are black-box models that cannot clearly explain or interpret the prediction results. The ViT-NeT proposed in this paper uses the ViT backbone to extract high-quality local and global features as well as attention information, and applies a new NeT designed to present the decision-making and image explanation for the classification process. Although existing fine-grained visual categorization methods have a trade-off between interpretability and performance, the proposed ViT-NeT achieves a newly demonstrated SOTA performance and excellent interpretability on the benchmark dataset without any trade-offs.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(2021R1A6A3A13039438) and (2022R1I1A3058128).

References

Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

Alaniz, S., Marcos, D., Schiele, B., and Akata, Z. Learning decision trees recurrently through communication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13518–13527, 2021.

Balestriero, R. Neural decision trees. *arXiv preprint arXiv:1702.07360*, 2017.

Balikas, G., Moura, S., and Amini, M.-R. Multitask learning for fine-grained twitter sentiment analysis. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pp. 63–71. Springer, 2016.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847, 2018.

Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., and Rudin, C. This looks like that: deep learning for interpretable image recognition. *arXiv preprint arXiv:1806.10574*, 2018.

Chen, H., Zhang, H., Boning, D., and Hsieh, C.-J. Robust decision trees against adversarial examples. In *International Conference on Machine Learning*, pp. 1122–1131. PMLR, 2019a.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Chen, Y., Bai, Y., Zhang, W., and Mei, T. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5157–5166, 2019b.

Dorogush, A. V., Ershov, V., and Gulin, A. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Du, R., Chang, D., Bhunia, A. K., Xie, J., Ma, Z., Song, Y.-Z., and Guo, J. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, pp. 153–168. Springer, 2020.

Dubey, A., Gupta, O., Raskar, R., and Naik, N. Maximum-entropy fine grained classification. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/0c74b7f78409a4022a2c4c5a5ca3ee19-Paper.pdf`.

Fu, J., Zheng, H., and Mei, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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

He, J., Chen, J.-N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., Wang, C., and Yuille, A. Transfg: A transformer architecture for fine-grained recognition. *arXiv preprint arXiv:2103.07976*, 2021.

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

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Itaya, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H., and Sugiura, K. Visual explanation using attention mechanism in actor-critic-based deep reinforcement learning. *arXiv preprint arXiv:2103.04067*, 2021.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

Ji, R., Wen, L., Zhang, L., Du, D., Wu, Y., Zhao, C., Liu, X., and Huang, F. Attention convolutional binary neural tree for fine-grained visual categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10468–10477, 2020.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

Khosla, A., Jayadevaprakash, N., Yao, B., and Li, F.-F. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011.

Kim, S., Jeong, M., and Ko, B. C. Lightweight surrogate random forest support for model simplification and feature relevance. *Applied Intelligence*, pp. 1–11, 2021.

Kontschieder, P., Fiterau, M., Criminisi, A., and Bulo, S. R. Deep neural decision forests. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1467–1475, 2015.

Krause, J., Deng, J., Stark, M., and Fei-Fei, L. Collecting a large-scale dataset of fine-grained cars. 2013.

Liu, C., Xie, H., Zha, Z.-J., Ma, L., Yu, L., and Zhang, Y. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11555–11562, 2020.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777, 2017.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

Luo, W., Yang, X., Mo, X., Lu, Y., Davis, L. S., Li, J., Yang, J., and Lim, S.-N. Cross-x learning for fine-grained visual categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Luo, W., Zhang, H., Li, J., and Wei, X.-S. Learning semantically enhanced feature for fine-grained image classification. *IEEE Signal Processing Letters*, 27:1545–1549, 2020. doi: 10.1109/LSP.2020.3020227.

Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Nauta, M., van Bree, R., and Seifert, C. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14933–14943, 2021.

Ribeiro, M. T., Singh, S., and Guestrin, C. ”why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144, 2016.

Samek, W., Wiegand, T., and Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2017.

Tanno, R., Arulkumaran, K., Alexander, D., Criminisi, A., and Nori, A. Adaptive neural trees. In *International Conference on Machine Learning*, pp. 6166–6175. PMLR, 2019.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wan, A., Dunlap, L., Ho, D., Yin, J., Lee, S., Petryk, S., Bargal, S. A., and Gonzalez, J. E. NBDT: neural-backed decision tree. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=mCLVeEpplNE.

Wang, Y., Morariu, V. I., and Davis, L. S. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., and Wang, L. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Zhang, Y., Cao, J., Zhang, L., Liu, X., Wang, Z., Ling, F., and Chen, W. A free lunch from vit: Adaptive attention multi-scale fusion transformer for fine-grained visual recognition. *arXiv preprint arXiv:2110.01240*, 2021.

Zheng, H., Fu, J., Mei, T., and Luo, J. Learning multi-attention convolutional neural network for fine-grained image recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5209–5217, 2017.

Zheng, H., Fu, J., Zha, Z.-J., and Luo, J. Learning deep bilinear transformation for fine-grained image representation. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/959ef477884b6ac2241b19ee4fb776ae-Paper.pdf.

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

Zhuang, P., Wang, Y., and Qiao, Y. Learning attentive pairwise interaction for fine-grained classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13130–13137, 2020.