Comparison between transformers and convolutional models for fine-grained classification of insects

Rita Pucci[1] *Vincent J. Kalkman*[1] *Dan Stowell*[1],[2]

[1] *Naturalis Biodiversity Center, Leiden(NL)*
[2] *Tilburg University, TSHD, Tilburg(NL)*
* *E-mail: rita.pucci@naturalis.nl*

Abstract: Fine-grained classification is challenging due to the difficulty of finding discriminatory features. This problem is exacerbated when applied to identifying species within the same taxonomical class. This is because species are often sharing morphological characteristics that make them difficult to differentiate. We consider the taxonomical class of Insecta. Accurate identification of insects is essential in biodiversity monitoring as they are one of the inhabitants at the base of many ecosystems. Citizen science is doing brilliant work of collecting images of insects in the wild giving the possibility to experts to create improved distribution maps in all countries. Today, we have billions of images that need to be automatically classified and deep neural network algorithms are one of the main techniques explored for fine-grained tasks. At the state of the art, the field of deep learning algorithms is extremely fruitful, so how to identify the algorithm to use? In this paper, we focus on Odonata and Coleoptera orders, and we propose an initial comparative study to analyse the two best-known layer structures for computer vision: transformer and convolutional layers. We compare the performance of T2TViT_14, a model fully transformer-base, EfficientNet_v2, a model fully convolutional-base, and ViTAEv2, a hybrid model. We analyse the performance of the three models in identical conditions evaluating the performance per species, per morph together with sex, the inference time, and the overall performance with unbalanced datasets of images from smartphones. Although we observe high performances with all three families of models, our analysis shows that the hybrid model outperforms the fully convolutional-base and fully transformer-base models on accuracy performance and the fully transformer-base model outperforms the others on inference speed and, these prove the transformer to be robust to the shortage of samples and to be faster at inference time.

1 Introduction

Fine-grained classification task aims to differentiate between classes that belong to the same superclass. In the biodiversity monitoring field, we talk about species as classes and order as a superclass. The species are usually defined by domain experts (taxonomists) based on morphology or molecular data. The species within an order are closely related and look-alike, i.e. many of them shared colours and characteristics, therefore the fine-grained tasks in this field are particularly challenging. In this paper, we focus on two orders of Insecta: Coleoptera and Odonata in Europe. In the order Coleoptera, there are four main suborders: Archostemata, Myxophaga, Adephaga, and Polyphaga, and more than 130,000 species in Europe [4]. The order of Odonata has two main suborders: Epiprocta, and Zygoptera and it is estimated that, in Europe alone, we have more than 200 subspecies. The Coleoptera and Odonata are among the oldest insects with many important roles in our environment but still, they are under-studied. One of the main reasons is the complexity involved in the identification process. In both orders, many species are phenotypically similar in appearance both within and between families. In Coleoptera, many species are small with discriminating characters often difficult to see. A further factor complicating identification is the within-species variability due to differences between life stages, sexes and regional or seasonal variation. In Fig. 1, we present some samples from the datasets of interest in this paper. In the first row, all five images are Coleoptera, three of which are from the same family of Coccinellidae but even if they resemble each other to an inexpert eye, they belong to different species. On the second row, we present samples from the order Odonata, as we note the subjects in the images all show a similar shape and dimension and sometimes similar colour, but they all belong to different species. The ability to identify the insects that inhabit the ecosystems is one of the main steps to understanding them. Despite its significance, the fine-grained task in biodiversity has posed two key challenges: 1) The inter-class variances are often extremely subtle, thus requiring highly discriminative representation for effective classification; 2) As the rarity of the species increases, there are fewer training samples per category, impeding the performance of large-data favoured methods. The conventional identification technique is to cross-validate the image with the regional field guides, online sources, and field experts. Since these identification methods are highly time-consuming and unaffordable for the common person, there is increasing interest in the investigation of new deep learning fine-grained methods for biodiversity monitoring. Early and fast identification techniques are crucial and the fast-developing of deep learning technologies in computer vision have shown impressive solutions to many real-world problems such as animal identification [24]. At the state of the art, the convolutional neural network (CNN) for computer vision is an algorithm based on an inductive bias of locality and shift invariance these two main features make CNN a highly effective deep learning algorithm in image classification. Recently, we see an increased interest in the application

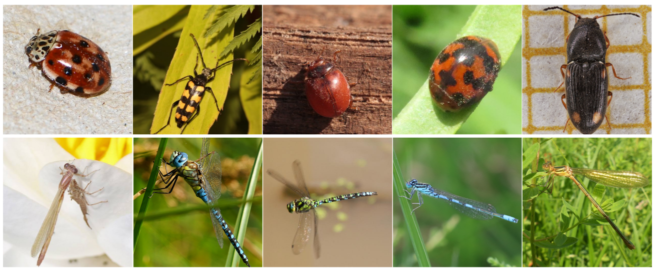


Fig. 1: Samples from datasets of Coleoptera and Odonata for fine-grained classification.

of transformers for the same tasks to which CNN was historically devoted. Vision transformer (ViT) [10] enables multi-head self-attention to capture long-range dependencies within an image and thus can extract diverse feature patterns for discriminative classification. Unfortunately, ViT is data-hungry and the lack of training data may impede its application in fine-grained tasks. With their pro and cons, both the convolutional and the transformer algorithms are good candidates for fine-grained tasks for insect images but to which extent?

In this paper, we are interested in understanding and evaluating the CNN and the ViT comparing and analysing them in the context of fine-grained tasks in biodiversity monitoring. We consider three of the main families of deep neural networks for computer vision: fully-convolutional, fully-transformer, and hybrid (based on both convolutional and transformer layers). For each of them, we select a model at the state of the art that obtains the best performance in image classification: EfficientNet_v2 [21](`EffNetv2`) for fully-convolutional, T2TViT_14 [29](`T2TViT14`) for fully-transformer, and ViTAEv2 [32](`ViTAEv2`) for the hybrid one. For training and validation, we consider datasets collected by citizen science and stored in Observation.org [1]. We evaluate the models on iNaturalist [3] and Artportalen [2] limited to Odonata and Coleoptera from Europe, which are collected from different communities of citizen science than training. The results are presented to address the fine-grained task at the species and the morph/sex levels.

2 Related work

In the domain of Insecta extensive work has been done to identify different species in different orders e.g. Lepidoptera, Coleoptera, Odonata, Orthoptera, and Hymenoptera. In particular, the application of deep learning algorithms such as CNN has seen increased popularity for the ability of automated feature extraction and high accuracy rate in fine-grained classification. CNN is now popularly used for insect identification and presents a wide range of models applied to classify Lepidoptera [5, 9, 22] which reach high performance in accuracy. Customized models are proposed for generic species from different orders in the class Insecta [18, 28], also specifically to classify bees in real time [7], Orthoptera for mobile application [6], and Odonata [23]. Even if we have a prolific application of dedicated CNN in fine-grained tasks for insects, these models are not robust in the identification of rare species, resulting in enormous limitations in practical applications. It is still an open challenge that requires investigation. Moreover, we do not observe equal interest in the application of transformer-based models in this task. An interesting comparison between very simple CNN and transformer-based algorithms for fine-grained tasks among species of different kingdoms identifies the ViT model as outperforming the CNN-based models [20]. A customized transformer model is proposed for insect pest recognition highlighting the need in integrating some of the CNN features into the transformer structure making the model focus more on global coarse-grained information rather than local fine-grained information [26].

Though a vast amount of work has been done in the domain of insect identification, to the best of our knowledge and extensive literature survey, we have not found any published research on a comparative evaluation of deep learning models from all three families of deep neural networks for fine-grained identification in biodiversity monitoring. Furthermore, there is no experimentation on the most modern models from computer vision for this task.

3 Methods

3.1 Models

We first define the three families of models that we are considering for image classification, and we select for each family the latest algorithm with the best performance for each family based on the ImageNet classification task. For a fair comparison, none of these models are specialised for fine-grained tasks.

Fully-convolutional: We consider models that are mainly based on the convolutional layers and fully connected layers. We are interested in models that are competitive with ViT in inference speed, and model size. We choose EfficientNet_ v2_ medium (hereafter named as `EffNetv2`) [21] for this family of models. The `EffNetv2` has the structure and connections optimised for speed, based on floating point operations per second (FLOPs), and for parameter efficiency and this model represents a good competitor to the transformer-based models. In particular, `EffNetv2` consists of convolutional-based layers [15, 16] to better utilise mobile or server accelerators. The CNN models are naturally equipped with intrinsic inductive bias, shift-invariance, and hierarchical structure to extract multi-scale features and locality. These are proper advantages in extracting representative features from images collected with smartphones. Even if CNN models are commonly used as the backbone of many image classification models at the state of the art, they are not well suited to model long-range dependency due to their structure focused on extracting local features from low level to high level progressively. This can affect the performance in the fine-grained tasks: these models are less inclined to identify relations among details of the subject. The details are typically the characteristics used by taxonomists to distinguish species.

Fully-transformer: Models based only on attention [25] are here referred to as fully-transformer models. The Vision Transformer (ViT) [10] is the first fully-transformer model applied for image classification. ViT demonstrates that transformers are promising for vision tasks. In fact, ViT is based on the self-attention mechanism which allows the model to capture global contextual information, enabling it to learn long-range dependencies and relationships between image tokens (patches). The self-attention mechanism weighs the importance of different tokens in the sequence when processing the input data. In this paper, we consider for comparison a recent evolution of ViT, the Token-2-Token ViT 14 [29] (`T2TViT14`) which uses a progressive tokenization module to aggregate neighbouring tokens into one token. In the first layer, a token is a patch of the image, while in the intermediate layers, a token is a patch of the feature maps. The model is able to extract local information reducing the length of the token iteratively. This architecture reduces the data hunger and boosts the performance relative to the vanilla ViT.

Hybrid: Finally we consider hybrid models, which create a collaboration between the convolutional and the transformer layers. In particular, we consider the `ViTAEv2` [32] which implements inductive bias and the scale-invariance properties into a transformer architecture. To obtain such a result, the algorithm exploits multiple parallel CNN layers for creating the scale invariance and inductive bias, and the transformer layers for creating long-range dependencies among the features extracted.

4 Experimental configuration

4.1 Pre-processing: Augmentation and data preparation

For a fair comparison, we implement the same training scheme for all three models. We set the image size as 224×224 and we apply augmentation methods: mixup [31], and cutmix [30] for all the models. We do not apply any balancing process and we evaluate the models on the validation split available for the datasets.

4.2 Dataset

We train all the tree models on the Coleoptera and Odonata datasets from Observation.org [1], the largest nature platform in the Netherlands for nature observation. Each of these datasets is split into train and test datasets to maximise the number of samples used for training. We validate the models on three datasets: the test split of the Observation.org datasets, the data from iNaturalist.org [3] -the global platform to record and organise nature-related findings- and from artportalen.org [2] -the Swedish nature conservation portal. In both orders, we consider only European species. The images are collected with mobile phones by citizen scientists. For the Coleoptera and Odonata datasets from Observation.org, we evaluate the models at the species level and morph level, we also distinguish among samples of morph imago (adults) with different sex (male, female, and unknown). Many species of Coleoptera and all species of Odonata present sexual dimorphism. With species, we refer to the taxonomical full name of the species which consists of the order (e.g. Odonata), the infraorder (e.g. Anisoptera), the family (e.g. Aeshnidae), the genus (e.g. *Aeshna*), and the species (e.g. *affinis*). With morph, we identify specific groups of insects that are all of the same species but differ in morphology e.g. Fig. 2 shows a Coleoptera and an Odonata species at different life stages.

Coleoptera dataset: the dataset, from Observation.org, consists of 849,296 images over 3,087 species. We split the dataset in train and test with the ratio of 80:20 samples per species (674,441:174,855 samples). The dataset is unbalanced with a minimum of 2 samples and a maximum of 11,523 samples per species. The dataset consists of species from 122 families, and we have samples from the Polyphaga and Adephaga suborders and 1,344 genera with a total of 3,087 species. In the dataset, there are samples of thirteen morphs: imago, imago brachypterous, imago macropterous, imago micropterous, unknown, gall, exuviae, deviant, larva/nymph, mine, egg, pupa, queen; and three sexes: male, female, and unknown (Tab. 1). We name this dataset Coleoptera_obs.

Odonata dataset: the Odonata dataset from Observation.org contains 628,189 images from 235 wild Odonata species. The ratio of train data and test data is roughly 80:20 per species (502,467:125,722 samples). The dataset is unbalanced with a minimum of 2 samples and a maximum of 19,754 samples per species. The dataset consists of species from both Epiprocta, in particular from Anisoptera, and Zygoptera infraorders. For the Anisoptera infraorder, we have samples from six families for a total of 153 species, and for the Zygoptera infraorder, we have five families for a total of 82 species, we name this dataset with Odonata_obs. For this dataset, information is available about morph and sex for further observation in the results. The dataset consists of samples of eight morphs: imago, unknown, fresh imago, exuviae, deviant, larva/nymph, prolarva, egg; and three sexes: male, female, and unknown (Tab. 1).

iNaturalist_Coleoptera: In the iNaturalist dataset [13], there are 236 species of Coleoptera available: among them, 87 species are recognised as European and available in the Coleoptera_obs dataset train split used in this paper. Our sub-sampled dataset consists of 4,350 samples with a balanced distribution of samples among species (each species has 50 samples). Hereafter, we name this dataset with Coleoptera_iNat.

iNaturalist_Odonata: In the iNaturalist dataset [14] 291 species are available. Among them, 58 are recognised as European and also available in the Odonata_obs dataset used for training. The selected 58 species from iNaturalist dataset consist of species from both Anisoptera and Zygoptera infraorders. The dataset has 2,900 samples, 50 samples for each species. Hereafter, we name this dataset with Odonata_iNat.

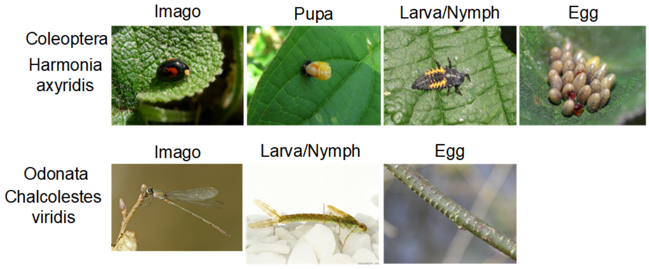


Fig. 2: Morphs of a Coleoptera (first row) and an Odonata (second row) species at different life phases.

Artportalen_Coleoptera: The Artportalen [11] consists of 3,426 species of Coleoptera. Among them, 1,574 are used to validate the models and are available in the train split of Coleoptera_obs. The dataset is unbalanced with a total of 118,464 samples. There are more than 400 species with less than 10 samples each and less than 30 species with more than 500 samples each. Hereafter, we name this dataset Coleoptera_art.

Artportalen_Odonata: The Artportalen [12] has 73 species from both Anisoptera and Zygoptera infraorders of which 69 species are available in the train split of Observation.org. The dataset consists of 55,680 samples and it is unbalanced with 12 species with less than 100 samples and 20 species with more than 1,000 samples. Hereafter, we name this dataset with Odonata_art.

4.3 *Model settings and hyperparameters*

All three selected models are pre-trained on ImageNet1k [8] and fine-tuned on the train split of the datasets. We modify the linear layer of the head of the original structures (the classifier layer) to be in line with the number of classes required for our datasets. Models are implemented on Pytorch Image Models (timm) [27] and executed on NVIDIA A40 GPU. For the models trained on Coleoptera_obs, we trained the model for a maximum of 90 epochs. Due to the high number of species and the low number of samples per species, we apply early stopping regularisation based on training loss to avoid overfitting, this is because we do not use a validation split. In this case, the `ViTAEv2` model stopped at 31 epochs, the `T2TViT14` model stopped at 89 epochs, and the `EffNetv2` stopped at 66 epochs. For the models trained on Odonata_obs, we trained the models for 61 epochs and we did not apply early stopping, because we did not observe overfitting behaviour with this dataset. With both datasets, we used a batch size of 32 samples, 5×10^{-4} as the learning rate, 0.065 weight decay, with AdamW [17] as the optimiser with cosine learning rate decay [19], and 10 warm-up epochs. Due to limited resources, we analysed only one run.

5 Experimental results

The three models are fine-tuned on the Coleoptera_obs and the Odonata_obs datasets train split. We evaluate the models on the Coleoptera and the Odonata datasets from Observation.org and from iNaturalist and Artportalen sub-datasets. To evaluate the models, we consider top-1 accuracy, hereafter named top-1, the model prediction

Table 1 Distribution of data at morph/sex level.

Morph and Sex	Coleoptera_obs	Odonata_obs
imago	130,990	76,025
imago male	5,963	25,293
imago female	5,948	17,756
imago unknown	119,079	32,976
unknown	31,321	34,675
imago macropterous	278	-
imago micropterous	9	-
imago brachypterous	2	-
deviant	2	134
fresh imago	-	12,124
exuviae	16	1,662
queen	2	-
gall	5	-
mine	92	-
larva/nymph	4,856	1,055
pupa	1,019	
prolarva	-	4
egg	87	42

with the highest probability must be exactly the expected answer; top-5 accuracy, hereafter named top-5, which considers any of our models' top 5 highest probability answers match with the expected answer. Moreover, we evaluate the range distribution of accuracy per species, the accuracy among species, the F1score, and the inference speed. We then examine the performance of the models considering the morph and sex in the case of insects at the morph imago.

Results based on species: We evaluate the three models computing an overall top-1 and top-5 averaging the results among the species. Tab. 2 shows the results obtained in the evaluation with the validation datasets. The Coleoptera_obs and Coleoptera_art datasets are heavily unbalanced with a high number of species and a low number of samples. For these datasets, the performance in all three metrics (top-1, top-5, and F1score) is in favour of `ViTAEv2` which shows to be robust on different distributions as proven by the fact that similar behaviour is observable also with the results on Coleoptera_iNat. The Odonata_obs and Odonata_art datasets also are unbalanced but they consist of a high number of samples and a low number of species. The models perform almost equally on both the Odonata_obs and Odonata_iNat datasets, while with Odonata_art the performance in top-1 accuracy is lower than 70%. Overall, the models have equal performance in top-5 accuracy with all three datasets, we take this as confirmation of the robustness of the models. With Odonata_obs and Odonata_iNat. The `ViTAEv2` outperforms the others in all three metrics for almost all the datasets. We can conclude that the `ViTAEv2` can be a better candidate compared to the other two models on the average species accuracy.

In the per-species accuracy (Fig. 3), the first line of plots shows the results obtained with the Coleoptera_obs and the Odonata_obs datasets, the second and third lines show the results obtained with the iNaturalist and Artportalen respectively. We observe that with both the Coleoptera_obs and the Odonata_obs datasets and in both top-1 and top-5 accuracy, `ViTAEv2` presents a lower amount of species with 0% accuracy than the other two considered models and a higher amount of species with accuracy higher than 80%. A similar behaviour is shown also for iNaturalist and Artportalen datasets, though with iNaturalist, both `ViTAEv2` and `T2TViT14` obtain accuracy higher than 90% for more species than `EffNetv2`, while with Artportalen, `ViTAEv2` outperforms the other models.

We now consider the relation between the amount of sample available per species in the train split and the accuracy top-1 obtained by the models with the test split. Fig. 4 shows the top-1 accuracy of the models per species, on the x-axis is the number of samples available in the train split for the species, and on the y-axis is the mean accuracy obtained. We observe that the `ViTAEv2` is the model able to predict species with less than 20 samples in train split within $10\% - 100\%$ top-1 per species, and all three models obtain top-1 higher than 50% for species with more than 20 samples in train split. For species with less than 100 samples in train split, the three models reach between $60\% - 100\%$ top-1. This behaviour is observed for the Coleoptera_obs and for the Odonata_obs which require more than 100 samples. We conclude that the ability of `ViTAEv2` to learn the singularity of each species based on details of the images is stronger than fully-convolutional and fully-transformer based models. That is shown by its ability to identify species even with few samples in training.

Table 2 The top-1, top-5, and F1score for the fine-grained task at species level. In the first block, the models are trained on the train split of the Coleoptera_obs dataset and tested on the test split of the Coleoptera_obs dataset and on the Coleoptera_iNat and Coleoptera_art datasets. The results are to be considered for the Odonata training in the second block of rows.

Dataset	Model	top-1	top-5	F1score
Coleoptera_obs	**`ViTAEv2`**	**89.8%**	**97.5%**	**89.53%**
	`EffNetv2`	88.0%	96.7%	87.78%
	`T2TViT14`	88.1%	96.7%	87.97%
Coleoptera_iNat	**`ViTAEv2`**	**83.6%**	**92.4%**	**81.26%**
	`EffNetv2`	78.6%	90.0%	78.56%
	`T2TViT14`	78.9%	90.6%	76.64%
Coleoptera_art	**`ViTAEv2`**	**90.4%**	**96.7%**	**90.44%**
	`EffNetv2`	88.8%	96.1%	88.55%
	`T2TViT14`	87.3%	96.2%	87.42%
Odonata_obs	**`ViTAEv2`**	**93.6%**	**98.7%**	93.29%
	`EffNetv2`	92.6%	98.5%	**94.30%**
	`T2TViT14`	91.5%	98.4%	93.65%
Odonata_iNat	**`ViTAEv2`**	81.4%	**91.2%**	**84.37%**
	`EffNetv2`	**82.4%**	**91.2%**	80.03%
	`T2TViT14`	79.8%	90.2%	74.56%
Odonata_art	**`ViTAEv2`**	**69.2%**	**84.8%**	**68.25%**
	`EffNetv2`	67.3%	82.9%	65.99%
	`T2TViT14`	66.7%	83.4%	65.72%

Inference speed: We evaluate the two taxa (Coleoptera and Odonata) separately because the difference in the output dimension impacts the number of parameters at the head layer thus at the computing time. The models' inference speed is presented in Tab. 3. On the Coleoptera_obs, all three models achieve a similar average inference speed per sample around $0.039sec$. On the Odonata_obs, we observe that the `T2TViT14` outperforms the other two models providing the prediction per sample on an average time of $0.009sec$. To give the reader the impact of the inference speed, to infer the prediction of the entire test split: with the Coleoptera_obs dataset, `T2TViT14` takes $2.19hrs$, `EffNetv2` takes $2.29hrs$, and `ViTAEv2` takes $2.39hrs$; with the Odonata_obs dataset, `T2TViT14` takes $18min51sec$, `EffNetv2` takes $24min5sec$, and `ViTAEv2` takes $38min45sec$.

Results based on Morph and sex: The second level of fine-grained task considered is morph/sex, as discussed in Sec. 4.2, both Coleoptera and Odonata appear in nature in different morphs due to the different stages in life. This aspect affects the dataset by increasing the granularity of the type of images that the models need to learn intra-species. It is worth noting that the models are trained to identify the species, so the results as to be interpreted as the top-1 species prediction in relation to the morph and sex. The Coleoptera and Odonata have different numbers of morphs (Tab. 1), Fig. 5 shows the top-1 achieved by each model for each morph or sex in the case of imago morph. We observe that `EffNetv2` and `ViTAEv2` both reach high performance with almost all the morphs in both the Coleoptera_obs and the Odonata_obs datasets. The `ViTAEv2` shows a similar behaviour as at the species level, in fact, it is able to well identify species in morphs that have a low number of samples. Moreover, for the imago morph, we evaluate the impact of sex on the performance of the models and we can conclude that all three models manifest a good performance with all three sexes.

Understanding the failures: In this section, we investigate the impact of misclassification to understand if they do make sense in biodiversity monitoring. We want to answer the question, are the models completely misunderstanding the species or there are some reasons that can justify the misclassification? To answer this question, we first consider the limitation of taxonomical datasets considered for this study. Datasets contain generic species refined to the actual species classification. This means that samples that are not identifiable by the taxonomist at the species level are labelled at the genus level with 'spec.' species. This peculiarity of the datasets

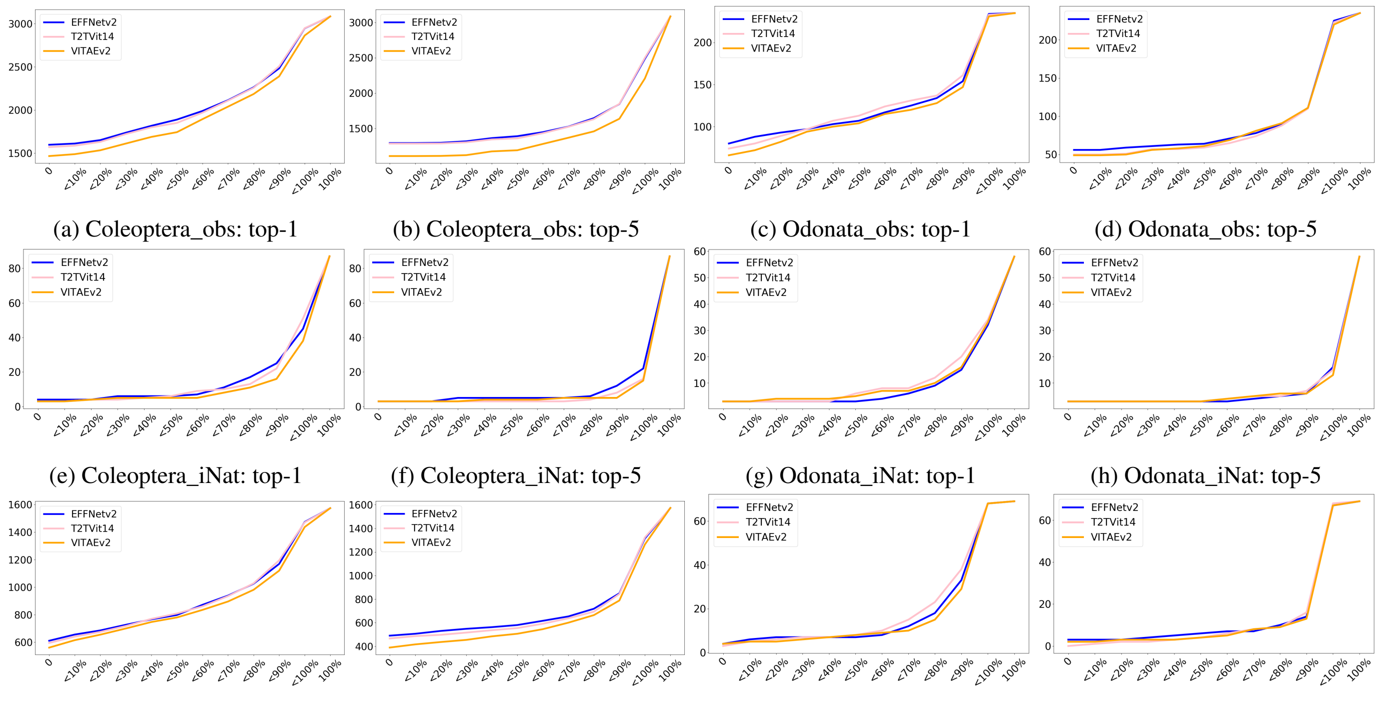
Table 3 Table shows the number of trainable parameters and the inference speed (sec/sample) of the models trained on Odonata_obs and on Coleoptera_obs.

model	Num Par	Coleoptera_obs	Odonata_obs
`ViTAEv2`	88M	0.041 (±0.002)	0.0185 (±0.0115)
`EffNetv2`	40M	0.039 (±0.003)	0.0115 (±0.0065)
`T2TViT14`	21M	**0.037** (±0.001)	**0.0090** (±0.0040)

generates some of the most common misclassifications in all three models. To prove this observation, we consider the species where more than 50% of the test data are misclassified. We observe that all these species are misclassified within the superspecies, or a species from the same genus or the generic genus. Fig. 6 shows samples of these types of misclassification obtained with all three models. In Fig. 6b, both the first two species, *Calopteryx virgo virgo* and *Cordulegaster boltonii algirica*, are confused with their superclasses, *Calopteryx virgo* and *Cordulegaster boltonii* respectively. In both Fig. 6a, 6b, *Cantharis paradoxa*, *Dytiscus circumsinstus*, and *Ischnura genei* are confused with other species that belong to the same genus. Finally in Fig. 6a, the *Monochamus sutor* is confused with the generic genus, the *Monochamus spec.* (which is a class in the dataset that represents a generic species within the genus *Monochamus*). These mistakes occur mostly with the species that are rare and so less represented in the datasets. Therefore, these misclassifications make sense and open new discussions on the proper use of these methods in ecology to exploit the possibility of using such models to help taxonomist to identify difficult images at the species level.

6 Discussion

Our results demonstrate that the introduction of transformers provides fast models that outperform the CNN models in accuracy at the species level and the morph and sex level. For the fine-grained species level, all three models show an overall good performance. As shown in Fig. 4, `ViTAEv2` proves to be more robust and performs well on rare species, it has fewer species with 0% top-1. We observe similar behaviour in the `T2TViT14` models even if with lower performance. The evaluation of the three models on iNaturalist and Artportalen sub-datasets, shows performances in line with the ones achieved with the Coleoptera_obs and the Odonata_obs datasets. These results confirm the robustness of the models and their good generalization to different distributions. The `EffNetv2` achieves competitive results that are slightly lower than the models based on



(a) Coleoptera_obs: top-1

(b) Coleoptera_obs: top-5

(c) Odonata_obs: top-1

(d) Odonata_obs: top-5

(e) Coleoptera_iNat: top-1

(f) Coleoptera_iNat: top-5

(g) Odonata_iNat: top-1

(h) Odonata_iNat: top-5

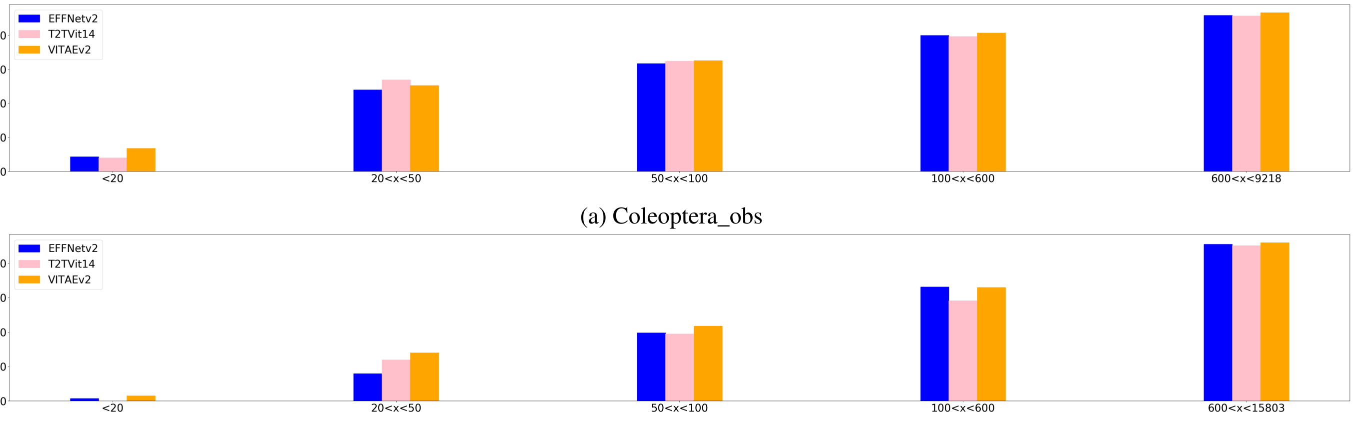
(i) Coleoptera_art: top-1

(j) Coleoptera_art: top-5

(k) Odonata_art: top-1

(l) Odonata_art: top-5

Fig. 3: The cumulative distribution of top-1 and top-5 accuracy per species. The x-axis is the range of accuracy, and on the y-axis is the number of species, i.e. in Fig.(a) the model `ViTAEv2` has accuracy $< 10\%$ for less than 1500 species and accuracy $> 90\%$ for more than 500 species.



(a) Coleoptera_obs

(b) Odonata_obs

Fig. 4: Performance in top-1 accuracy obtained at species level grouped by the number of training samples for each species.

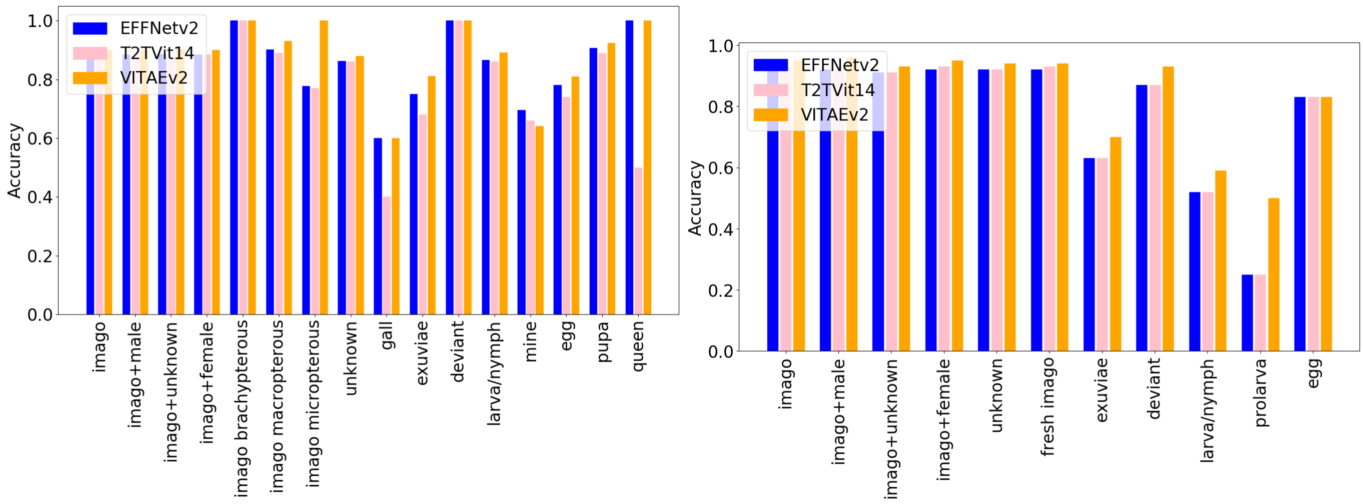
transformers. It is worth noting the analysis of the fine-grained task at morph/sex level. The results show the ability of the models to extract generalised representations for the species considering the intra species dimorphism among the different stages in life. We need to point out that the models achieve high performance for almost every morph and sex (Fig. 5) but we observe that for morphs such as gall and queen in Coleoptera, the `ViTAEv2` and `EffNetv2` outperform the `T2TViT14`, and for prolarva in Odonata, only `ViTAEv2` shows a good performance in classification. These results together with the low number of 0% top-1, discussed before, make us identify the `ViTAEv2` as a model able to deal with the shortage in data for the rare species and morphs. The inference speed gives a completely different evaluation of the models presenting the `T2TViT14` as the faster model compare to the other two models. Finally, we need to consider the end use of these models: if the performance and the robustness are the features mainly required, the `ViTAEv2` and `EffNetv2` models are the most suitable for the fine-grained tasks with a preference for the `ViTAEv2`; if the focal point is for public use so the inference speed is to be considered, the `T2TViT14` demonstrated to achieve good performance faster than the others.

7 Conclusions

In this paper, we investigated three state of the art models for fine-grained tasks in the taxonomic domain. The focus was on the comparison among fully-convolutional, fully-transformer, and hybrid, to depict the strengths and the weakness of three families. Our analysis shows different trade-offs between performance and inference speed. Hybrid models can identify species well even in case a low number of samples is available for training but these models demand a long inference time. Fully-convolutional models obtain good performance but we observe a lower performance with rare species compared to the hybrid models. Finally, the fully-transformer models are slightly less performant than the hybrid ones, but these models are faster compared to the other two models considered. We conclude that all three choices remain as candidates for more corroborating studies and for future deployment.

Acknowledgement

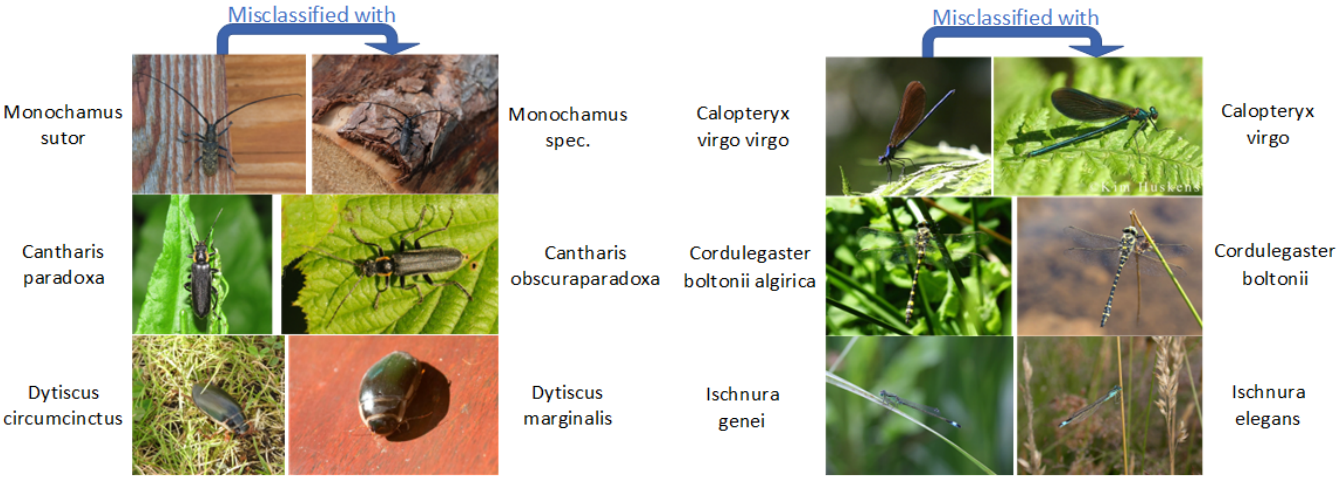
This research was supported by the EU Horizon projects MAMBO and TETTRIs.



(a) Coleoptera_obs

(b) Odonata_obs

Fig. 5: The performance in top-1 accuracy in species prediction obtained for each morph.



(a) Coleoptera_obs misclassified

(b) Odonata_obs misclassified

Fig. 6: Some of the most common failures of all three models.

8 References

1 Observation.org. https://observation.org/. Accessed: 2023-06-01.

2 artportalen.org. https://www.artdatabanken.se/tjanster-och-miljodata/artportalen/. Accessed: 2023-06-01.

3 inaturalist.org. https://www.inaturalist.org/. Accessed: 2023-06-01.

4 Paolo Audisio, Miguel-Angel Alonso Zarazaga, Adam Slipinski, Anders Nilsson, Josef Jelinek, Augusto Vigna Taglianti, Federica Turco, Carlos Otero, Claudio Canepari, David Kral, et al. Fauna europaea: Coleoptera 2 (excl. series elateriformia, scarabaeiformia, staphyliniformia and superfamily curculionoidea). *Biodiversity Data Journal*, (3), 2015.

5 Qi Chang, Hui Qu, Pengxiang Wu, and Jingru Yi. Fine-grained butterfly and moth classification using deep convolutional neural networks. *Rutgers University: New Brunswick, NJ, USA*, 2017.

6 Piotr Chudzik, Arthur Mitchell, Mohammad Alkaseem, Yingie Wu, Shibo Fang, Taghread Hudaib, Simon Pearson, and Bashir Al-Diri. Mobile real-time grasshopper detection and data aggregation framework. *Scientific reports*, 10(1):1–10, 2020.

7 Jerzy Dembski and Julian Szymański. Bees detection on images: study of different color models for neural networks. In *Distributed Computing and Internet Technology: 15th International Conference, ICDCIT 2019, Bhubaneswar, India, January 10–13, 2019, Proceedings 15*, pages 295–308. Springer, 2019.

8 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

9 Weiguang Ding and Graham Taylor. Automatic moth detection from trap images for pest management. *Computers and Electronics in Agriculture*, 123:17–28, 2016.

10 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010.

11 GBIF.Org User. Occurrence download, 2023. URL https://www.gbif.org/occurrence/download/0014398-230530130749713.

12 GBIF.Org User. Occurrence download, 2023. URL https://www.gbif.org/occurrence/download/0014782-230530130749713.

13 GBIF.Org User. Occurrence download, 2023. URL https://www.gbif.org/occurrence/download/0022763-230530130749713.

14 GBIF.Org User. Occurrence download, 2023. URL https://www.gbif.org/occurrence/download/0022742-230530130749713.

15 Suyog Gupta and Berkin Akin. Accelerator-aware neural network design using automl. *arXiv preprint arXiv:2003.02838*, 2020.

16 Suyog Gupta and Mingxing Tan. Efficientnet-edgetpu: Creating accelerator-optimized neural networks with automl. *Google AI Blog*, 2(1), 2019.

17 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

18 Suchang Lim, Seunghyun Kim, and Doyeon Kim. Performance effect analysis for insect classification using convolutional neural network. In *2017 7th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pages 210–215. IEEE, 2017.

19 Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

20 Yingshu Peng and Yi Wang. Cnn and transformer framework for insect pest classification. *Ecological Informatics*, 72:101846, 2022.

21 M Tan and QV Le. Efficientnetv2: Smaller models and faster training. arxiv 2021. *arXiv preprint arXiv:2104.00298*.

22 Hari Theivaprakasham. Identification of indian butterflies using deep convolutional neural network. *Journal of Asia-Pacific Entomology*, 24(1):329–340, 2021.

23 Hari Theivaprakasham, S Darshana, Vinayakumar Ravi, V Sowmya, EA Gopalakrishnan, and KP Soman. Odonata identification using customized convolutional neural networks. *Expert Systems with Applications*, 206:117688, 2022.

24 Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

25 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

26 Qi Wang, JianJun Wang, Hongyu Deng, Xue Wu, Yazhou Wang, and Gefei Hao. Aa-trans: Core attention aggregating transformer with information entropy selector for fine-grained visual classification. *Pattern Recognition*, 140:109547, 2023.

27 Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

28 Denan Xia, Peng Chen, Bing Wang, Jun Zhang, and Chengjun Xie. Insect detection and classification based on an improved convolutional neural network. *Sensors*, 18 (12):4169, 2018.

29 Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021.

30 Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

31 Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

32 Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, pages 1–22, 2023.