Making a Bird AI Expert Work for You and Me

Dongliang Chang, Kaiyue Pang, Ruoyi Du, Yujun Tong, Yi-Zhe Song, *Senior Member, IEEE*, Zhanyu Ma, *Senior Member, IEEE*, and Jun Guo

***Abstract*—As powerful as fine-grained visual classification (FGVC) is, responding your query with a bird name of "Whip-poor-will" or "Mallard" probably does not make much sense. This however commonly accepted in the literature, underlines a fundamental question interfacing AI and human – what constitutes transferable knowledge for human to learn from AI? This paper sets out to answer this very question using FGVC as a test bed. Specifically, we envisage a scenario where a trained FGVC model (the AI expert) functions as a knowledge provider in enabling average people (you and me) to become better domain experts ourselves. Assuming an AI expert trained using expert human labels, we anchor our focus on asking and providing solutions for two questions: (i) what is the best transferable knowledge we can extract from AI, and (ii) what is the most practical means to measure the gains in expertise given that knowledge? We propose to represent knowledge as highly discriminative visual regions that are expert-exclusive and instantiate it via a novel multi-stage learning framework. A human study of 15,000 trials shows our method is able to consistently improve people of divergent bird expertise to recognise once unrecognisable birds. We further propose a crude but benchmarkable metric TEMI and therefore allow future efforts in this direction to be comparable to ours without the need of large-scale human studies.**

***Index Terms*—Fine-grained visual classification, AI for enriching human knowledge, visual attention, model interpretability.**

I. Introduction

AI IS great – arguably the debate is on how it ultimately benefits mankind. Progress on computer vision has predominately followed the "Human for AI" trend, where human data are used to train AI models that replace humans in some capacity. In this paper, we are interested in the complete opposite direction, i.e., "AI for Human", and ask the question "can trained AI models help to enrich human knowledge instead?".

We pick the problem of fine-grained visual classification (FGVC) as a test bed on this quest. FGVC is a good fit as one of the few areas in computer vision where AI agents (We call FGVC and AI Expert interchangeably throughout the paper) can already reasonably *replace* human experts [5], [6], [17], [39], [42], [73], [93], [69], [78], [89], e.g., in identifying species of birds [81], models of cars and aircrafts [45], [56], and tell one flower from another [59]. The question then becomes – can the expert knowledge [47], [48], [49] learned by AI be transferred across to an average human, so that "you and me" become experts too? i.e., those that can tell that the eight birds in Fig. 1(a) are in fact from different species.

Fig. 1(a) illustrates the ambition of this paper – to complete this three-way transfer cycle among human expert, AI expert and average human (you and me). The link where human experts provide labels to train an AI bird expert is the known part and precisely what FGVC in its conventional form strives for. Key for this paper is on how to make the remaining two connections: (i) how to extract *knowledge* from AI that is digestible to a human (like a book), and (ii) how can we *measure* the progress on "you and me" becoming more expert-like using that *knowledge*.

On making the first link, we first stand with past works [9], [55], [62] on the lack of interpretability of fine-grained expert labels to an average human (Fig. 1(b)). As such, they do not constitute good "knowledge" in our context, e.g., telling me the top left bird in Fig. 1(b) is a "California Gull" probably does not say much. Our key innovation here is resorting to the highly discriminative regions that experts exclusively attend to as *transferable* knowledge (bottom row of Fig. 1(b)). This echoes well with psychological findings on the importance of using visual highlights for novices to learn in complex visual tasks [29], [36], [66].

To form the second connection and therefore close the loop, we literally take inspiration from a "book" – an expert bird guide in this context. More specifically, we present knowledge extracted from the AI expert as a bird guide to a human. The idea is then a *better bird guide* (i.e., knowledge) will result in "you and me" becoming *more expert-like*. We therefore take the degree to which the human has improved in being able to tell different species of birds as a measure of how good the extracted knowledge actually is.

It follows that we define knowledge as highly discriminative visual regions that are exclusively attended by domain experts, i.e., what parts of a photo experts focus on upon recognition. More specifically, we represent this expert attention as

Manuscript received 15 September 2022; revised 23 March 2023; accepted 3 May 2023. Date of publication 9 May 2023; date of current version 5 September 2023. This work was supported in part by Beijing Natural Science Foundation under Grant Z200002, in part by the National Natural Science Foundation of China (NSFC) under Grants U19B2036 and 62225601, in part by the Youth Innovative Research Team of BUPT under Grant 2023QNTD02, in part by MoE-CMCC "Artifical Intelligence" under Grant MCM20190701, in part by scholarships from China Scholarship Council (CSC) under Grants CSC 202006470036, and 202206470055, and in part by BUPT Excellent Ph.D. Students Foundation under Grants CX2020105 and CX2022152. Recommended for acceptance by X. Bai. *(Corresponding author: Zhanyu Ma.)*

Dongliang Chang, Ruoyi Du, Yujun Tong, Zhanyu Ma, and Jun Guo are with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: changdongliang@bupt.edu.cn; duruoyi@bupt.edu.cn; tongyujun@bupt.edu.cn; mazhanyu@bupt.edu.cn; guojun@bupt.edu.cn).

Kaiyue Pang and Yi-Zhe Song are with the SketchX, CVSSP, University of Surrey, GU2 7XH Guildford, U.K. (e-mail: kaiyue.pang1993@gmail.com; y.song@surrey.ac.uk).

Codes and all details on the human study are available at: https://github.com/PRIS-CV/Making-a-Bird-AI-Expert-Work-for-You-and-Me.

Digital Object Identifier 10.1109/TPAMI.2023.3274593

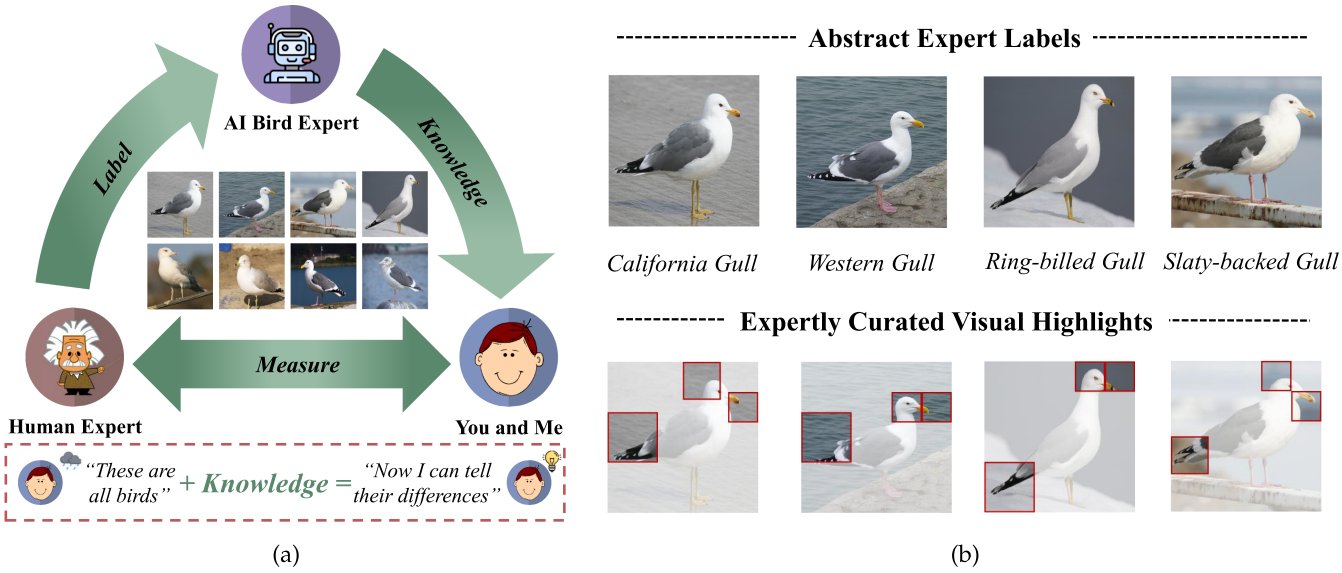

Fig. 1. AI Bird Expert Enriches Human Bird Knowledge. (a) By retreating from the common goal of a FGVC model in pursuing better expert label predictions, we envision a human-centric FGVC endeavour with a three-way cycle for human knowledge consumption and propose a first solution. (b) Capitalising on knowledge in visual form (instead of abstract label inherent to a FGVC model), we show positive human feedback in digesting it towards better recognition.

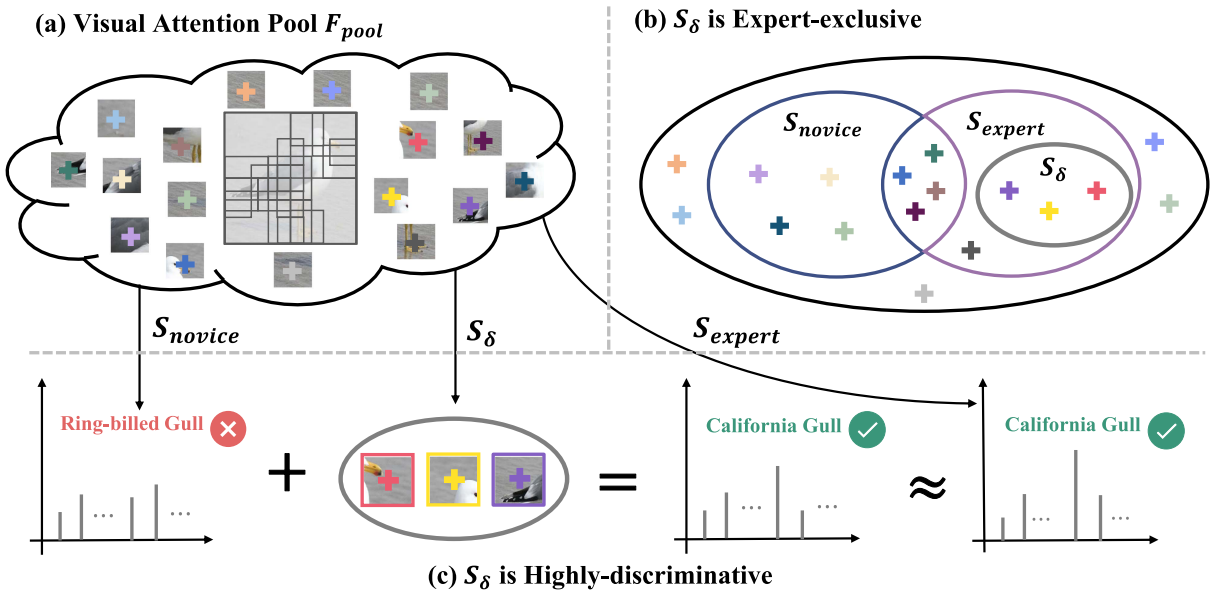

Fig. 2. Schematic illustration of how to obtain expert-exclusive but highly discriminative visual regions S_δ via our approach (Section II). (a) We assume the functioning of a visual classification model as an attention sampling process from a pre-defined pool of visual regions F_{pool} (global and local image patches). (b) We define the regions most effective for human knowledge consumption as a subset of visual attentions of domain experts *only* S_δ. We model regions attended by domain novices S_{novice} with the best possible fine-grained amateur textural descriptions of an image respectively – see Section III-D. (c) Our final refined elections of S_δ from F_{pool} are that of the most discriminative.

an optimal subset from a mixed pool of potentially discriminative regions (Fig. 2(a)) that leads to maximum recognisability (Fig. 2(c)). Our goal is then to eliminate non-expert ones that are shared between experts and novices so that expert-exclusive parts (knowledge) can be identified (Fig. 2(b)). In accounting for novice knowledge, we show that fine-grained image caption works best amongst alternatives (e.g., human scribbles or annotated bubble regions) (Section III-D). Taken together, our technical solution is a multi-stage learning framework that (i) first conducts fine-grained representation learning in the visual domain, (ii) followed by associating human caption onto corresponding image regions, and (iii) distilling cross-modal attention differences to account for expert-exclusive knowledge.

On measuring the efficacy of our bird guide (i.e., knowledge learned from AI), we conduct a large-scale human study with a total of 15,000 trials on a fine-grained bird dataset [81]. Results show, of the 407 trials that participants initially failed fine-grained bird recognition, an average

53.39% later successfully reverted their decisions, after being presented with our bird guide. We provide further analysis showing our approach is not constrained to work with bird species only, and improves conventional FGVC performance when exploited as a way to achieve explicit discriminative localisation.

A. Visual Attention as Knowledge

Our work can be seen as a general extension to the existing bulky literature of FGVC (a most recent survey at [87]), where we re-envisage the FGVC functionality from better label classification accuracy to that of providing useful knowledge for human consumption. It brings out an important question of whether current FGVC methodologies in the traditional benchmarking sense have in fact learned any fine-grained knowledge – the exact implication of that we however leave to future work. From a knowledge dissemination perspective, the way how we reason and dissect a FGVC model also seems to resonate with recent literature on generating FGVC visual explanations [8], [11], [38], [39], [68], [80], [44], [58].

A closer inspection however reveals the fundamentally different purposes they each serve. The goal of the existing works is on machine explainability, i.e., looking into the pixels responsible for a model's decision and judging whether they align with human intuition or not (e.g., by trying to make sense of visualised attention maps). We however take a human-centric view and only care if whatever extracted information can be instilled into our very brain as *transferable knowledge*. More precisely, existing works generally present a pixel selection function h(·) that either explains the decision of the black-box FGVC model in the form of post-hoc visualisation p(y | h(x)) [8], [68], [80] or making FGVC an explainable model itself p(h(x)|y) [11], [38], [39]. As a result, h(x) inevitably contains many visual cues that most human novices can already perceive. Our solution instead models p(y|h(x\x_human)), i.e., we take into account of human (non-expert) prior knowledge of an image and exclude them from our knowledge base to ensure a well-defined expert representation. We verify the importance of doing so in Section III-C.

As convincing as these arguments may sound, for many readers, our results would not appear too different to the fancy colour maps/bounding boxes that is commonplace today (Fig. 8). We understand that – for too long, people are tired of and became oblivious to the often cherry-picked qualitative demonstrations. The adoption of humans subjects to perform perceptual evaluation also furthers the concern – that to what extent a proposed method really progresses transferable knowledge extraction when human subjectivity is excluded.

B. Quantifiable Metric for Knowledge Transfer

Realising this burning problem in identifying the true efficacy of visual attentional knowledge, we first resort to the very recent work of [3], which investigates the transparency and reliability over a comprehensive checklist of existing attention-based model explanation mechanisms. We compare with their best reported method CVE [28] and show that how CVE fails to offer similarly good performance (Fig. 8) as ours for extracting human transferable knowledge, confirming the need for conceptualising a new type of model attention.

We then focus on the practical limitations of leveraging humans as the only method to perform perceptual studies – that (i) it requires extensive effort in time and financial resources, and perhaps more importantly, (ii) it renders complete reproducibility an infeasible task. This effort importantly paves the way for a more sustainable development for this "AI for Human" stream of work, in making future progress measurable without significant efforts in running large-scale human studies. For that, and as a last contribution, we propose Transferable Effective Model Attention (TEMI), a crude but quantifiable metric that aims at replacing human studies in measuring the AI $\rightarrow$ Human knowledge transfer (Section III-E). We attest to integrity of TEMI by (i) showing a strong correlations between TEMI and raw human study data, and (ii) demonstrating its designed behaviour holds for a large body of popular attention-based model visualisation methods.[1]

The general design principle of TEMI is that a higher score should be obtained when the input visual attentions indeed correspond to regions that are (i) noticeably different (i.e., specificity) and (ii) significantly more discriminative (i.e., improvability), when compared with those commonly perceived by domain novices. More specifically, we model specificity in a retrieval setup [16], [18], [50] asking to what extent our visual knowledge can no longer be backtracked to the image source it once extracted from, and improvability as a classification task where recognition gain is seen as a way to measure how much expert-exclusive discriminative information is contained in that same knowledge. We showcase some TEMI results for 16 popular attention-based model visualisation methods and some of their qualitative examples in Table III and Fig. 8. Readers can examine these results to better understand why conventional visualisation models do not stand up to the test of human knowledge consumption and why ours do.

II. Methodology

In the traditional FGVC setting, given an image x and its fine-grained label y (e.g., bird species name), a deep feature extractor $F(\cdot)$ will first process x into a set of feature pool $F_{pool} = \{f_1, f_2, \ldots, f_N\}$, representing a total size of N visual features covering a diverse location and scale of visual regions. A classifier $Cls(\cdot)$ is then appended upon the rich visual information provided in F_{pool} and optimised to predict y under cross-entropy classification objective. The composition of $F(\cdot)$ and $Cls(\cdot)$ is therefore what we often regard as an AI-enabled domain expert.

Our goal is to extract highly discriminative visual regions that experts exclusively attend to in classifying a fine-grained image. Denoting the visual attentions of experts and novices as S_{expert} and S_{novice}, this equates to learning an *attention re-sampling operation* S_δ on F_{pool} that can successfully bridge the gap between S_{expert} and S_{novice} in expert label prediction

[1] An online leader-board will be maintained and continuously updated at https://www.dongliangchang.cn/Making/leaderboard.html. Readers are encouraged to suggest new baselines to be included.

(Fig. 2). Putting it formally:

$$Cls(S_{expert} \odot F_{pool}) \approx Cls((S_{novice} + S_\delta) \odot F_{pool})$$
$$\text{s.t. } S_\delta \subseteq S_{expert} \setminus (S_{expert} \cap S_{novice}) \quad (1)$$

We model S_{expert}, S_{novice}, S_δ as a learnable probability row vector ($\mathbb{R}^N$) in practice, i.e., $\sum_{i=1}^{N} s_*^i = 1$. We shall now detail below how to obtain each component in (1).

A. Stage I: Visual Learning for $F_{pool}, S_{expert}, Cls(\cdot)$

Obtaining F_{pool}: Though obtaining F_{pool} is not restricted to one specific method, it does need some careful consideration given F_{pool} will be subjected to *bi-modal* sampling from both S_{expert} and S_{novice}. We find that the trivial workflow of learning F_{pool}, S_{expert} and $Cls(\cdot)$ in end-to-end fashion will bias $F(\cdot)$ towards S_{expert} and makes it incompatible to work with S_{novice} later. For this, we propose a simple yet effective solution by decoupling the learning of F_{pool} with that of $S_{expert}, Cls(\cdot)$. Specifically, given an image x, we divide it into 1×1, 2×2,..., and $k \times k$ uniform image blocks and use $F(\cdot)$ to extract feature for each local block to build our visual feature pool F_{pool}. We introduce an auxiliary classifier $Aux(\cdot)$ to guide the learning of F_{pool} for ground-truth label predictions. F_{pool} is then fixed with $Aux(\cdot)$ scrapped after this stage.

Obtaining $S_{expert}, Cls(\cdot)$. We compute S_{expert} by first conducting self-attention (SA)[2] on F_{pool} to better capture the long-term visual spatial dependency. We then append one fully-connected (FC) layer normalised with Softmax to simulate expert visual attention upon recognising a fine-grained object:

$$S_{expert} = \text{Softmax}(\text{FC}(\text{SA}(\Gamma(F_{pool})))) \quad (2)$$

where $\Gamma(\cdot)$ is a `stop_gradient` operation that forbids gradient flowing through the variable it functions on, which we will apply throughout. Denoting F_{expert} as $S_{expert} \odot F_{pool}$, we optimise $\{S_{expert}, Cls(\cdot)\}$ in the multi-label classification formulation:

$$L_{vision} = \text{Cross\_Entropy}(Cls(F_{expert}), y) \quad (3)$$

B. Stage II: Visual Grounding for S_{novice}

To bypass the otherwise fatal lack of human novice annotations on their perceivable visual regions of an image, we exploit the existing human fine-grained image caption dataset [64] to model S_{novice}. Given an image, ten single sentence visual descriptions are collected from different crowdsourced workers and we use their aggregate[3] c as a summary of the best possible visual perceptive zones from human novices. The question is how to ground human language input c to the visual representation of S_{novice}? We first process c with an off-the-shelf pre-trained language model Bert$(\cdot)$ [19] to get its semantic embedding f_c and append a multi-layer perceptron Mlp$(\cdot)$ aiming to project f_c to an embedding space compatible with F_{pool}. S_{novice} is then formulated as the broadcast element-wise cosine similarity between Mlp(f_c) and F_{pool}:

$$\begin{aligned} S_{novice} = \text{Softmax}(&\cos(\text{Mlp}(\Gamma(f_c)), \Gamma(f_1)), \\ &\cos(\text{Mlp}(\Gamma(f_c)), \Gamma(f_2)), \ldots, \\ &\cos(\text{Mlp}(\Gamma(f_c)), \Gamma(f_N))) \end{aligned} \quad (4)$$

Since the role of S_{novice} is to ensure human intentions expressed in language transfer visually, we require the training objective to maximise the cross-modal feature-wise mutual information $MI(\text{Mlp}(f_c), F_{novice})$, where $F_{novice} = S_{novice} \odot F_{pool}$.

Noise Contrastive Learning: Mutual information is notoriously intractable to optimise, where we resort to noise contrastive estimation [31] as a surrogate loss function. In particular, we implement it as InfoNCE [61] due to its wide adoption in the weakly-supervised visual grounding literature [30], [83], [83], [88]. InfoNCE is manifested in the popular cross-entropy fashion and measures how well the model can classify one positive representation amongst a set of unrelated negative samples:

$$L_{ground} = \sum_{i=1}^{bs} -\log \frac{F_{novice,i} \cdot \text{Mlp}(\Gamma(f_{c,i}))}{\sum_{j=1}^{bs} F_{novice,i} \cdot \text{Mlp}(\Gamma(f_{c,j}))} \quad (5)$$

with some slight abuse of notations, we use $F_{novice,*}$ and $f_{c,*}$ as the batch alternatives to F_{novice} and f_c. bs is the size of samples we use for contrastive learning with always 1 positive and $bs - 1$ negatives.

C. Stage III: Knowledge Distillation for S_δ

Recall the two key traits of S_δ we defined conceptually. S_δ is first expert-exclusive visual attention on F_{pool}. This gives us the important prior information of the element-wise importance of F_{pool} for S_δ: S_δ attend to a subset of visual regions in S_{expert} that is disjoint with S_{novice}, i.e., S_δ corresponds to an attention re-sampling operation from the non-zero entry in $\max(S_{expert} - S_{novice}, 0)$. Similar to (2), we model S_δ with feature-wise self-attention followed by one FC layer for output normalisation:

$$\begin{aligned} S_\delta = \text{Softmax}(&\text{FC}(\text{SA} \\ &(\Gamma(F_{pool} \odot \max(S_{expert} - S_{novice}, 0))))) \end{aligned} \quad (6)$$

The second trait of S_δ is being highly discriminative that bridges the recognition gap between S_{expert} and S_{novice}. Denoting the visual feature attended by S_δ as $F_\delta = F_{pool} \odot S_\delta$, we portray the learning as a process of knowledge distillation, where the student (S_δ) tries to distil expert-exclusive knowledge from the teacher (S_{expert}):

$$\begin{aligned} L_{distil} = &\text{KL}(\\ &Cls(\Gamma(F_{expert}/t)) \;||\; Cls((\Gamma(F_{novice}) + F_\delta)/t)) \end{aligned} \quad (7)$$

where KL$(\cdot)$ is the Kullback-Leibler divergence between two distributions and t is the temperature hyperparameter [34], [94] balancing the quality (sharpness) of the knowledge distilled, with smaller t corresponding to fewer coverage of teacher's knowledge base and larger t risking over smoothing out teacher's focus. We set $t = 5$ throughout.

[2]We implement SA with the popular Scaled Dot-Product Attention [79]: $SA(F_{pool}) = \text{Softmax}(\frac{Q_{pool} K_{pool}^T}{\sqrt{d_{pool}}}) V_{pool}$.

[3]To get image caption aggregate from different human visual descriptions, we use TextBlob [1] to extract registered noun phrases from each human and combine them into one caption by eliminating the duplicates.

1) Inference Without Reliance on c: There is one shortcoming in (6) when practically deployed: S_δ relies on the outcome of S_{novice} that requires the fine-grained human language description of an image (c) from the user. We argue it's a big ask of user to offer the same level of descriptive comprehensiveness like those we use for training: "This bird has a yellow, long, pointy beak, grayish feathers and grayish feathers, with white on the crown and black on the wingbars". We provide a simple solution to address this. A post-hoc approach is adopted that learns how to produce similar expert-exclusive discriminative visual attentions as with S_δ directly from F_{pool}:

$$\hat{S}_\delta = \text{Softmax}(\text{FC}(\text{SA}(\Gamma(F_{pool}))))$$
$$\hat{F}_\delta = \hat{S}_\delta \odot F_{pool} \quad L_{posthoc} = \text{MMD}(\hat{F}_\delta, \Gamma(F_\delta)) \tag{8}$$

We choose Maximum Mean Discrepancy (MMD) as a discrepancy metric for its ability to distinguish between two distributions with finite samples:

$$\text{MMD}^2(\hat{F}_\delta, F_\delta) = \frac{1}{\binom{N}{2}} \sum_{i \neq i'} k(\hat{F}_{\delta,i}, \hat{F}_{\delta,i'}) - \frac{1}{\binom{N}{2}} \sum_{i \neq j} k(\hat{F}_{\delta,i}, F_{\delta,j}) + \frac{1}{\binom{N}{2}} \sum_{j \neq j'} k(F_{\delta,j}, F_{\delta,j'}) \tag{9}$$

where $k(x, x') = exp(||x - x'||^2/\gamma)$ with γ as a bandwidth hyperparameter. We show by replacing S_δ with $\hat{S}_\delta$ only brings marginal performance downgrading in our empirical evaluations.

D. Incorporating S_δ as FGVC Booster

S_δ also provides an answer to the FGVC debate on what is the best way to achieve discriminative localisation [18], [40], [51], [86], [92]. Our speculation is that if S_δ has successfully encoded expert-exclusive visual attentions, visual regions δ corresponding to S_δ are then already locally discriminative with expert endorsement. There are of course many ways to embed δ as explicit localisation information into existing FGVC frameworks, of which we choose perhaps the most intuitive of seeing δ as another pixel space input together with the original image x. We find such simple solution suffices to bring improvements on top of many existing FGVC frameworks. To avoid ambiguity with existing notations, we abstract a FGVC solver into two parts of feature extractor $E(\cdot)$ and classifier $C(\cdot)$. We now formulate a new generic way of FGVC training with both x and δ as input:

$$L_{ce} = \text{Cross\_Entropy}(C(E(x) + \sum_{i=1}^{N_{top}} E(\delta^i)), y) \tag{10}$$

where in practice, we set $N_{top} = 1$, i.e., we only select the single most discriminative region in δ as one more image input for explicit localisation.

III. Experiments

Our main experiments are conducted on the CUB-Bird-200 dataset [81], which contains 11,877 images from a label categorisation of 200 bird species. We first show how the learned knowledge in the visual form of expert-exclusive discriminative regions $S_\delta / \hat{S}_\delta$ can help people with divergent levels of bird expertise towards better recognising their once unrecognisable birds. We then confirm S_δ is indeed attending to visual regions exclusive to domain experts and only with such type of visual feedback (*versus* $\{S_{novice}, S_{expert}\}$) can enable practically more interpretative and digestible knowledge to human participants. Given the lack of suitable baselines from existing FGVC research (elaborated in Section I-A), we move on to conduct ablation on our key technical choice of adopting image caption aggregate to represent the otherwise hard-to-quantify domain novice visual attentions. Finally, and perhaps most importantly, the human and financial cost of large-scale human studies can be largely off-putting for the future endeavours on this line of research. We therefore design an automatic human-like evaluation metric for measuring the efficacy of knowledge consumption, which is empirically shown to strongly correlate with the results from real large-scale human studies (Section III-E). We wrap up the experiments by providing further analysis on S_δ.

Implementation Details: We extract image features using pre-trained ResNet50, and text features using pre-trained Bert [19] in all experiments. The input images are resized to 300×300 and then randomly cropped to 224×224 during training. We always centre crop the image for model inference. At each stage, we adopt SGD optimiser with a momentum of 0.9. We set the learning rate for pre-trained feature extraction layers as a fixed value of $1e-4$. We however take a Cyclic Cosine Annealing [53] strategy for the fully connected layers with learning rate initially set to 0.01 and finally decayed to 0. We train the model for 100 epochs with weight_decay value fixed as $5e-4$. Please see our release code for details.

A. Data & Participant Setup

We recruited 200 participants across different ages, genders and education levels, where each of whom was expected to complete a questionnaire with 300 bird recognition tasks. In each task trial, the participant was given a query bird image and five gallery bird images, and asked to select the *only* image in the gallery that he/she believed to belong to the same bird sub-class with the query (Fig. 4). The difficulty of the task then lies in the similarity level between the query and gallery images, where we define three challenge levels based on the biological bird stratification of Order-Family-Species: (i) *Easy:* gallery samples are manifested in different bird orders. (ii) *Medium:* gallery samples are from different bird families but all belonging to the same bird order. (iii) *Hard:* gallery samples only differ in the finest species level, i.e., they come from the same bird order and family. We assign different score for correct answer to question of different difficulty (0.5, 1, 1.5 point for easy, medium and hard respectively). This means when the 300 tasks are decomposed into three subsets of [90, 120, 90] for each challenge level (the setting we adopt), the full mark would be 300 points. We plot the normalised scoring histogram by counting the number of people falling in each of the 10 discretised bins and observe an intuitive Gaussian-like bell shape distribution. Since our goal is to simulate a study covering fairly for people with divergent

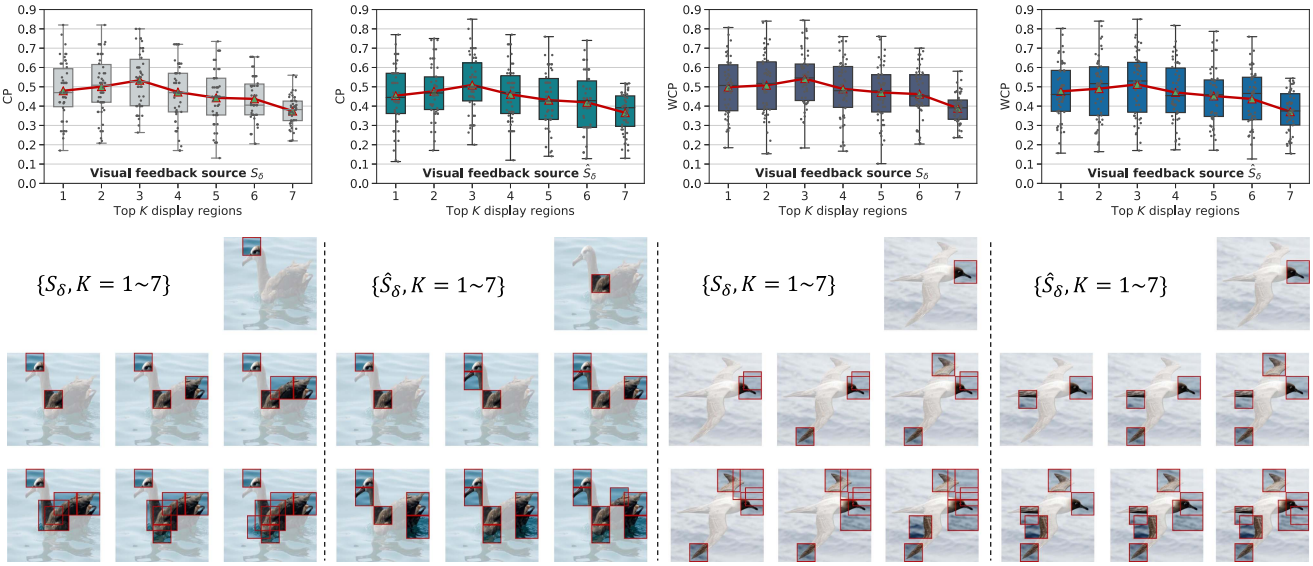

Fig. 3. Qualitative and quantitative evidence of $S_\delta/\hat{S}_\delta$. Top row: box plot to demonstrate the efficacy of $S_\delta/\hat{S}_\delta$ in helping people reverse their failed decision for bird recognition. Green triangle represents the mean performance. Bottom two rows: sample illustration of top K visual regions $S_\delta/\hat{S}_\delta$ attend to.

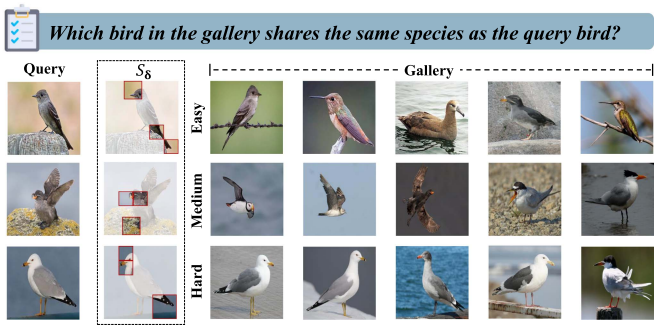

Fig. 4. Sample questionnaire for measuring the efficacy of AI-empowered knowledge by simulating it as a book guide. In data and participant setup stage, S_δ is not shown.

levels of bird expertise, we form three representative population groups by randomly selecting [15, 20, 15] participants from the first three bins (worst scorers), the fourth to seventh bin (medium scorers) and the last three bins (best scorers) respectively. These 50 participants and their bird recognition cases are then setting our basis for the experiments later. Note that the entire process of the human study was completely anonymous and only made use of a randomly allocated ID that allows different trials to be grouped by participant. Each participant was made aware prior to volunteer to participate that the data will be released without any personal information as a part of a research publication, and this data will only be used for research purposes.

B. S_δ is Your Expertly Curated Bird Guide

Experimental Method: We now take the failed recognition cases from the 50 participants at the setup stage and examine the efficacy of S_δ on improving their recognisability. We also test the performance of $\hat{S}_\delta$, a practical alternative to S_δ when there is lack of fine-grained visual description of an image. We follow the similar "query-gallery" experimental procedure, with just the difference that the query is now highlighted with the knowledge provided by S_δ (Fig. 4). By paying extra attention to the AI-empowered knowledge, participants are required to re-make their decision of selecting the target image from the gallery that shares the same sub-class with the query. Our evaluation metric is twofold: average human correction percentage (*CP*) and average weighted human correction percentage (*WCP*). The former calculates the percentage of cases (2407 in total) that one human participant has successfully reverted their erroneous recognition under the guidance of S_δ, where the latter corresponds to a weighted version that assesses the correction rate for cases in each challenge level first and weight it with the corresponding challenged point. Lastly, we rank the visual attentions of S_δ in descending order and always present the Top K visual regions to human participants. We experiment with different K values of [1, 2, 3, 4, 5, 6, 7] – we find $K = 7$ already brings notable degenerate performance in our pilot study as people tend to feel uncomfortable and fail to focus when faced with too many visual cues.

Rule of Familiarity: To mitigate the effect of participants intentionally altering their decision based on past failures, and keeping their decisions unchanged because of cognitive inertia [2], [57], three strategies were devised: (i) we repeated the tests of successful recognition cases and interleaved them between failed ones, while not including these results when computing our statistics. The purpose is then to avoid reminding a participant "you were once wrong and re-decide now with the additional help from the knowledge provided". This could jeopardise the fairness of our evaluation – it is not knowledge that actually works, but people are asked to make a second choice. (ii) we enforced the participants to have at least have 24-hour time gap between undergoing any two different purposed experiments.

Since every participant might be asked to proceed with the recognition task on the same case more than once under the different types of knowledge guidance, comparisons between these knowledge could only be deemed as near independent when conducted spanning a time period [24]; (iii) we adopted the common strategy in mitigating familiarity by randomising the ordering of individual trials [75].

We also conducted a pilot study to confirm the removal of familiarity bias, where we asked the same participants to redo the questions that they previously failed on after 24 hours in a randomised order. The hypothesis is therefore, if familiarity bias is successfully eliminated, the average correction rate would be close to random guess $(1/5)$. Our result shows an average correction rate of $21.32 \pm 5.41\%$ off the 2407 trials from 50 participants, supporting the hypothesis. Our 24-hour gap rule in removing familiarity bias also finds itself in the famous Ebbinghaus Forgetting Law Curve [24] in psychology, which suggests that people on average have only 33.7% of their memory left after a one-day gap.

Results: We graphically describe CP and WCP for each human participant via the box plot of five-number quartile summary [77] in Fig. 3. Following observations can be made: (i) Under the metric of both CP and WCP, S_δ is able to provide the best performance when used to display its top 3 visual attentions with mean values, 53.39% and 54.24%. This provides compelling evidence that our learned S_δ is indeed extracting knowledge from an AI agent in a way that guides people towards better recognising a unknown bird. We further calculate the mean CP (mCP) and WCP (mWCP) values for people from three different scorer groups at setup stage, where the small differences among groups (52.63%, 52.91%, 54.91%@mCP, 52.96%, 53.88%, 56.53%@mWCP) confirm that S_δ is friendly and effective to users with divergent levels of bird expertise. (ii) Difference between S_δ and $\hat{S}_\delta$ is marginal in the eyes of human participants. This is an important message indicating that our framework can stand up to the fatal but common lack of per-image fine-grained language descriptions with little performance sacrifice. (iii) WCP values are slightly larger compared with those of CP. This is expected. Given $S_\delta/\hat{S}_\delta$ is designed to offer the most subtle expert-exclusive visual cues for successful fine-grained recognition, it naturally works better for solving harder cases with more bonus points (California Gull *versus* Western Gull) compared with that for easier ones (Gull *versus* Flamingo). (iv) There seems to exist a safety value (K<7) of how many visual regions to display and when the threshold is violated, humans start to show general failure in digesting the visual knowledge provided by $S_\delta/\hat{S}_\delta$. In line with psychological findings [29], [66], we ascribe such phenomenon to the fact that redundant visual distractors superimposed upon the most *compact* visual highlights can be very detrimental for people to gain attentional expertise in practice. Interestingly, we also demonstrate some common visualisation results of existing FGVC works in Fig. 6. We can see how they generally cover the full attention map of a bird and correspond roughly to a K = 7 scenario (or even worse!) under $S_\delta/\hat{S}_\delta$ – indicating their inherent unsuitability for human consumption as argued in Section I-A. We also consult our human participants on why they perform drastically poorer when K grows over a certain value and their response is unanimous: "we don't know how to make sense from the knowledge manifested in crowded and cluttered visual regions.

C. S_δ Works Because it is Expert-Exclusive

In this section, we conduct deeper probe on S_δ. Our goal is to show that S_δ is indeed distilling unique visual attentions from experts (S_{expert}) that are not shared by domain novices (S_{novice}), and this very property of S_δ consequently helps human participants to better recognise a bird. Below is detailed analysis.

We first adopt Intersection over Union (IoU) to measure the correlation between the top K rankings of two visual attention sequences – if $\text{IoU}_{\boldsymbol{K}}(\boldsymbol{S_{novice}}, \boldsymbol{S_{expert}})$ is significantly larger than $\text{IoU}_{\boldsymbol{K}}(\boldsymbol{S_{novice}}, \boldsymbol{S_\delta})$ before K grows impractically large, we know S_δ has successfully extracted the exclusive parts from S_{expert}. Results in Fig. 5(a) and (b) confirm that S_δ indeed shares negligible (≤ 0.01) attentional overlap with S_{novice} for K up to 20, in a stark contrast with the strong correlated interplay between S_{expert} and S_{novice}. To shed further light on the importance of refining S_{expert} to S_δ and its practical meaning as a form of knowledge to human participants, we calculate the expert label prediction accuracy $\text{Acc}_{\boldsymbol{K}}(\boldsymbol{\delta})$ and $\text{Acc}_{\boldsymbol{K}}(\boldsymbol{expert})$[4] with the combined visual cues from $\{S_{novice}, \sum_{i=1}^{K} S_\delta^i\}$ and $\{S_{novice}, \sum_{i=1}^{K} S_{expert}^i\}$ respectively. Our intuition is that if S_δ provides practically more useful complementary knowledge to what human already knows, $\text{Acc}_K(\delta)$ should obtain a considerably satisfying performance at a much smaller K than that of $\text{Acc}_K(expert)$. In other words, knowledge encoded in S_δ is more condensed and effective for human to digest because of its nature of expert-exclusive. Fig. 5(c) shows this is exactly the case where ≈91.40% of label prediction performance is retained with only *one* best visual attention in S_δ and up to ≈94.05% when K = 3. We also repeat the "query-gallery" experiment in Section III-B and aim to figure out to what extent can S_{novice} and S_{expert} improve people's bird recognisability in a human study. By examining their performance under both mCP and mWCP and comparing them with S_δ (40.02% and 47.05% *versus* 53.39% @mCP, 39.56% and 45.51% *versus* 54.24% @mWCP), we can fairly conclude that the hypothesis of fine-grained visual knowledge being expert-exclusive does matter for practically more effective human consumption.

D. Good S_δ Solver Needs Human Language Input

A critical part of our framework at *design-level* is how to define and quantify S_{novice} with data at hand. Given the rich visual elements of an image and the subjective nature of human vision on their relative importance, deciding the best form of representing domain novice visual attentions becomes indeed an art of choice. Our proposed method advocates the use of

[4]To obtain $\text{Acc}_K(\delta)$ ($\text{Acc}_K(expert)$) on different K values, we first work out normalised mean feature representation $\frac{1}{2}(f(F_{novice}) + \sum_{i=1}^{K} \lambda S_\delta^i F_\delta^i)$, s.t. $\lambda = 1/\sum_{i=1}^{K} S_\delta^i$ before feeding it into classifier.

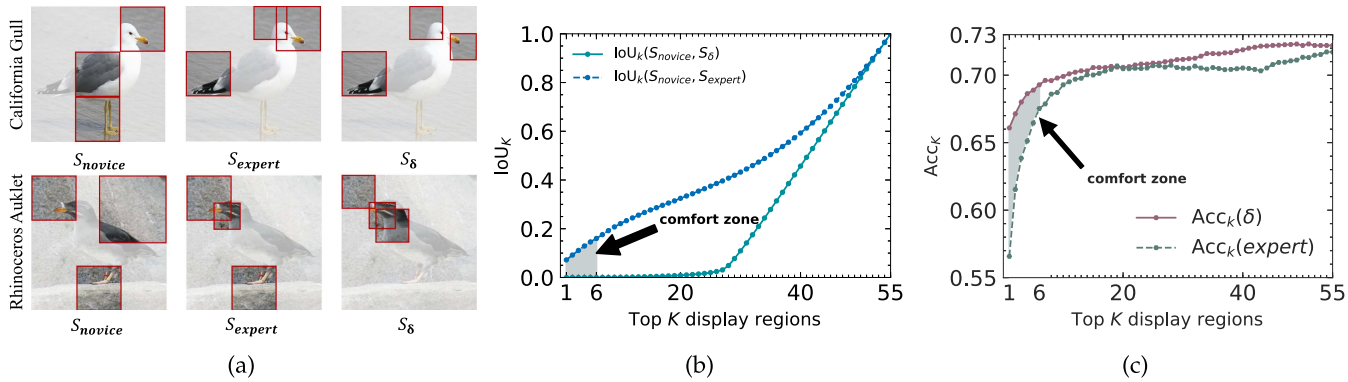

Fig. 5. Comparisons between S_{novice}, S_{expert} and S_{δ}. (a): Exemplified comparisons among the Top 3 visual attentions encoded. (b)(c): Understanding our learned S_{δ} from two different aspects. Comfort zone: maximum number of visual regions for display that we find humans can practically make sense of (K<7). More details in text.

TABLE I
HOW TO REPRESENT S_{novice} WITH DIFFERENT CONCEPTUALISATIONS? WE LEVERAGE HUMAN ANNOTATED BUBBLES [18] AND DRAWINGS [85] AS ALTERNATIVES OF CAPTION AGGREGATE (OURS) TO REPRESENTING S_{novice} FOR ABLATION STUDY IN SECTION III-D

Raw Image	Bubbles [18]	Drawing [85]	Caption Aggregate	Raw Image	Bubbles [18]	Drawing [85]	Caption Aggregate
			This bird has a yellow-breasted, a black cheek patch, and a white superciliary, ...				This bird has a white breast, a brown tail and long wings, and a small yellow beak, ...
			This bird has long legs, short tan wings, a small bill, and grayish-brown belly feathers, ...				This bird has a yellow chest, a black pointed beak, and a orange-yellow chin, ...

human fine-grained caption aggregate of an image for learning S_{novice}, where we compare it with several competitors below.

Competitors: We include three competitors for different conceptualisations of S_{novice}: (1) *Discriminative bird bubbles:* These annotated bird circular regions (Table I), namely "bubbles" [18], are collected via a novel online game aiming to reveal the most discriminative parts of a bird image. We aggregate the available bubbles of an image from multiple players and use their mean ImageNet pre-trained feature representation to learn S_{novice}. (2) *Human bird drawings:* CUB-200-Painting [85] is an extension of CUB-200-2011, which contains diverse human drawing forms (Table I) aiming to visually interpret a fine-grained bird species, including watercolors, oil paintings, sketches and cartoons. We aggregate the human drawings under one bird species and use their mean ImageNet pre-trained feature representation to learn S_{novice}. (3) *Junior bird expert:* We also model S_{novice} as a beginner-level bird specialist that can differentiate between 13 bird subclasses at order level [9] – instead of the finer recognition of 200 subclasses at species level required towards a bird expert. For this, we train a 13-way classification model and adopt it (like ImageNet pre-trained feature) to learn S_{novice}.

Results: We follow the same "query-gallery" experimental procedures in Section III-B to evaluate the knowledge efficacy of S_{δ} provided by the different realisations of S_{novice} as described above. We report the result in Table II and confirm the significance of our technical choice of using fine-grained caption aggregate to represent what domain novice can perceive from an image. Interestingly, the worst choice of S_{novice} (Bubble) still outperforms S_{expert} and S_{novice} (48.40% *versus* 47.05% and 40.02% @mCP, 46.98% *versus* 45.51% and 39.56% @mWCP), stressing again the importance of our expert-exclusive modelling.

TABLE II
PERFORMANCE COMPARISONS (%) BETWEEN DIFFERENT REALISATIONS OF S_{novice} FOR HUMAN KNOWLEDGE CONSUMPTION

–	**Ours**	Bubble [18]	Drawing [85]	Beginner [9]
mCP	53.39	48.40	49.07	50.04
mWCP	54.24	46.98	49.69	51.68

E. *TEMI: Human-Like Evaluation* Without *Humans*

Thus far, our evaluation method remains completely reliant on a large-scale human study. While seemingly a natural choice, this in hindsight comes with a few important limitations. First, relying on human judgements introduces subjective bias, that despite our best efforts (Section III-B), are impossible to completely eliminate, potentially leaving room for misinterpretation. Then there is the labour, time and financial costs associated that prevents a similar-scale study to be implemented in practice. Both combined, the need for large-scale human participation acts as a strong barrier for consequent works to be compared, and therefore largely prevents the steady progression of this "AI for Human" stream of work.

For that, we develop a crude but automatic evaluation metric, Transferable Effective Model Attention (TEMI). The main benefit is therefore that progression in this new field of "AI for Human" becomes quantifiable without the need for large-scale human studies. The guiding principle when designing TEMI is that it should simulate the efficacy of S_δ for knowledge transfer, similar to that off human performance in real-world probes. TEMI can be directly calculated from our trained model and generally applicable to any attention map inputs – just substitute S_δ. Below is a more detailed description and analysis.

Just like how we conceptually characterise S_δ as both highly discriminative and expert exclusive, TEMI comprises of two individual components that quantitatively measure the actual success on those two aspects and further combine both into a unified metric – (i) *improvability:* how much S_δ has *improved* the recognisability over S_{novice}, and (ii) *specificity:* how differently are the visual regions S_δ attended to when compared with those of S_{novice}?

Improvability: Assuming the classification accuracy obtained by experts, novices with and without the guidance of visual attention knowledge are $Acc(expert)$, $Acc(\delta + novice)$ and $Acc(novice)$, respectively, improvability is defined in the form of a *classification setup* as follows:

$$Improvability = \frac{Acc(\delta + novice) - Acc(novice)}{Acc(expert) - Acc(novice)} \quad (11)$$

Specificity: We evaluate the specificity of the visual regions that S_δ attends to in a *retrieval setup*. Given the visual features F_δ, we aim to find out whether S_δ as query can retrieve the corresponding S_{novice} from a list of candidates (16 in our setting) – if not, we can then confidently confirm that S_δ is indeed containing significantly different visual attentions compared with that in S_{novice}. Since the learning of S_{novice} is only weakly supervised from the human caption aggregate c, we in practice replace S_{novice} with c to form the gallery and find this allows for a more accurate characterisation. Formulating the retrieval accuracy as a metric of recall ($R@1(\cdot)$), we define specificity as:

$$Specificity = 1 - \frac{R@1(\delta)}{R@1(novice)} = \frac{R@1(novice) - R@1(\delta)}{R@1(novice)} \quad (12)$$

Note that we add the regularisation term $R@1(novice)$ to the equation instead of the more intuitive alternative $1 - R@1(\delta)$. This is because there is simply no guarantee that S_{novice} can always retrieve its corresponding text (i.e., $R@1(novice) = 1.0$). We then use the harmonic mean of the two factors to calculate TEMI:

$$TEMI = 2 \times \frac{Improvability \times Specificity}{Improvability + Specificity} \quad (13)$$

TEMI is bounded in [0, 1] (or equivalently [0%, 100%]), and achieves its upper bound when both improvability and specificity reach their optimal value of 1, i.e., $Acc(\delta + novice) = Acc(expert)$ and $R@1(\delta) = 0$. In contrast, TEMI is 0 when either of the two factors stays at their worst with a value of 0.

Results: We did two things to verify TEMI. We first compare TEMI with the results from the real-world human simulations as those in Section III-B and show a strong correlation between the two. We then calculate TEMI for various model attention maps presented in the existing FGVC systems and provide some qualitative demonstrations for readers to connect TEMI with tangible visual evidence.

CVPR18 [86] CVPR18 [91] CVPR19 [12] CVPR20 [39] ICCV21 [38] ICCV21 [63]

Fig. 6. Visualisations of the typical supporting regions for the classifiers in the existing FGVC works. Examples shown are from direct copy-and-paste of the original paper.

In Fig. 7, we plot the results of mCP and TEMI for two cases: (i) Fix S_{novice} while changing S_δ (NTS (ECCV18 [91]), CAL (ICCV21 [63]), INTER (CVPR20 [39]), CVE [28], S_{expert}); (ii) Fix S_δ while changing S_{novice} (Bubble [18], Drawing [85], Beginner [9]). Strong correlation can be observed in Fig. 7(a), indicating that TEMI is indeed a good substitute of the once cumbersome and costly human study evaluation method. Researchers can now fixate on designing a better representation for S_{novice} or S_δ and coarsely evaluate any executions on-the-fly with TEMI, while only leverages human study at the final idea verdict stage for more rigorous justification. As a sanity check, we show further in Fig. 7(b)–(d) that TEMI is not constrained to align well only with the mCP results from the specific population represented by our 50 human participants. Specifically, we simulate three scenarios: (i) Random: we randomly select 40 participants out of the total 50 and calculate the mCP accordingly (repeat five times). (ii) Expertise/Gender: since we know the bird expertise level and the gender of each participant during the main experiment preparation stage, we re-sample the 50 participants into five sub-groups, with each comprises a dramatically different expertise/gender distribution, e.g., bell-shaped/long-tailed/uniform. TEMI remains valid.

It is interesting to see how CVE, a specifically designed, previously best reported visual attention based explanation method is dramatically worse than Ours for human knowledge consumption (49.21% *versus* 53.39% @mCP, 48.41% *versus* 54.24% @mWCP), but also noticeably better than S_{expert} (47.05% @mCP, 45.51% @mWCP). CVE exposes visually decisive regions by contrastively explaining an expert classifier with respect to a distractor image, while ours can also been as a contrastive solution by differencing the novice part from the visual attentional process upon expert classification. This shows it is important to distil a refined version of expert knowledge for transferability to humans and how to define the exact meaning of that distillation is the key question for researchers to answer (i.e., $S_\delta >$ CVE).

We further calculate TEMI in Table III for 16 types of attention maps observed in the existing literature, including the traditional FGVC works set out to pursue new state-of-the-art performance [12], [22], [23], [54], [63], [76], [86], [91] and works that aim to provide a generally better probe for model

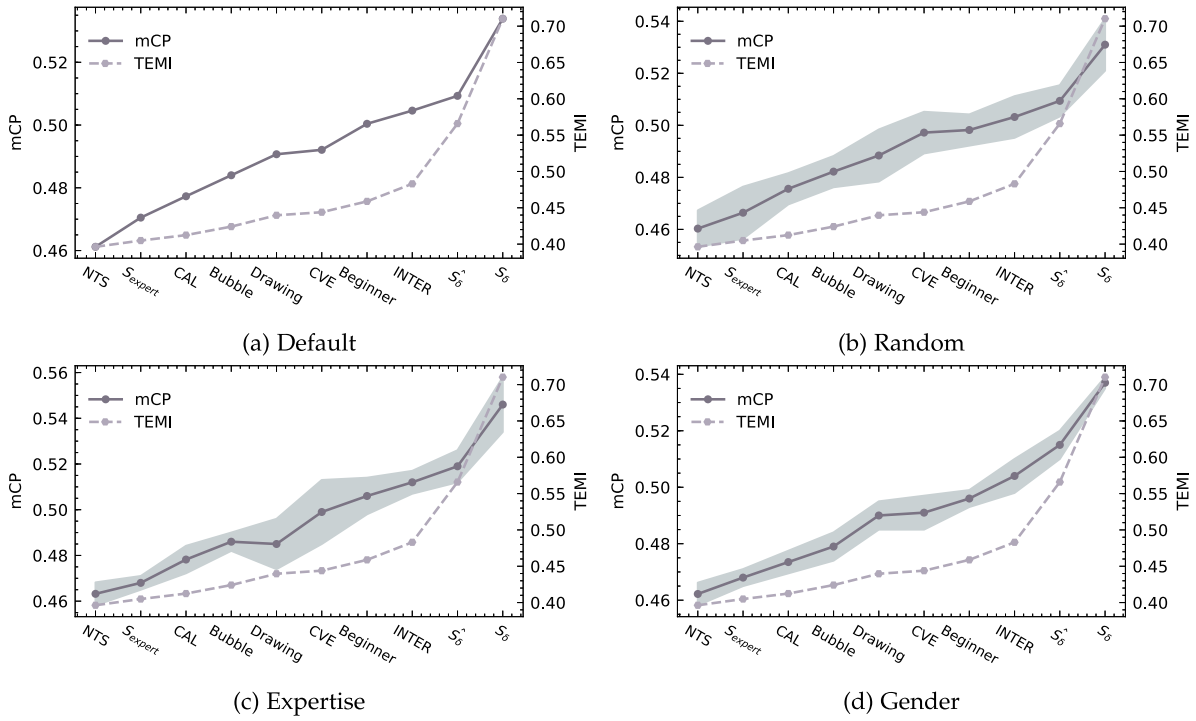

Fig. 7. Quantitative comparison between mCP and TEMI for measuring the efficacy of transferable knowledge. mCP is the mean correction percentage for human participants to reverse their once erroneous decision provided with the visual attention knowledge. TEMI is our proposed discriminative metric aiming to simulate mCP performance to bypass the critical need of human study evaluation. (a) the default setting of leveraging the results from all 50 human participants to calculate mCP; (b-d) to ensure TEMI results also align well with a diverse background of population other than the fixed 50 participants we recruited, we re-sample the human study results in different directions. More details in text.

TABLE III

COMPARISONS OF TEMI SCORES FOR THE VISUAL ATTENTION MAPS OBTAINED BY OUR PROPOSED METHOD AND THOSE PRESENTED IN THE 16 EXISTING RELEVANT WORKS. WE DEMONSTRATE ALL STATS NEEDED TO CALCULATE IMPROVABILITY AND SPECIFICITY ((11) & (12)), BOTH OF WHICH ARE THEN USED TO REPRESENT TEMI (13). TOWARDS FAIR EVALUATION, WE RE-IMPLEMENT THE 16 EXITING METHODS FROM THEIR PUBLICLY RELEASED CODES ON OUR PRE-PROCESSED CUB BIRD DATA WITH SAME TRAINING STRATEGY. FOR METHODS THAT PURSUE OPTIMAL CLASSIFICATION PERFORMANCE WITHOUT A SPECIFIED ATTENTION VISUALISATION MECHANISM, WE RESORT TO GRADCAM [68] AND TRANSRELEVANCE [10] AS TWO GENERIC WAYS FOR CNN AND TRANSFORMER TYPE OF MODELS, WHICH COVERED MOST NETWORK BACKBONE CHOICES ACROSS DISCIPLINES IN COMPUTER VISION WORLD NOWADAYS. MORE DETAILS IN TEXT

	Method	Acc				R@1				Improvability	Specificity	TEMI
		S_{expert}	S_{novice}	S_δ	$S_{novice+\delta}$	S_{expert}	S_{novice}	S_δ	$S_{novice+\delta}$			
	Upper-bound	79.11	56.04	n/a	79.11	40.71	57.28	n/a	40.71	100.0	100.0	100.0
	Lower-bound				56.04				57.28	0.00	0.00	0.00
Traditional FGVC	DFL (CVPR18 [86])	79.11	56.04	71.83	69.77	40.71	57.28	35.15	54.26	59.53	38.63	46.86
	NTS (ECCV18 [91])			79.01	71.53			41.18	55.05	67.16	28.11	39.63
	PC (ECCV18 [23])			79.03	71.91			40.31	55.10	68.80	29.63	41.40
	DCL (CVPR19 [12])			65.64	70.98			36.58	51.03	64.77	36.14	46.39
	CrossX (ICCV19 [54])			78.56	70.14			41.22	55.53	61.14	28.04	38.44
	CAL (ICCV21 [63])			78.91	71.52			40.23	54.97	67.11	29.77	41.24
	PMG (TPAMI21 [22])			78.77	71.69			40.51	55.81	67.85	29.28	40.90
	DeiT-B (ICML21 [76])			64.60	69.74			35.66	53.14	59.40	37.74	46.16
Model explanation	Lime (SIGKDD16 [65])	79.11	56.04	76.63	71.22	40.71	57.28	38.35	55.06	65.81	33.05	44.00
	IEBB (ICCV17 [26])			69.62	71.21			36.50	53.66	65.77	36.28	46.76
	IntegratedGrad (ICML17 [74])			62.18	67.89			35.73	52.29	51.38	37.62	43.44
	SmoothGrad (arXiv17 [70])			58.14	66.98			34.50	51.41	47.44	39.77	43.27
	IBA (ICLR19 [67])			77.39	70.06			39.45	54.60	60.79	31.13	41.17
	CVE (ICML19 [28])			69.91	69.97			37.18	54.21	60.38	35.09	44.39
	INTER (CVPR20 [39])			65.95	70.06			34.33	53.78	60.79	40.01	48.30
	PathwayGrad (CVPR21 [43])			63.37	69.29			39.95	55.22	57.45	30.25	39.64
S_{novice}	Bubble (TPAMI 15 [18])	79.11	56.86	63.08	68.17	56.50	72.13	45.89	68.93	50.83	36.38	42.41
	Drawing (CVPR 20 [85])		57.11	65.53	69.79	55.57	71.09	45.82	69.31	57.64	35.55	43.97
	Beginner (CVPR 21 [9])		58.32	66.87	71.22	58.87	72.56	46.15	71.21	62.05	36.40	45.88
Ours	S_{expert}	79.11	56.04	79.11	71.60	40.71	57.28	40.71	55.20	67.45	28.93	40.49
	$\hat{S}_\delta$			65.60	70.34			27.46	54.64	61.99	52.06	56.59
	S_δ			67.76	72.31			16.30	53.54	**70.52**	**71.54**	**71.03**

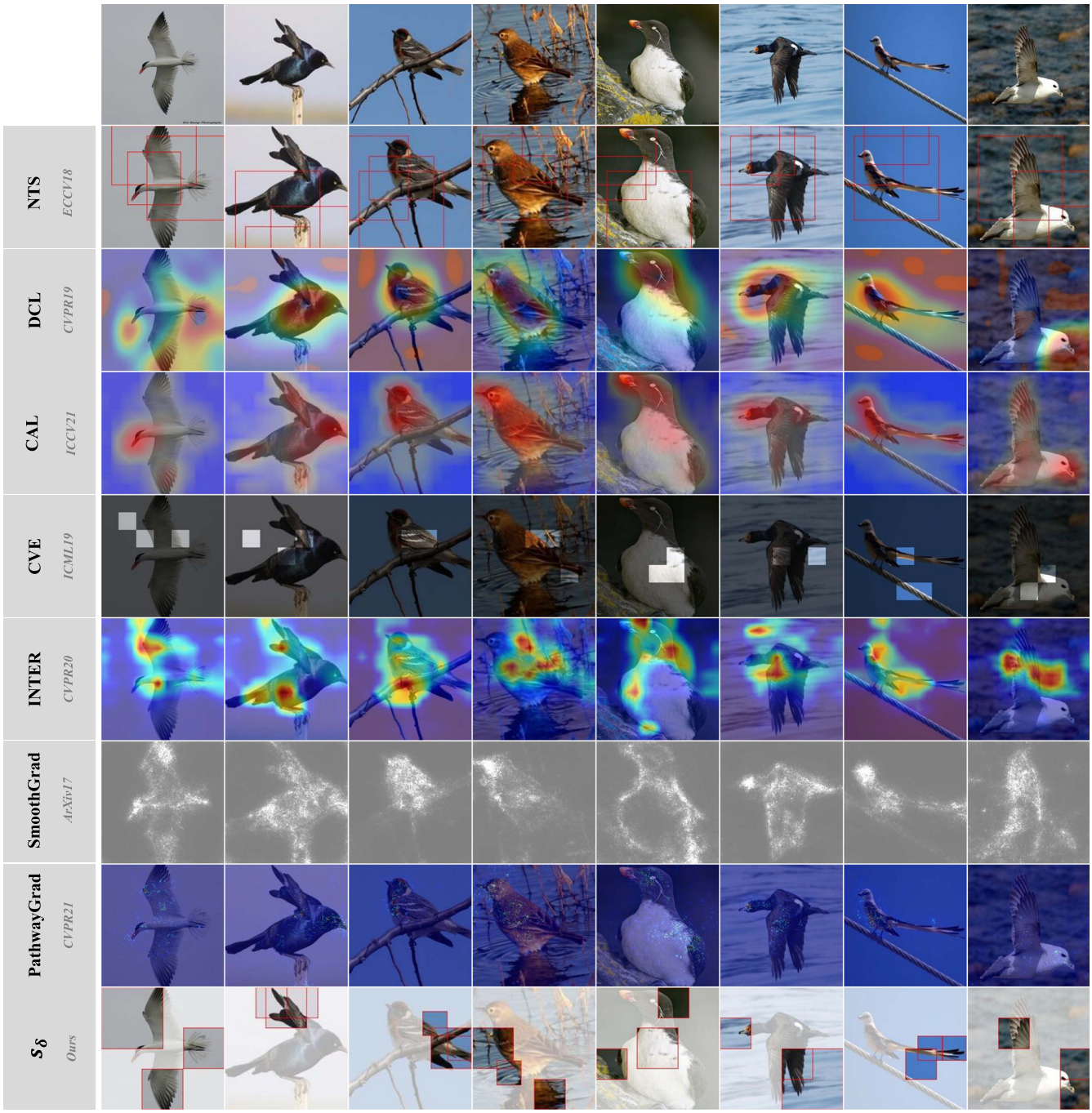

Fig. 8. Qualitative comparisons of the recognition decisive regions between our proposed S_δ and other methods. For NTS, DCL, CAL, we leverage GradCAM [68] to detect model attentions in the input image space with a unified approach. For CVE, INTER, SmoothGrad, PathwayGrad, we re-implement their public released code on our pre-processed CUB bird data.

explanation [26], [28], [39], [43], [65], [67], [70], [74]. For works that do not specify a concrete model visualisation approach, we adopt GradCAM [68] and TransRelevance [10] as two generic ways for CNN and Transformer type of models respectively. For fair comparison, we process the saliency/heatmap type of visualisations wherever applicable (as those in Fig. 8) into the same form of bounding boxes just like S_δ. We do so by assigning each local pixel block with the bounding box that has the largest normalised area intersection ratio. A leaderboard of TEMI applied to state-of-the-art attention models is maintained at: https://www.dongliangchang.cn/Making/leaderboard.html. Readers are encouraged to suggest new entries.

We can make the following observations: *Observation 1:* While the performance of improvability vary, what really separate these baselines apart are the dramatically different specificity scores, which ultimately leads to the conclusion of the superiority of our proposed method under TEMI (71.03

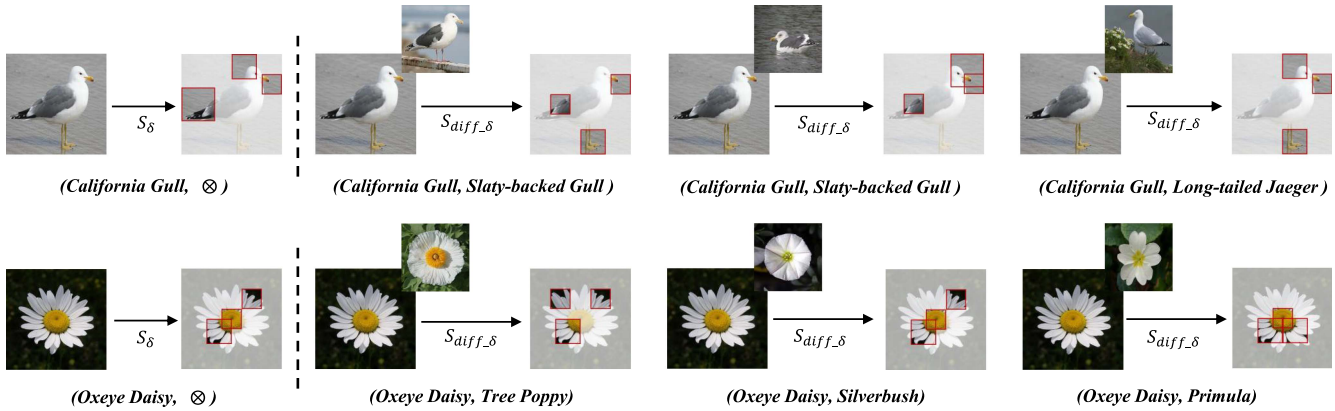

Fig. 9. Comparative guidance from two different S_δs. We explore the possibility of providing comparative and consequently even finer visual attentional guidance. Such guidance then adaptively characterises the key recognisable trait of one image instance by additionally considering the separability to another reference input. Here we showcase two examples with their Top 3 attended visual regions from S_δ and different instantiations of S_{diff_δ}.

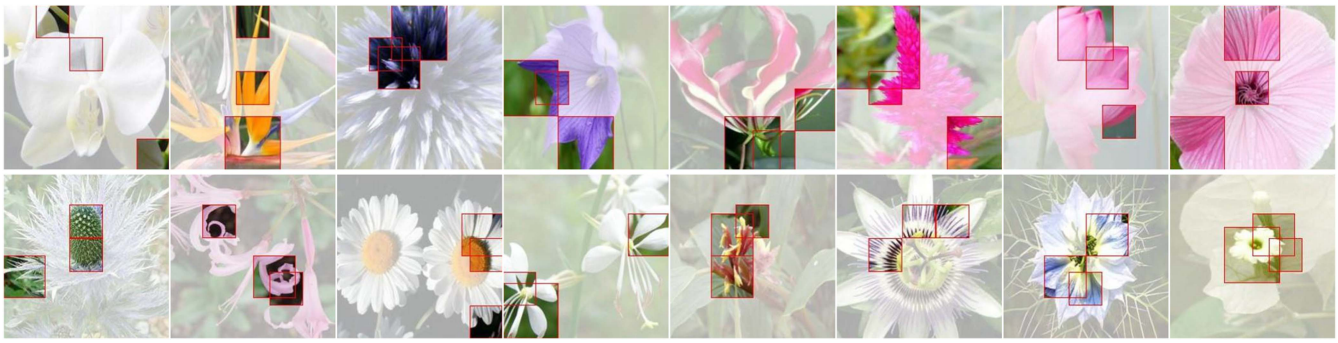

Fig. 10. S_δ works beyond birds. We examine the efficacy of our proposed method in helping people like you and me towards better recognition of flower types. We demonstrate some examples of typical Top 3 visual attentions of S_δ when trained on fine-grained flower dataset Oxford-102 [59].

versus the second best 48.30 (INTER)). This is not surprising. On the one hand, these fancy attention visualisations we all got accustom to in the existing FGVC literature have, almost without exception, comprised of recognition-focused information gained from experts – they must in a sense in order to achieve expert-level recognition performance. On the other, they are generally less digestible as transferable human knowledge due to poor specificity, i.e., they are not showcasing novice perceivable visual hints that they can specifically attend to and gain knowledge from. Qualitative evidence in Fig. 6 tells the same story, where only our proposed S_δ exhibits both local and global visual explanations, e.g., all top three visual attentions in S_δ can repeatedly focus on one visual trait as shown in the second column, a phenomenon we fail to find in other competitors. *Observation 2:* The progressively better FGVC quantitative benchmarks in recent years (i.e., from DFL to DeiT-B) do not guarantee more efficient knowledge transfer. On the contrary, they perform worse than the different types of works from model interpretability/explainability field most of the times (i.e., from Lime to PathwayGrad), works that usually sacrifice finer discriminative performance for explainability tradeoff. This suggests that for transferring knowledge from a visual classification expert, the key ask facing researchers is perhaps not to further the endeavours in refining expert performance, but exploring ways to explain and present the already superior expert knowledge to human. *Observation 3:* Methods achieve relatively good TEMI scores (e.g., DFL, DCL, IEBB, CVE, INTER), once again, are those of significantly higher specificity scores (> 35), even at the expense of very mediocre improvability performance (e.g., DFL: 59.53). Some closer inspection may reveal why more "specific" visual regions have been carved out. DFL decomposes the classification problem into the learning of a bank of convolutional filters. Filters with only strong activations are highlighted, which means *not all* visual attentions corresponding to model decisions are selected, hence specific. Similarly, DCL introduces destruction and construction learning that forces model to be only able to attend to a subset of visual receptive field, INTER discovers object part segments and only identify a subset of them that are critical for classification, both are then naturally specific because of the "subset" operation. *Observation 4:* choice of characterising S_{novice} still matters under TEMI measurements. The advantage of our proposed method simply disappears when S_{novice} is not executed as fine-grained image caption aggregate. The side effect of replacing S_{novice} with a learning-based $\hat{S}_{novice}$ under TEMI is also more significant than one may reckon. This suggests the lack of image caption information during test-time remains an open case that is largely unsolved.

IV. Further Analysis on S_δ

Comparative Guidance From Two Different S_δ: We have described how to obtain S_δ from multi-class recognition perspective. We further provide a pilot study here on another common request in practical scenarios, where end users ask the expert system to offer *instance-level comparative* answer on what makes one image different over another reference image. Specifically, given a query-reference image pair $\{x_q, x_r\}$ and their corresponding expert-exclusive visual attentions $\{S_{q\_\delta}, S_{r\_\delta}\}$, our objective is to find S_{diff_δ} that uniquely attends to visual regions belonging to x_q, while not being shared by x_r. Solving it seems straightforward by calculating $S_{diff\_\delta} = \max(S_{q\_\delta} - S_{r\_\delta}, \Delta_{threshold})$. Yet in hindsight, there is a problem: S_{q_δ} and S_{r_δ} are acting upon two different feature pools $\{F_{q\_pool}, F_{r\_pool}\}$, which naturally raises the concern of whether they are subtractable at the first place. We quantify the comparability ($m \in \mathbb{R}^N$) between S_{q_δ} and S_{r_δ} by measuring the similarity of the visual features they attend to. We obtain the global feature affinity matrix A between F_{q_pool} and F_{r_pool}[5] before defining m as a symmetric max operation on both rows and columns of A. With now m calibrating the two image instances, we obtain the final formulation of S_{diff_δ} as

$$S_{diff\_\delta} = \max((S_{q\_\delta} - S_{r\_\delta}) \odot \text{Softmax}(m), \Delta_{threshold}) \tag{14}$$

where $\Delta_{threshold}$ is a positive value that reflects the minimum attention discrepancy we permit (we set $\Delta_{threshold} = 0$ in practice). We show some qualitative evidence in Fig. 9, where S_{diff_δ} is able to dynamically adjust its visual attentions on one image input according to different referenced images. These adjustments meanwhile align well with our intuitions, see how S_{diff_δ} attends to the bird leg when that is visible in both query (California Gull) and reference (Slaty-backed_Gull) image, and knowingly diverts to other discriminative parts when the reference bird is under water with legs becoming visually inaccessible.

S_δ Works Beyond Birds: We repeat the learning process on the fine-grained flower dataset [59] and conduct the same human study pipeline as with birds. The mCP performance of 69.39% when displaying Top 3 visual region from S_δ confirms its efficacy as digestible knowledge in helping human participants to better recognise an unknown flower type. We show some visualisations of S_δ in Fig. 10.

S_δ Improves Fine-Grained Visual Analysis Performance: We embed S_δ into existing FGVC frameworks aiming to help model better localise to the most discriminative visual regions (detail in Section II-D). We confirm in Table IV(a) that S_δ is indeed a promising universal FGVC booster regardless of the base models built upon. Notably, our result also improves over CVL and PMA, two methods that have already specifically hedged their bets on the fine-grained textural information for more discriminative attention modelling. Similar with FGVC, we also can see in Table IV(b) and (c) that S_δ also positively impacts fine-grained visual retrieval and fine-grained visual localisation tasks.

[5] $A_{i,j}$ is the cosine similarity between $F^i_{q\_pool}$ and $F^j_{r\_pool}$.

TABLE IV
S_δ Improves Existing Fine-Grained Visual Analysis Methods When Exploited for Providing Discriminantive Localisation Information. ‡: Re-Implementation of the Publicly Available Released Code. §: Methods That Use Fine-Grained Text Descriptions as Side Information

Method	CUB-Bird-200		Oxford-Flower-102	
	Baseline‡	**Ours**	Baseline‡	**Ours**
B-CNN (ICCV15 [52])	87.16	**87.78**	96.77	**97.35**
NTS (ECCV18 [91])	87.02	**87.54**	95.84	**96.51**
PC (ECCV18 [23])	86.71	**87.34**	97.03	**97.42**
DCL (CVPR19 [12])	87.31	**87.86**	96.49	**97.12**
CrossX (ICCV19 [54])	87.36	**87.84**	97.06	**97.36**
PMG (ECCV20 [21])	89.62	**89.74**	98.02	**98.21**
DeiT-B (ICML21 [76])	90.04	**90.15**	98.16	**98.28**
ViT-B-16 (ICLR21 [20])	90.33	**90.42**	99.04	**99.17**
TransFG (AAAI22 [32])	91.72	**91.78**	99.65	**99.67**
CVL§ (CVPR17 [33])	86.74	**87.02**	96.87	**97.23**
PMA§ (TIP20 [71])	88.70	**88.92**	97.12	**97.65**

(a) Fine-grained visual classification (acc(%)).

Method	CUB-Bird-200		Oxford-Flower-102	
	Baseline‡	**Ours**	Baseline‡	**Ours**
HashNet (ICCV17 [7])	14.76	**20.52**	34.96	**36.31**
DPN (IJCAI20 [25])	26.68	**30.07**	27.65	**29.29**
Ortho (NeruIPS21 [35])	32.25	**37.03**	36.82	**37.98**

(b) Fine-grained visual retrieval (mAP@all(%)).

Method	CUB-Bird-200		Oxford-Flower-102	
	Baseline‡	**Ours**	Baseline‡	**Ours**
ADL (ICCV19 [14])	45.53	**46.74**	–	–
TS-CAM (ICCV21 [27])	54.50	**65.32**	–	–
TRT (BMVC22 [72])	75.60	**76.30**	–	–

(c) Fine-grained visual localisation (Top-1 Loc(%)). As the Oxford-Flower-102 dataset lacks bounding box annotations, evaluation becomes infeasible.

Bold indicates the best results.

V. Discussion

Retrieval-Based Human Study: While there is evidence showing human participants are able to tell the fine-grained differences under our provided visual guidance, they do not correlate well with the common intuition that successful recognition is to be able to spell out the names (labels)! Admittedly this can pose itself as a potential limitation for some. We argue however that this might already approach the best practical scenario one could get. Asking humans to link a biological name ("Whip-poor-will") to a specific (bird) image itself is a hard task, which requires extensive professional training and in turn makes human study less feasible without considerable budget and resources. Our final solution therefore excludes the name labelling process from human study so to make it accessible for all like you and me. We propose a query-gallery retrieval approach that simulates the human visual recognition process under a book guide – success is counted when visually fine-grained differences are spotted and images of same species are recognised. We note that such human study approach is also reflected in some existing FGVC works with a particular focus on studying human-machine visual interaction [16], [18].

Image Captioning and its Limitation on Applying to Other Domains: So far, we mainly revolve our empirical analysis around a bird domain expert and show some initial evidence on flowers in Fig. 10. We have attempted to include more daily fine-grained

TABLE V
REPRESENTING S_{novice} VIA OFF-THE-SHELF LARGE-SCALE PRE-TRAINED VISUAL-LANGUAGE FOUNDATION MODELS. DESPITE OF OUR EFFORTS TO INTRODUCE A POST-HOC APPROACH IN SECTION II-C1 AND IMPORTANTLY ALLOWS OUR MODEL TO PERFORM WELL EVEN WITHOUT HUMAN IMAGE CAPTIONING DURING DEPLOYMENT STAGE, SUCCESSFUL LEARNING IN OUR CASE STILL CRUCIALLY HINGES ON THE EXPENSIVE ANNOTATIONS OF PER-IMAGE FINE-GRAINED CAPTION AGGREGATE. ONE CONSEQUENCE IS THE LIMITED APPLICATION TO OTHER IMAGE DOMAINS WHERE SUCH ANNOTATIONS ARE NOT AVAILABLE. WE EXPLORE HERE WHETHER IT IS POSSIBLE TO LEVERAGE THE POWERFUL PRE-TRAINED FOUNDATIONAL MODELS TO BYPASS THIS ISSUE AND GET A NEGATIVE ANSWER

Raw Image	Caption Aggregate [64]	BLIP [46]	ViT-GPT2 [60]	Raw Image	Caption Aggregate [64]	BLIP [46]	ViT-GPT2 [60]
	-	there is a cupcake with chocolate frosting on top of it	a cake with a white frosting on top		-	soccer player in blue uniform playing soccer on a field	a soccer player kicking a soccer ball
	This bird has a yellow-breasted, a black cheek patch, and a white superciliary, ...	a close up of a yellow bird perched on a branch of a tree	a small bird sitting on a branch		This flower has a central white blossom surrounded by large pointed red petals ...	there is a picture of a flower that is in the middle of a flower	a flower in a flower pot on a table
	-	a picture taken from the front of a car	a car parked in a parking lot		-	a close up of a plane flying through the air	a large jetliner flying through a blue sky

object categories to entertain a wider audience but failed. The first problem confronting us is the lack of fine-grained image caption aggregates to optimise S_{novice} for the relevant open benchmarks (e.g., car [45], airplane [56]). To address this, we have leveraged large-scale pre-trained vision-language models to generate captions for each image [46], [60]. These models built upon large-scale internet data are believed to be capable of generating detailed captions for complex visual scenes (the first row of Table V). However, as we show in the last two rows of Table V, they still struggle hard in providing the object descriptions of similar granularity level compared to the caption aggregates commonly observed in this work. We therefore leave this issue to a more systematic study in future work. Prompt/transfer learning or test time chain of thoughts are two promising ways forward to enable existing GPT-style multi-modal foundation models [4], [37] for finer visual description generations.

Connections to Existing Related Fields: i) *FGVC:* FGVC has been an extensively studied problem, seeing itself with a mass of literature – please refer to the latest TPAMI FGVC survey [87]. Investment so far however is mostly uni-directional, aiming to elevate the performance upper bound of a classification model on standard benchmarks. Only until very recently, the such single-minded pursuit of better expert label prediction through expensive large-scale training is formally challenged [9], [15], [82]. One main line of arguments is to question the practicality of expert labels, which are generally less interpretable to domain non-experts. Relevant efforts are therefore made towards more democratically digestible FGVC models including extending single-label prediction ("flamingo") to predictions of label hierarchy ("bird" ⇒ "Phoenicopteriformes" ⇒ "Phoenicopteridae" ⇒ "flamingo") [9], replacing expert labels with Wikipedia text paragraphs [15], or easing the understanding of expert label via decision tree [82]. Our work shares a similar spirit in re-purposing FGVC models beyond expert label predictions, but does so by posing a completely different question: how AI expert helps you and me become better experts ourselves, i.e., can expertly understanding encoded in AI be transferable for human consumption? We have given a successful first stab to this question. ii) *Attention Model & Explainable Computer Vision:* AI needs explainability. Despite the great stride AI has made, there are still limited understanding and insights on how AI actually works, which is fundamental to achieving AI trust in an increasingly AI-governed world. Explainable Computer vision is a field set up to answer such calls and has gained traction in recent years. Mainstream approaches have been dichotomised into providing post-hoc model explanations [8], [68], [80] or making models intrinsically more explainable themselves [11], [38], [39]. Our framework of learning an optimal subset from the visual attention pool belongs to the latter type and is one of its most representative instantiations – explainable models via compositional (attentional) reasoning. In other words, the way how our model is built is explainable in nature because it can attribute its decision to its constituent parts. However instead of purely data-driven attention learning as in past works [38], [39], [63], we are unique in that we explicitly consider the prior knowledge of non-expert humans and exclude it from our reasoning chain. iii) *Machine Teaching.* Our work is also closely related to the field of machine teaching [13], [41], [55], [84]. In a typical machine teaching setting, participants are first shown an image that they most likely find alien and asked to label it. Once the labelling is finished, the correct answer will be given and the process continues until the system believes the participants have learned the visual concept. The goal of machine teaching often resembles that of active learning, as to identify the optimal sample subset from an image gallery and simultaneously decide their best presentation order (with replacement). Such learning again is to gear a participant towards spelling out the right biological name like "Geococcyx", in a stark contrast to this work, where focus is on visual differentiation between "Geococcyx" and "Corythaixoides" without the unnecessarily extra efforts on memorising them linguistically. Visual attention of a FGVC model [12], [86], [90], [91] has also been exploited for machine learning and is believed to help participants to grasp a visual concept faster and better. The real utility of these FGVC-initiated visual regions is however less feasible to quantify. Their role is designed to be auxiliary, to the main task of finding the most informative image sequence.

VI. CONCLUSION

Results from a large-scale human study suggest that we can indeed obtain transferable knowledge from a FGVC model (i.e., AI), that improves human's ability of distinguishing between fine-grained objects. This is made possible by our proposal of

representing knowledge as visual regions attended exclusively by domain experts, and a novel multi-stage cross-modal learning framework as an implementation. We also propose TEMI, a quantifiable metric that is proven to be successful in crudely evaluating the effect of AI $\rightarrow$ Human knowledge transfer. We hope this crucially enables future investments on the stream of works to become benchmarkable without resorting to extensive human studies.

REFERENCES

[1] TextBlob: Simplified text processing, 2017. [Online]. Available: https://textblob.readthedocs.io/

[2] R. P. Abelson, E. E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, and P. H. Tannenbaum, *Theories of Cognitive Consistency: A Sourcebook*. Chicago, IL, USA: Rand McNally, 1968.

[3] A. R. Akula and S.-C. Zhu, "Attention cannot be an explanation," 2022, *arXiv:2201.11194*.

[4] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," 2022, *arXiv:2204.14198*.

[5] T. Berg and P. Belhumeur, "POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 955–962.

[6] I. Biederman, S. Subramaniam, M. Bar, P. Kalocsai, and J. Fiser, "Subordinate-level object classification reexamined," *Psychol. Res.*, vol. 62, pp. 131–153, 1999.

[7] Z. Cao, M. Long, J. Wang, and P. S. Yu, "HashNet: Deep learning to hash by continuation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5609–5618.

[8] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–19.

[9] D. Chang, K. Pang, Y. Zheng, Z. Ma, Y.-Z. Song, and J. Guo, "Your "Flamingo" is my "bird": Fine-grained, or not," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11471–11480.

[10] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 782–791.

[11] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8928–8939.

[12] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5152–5161.

[13] Y. Chen, A. Singla, O. Mac Aodha, P. Perona, and Y. Yue, "Understanding the role of adaptivity in machine teaching: The case of version space learners," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1483–1493.

[14] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4256–4271.

[15] S. Choudhury, I. Laina, C. Rupprecht, and A. Vedaldi, "The curious layperson: Fine-grained image recognition without expert labels," 2021, *arXiv:2111.03651*.

[16] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1153–1162.

[17] J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei, "Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3450–3457.

[18] J. Deng, J. Krause, M. Stark, and L. Fei-Fei, "Leveraging the wisdom of the crowd for fine-grained recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 666–676, Apr. 2016.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[20] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.

[21] R. Du et al., "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 153–168.

[22] R. Du, J. Xie, Z. Ma, D. Chang, Y.-Z. Song, and J. Guo, "Progressive learning of category-consistent multi-granularity features for fine-grained visual classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9521–9535, Dec. 2022.

[23] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1–17.

[24] H. Ebbinghaus, "Memory: A contribution to experimental psychology," *Ann. Neurosci.*, vol. 20, pp. 155–156, 2013.

[25] L. Fan, K. W. Ng, C. Ju, T. Zhang, and C. S. Chan, "Deep polarized network for supervised learning of accurate binary hashing codes," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, Art. no. 115.

[26] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3449–3457.

[27] W. Gao et al., "TS-CAM: Token semantic coupled attention map for weakly supervised object localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2866–2875.

[28] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 2376–2384.

[29] E. R. Grant and M. J. Spivey, "Eye movements and problem solving: Guiding attention guides thought," *Psychol. Sci.*, vol. 14, pp. 462–466, 2003.

[30] T. Gupta, A. Vahdat, G. Chechik, X. Yang, J. Kautz, and D. Hoiem, "Contrastive learning for weakly supervised phrase grounding," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 752–768.

[31] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," *J. Mach. Learn. Res.*, vol. 13, pp. 307–361, 2012.

[32] J. He et al., "TransFG: A transformer architecture for fine-grained recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 852–860.

[33] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7332–7340.

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

[35] J. T. Hoe, K. W. Ng, T. Zhang, C. S. Chan, Y.-Z. Song, and T. Xiang, "One loss for all: Deep hashing with a single cosine similarity based learning objective," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 24286–24298.

[36] B. Hommel, C. S. Chapman, P. Cisek, H. F. Neyedli, J.-H. Song, and T. N. Welsh, "No one knows what attention is," *Attention Percep. Psychophys.*, vol. 81, pp. 2288–2303, 2019.

[37] S. Huang et al., "Language is not all you need: Aligning perception with language models," 2023, *arXiv:2302.14045*.

[38] S. Huang, X. Wang, and D. Tao, "Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 600–609.

[39] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8659–8669.

[40] R. Ji et al., "Attention convolutional binary neural tree for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10465–10474.

[41] E. Johns, O. Mac Aodha, and G. J. Brostow, "Becoming the expert-interactive multi-class machine teaching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2616–2624.

[42] S. Joung, S. Kim, M. Kim, I.-J. Kim, and K. Sohn, "Learning canonical 3D object representation for fine-grained recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1015–1025.

[43] A. Khakzar, S. Baselizadeh, S. Khanduja, C. Rupprecht, S. T. Kim, and N. Navab, "Neural response interpretation through the lens of critical pathways," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13523–13533.

[44] S. Kim, J. Nam, and B. C. Ko, "ViT-NeT: Interpretable vision transformers with neural tree decoder," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 11162–11172.

[45] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2013, pp. 554–561.

[46] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[47] Z. Li, Y. Sun, L. Zhang, and J. Tang, "CTNet: Context-based tandem network for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9904–9917, Dec. 2022.
[48] Z. Li, H. Tang, Z. Peng, G.-J. Qi, and J. Tang, "Knowledge-guided semantic transfer network for few-shot image recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 6, 2023, doi: 10.1109/TNNLS.2023.3240195.
[49] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2070–2083, Sep. 2019.
[50] Z. Li, J. Tang, L. Zhang, and J. Yang, "Weakly-supervised semantic guided hashing for social image retrieval," *Int. J. Comput. Vis.*, vol. 128, pp. 2265–2278, 2020.
[51] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1666–1674.
[52] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.
[53] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–16.
[54] W. Luo et al., "Cross-X learning for fine-grained visual categorization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8241–8250.
[55] O. Mac Aodha, S. Su, Y. Chen, P. Perona, and Y. Yue, "Teaching categories to human learners with visual explanations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3820–3828.
[56] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*.
[57] W. J. McGuire, "The current status of cognitive consistency theories," in, *Cognitive Consistency: Motivational Antecedents and Behavioral Consequents*. New York, NY, USA: Academic, 1966.
[58] M. Nauta, R. Van Bree, and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14928–14938.
[59] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis. Graph. Image Process.*, 2008, pp. 722–729.
[60] NLP Connect, "vit-gpt2-image-captioning," Hugging Face, 2022, doi: 10.57967/hf/0222.
[61] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv: 1807.03748*.
[62] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg, "From large scale image categorization to entry-level categories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2768–2775.
[63] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1005–1014.
[64] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 49–58.
[65] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.
[66] B. Roads, M. C. Mozer, and T. A. Busey, "Using highlighting to train attentional expertise," *PLoS One*, vol. 11, 2016, Art. no. e0146266.
[67] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–18.
[68] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[69] Y. Shu, B. Yu, H. Xu, and L. Liu, "Improving fine-grained visual recognition in low data regimes via self-boosting attention mechanism," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 449–465.
[70] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv: 1706.03825*.
[71] K. Song, X.-S. Wei, X. Shu, R.-J. Song, and J. Lu, "Bi-modal progressive mask attention for fine-grained recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 7006–7018, 2020.
[72] H. Su, Y. Ye, Z. Chen, M. Song, and L. Cheng, "Re-attention transformer for weakly supervised object localization," in *Proc. Brit. Mach. Vis. Conf.*, 2022, Art. no. 70.
[73] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 834–850.
[74] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.
[75] L. Thabane et al., "A tutorial on pilot studies: The what, why and how," *BMC Med. Res. Methodol.*, vol. 10, pp. 1–10, 2010.
[76] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[77] J. W. Tukey et al., *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley, 1977.
[78] G. Van Horn, R. Qian, K. Wilber, H. Adam, O. Mac Aodha, and S. Belongie, "Exploring fine-grained audiovisual categorization with the SSW60 dataset," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 271–289.
[79] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[80] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, and S. Behnke, "Interpretable and fine-grained visual explanations for convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9089–9099.
[81] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200–2011 dataset," California Inst. Tech., California, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
[82] A. Wan et al., "NBDT: Neural-backed decision tree," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–12.
[83] L. Wang, J. Huang, Y. Li, K. Xu, Z. Yang, and D. Yu, "Improving weakly supervised visual grounding by contrastive knowledge distillation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14085–14095.
[84] P. Wang, K. Nagrecha, and N. Vasconcelos, "Gradient-based algorithms for machine teaching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1387–1396.
[85] S. Wang, X. Chen, Y. Wang, M. Long, and J. Wang, "Progressive adversarial networks for fine-grained domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9210–9219.
[86] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4148–4157.
[87] X.-S. Wei et al., "Fine-grained image analysis with deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8927–8948, Dec. 2022.
[88] F. Xiao, L. Sigal, and Y. Jae Lee, "Weakly-supervised visual grounding of phrases with linguistic structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5253–5262.
[89] L. Yang et al., "Dynamic MLP for fine-grained image classification by leveraging geographical and temporal information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10935–10944.
[90] X. Yang, Y. Wang, K. Chen, Y. Xu, and Y. Tian, "Fine-grained object classification via self-supervised pose alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7389–7398.
[91] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 438–454.
[92] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.
[93] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5219–5227.
[94] H. Zhou et al., "Rethinking soft labels for knowledge distillation: A Bias–variance tradeoff perspective," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–15.

Dongliang Chang received the MEng degree in Internet of Things engineering from the Lanzhou University of Technology, China, in 2019. He is currently working toward the PhD degree with the Beijing University of Posts and Telecommunications. His research interest lies at the intersection of deep learning and computer vision, with a specific focus on fine-grained visual understanding.

Kaiyue Pang received the dual bachelor's (first class honours) degree from the Queen Mary University of London and the Beijing University of Posts and Telecommunications, in 2016, and the PhD degree from the Queen Mary University of London, in 2020. He is currently a researcher with SketchX AI. He was previously a research fellow of computer vision and machine learning with the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. He has maintained a track record of more than fifteen publications on human sketch data analysis and fine-grained visual understanding, mostly at flagship computer vision and machine learning conferences and journals, including CVPR, ICCV, ECCV, *International Journal of Computer Vision*, IEEE Transactions on Image Processing. His research marries advances in machine learning with insights from computer vision, and leads to applications that often improve human well-beings in various real-world visual activities.

Ruoyi Du received the BEng degree in telecommunication with management from the Beijing University of Posts and Telecommunications (BUPT), in 2020, where he is currently working toward the PhD degree. His research interests include pattern recognition and computer vision.

Yujun Tong received the BEng degree with a major on telecommunication with management from the Beijing University of Posts and Telecommunications, in 2021, where he is currently working toward the PhD degree. His research interests include pattern recognition and computer vision.

Yi-Zhe Song (Senior Member, IEEE) received the bachelor's (first class honours) degree from the University of Bath, in 2003, the MSc (with Best Dissertation Award) degree from the University of Cambridge, in 2004, and the PhD degree in computer vision and machine learning from the University of Bath, in 2008. He is a chair professor of computer vision and machine learning, and director of SketchX Lab, Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. He is an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI), and *Frontiers in Computer Science – Computer Vision*. He was a program chair for British Machine Vision Conference (BMVC) 2021, and regularly serves as area chair (AC) for flagship computer vision and machine learning conferences, most recently at CVPR'22 and ICCV'21. He is a fellow of the Higher Education Academy, as well as full member of the EPSRC Review College.

Zhanyu Ma (Senior Member, IEEE) received the PhD degree in electrical engineering from the KTH Royal Institute of Technology, Sweden, in 2011. He is currently a professor with the Beijing University of Posts and Telecommunications, Beijing, China, since 2019. From 2012 to 2013, he was a postdoctoral research fellow with the School of Electrical Engineering, KTH. He has been an associate professor with the Beijing University of Posts and Telecommunications, Beijing, China, from 2014 to 2019. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in computer vision, multimedia signal processing.

Jun Guo received the BEng and MEng degrees from the Beijing University of Posts and Telecommunications (BUPT), China, in 1982 and 1985, respectively, and the PhD degree from the Tohuku-Gakuin University, Japan, in 1993. At present, he is a professor and a vice president with BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and bioinformatics. He has published more than 200 papers on the journals and conferences including the *Science*, *Nature Scientific Reports*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Pattern Recognition*, AAAI, CVPR, ICCV, SIGIR, etc.