SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

March 2023, Vol. 66 132107:1–132107:16
https://doi.org/10.1007/s11432-022-3586-y

Siamese transformer with hierarchical concept embedding for fine-grained image recognition

Yilin LYU[1,2], Liping JING[1,2*], Jiaqi WANG[1,2], Mingzhe GUO[1,2], Xinyue WANG[3] & Jian YU[1,2]

[1]*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044*, China;*
[2]*Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing* 100044*, China;*
[3]*Alibaba Group, Beijing* 100102*, China*

Received 25 January 2022/Revised 18 May 2022/Accepted 16 August 2022/Published online 31 January 2023

Abstract Distinguishing the subtle differences among fine-grained images from subordinate concepts of a concept hierarchy is a challenging task. In this paper, we propose a Siamese transformer with hierarchical concept embedding (STrHCE), which contains two transformer subnetworks sharing all configurations, and each subnetwork is equipped with the hierarchical semantic information at different concept levels for fine-grained image embeddings. In particular, one subnetwork is for coarse-scale patches to learn the discriminative regions with the aid of the innate multi-head self-attention mechanism of the transformer. The other subnetwork is for finer-scale patches, which are adaptively sampled from the discriminative regions, to capture subtle yet discriminative visual cues and eliminate redundant information. STrHCE connects the two subnetworks through a score margin adjustor to enforce the most discriminative regions generating more confident predictions. Extensive experiments conducted on four commonly-used benchmark datasets, including CUB-200-2011, FGVC-Aircraft, Stanford Dogs, and NABirds, empirically demonstrate the superiority of the proposed STrHCE over state-of-the-art baselines.

Keywords fine-grained image recognition, transformer, hierarchical concept embedding, adaptive sampling, Siamese network

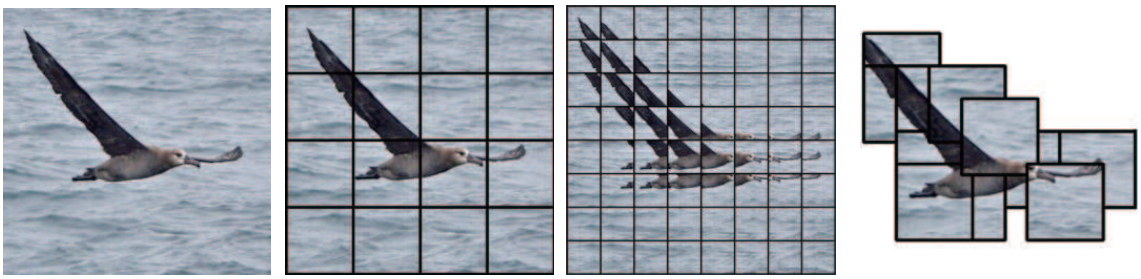
Citation Lyu Y L, Jing L P, Wang J Q, et al. Siamese transformer with hierarchical concept embedding for fine-grained image recognition. Sci China Inf Sci, 2023, 66(3): 132107, https://doi.org/10.1007/s11432-022-3586-y

1 Introduction

Fine-grained image recognition is one of the most important and challenging topics in the field of computer vision. Compared to generic image recognition, fine-grained image recognition focuses on distinguishing categories of subordinate concepts (e.g., species of birds [1, 2], variants of aircraft [3], and breeds of dogs [4]). These subordinate concepts are inherently from a hierarchy with different levels, for instance, the Caltech-UCSD birds dataset with 200 bird classes from a concept hierarchy of 37 families, 122 genera, and 200 species [5]. In this case, the images of different subordinate concepts (e.g., birds of different species) may be very similar because of the fine granularity, while the images belonging to the same subordinate concepts usually have large differences in illumination, size, posture, and background. Such low inter-class variance and high intra-class variance bring more difficulties in determining discriminative features and the correlations among them, which are the key factors for distinguishing the subtle differences between subordinate concepts.

Most efforts in the fine-grained community focus on localizing the discriminative regions either by exploiting bounding box/part annotations [6–9] or more prevalently, in a weakly-supervised manner [10–15]. To capture the correlations among regions and incorporate them into the discriminative region selection process, graph-based models were designed to discover the discriminative region group rather than individually extracting the regions [16,17]. Their main idea is to establish the correlative relationship through a grouping module and combine the module with an existing deep representation network (e.g., CNN architecture), which leads to a complicated learning model and requires high computational complexity.

* Corresponding author (email: lpjing@bjtu.edu.cn)



(a) (b) (c) (d)

Figure 1 (Color online) Illustration of different patching approaches. Given an input image (a), the transformer splits it into a sequence of nonoverlapping patches (b), which harms the local structure. A recent study alleviates this issue by generating overlapping sequences (c) but leads to high computational cost. Our proposed APS mechanism (d) adaptively samples patches from the discriminative regions without increasing the sequence length. Each square represents an image patch.

An alternative strategy considering correlations among regions is a transformer, which was originally proposed for natural language processing tasks and is popular in various computer vision tasks (such as image recognition [18], object detection [19,20], and segmentation [21]). The transformer treats an input image as a sequence of nonoverlapping image patches with global self-attention modeling for discriminative feature representation learning [18,22], as shown in Figure 1(b). Compared with the stacked convolution layer-based encoder, although spatial resolution is not downsampled, the transformer harms the local structure among discriminative patches, which are subtle but crucial in fine-grained recognition. To mitigate this issue, researchers have attempted to generate more "words" (i.e., overlapping patches) with sliding windows [23, 24], as shown in Figure 1(c), but this approach substantially increases the patch sequence size, resulting in high computational cost. Meanwhile, the existing methods ignore the hierarchical structure among fine-grained concepts. Actually, a concept hierarchy contains rich semantic information, which benefits fine-grained recognition [5, 25, 26].

In this paper, thus, we propose a Siamese transformer with hierarchical concept embedding for fine-grained image recognition, referred to as STrHCE. The proposed Siamese transformer contains two subnetworks, one for coarse-scale patches and the other for finer-scale patches, while they have the same configuration with the same parameters. Each subnetwork has a transformer architecture equipped with hierarchical concept semantic embedding (HCE). Its purpose is to encode the level-aware hierarchical information into fine-grained image embeddings. To the best of our knowledge, this work is the first to combine hierarchical semantic information with embedding learning, which then mutually reinforces each other throughout the learning process of the model. We conceptually and empirically demonstrate the superiority of this model below. In STrHCE, the subnetwork for coarse-scale patches is applied to the original fine-grained image patch sequence to determine the discriminative regions with the aid of the innate multi-head self-attention mechanism of the transformer. The other subnetwork is fed with patch sequences that are adaptively sampled from the discriminative parts rather than redundant information or background parts. Specifically, an adaptive patch sampling (APS) mechanism is proposed to approximate the attention of the original patches to grasp the attended regions and then adaptively sample the finer-scale patches to generate the patch sequence (with an equal length to the original sequence, as shown in Figure 1(d)), which provides subtle yet discriminative visual cues for fine-grained image recognition. STrHCE connects two subnetworks through a score margin adjustor (SMA) module, which takes the prediction from the original coarse-scale patches as references, concentrates on the most discriminative sampled finer-scale patches, and enforces them to generate more confident predictions.

The main contributions of this work can be summarized as follows.

• A Siamese transformer framework is proposed for fine-grained image recognition, which considers coarse-scale and finer-scale image patches to obtain confident predictions with the most discriminative patches.

• The transformer architecture is explicitly incorporated with the semantic structure information on a concept hierarchy so that the bidirectional semantic information can be flexibly captured and then consistently benefit the learning of fine-grained embeddings.

• An APS mechanism is proposed to adaptively sample the attended patches and feed the patch sequence into the Siamese transformer, which can preserve the subtle discriminative visual cues of the fine-grained objects.

• Extensive experiments are conducted on four commonly-used fine-grained benchmark datasets (CUB-200-2011, FGVC-Aircraft, Stanford Dogs, and NABirds) and demonstrate that the proposed method outperforms state-of-the-art methods.

2 Related work

In this section, we briefly review the existing fine-grained image recognition studies.

Similar to generic image recognition, learning a discriminative feature representation is essential for fine-grained image recognition. With the success of CNN architecture in computer vision, a series of CNN-based fine-grained image representation methods have been proposed. Lin et al. [27] first introduced bilinear pooling to capture the pairwise feature interactions and model the subtle differences among fine-grained images. To handle the high-dimensionality issue caused by the bilinear features, researchers tried to learn a compact form of bilinear representation [28–30]. To obtain robust representation, Li et al. [31] proposed a matrix power normalization for exploiting the geometry of bilinear matrices. To sufficiently investigate the difference between each pair of images, Zhuang et al. [32], simulating the comparison procedure of human beings, mined the contrastive clues through a mutual feature vector. To make the channel more discriminative, Gao et al. [33] tried to discover the channel-wise complementary clues by modeling the relationships between various channels.

The aforementioned methods mostly focus on the image-level features when learning the discriminative representation. However, the subtle differences among fine-grained images usually lie in partial regions rather than the entire image. In the literature, discriminative region localization and adaptive image amplification are two main strategies for enhancing the fine-grained representation.

An easy way for localizing the discriminative regions is taking the bounding box or part annotation [6–9] as supervised information to train model. However, collecting such information is expensive. Consequently, a variety of studies determine the discriminative parts in a weakly-supervised manner [10–15]. Among these studies, one direction is to learn to focus on subtle yet discriminative parts. Fu et al. [10] recurrently learned attention maps in multiple scales and constructed region-based feature representations in a mutually reinforcing manner. Zheng et al. [11] learned multiple consistency attention maps in a single scale through a channel grouping module. He et al. [13] enforced two spatial constraints to select the distinguished parts. Later, modern techniques such as multi-agent cooperative learning [14] and reinforcement learning [12, 15] were adopted to design an effective attention model for extracting discriminative regions.

The above studies try to localize the discriminative regions independently but ignore the inherent correlation among regions. In fact, early research on visual vocabulary has shown that ignoring correlation among visual words can result in unstable local features and further reduce the recognition performance [34]. To alleviate this issue in the fine-grained recognition task, Wang et al. [16] designed a discriminative feature strengthening subnetwork to explore the internal spatial correlation among regions. Later, they promoted the performance through a graph propagation subnetwork to characterize the region correlations in a criss-cross way [17]. In these models, the extra subnetwork is usually combined into the CNN network sitting on the top layer, which leads to a complicated learning model and suffers from high computational complexity.

In this paper, we propose to leverage the inherent global perception ability of the transformer to capture the correlations of different discriminative regions throughout the entire model, rather than adding an extra subnetwork. Our work is related to existing fine-grained recognition methods based on the transformer architecture such as [24, 35], but differs in three aspects: (1) A Siamese transformer framework is proposed for coarse-scale patches and finer-scale patches to produce confident predictions. (2) Compared with [24], much smaller size of overlapping patch sequences are fed into individual transformers, which are adaptively sampled by the proposed APS mechanism, guaranteeing the efficiency of the training transformer. The redundant category-irrelevant information can almost be eliminated without being limited by the space constraints of the bounding box, while Ref. [35] required multiple stages to extract the bounding boxes of the object and parts and then input them to the network separately. (3) Hierarchical structure information is integrated as extra tokens in each layer to capture the semantic correlations among subordinate concepts.

To investigate the finer-scale discriminative regions, a series of models have been proposed to crop and amplify the attended parts [6, 7, 9, 10, 36, 37]. Although these models have achieved promising per-

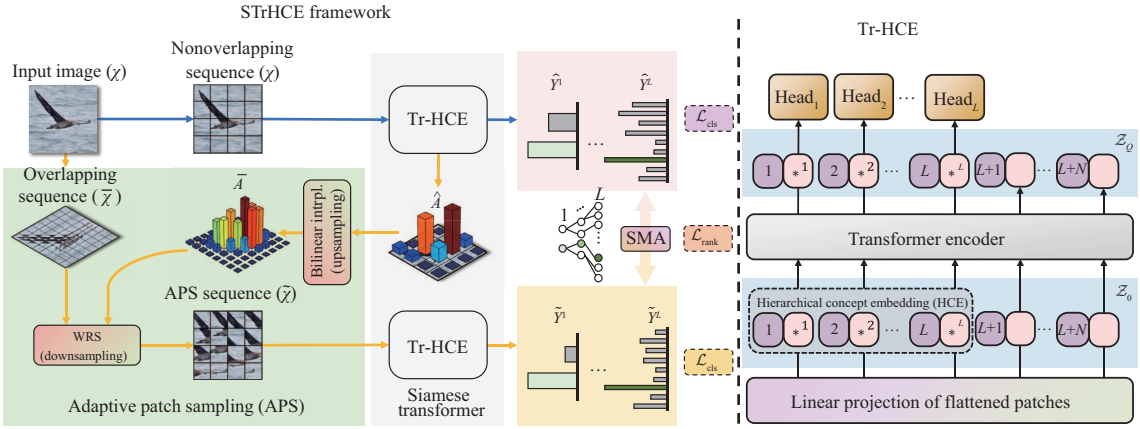


Figure 2 (Color online) Overview of the proposed Siamese transformer with hierarchical concept embedding (STrHCE) framework. The input image is first split into a sequence of nonoverlapping patches and fed into the coarse-scale transformer subnetwork equipped with hierarchical concept embedding (Tr-HCE), which encodes the hierarchical semantic information at different concept levels. The coarse-scale subnetwork is applied to the original fine-grained image patch sequence to determine the discriminative regions. Next, an APS mechanism is applied to generate fine-scale patch sequences by adaptively sampling discriminative patches according to attention information using the weighted reservoir sampling algorithm (WRS). The fine-scale patch sequences are then fed to the fine-scale subnetwork. Finally, the two subnetworks are connected through a score margin adjustor (SMA) module to enhance each other.

formance, they amplified the attended regions with the same resolutions, ignoring the importance of different regions. Recently, Zheng et al. [38] introduced an attention-based non-uniform sampling strategy for selecting different channels to preserve the corresponding fine-grained details with high resolution. Similarly, according to the class response map, Ding et al. [39] proposed a selective sparse sampling technique to highlight informative regions while preserving the surrounding context information. Unlike the above studies focusing on the attention of the top layer, our proposed APS technique leverages the attention information propagated from the input layer to the highest layer, which can provide more focused attention regions than the raw attention [40]. Meanwhile, our proposed method can obtain attention maps during model training, while these methods must introduce an extra module to calculate attention weights.

The rich semantic information in a concept hierarchy benefits fine-grained recognition [5, 25, 26]. For instance, Chen et al. [5] learned a hierarchical semantic embedding for each concept level and regularized the subordinate predictions by the superordinate predictions. He et al. [26] used a hierarchical sampling-based triplet network and a tree classifier to classify hierarchical features from coarse to fine. However, these methods are restricted to only transferring knowledge from the high level to the lower level and lack the interaction between hierarchical semantic information and image embeddings, while our proposed method mitigates these issues. Next, we provide the proposed fine-grained image recognition framework in detail.

3 STrHCE framework

In this section, we introduce the STrHCE, which simultaneously considers coarse/finer-scale image representation and concept hierarchy information among subordinate categories for fine-grained image recognition. The STrHCE framework comprises two subnetworks in a Siamese-like architecture [41] that share the same module, a transformer equipped with hierarchical concept embedding (Tr-HCE). The inputs of these two subnetworks differ, one being for coarse-scale (nonoverlapping) patch sequences and the other being for finer-scale (overlapping) patch sequences obtained by the proposed APS module. These two subnetworks are connected by an SMA module.

An overview of the proposed STrHCE framework is illustrated in Figure 2. Given an input image, we first process it into a sequence of nonoverlapping flattened patches as the input of the coarse-scale subnetwork, and transform it by the proposed Tr-HCE module into L image representations (one representation for each level of a concept hierarchy) and an attention map. To enhance the discriminative regions, the attended parts with high attention values are sampled by the APS module and taken as the input of the finer-scale Tr-HCE subnetwork. The overlapping APS patches preserve more subtle discriminative visual

cues of the fine-grained object and partially prevent the corruption of local structure in vision transformer (ViT) [18]. Finally, these two Tr-HCE subnetworks affect each other through the SMA module to enforce the discriminative regions, generating more confident predictions on subordinate concepts.

3.1 Tr-HCE

Subordinate concepts of fine-grained images are usually from a hierarchy with different levels of concepts. The nodes closer to the root of the hierarchy (i.e., high-level) refer to more abstract concepts (i.e., general superordinate concepts), while the nodes closer to the leaves (i.e., low-level) refer to more specific concepts (i.e., basic subordinate concepts). The categories of fine-grained images are usually located in the leaves. This concept hierarchy benefits fine-grained recognition [5, 26]. The main idea is to transfer knowledge learned from the high level to the lower level. Actually, a large body of research in cognitive science [42] has shown that infants learn the superordinate concepts much earlier than the subordinate concepts; meanwhile, the subordinate concepts are often affected prior to superordinate concepts during the progressive loss of knowledge in semantic dementia. Thus, the researchers of cognitive science suggest that, when building a semantic memory model, the subordinate and superordinate concepts should be stored transparently at different levels of a hierarchy. Inspired by this, we propose to equip the transformer with the novel HCE to encode the level-aware hierarchical information and incorporate it with the fine-grained image embeddings, which enables them to reinforce each other during the learning process of the entire network. By leveraging the inherent global perception ability of the transformer, the bidirectional semantic correlations among subordinate and superordinate concepts can be flexibly captured. Moreover, the fine-grained image embeddings can also consistently interact with the semantic information at different concept levels. This paradigm is obviously superior to the existing method in which only the correlations from superordinate concepts to subordinate concepts are encoded [5, 26].

Assume that there is an L-level concept hierarchy in a particular fine-grained image set. For example, CUB-200-2011 has a four-level concept hierarchy with orders, families, genera, and species, similar to Stanford Dogs, NABirds, etc. Notably, such concept hierarchies can be obtained from the literature on taxonomy or conveniently retrieved from Wikipedia; thus, this structured information can be easily exploited. Given a 2D image $X \in \mathbb{R}^{H\times W\times C}$ (H, W, and C are the height, width, and the number of channels, respectively), we feed it to the coarse-scale subnetwork branch (i.e., a transformer equipped with HCE). The image is first preprocessed into a sequence of N nonoverlapping patches, denoted by $\mathcal{X} = [X^1, \ldots, X^i, \ldots, X^N]$, where $X^i \in \mathbb{R}^{P\times P\times C}$ indicates the i-th patch, (P, P) is the size of a patch, and $N = WH/P^2$ is the number of patches which is an effective input sequence length for the transformer. Following the standard ViT setting [18], each patch in the sequence $\mathcal{X}$ is linearly mapped into a D-dimension feature space with a trainable projection $\mathcal{E}$. Here, all layers in the transformer use constant latent vector size D. To exploit the semantic information of concept hierarchy, for each input image, L trainable concept-level embeddings $\{X^1_{\mathrm{HCE}}, \ldots, X^l_{\mathrm{HCE}}, \ldots, X^L_{\mathrm{HCE}}\}$ (X^l_{HCE} is for the l-th level of hierarchy) are introduced and concatenated into the patch embeddings. Meanwhile, a learnable position embedding $\mathcal{E}_{\mathrm{pos}}$ is added to the sequence of embedded patches. As input to the transformer encoder, the resulting sequence of embedding vectors can be formulated as

$$\mathcal{Z}_0 = [X^1_{\mathrm{HCE}}, X^2_{\mathrm{HCE}}, \ldots, X^L_{\mathrm{HCE}}, X^1\mathcal{E}, X^2\mathcal{E}, \ldots, X^N\mathcal{E}] + \mathcal{E}_{\mathrm{pos}}, \tag{1}$$

where $\mathcal{E} \in \mathbb{R}^{(P^2\cdot C)\times D}$ is the weight of linear projection and $\mathcal{E}_{\mathrm{pos}} \in \mathbb{R}^{(N+L)\times D}$ is the learnable position embedding. For the transformer encoder [18] with Q layers of multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks (each block is processed by Layernorm (LN)), the output of the q-th layer can be written as follows:

$$\mathcal{Z}'_q = \mathrm{MSA}(\mathrm{LN}(\mathcal{Z}_{q-1})) + \mathcal{Z}_{q-1}, \quad q = 1, \ldots, Q; \tag{2}$$

$$\mathcal{Z}_q = \mathrm{MLP}(\mathrm{LN}(\mathcal{Z}'_q)) + \mathcal{Z}'_q, \qquad q = 1, \ldots, Q. \tag{3}$$

Unlike ViT, which only uses the first token of the last encoder layer (i.e., $\mathcal{Z}^1_Q$) as the representation to obtain the final classification result, our proposed Tr-HCE sufficiently exploits the first L tokens (i.e., $\mathcal{Z}^1_Q, \ldots, \mathcal{Z}^l_Q, \ldots, \mathcal{Z}^L_Q$) as the representation at different levels in a concept hierarchy. Each token $\mathcal{Z}^l_Q$ ($l = 1, \ldots, L$) is forwarded into the corresponding classifier head $\mathcal{H}^l$ ($l = 1, \ldots, L$) to obtain the predicted

score vectors $\hat{Y}^l \in \mathcal{R}^{K^l}$ (K^l is the number of concepts at the l-th level of hierarchy, and $l = 1, \ldots, L$) using

$$\hat{Y}^l = \mathcal{H}^l(\mathcal{Z}_Q^l), \quad l = 1, \ldots, L. \tag{4}$$

Note that the number of concept levels (L) is usually not large, e.g., less than 4 in our experimental datasets. Actually, the multi-head classifier can be trained in parallel; thus, the total computational complexity will not greatly increase.

Because of the multi-head self-attention mechanism of the transformer, the correlations across different hierarchy levels can be captured well, which benefits semantic space learning and makes the prediction less ambiguous. Thus, it should reasonably represent the fine-grained images in the learned semantic space so that the discriminative regions are highlighted to improve the final prediction.

3.2 Adaptive patch sampling

For the fine-grained image recognition task, the semantic information encoded in fine-grained objects is much denser and more difficult to detect than that in generic objects. To capture the main characteristics of fine-grained objects, more visual "words" related to the subtle but discriminative regions are intuitively required, while visual "words" related to the category-irrelevant regions should be as few as possible. The previous methods usually uniformly generate patches from every region [18,22,24]. Obviously, they cannot handle fine-grained images very well. Thus, we propose to adaptively sample the finer-scale patches from the discriminative regions with the aid of attention weights obtained from the Tr-HCE.

Suppose there are V self-attention heads. The attention weights of $L+N$ hidden features in all Tr-HCE layers can be written as

$$\mathcal{A}_q = [A_q^1, A_q^2, \ldots, A_q^v, \ldots, A_q^V], \quad q = 1, \ldots, Q; \tag{5}$$

$$A_q^v = [A_q^{v_1}; A_q^{v_2}; \cdots; A_q^{v_{L+N}}], \quad v = 1, \ldots, V. \tag{6}$$

To conveniently analyze the attention weights with the multi-head setup, following [40], we average all heads and output single attention for each token in every layer by

$$A_q = \sum_{v=1}^{V} A_q^v, \quad q = 1, \ldots, Q. \tag{7}$$

Leveraging the attention propagated from the input layer to the highest layer can provide more complete information than merely focusing on a certain layer [40]. Therefore, to sufficiently capture the attention scores of final predictions to the input tokens, the raw attention weights in all layers are recursively multiplied together through

$$A = \prod_{q=1}^{Q} (A_q + I) \in \mathbb{R}^{(L+N)\times(L+N)}, \tag{8}$$

where L is the number of levels in the concept hierarchy, N is the number of input patches, and I is an identity matrix to account for the residual connection in the transformer. As the attention of the lowest-level concepts contains discriminative cues that are most conducive for fine-scale patch learning and bear the greatest responsibility for the fine-grained recognition performance, we retain the attention scores in the L-th level of hierarchy by collecting the last N elements of the L-th row in A, i.e., $A(L, L+1 : L+N)$. By reshaping these N elements to 2D, we obtain the attention matrix

$$\hat{A} = \text{Reshape}(A(L, L+1 : L+N)) \in \mathbb{R}^{N_H \times N_W}, \tag{9}$$

where $N_H = H/P$ and $N_W = W/P$ indicate the number of patches in each column and row, respectively. Meanwhile, a min-max normalizing operation is applied to rescale the weights and obtain the final importance scores. Compared with the attention computing methods in [38, 39], $\hat{A}$ promotes the importance scores of each patch to be more accurately evaluated, so that it is much more reasonable to extract the discriminative regions.

As recently demonstrated [24], the attention matrix $\hat{A}$ provides precious clues on discriminative regions. He et al. selected the patches with maximum weight in different attention heads to represent the local information of fine-grained objects. However, as shown in [39], the contextual information near the informative region plays an important role in fine-grained image recognition. Thus, in this work, we

propose an APS mechanism to generate a finer-scale patch sequence according to the importance scores ($\hat{A}$) of the original patches.

We aim to generate more tokens (i.e., patches) for the discriminative regions and avoid high computational complexity. To this end, we first construct $\bar{N}$ overlapping patches ($\bar{\mathcal{X}}$) through a sliding window of step size S (as shown in the left part of Figure 2), where

$$\bar{N} = \bar{N}_H \times \bar{N}_W = \left\lfloor \frac{H+S-P}{S} \right\rfloor \times \left\lfloor \frac{W+S-P}{S} \right\rfloor. \tag{10}$$

To focus on the discriminative regions in this overlapping patch sequence, the bilinear interpolation technique is introduced to upsample the learned attention matrix $\hat{A}$ and construct the new attention matrix $\bar{A}$ to fit the size of the overlapping patch sequence. Mathematically, the (i,j)-th element of $\bar{A}$ can be calculated according to the nearest four inputs in $\hat{A}$ by a linear map, which can be written as

$$\bar{A}(i,j) = \sum_{\alpha,\beta=0}^{1} |1-\alpha-\{i/\mu_H\}||1-\beta-\{j/\mu_W\}|\hat{A}(m,n), \tag{11}$$

where $\alpha,\beta = \{0,1\}$, $m = \lfloor i/\mu_H \rfloor + \alpha$, $n = \lfloor j/\mu_W \rfloor + \beta$, and μ_H and μ_W are upsampling factors on a column and row, which are set as $\bar{N}_H/N_H$ and $\bar{N}_W/N_W$, respectively. $\lfloor\cdot\rfloor$ and $\{\cdot\}$ denote the integral and fractional parts, respectively.

Considering the upsampled attention weights $\bar{A}$, we select the informative and discriminative overlapping patches through weighted reservoir sampling (WRS) [43]. WRS is a streaming algorithm for sampling a subset of items from a collection of items without replacement based on the weights associated with each item. To ensure that the Tr-HCE works on the sampled patches, we define a weighted reservoir sample as

$$T = [t_1, t_2, \ldots, t_i, \ldots, t_N], \tag{12}$$

where N ($N < \bar{N}$) is equal to the size of the original patch sequence and each t_i is considered the sampling index of the overlapping patch. For notational simplicity, we view $\bar{A}$ as a 1D vector that can be indexed by t_i. Then, each weighted reservoir sample should follow the probability distribution:

$$p(T \mid \bar{A}) = \frac{\bar{A}_{t_1}}{Z} \frac{\bar{A}_{t_2}}{Z-\bar{A}_{t_1}} \cdots \frac{\bar{A}_{t_N}}{Z-\sum_{i=1}^{N-1}\bar{A}_{t_i}}, \tag{13}$$

where $Z = \sum_{i=1}^{N} \bar{A}_i$. Once T is determined, a new patch sequence $\tilde{\mathcal{X}}$ will be constructed with finer-scale overlapping patches, denoted by the APS sequence

$$\tilde{\mathcal{X}} = [\bar{\mathcal{X}}^{t_1}, \bar{\mathcal{X}}^{t_2}, \ldots, \bar{\mathcal{X}}^{t_i}, \ldots, \bar{\mathcal{X}}^{t_N}] \in \mathbb{R}^{N\times P\times P\times C}, \tag{14}$$

where $\bar{\mathcal{X}}^{t_i}$ is the sampled i-th patch from the intermediate overlapping patch sequence $\bar{\mathcal{X}}$.

Note that compared to existing studies [23,24] that significantly increase the scale of input patches, we retain the size of overlapping patch sequence $\tilde{\mathcal{X}}$ consistent with original patch sequence $\mathcal{X}$, which leads to lower computational cost. Moreover, with the aid of the importance score which aggregates the attention information of all transformer layers, the proposed APS sequence can naturally grasp the attended regions and characterize them with sufficient visual words. Meanwhile, the category-irrelevant regions can be effectively removed from the finer-scale representation. The constructed APS sequence will be forwarded to the second Tr-HCE to further determine the subtle but discriminative parts. The predicted score vector of this subnetwork (obtained by applying (1)–(4) on $\tilde{\mathcal{X}}$) is defined as $\tilde{Y}^l$ ($l = 1, \ldots, L$, and L is the number of levels of concept hierarchy). Note that two Tr-HCE subnetworks share the same configuration including all hyperparameters and weights, so we call the entire framework as Siamese transformer.

3.3 Overall model

The Tr-HCE subnetwork for the coarse-scale patch sequence $\mathcal{X}$ and that for the finer-scale patch sequence $\tilde{\mathcal{X}}$ are trained through the same classification loss on all levels of a concept hierarchy, which is formulated as

$$\mathcal{L}_{\mathrm{CLS}} = \sum_{l=1}^{L} \lambda^{L-l}[\mathcal{L}_{\mathrm{CE}}(\hat{Y}^l, Y^l) + \mathcal{L}_{\mathrm{CE}}(\tilde{Y}^l, Y^l)], \tag{15}$$

where Y^l is the ground truth one-hot vector at the l-th level and the cross-entropy loss is adopted because of its efficiency and effectiveness. $\lambda \in (0, 1]$ is a hyperparameter for balancing the influence of each concept level. As we know, the lower the concept level is, the greater the impact on the final fine-grained concept predicted result will be. Thus, we introduce the exponential value $L-l$ for λ, and λ^{L-l} will increase with the hierarchical level l.

Thus far, the coarse-scale subnetwork and fine-scale subnetwork have been trained. As a reminder, a Siamese network usually requires a similarity measure to connect each branch so that they can enhance each other [41]. In our context, we desire to not only connect two Tr-HCE subnetworks but also impose the constraint that a fine-scale subnetwork holding more discriminative information can produce more confident predictions. Consequently, an SMA module is designed to guarantee more confident predictions and, more importantly, provide a way to transfer residual information between each branch. To be specific, let t denote the index of the correct category label for the input image. Then, $\hat{Y}_t^L$ and $\tilde{Y}_t^L$ indicate the predicted probability of the ground truth label from the two subnetworks. SMA aims to enlarge $\tilde{Y}_t^L$ to a certain extent by

$$\mathcal{L}_{\text{SMA}} = \max\{0, \hat{Y}_t^L - \tilde{Y}_t^L + \epsilon\}, \tag{16}$$

where ϵ is a hyperparameter for the margin. This term will enforce $\tilde{Y}_t^L > \hat{Y}_t^L + \epsilon$ in training, which takes the prediction from the original coarse-scale and nonoverlapping patches as references, concentrates on the most discriminative sampled finer-scale and overlapping patches, and finally outputs more confident predictions.

Overall, the proposed STrHCE framework can be modeled by

$$\mathcal{L} = \mathcal{L}_{\text{CLS}} + \gamma \mathcal{L}_{\text{SMA}}, \tag{17}$$

where γ is a hyperparameter to trade off the classification and margin adjusting terms. After obtaining the trained STrHCE model, the final predicted label can be obtained by

$$t = \arg\max_k \{\tilde{Y}_k^L|_{k=1}^{K^L}\}, \tag{18}$$

where K^L is the number of fine-grained subordinate concepts.

STrHCE is the first work to leverage a transformer with HCE and APS so that these modules can be seamlessly integrated and trained in an end-to-end manner. Thus, STrHCE should benefit from these modules and the entire framework to output promising performance. Extensive experiments in Section 4 confirm this point.

4 Experiments

In this section, a series of experiments are conducted to evaluate the proposed framework STrHCE using ablation studies and comparisons with state-of-the-art methods. The source code is available at the website[1)].

4.1 Datasets

Four commonly-used fine-grained benchmark datasets with a concept hierarchy are used as experimental datasets, including CUB-200-2011 [1], FGVC-Aircraft [3], Stanford Dogs [4] and NABirds [2]. Among these datasets, CUB-200-2011 contains a three-level concept hierarchy by grouping 200 bird species into 122 genera and further to 37 families following [5]. Stanford Dogs has a two-level concept hierarchy that is constructed by grouping 120 dog breeds into 72 genera. FGVC-Aircraft and NABirds originally contained the corresponding concept hierarchy, where a three-level hierarchy groups 100 variants of aircraft into 70 families and further to 30 manufacturers, and a two-level hierarchy groups 555 bird species into 404 upper-level categories. The detailed statistics with the number of subordinate concepts (i.e., the final classes), the number of levels in the concept hierarchy, and the size of training/testing data are summarized in Table 1.

1) https://github.com/lvyilin/STrHCE.

Table 1 Statistics of benchmark datasets

Dataset	#Class	#Level	#Training	#Testing
CUB-200-2011	200	3	5994	5794
FGVC-Aircraft	100	3	6667	3333
Stanford Dogs	120	2	12000	8580
NABirds	555	2	23929	24633

Table 2 Contribution of different modules to accuracy (%) at different concept levels. "−" denotes removing the specific module. The backbone is DeiT-S-16/ViT-B-16.

Method	l_1:family	l_2:genus	l_3:species
Baseline	96.3/98.1	92.5/94.9	87.3/90.3
STrHCE$_{-\mathrm{SMA}-\mathrm{HCE}}$	96.6/98.2	93.4/95.2	88.6/91.6
STrHCE$_{-\mathrm{HCE}}$	96.7/98.2	93.6/95.6	88.8/91.7
STrHCE	96.8/98.5	94.2/95.9	89.2/91.9

4.2 Implementation details

In all our experiments, all images are resized to 448×448. Following [22, 24], we adopt AutoAugment and Random Erasing [44] for data augmentation. The patch size P of the input sequence is set to 16; thus, there are 768 patches in the coarse-scale sequence. Unless otherwise stated, the step size S in (10) is set to 8 (thus, $\bar{N}$ is 3025) and the parameter λ in (15) is set to 0.8. We will discuss the effects of different parameter schemes in Subsection 4.3. The margin ϵ in (16) and the trade-off parameter γ in (17) are empirically set to 0.05 and 1, respectively. For fair comparison with other state-of-the-art methods, we choose DeiT-S-16 [22] as the main transformer backbone, which shares the same architecture with ViT [18], but the number of parameters is approximate to ResNet50 [45] (22 M for DeiT-S-16 and 26 M for ResNet50) and is reduced by half compared to ViT-S-16 (48.8 M). We load the network weight pretrained on ImageNet-1k from the PyTorch image models library[2)], and then fine-tune with each experimental fine-grained benchmark dataset. Other backbones, such as DeiT-B-16 and ViT-B-16, are also adopted to make a fair comparison with the baselines that use extra data or complicated computation modules. For example, He et al. [24] used ViT-B-16 as the backbone of TransFG and pretrained it on a larger dataset, ImageNet-21k. The SGD optimizer with a momentum of 0.9 and a weight decay of 1E−4 is adopted to train the deep model. The initial learning rate is set to 1E−4 for Stanford Dogs and 1E−3 for the three other datasets, and it is multiplied by 0.1 every 30 epochs. All experiments are implemented with PyTorch and performed on four NVIDIA RTX 2080 Ti GPUs.

4.3 Ablation studies

A series of ablation experiments are conducted to demonstrate the effectiveness of each module in STrHCE and the corresponding hyperparameters. Unless otherwise stated, the commonly-used CUB-200-2011 dataset is chosen, and the DeiT-S-16 backbone is used due to its lower computational complexity.

4.3.1 *Contributions of different modules*

We evaluate the contributions of different modules (APS, SMA, and HCE) in the proposed STrHCE framework. In this experiment, DeiT-S-16 and ViT-B-16 are taken as the baseline. We test different variants of STrHCE (STrHCE$_{-\mathrm{HCE}}$ indicates the Siamese transformer without HCE but with APS and SMA, and STrHCE$_{-\mathrm{SMA}-\mathrm{HCE}}$ denotes the Siamese transformer without the SMA and HCE but with APS) to evaluate the contribution of each module. We present the recognition accuracies of all levels for comprehensive comparisons (three levels for the CUB-200-2011 dataset).

Table 2 shows that the Siamese transformer framework benefits fine-grained image recognition; i.e., STrHCE$_{-\mathrm{SMA}-\mathrm{HCE}}$ outperforms the baseline for each concept level. This result confirms that the APS module leverages the attention weights derived from the coarse-scale subnetwork, which can retain more subtle but discriminative information while eliminating redundant information. By adding the SMA module, STrHCE$_{-\mathrm{SMA}-\mathrm{HCE}}$ obtains further improvement; therefore, the information transferred by the SMA module can facilitate the discrimination of fine-grained cues as being constrained to produce more confident predictions by taking the output from original patches as references. The last line, i.e., the overall

2) https://github.com/rwightman/pytorch-image-models.

Table 3 Comparison of different sampling mechanisms on the CUB-200-2011 dataset. The baseline is DeiT-S-16.

Method	Acc. (%)
Baseline	87.3
Baseline ($S = 14$)	87.7
Baseline ($S = 12$)	88.4
Baseline ($S = 10$)	88.4
Selective sampling [39]	88.1
Non-uniform sampling [38]	88.6
Our APS	89.2

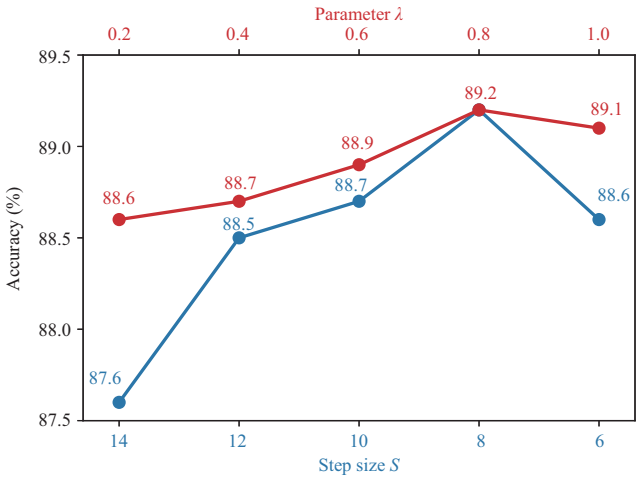


Figure 3 (Color online) The effects of step size S (blue line) and parameter λ (red line) on STrHCE in terms of accuracy (for CUB-200-2011 dataset).

proposed STrHCE framework, empirically benefits from the HCE. For the bottom two concept levels (l_2 genus and l_3 species), STrHCE achieves sufficient performance gain compared to the model without HCE. As mentioned in Subsection 3.1, more specific concepts indicate more difficult distinguishability; thus, the proposed STrHCE integrated with three modules (APS, SMA, and HCE) is much more reasonable and effective for handling a fine-grained image recognition task.

4.3.2 *Effect of sampling methods*

As shown in Table 2, the proposed STrHCE works well with APS to generate a finer-scale patch sequence (the second line vs. the first line). To further demonstrate the effectiveness of our proposed APS, we evaluate the performance of simply reducing the step size S to 14/12/10. Note that these experiments are conducted without the Siamese architecture, and it is impractical to set the step size to less than 10 due to the GPU memory limitations. For example, when $S = 8$, every single sample requires approximately 14 G of GPU memory. Table 3 shows that reducing the step size can remarkably improve performance, which indicates that preserving the local structure benefits fine-grained recognition. Moreover, the literature contains several sampling methods for extracting important information from fine-grained images, such as selective sampling [39] and non-uniform sampling [38]. To show the superiority of APS, we replace the APS module with these previous sampling methods while retaining other modules. As shown in Table 3, all sampling mechanisms obtain accuracy improvements when cooperating with our proposed Siamese transformer framework. These results indicate that determining the discriminative regions is essential for fine-grained recognition. As expected, the proposed APS outperforms the existing sampling methods. The main reason, we believe, is that APS not only retains more discriminative regions but also prevents the sequentialization process from corrupting the local structure of discriminative regions.

4.3.3 *Effect of step size S*

Step size S (in (10)) controls the number of overlapping patches in the intermediate overlapping sequence $\bar{\mathcal{X}}$, which further affects the generation of APS sequence $\tilde{\mathcal{X}}$. To demonstrate its effectiveness, several values are set for S, and the final results are shown in Figure 3 (blue line). As shown, STrHCE works

best at $S = 8$ (which is exactly half of the patch size $P = 16$), while the performance will worsen when S is set to a small or large value. This result makes sense because a small step size leads to more patches for a particular region (i.e., the most informative region), which reduces the information from other regions (i.e., the context regions). Meanwhile, a large step size means degenerating to the original ViT scheme (equal when $S = P$), which may not guarantee sufficient preservation of local structure information. Similar results can be observed on other datasets; thus, S is set to 8 in the following experiments.

4.3.4 *Effect of parameter* λ

The parameter λ (in (15)) is used to balance the influence of each concept level in HCE. To verify its effectiveness, several experiments have been conducted by varying the λ value (in the range of (0, 1]) while fixing other parameters. The results are shown in Figure 3 (red line). As λ increases (from 0.2^2 to 0.8^2 at the highest concept level, from 0.2 to 0.8 at the second concept level, and always 1 at the third concept level), the final prediction results continually improve. This result verifies that introducing more hierarchical semantic information helps fine-grained category learning. However, if all concept levels are treated equally ($\lambda = 1$), the performance will worsen, possibly because superordinate concepts have too much influence on subordinate concepts, thus, hindering their learning. We set λ to 0.8 in the following experiments.

4.4 Comparison with state-of-the-art methods

We compared the proposed STrHCE with state-of-the-art methods on four experimental datasets with a concept hierarchy. STrHCE is combined with different backbones for a fair comparison. Among these backbones, ViT-B-16 is pretrained on ImageNet-21k, and the corresponding method is marked with "†", while other backbones are pretrained on ImageNet-1k.

Specifically, for the CUB-200-2011 dataset, as shown in the third column of Table 4 [46–52], our proposed STrHCE with the ViT-B-16 backbone outperforms all state-of-the-art methods. Compared with GCL [17] and CDL [16], which learn the correlations among patches using an extra module, STrHCE obtains better performance by leveraging the self-attention mechanism in the transformer (similar to ViT and TransFG). Our method is superior to TASN [38] and S3N [39], which exploit discriminative regions using non-uniform sampling or selective sampling but cannot eliminate the redundant category-irrelevant information. MGE-CNN [36] and WS-DAN [48] are three-stage frameworks for fine-grained image recognition. MGE-CNN sequentially learns three ResNet-101 networks as different granularity-specific experts, and WS-DAN augments the data with an extra attention-cropping and attention-dropping pipeline. PMG [50] progressively learns the discriminative features in four granularities. Our method outperforms these methods because of the proposed Siamese transformer framework with the aid of DeiT-B-16 and ViT-B-16 backbones. STrHCE outperforms the existing hierarchical semantic embedding method (HSE) [5] by 1.1% with the DeiT-S-16 backbone, which further confirms the effectiveness of our framework. Compared to the recent transformer-based methods DeiT [22], ViT [18], Conde et al. [35], and TransFG [24], STrHCE consistently performs better with the same backbone on all benchmark datasets we conducted. Among these methods, TransFG [24] uses a part selection module to select the patches with maximum weight in different attention heads to represent the local information of fine-grained objects. A contrastive loss is adopted to expand the fine-grained category boundaries. Conde et al. [35] contains a three-stage framework that must extract the bounding boxes of the object and multiple parts and then inputs them into the network. We believe that our model surpasses these models mainly because it preserves more discriminative details with the local structure and explicitly exploits the concept hierarchy information, which the other models neglect.

For the FGVC-Aircraft dataset, the results are listed in the fourth column of Table 4, which shows that transformer-based methods (DeiT and ViT) did not obtain remarkable results. We believe that this result is mainly due to the important role played by the locality in an image in determining the discriminative regions of aircraft [3]. In this case, CNN-based methods benefit more from the locality inductive bias, including the best baseline, API-Net [32]. Nevertheless, the proposed STrHCE method obtains competitive performance and 1.4% (92.2% vs. 90.8%)/2.9% (92.1% vs. 89.2%) improvement compared to the baseline DeiT-S-16/ViT-B-16 due to considering the local structure in the APS module. STrHCE also achieves the best results based on the DeiT-B-16 backbone, which is on a par with PMG [50] using four-level granular information.

Table 4 Comparison with state-of-the-art methods on four commonly-used fine-grained benchmark datasets[a)]

Method	Backbone	CUB Acc. (%)	Airs Acc. (%)	Dogs Acc. (%)	NABirds Acc. (%)
RA-CNN [10]	VGGNet-19	85.3	–	87.3	–
MA-CNN [11]	VGGNet-19	86.5	89.9	–	–
NTS-Net [14]	ResNet-50	87.5	91.4	–	–
Cross-X [46]	ResNet-50	87.7	92.6	88.9	86.4
GCL [17]	ResNet-50	88.3	93.2	–	–
CDL [16]	ResNet-50	88.4	–	–	–
FDL [47]	DenseNet-161	89.1	91.3	84.9	–
TASN [38]	ResNet-50	87.9	–	–	–
S3N [39]	ResNet-50	88.5	92.8	–	–
MGE-CNN [36]	ResNet-101	89.4	–	–	88.6
WS-DAN [48]	InceptionV3	89.4	93.0	–	–
Ge et al. [49]	GoogleNet	90.4	–	–	–
PMG [50]	ResNet-50	89.6	**93.4**	–	–
CIN [33]	ResNet-101	88.1	92.8	–	–
iSQRT-COV [51]	ResNet-101	88.7	91.4	–	–
API-Net [32]	DenseNet-161	90.0	–	90.3	88.1
MG-CNN [25]	VGGNet-19	83.0	–	–	–
PA-CNN [52]	VGGNet-19	87.8	91.0	–	–
HSE [5]	ResNet-50	88.1	–	–	–
DeiT [22]	DeiT-S-16	87.3	90.8	91.8	85.5
ViT† [18]	ViT-B-16	90.3	89.2	91.7	89.9
Conde et al. [35]	ViT-B-16	91.0	–	93.2	–
TransFG† [24]	ViT-B-16	<u>91.7</u>	–	92.3	<u>90.8</u>
STrHCE	DeiT-S-16	89.2	92.2	92.1	88.7
STrHCE	DeiT-B-16	90.0	**93.4**	**94.0**	90.0
STrHCE†	ViT-B-16	**91.9**	92.1	<u>93.8</u>	**91.5**

a) Best results are marked in bold while the second best results are underlined.

The Stanford Dogs dataset is more challenging because of the larger intra-class variance [4]. As shown in the fifth column of Table 4, the transformer-based methods obviously achieve better results than the CNN-based methods. Specifically, STrHCE outperforms TransFG by 1.5% on the ViT-B-16 backbone. In addition, STrHCE outperforms the best baseline [35] by 0.6% on the ViT-B-16 backbone and achieves a new state-of-the-art by 94.0% based on the DeiT-B-16 backbone.

The NABirds dataset is larger in scale and contains more categories compared to the other three datasets. The results on the NABirds dataset in the last column of Table 4 further demonstrate similar statistics. In particular, STrHCE outperforms TransFG by 0.7% and achieves a new state-of-the-art by 91.5%.

To obtain more insight into the performance of our proposed model, we visualize the confusion matrices of different baselines and STrHCE among the most confusing classes at different conceptual levels. Here, we define the most confusing classes as a subset of the classes of the maximal sum of misclassifications. To eliminate the influence of the choice of classes on the results, we use a proxy model, e.g., ResNet50, to construct the sets of the most confusing classes. Specifically, we construct an undirected graph where each node represents a class at a certain conceptual level, and each edge represents the number of misclassifications between two classes. Then, the most confusing classes can be obtained by solving a maximum-weight connected subgraph problem. We experiment on the CUB-200-2011 dataset. As revealed in Figure 4, STrHCE shows inspiring performance in the most confusing classes; e.g., among the most difficult-to-distinguish species, the number of correctly classified samples is 116, compared to 100/102/113 for API-Net/ViT-B-16/TransFG, respectively. This comparison further verifies the effectiveness of our Tr-HCE for modeling a concept hierarchy. Meanwhile, cooperating with APS and Siamese architecture can substantially improve a model's ability to identify confusing classes, which is essential for fine-grained recognition.

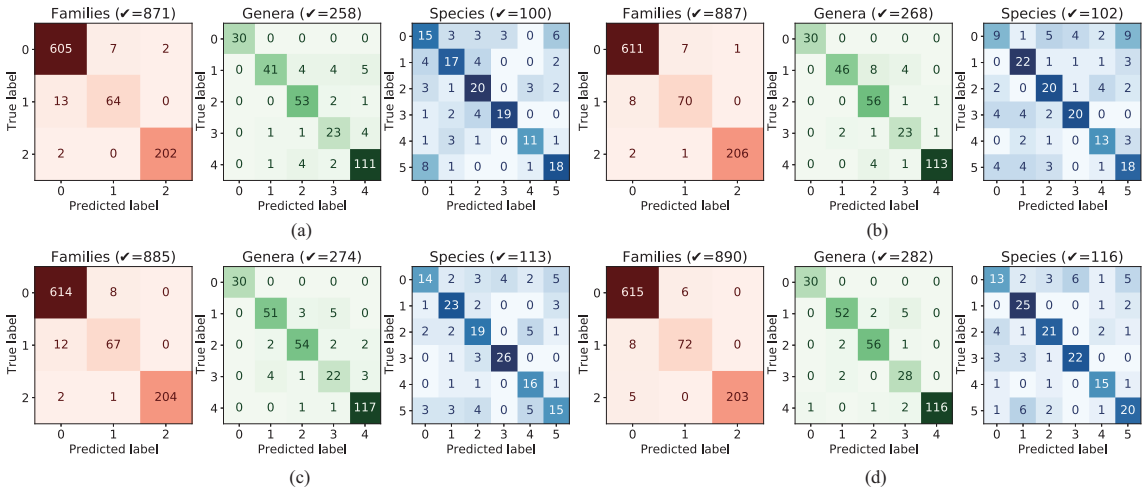


Figure 4 (Color online) Confusion matrices of different baselines and the proposed method among the most confusing classes at different conceptual levels. The most confusing classes are defined as a subset of the classes of the maximal sum of misclassifications, which is obtained by a proxy model, e.g., ResNet50. The experiment is conducted on the CUB-200-2011 dataset. "✓" represents the number of samples correctly classified, i.e., the sum of diagonal elements. (a) API-Net; (b) ViT-B-16; (c) TransFG; (d) STrHCE.

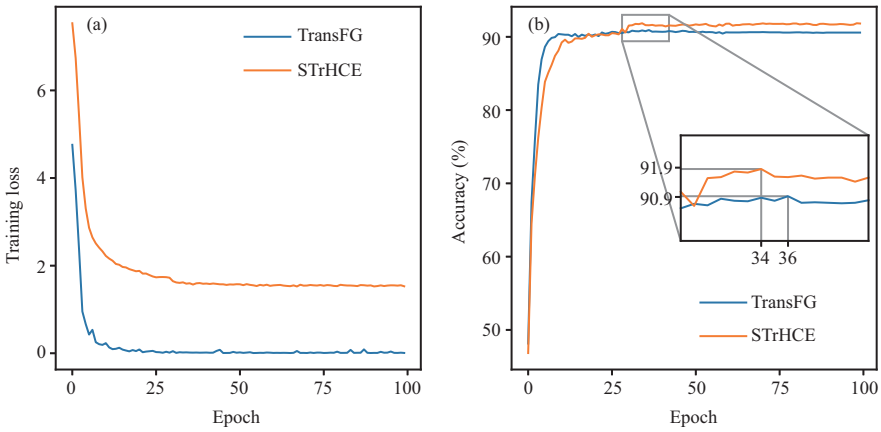


Figure 5 (Color online) Comparison of the training loss (a) and validation accuracy (b) of TransFG and the proposed STrHCE on the CUB-200-2011 dataset. The epoch that achieves the best accuracy for each method is highlighted in (b).

4.5 Time complexity analysis

The main computational component of STrHCE is the transformer block with complexity $O((L+N)^2D+(L+N)D^2)$, where L is the number of concept levels (less than 4), N is the number of patches, and D is the latent feature size. To demonstrate the efficiency of our algorithm, we use TransFG for comparison with STrHCE, because TransFG does not require multiple rounds of input and is obviously more efficient than [35]. In terms of framework, STrHCE contains a Siamese transformer architecture, while TransFG comprises one transformer. STrHCE requires twice as much running time as TransFG on the surface, but not in practice for two reasons: (1) in STrHCE, each transformer takes N nonoverlapping patches as input, which is more efficient than TransFG with more overlapping patches (e.g., 784 vs. 1369 when $S = 12$); and (2) the Siamese architecture benefits the convergence of transformer model training. Figure 5(a) shows TransFG and STrHCE have similar convergence curves. However, as shown in Figure 5(b), TransFG takes 36 epochs to obtain the best accuracy of 90.9%, while STrHCE only needs 34 epochs and outperforms TransFG by 1.0%. The running times of TransFG and STrHCE are 2.79 and 2.49 h, respectively. As expected, STrHCE benefits from the entire framework, thus, obtaining promising efficiency as well as performance.

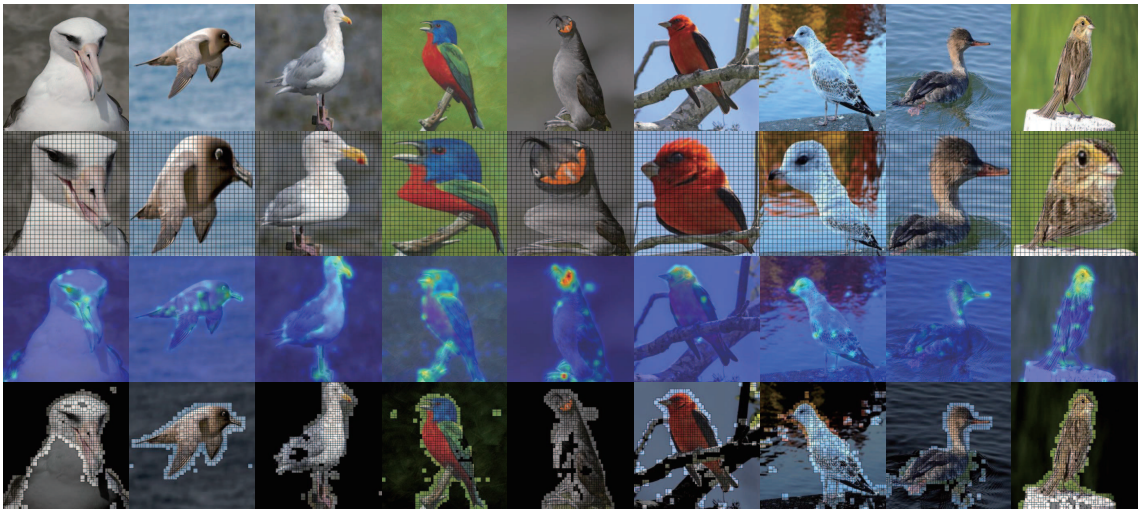


Figure 6 (Color online) Visual comparison of the input patch sequences produced by APS and Non-uniform sampling. The first row shows the original image. The second row shows the patch sequence of Non-uniform sampling. Each square represents a sampled image patch that is constrained to align in a grid layout. By leveraging the learned attention weights shown in the third row, the APS patch sequence in the fourth row can eliminate redundant category-irrelevant information without distorting the original image (unsampled regions are shaded). Best viewed with zooming in.

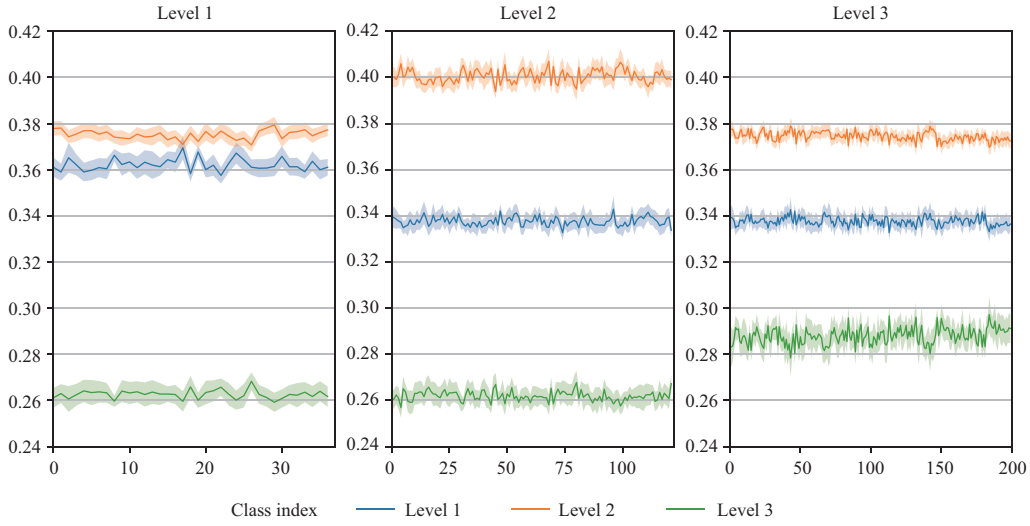


Figure 7 (Color online) Correlations among different concept levels. Each subfigure indicates one concept level, similar to each line. Each line records the mean of attention values between two corresponding levels in the corresponding classes, and the shaded area represents their standard deviation.

4.6 Qualitative analysis

To investigate how the learned attention values affect the generation of APS patches and to demonstrate the effectiveness of APS, we randomly select nine images from CUB-200-2011 dataset, and then visualize the attention map and the corresponding APS sequence. We also make a comparison with non-uniform sampling sequence [38] with the same settings as described in Subsection 4.3.2. As shown in Figure 6, the patch sequences of non-uniform sampling are constrained to align in a grid layout, and the resulting image may be severely distorted. By leveraging the learned attention weights in the third row, we can focus on the discriminative regions of the fine-grained object. We then plot the APS sequence by remapping the selected patches to the original image. As shown in the last row, each highlighted square represents an APS patch and the shaded area indicates that the corresponding regions are not sampled. Leveraging the attention information, the generated APS sequence can contain more patches for discriminative regions and remove irrelevant regions, while retaining the shape.

In STrHCE, a concept hierarchy is exploited as prior knowledge to integrate the correlations among

concepts at different levels. To verify how these correlations are embedded in STrHCE, we visualized the attention values (mean and standard deviation) in $A(1:L, 1:L)$ (Eq. (8)) for the training images from each class (there are 200 classes in total). The mean of the correlations between levels $l1$ and $l2$ along the k-th class is computed using $\frac{1}{n^k}\sum_i A^{(i,k)}(l1, l2)$ $(l1, l2 \in \{1, 2, 3\})$, and the corresponding standard deviation is also demonstrated, where n^k is the number of images belonging to the k-th class. As shown in Figure 7, the correlations among concept levels are relatively stable across different classes, which indicates that the proposed STrHCE method can model the inherent semantic prior knowledge.

5 Conclusion

In this paper, we propose an STrHCE for fine-grained image recognition. STrHCE leverages the global perception ability of transformers in a Siamese-like architecture to capture the correlations among the discriminative regions and weaken the effect of category-irrelevant regions. It simultaneously considers the local structure of discriminative regions and exploits the prior knowledge of concept hierarchy. Extensive experiments demonstrate the superiority of STrHCE over state-of-the-art baselines. In this paper, STrHCE focuses on handling fine-grained datasets with balanced classes. Exploring the performance of STrHCE on long-tailed fine-grained datasets will be interesting. Additionally, exploring how to simplify the two networks while achieving equally good results has promise.

Acknowledgements This work was partly supported by National Key Research and Development Program of China (Grant No. 2020AAA0106800), Beijing Natural Science Foundation (Grant Nos. Z180006, L211016), National Natural Science Foundation of China (Grant No. 62176020), CAAI-Huawei MindSpore Open Fund, and Chinese Academy of Sciences (Grant No. OEIP-O-202004).

References

1 Welinder P, Branson S, Mita T, et al. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010
2 Horn G V, Branson S, Farrell R, et al. Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 595–604
3 Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft. 2013. ArXiv:1306.5151
4 Khosla A, Jayadevaprakash N, Yao B, et al. Novel dataset for fine-grained image categorization. In: Proceedings of the 1st Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition, 2011
5 Chen T, Wu W, Gao Y, et al. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In: Proceedings of the 26th ACM International Conference on Multimedia, Seoul, 2018. 2023–2031
6 Zhang N, Donahue J, Girshick R B, et al. Part-based R-CNNs for fine-grained category detection. In: Proceedings of the 13th European Conference on Computer Vision, Zurich, 2014. 8689: 834–849
7 Lin D, Shen X, Lu C, et al. Deep LAC: deep localization, alignment and classification for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 1666–1674
8 Krause J, Jin H, Yang J, et al. Fine-grained recognition without part annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 5546–5555
9 Zhang H, Xu T, Elhoseiny M, et al. SPDA-CNN: unifying semantic part detection and abstraction for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 1143–1152
10 Fu J, Zheng H, Mei T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 4476–4484
11 Zheng H, Fu J, Mei T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, Venice, 2017. 5219–5227
12 Li Z, Yang Y, Liu X, et al. Dynamic computational time for visual attention. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, 2017. 1199–1209
13 He X, Peng Y. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, 2017. 4075–4081
14 Yang Z, Luo T, Wang D, et al. Learning to navigate for fine-grained classification. In: Proceedings of the 15th European Conference on Computer Vision, Munich, 2018. 11218: 438–454
15 He X, Peng Y, Zhao J. Which and how many regions to gaze: focus discriminative regions for fine-grained visual categorization. Int J Comput Vis, 2019, 127: 1235–1255
16 Wang Z, Wang S, Zhang P, et al. Weakly supervised fine-grained image classification via correlation-guided discriminative learning. In: Proceedings of the 27th ACM International Conference on Multimedia, Nice, 2019. 1851–1860
17 Wang Z, Wang S, Li H, et al. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, 2020. 12289–12296
18 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. In: Proceedings of the 9th International Conference on Learning Representations, Vienna, 2021
19 Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. In: Proceedings of the 16th European Conference on Computer Vision, Glasgow, 2020. 12346: 213–229
20 Zhu X, Su W, Lu L, et al. Deformable DETR: deformable transformers for end-to-end object detection. In: Proceedings of the 9th International Conference on Learning Representations, Vienna, 2021
21 Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2021. 6881–6890

22 Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning, 2021. 10347–10357
23 He S, Luo H, Wang P, et al. TransReID: transformer-based object re-identification. In: Proceedings of IEEE/CVF International Conference on Computer Vision, Montreal, 2021. 14993–15002
24 He J, Chen J, Liu S, et al. TransFG: a transformer architecture for fine-grained recognition. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence, the 34th Conference on Innovative Applications of Artificial Intelligence, and the 12th Symposium on Educational Advances in Artificial Intelligence, 2022. 852–860
25 Wang D, Shen Z, Shao J, et al. Multiple granularity descriptors for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 2399–2406
26 He G, Li F, Wang Q, et al. A hierarchical sampling based triplet network for fine-grained image classification. Pattern Recognit, 2021, 115: 107889
27 Lin T, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1449–1457
28 Gao Y, Beijbom O, Zhang N, et al. Compact bilinear pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 317–326
29 Kong S, Fowlkes C C. Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 7025–7034
30 Wei X, Zhang Y, Gong Y, et al. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification. In: Proceedings of the 15th European Conference on Computer Vision, Munich, 2018. 11207: 365–380
31 Li P, Xie J, Wang Q, et al. Is second-order information helpful for large-scale visual recognition? In: Proceedings of the IEEE International Conference on Computer Vision, Venice, 2017. 2089–2097
32 Zhuang P, Wang Y, Qiao Y. Learning attentive pairwise interaction for fine-grained classification. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, 2020. 13130–13137
33 Gao Y, Han X, Wang X, et al. Channel interaction networks for fine-grained image categorization. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, 2020. 10818–10825
34 Zhang S, Huang Q, Hua G, et al. Building contextual visual vocabulary for large-scale image applications. In: Proceedings of the 18th ACM International Conference on Multimedia, Firenze, 2010. 501–510
35 Conde M V, Turgutlu K. Exploring vision transformers for fine-grained classification. 2021. ArXiv:2106.10587
36 Zhang L, Huang S, Liu W, et al. Learning a mixture of granularity-specific experts for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul, 2019. 8330–8339
37 He X, Peng Y. Only learn one sample: fine-grained visual categorization with one sample training. In: Proceedings of ACM International Conference on Multimedia, Seoul, 2018. 1372–1380
38 Zheng H, Fu J, Zha Z, et al. Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 5012–5021
39 Ding Y, Zhou Y, Zhu Y, et al. Selective sparse sampling for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul, 2019. 6598–6607
40 Abnar S, Zuidema W H. Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 4190–4197
41 Chicco D. Siamese Neural Networks: An Overview. New York: Springer, 2021. 73–94
42 O'Connor C M, Cree G S, McRae K. Conceptual hierarchies in a flat attractor network: dynamics of learning and computations. Cogn Sci, 2009, 33: 665–708
43 Efraimidis P S, Spirakis P G. Weighted random sampling with a reservoir. Inf Process Lett, 2006, 97: 181–185
44 Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, 2020. 13001–13008
45 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 770–778
46 Luo W, Yang X, Mo X, et al. Cross-X learning for fine-grained visual categorization. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul, 2019. 8241–8250
47 Liu C, Xie H, Zha Z, et al. Filtration and distillation: enhancing region attention for fine-grained visual categorization. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, 2020. 11555–11562
48 Hu T, Qi H. See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification. 2019. ArXiv:1901.09891
49 Ge W, Lin X, Yu Y. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 3034–3043
50 Du R, Chang D, Bhunia A K, et al. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In: Proceedings of the 16th European Conference on Computer Vision, Glasgow, 2020. 12365: 153–168
51 Li P, Xie J, Wang Q, et al. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 947–955
52 Zheng H, Fu J, Zha Z J, et al. Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. IEEE Trans Image Process, 2020, 29: 476–488