Analogous to Evolutionary Algorithm: Designing a Unified Sequence Model

Jiangning Zhang[1*] **Chao Xu**[1] **Jian Li**[2] **Wenzhou Chen**[1] **Yabiao Wang**[2]
Ying Tai[2] **Shuo Chen**[3] **Chengjie Wang**[2] **Feiyue Huang**[2] **Yong Liu**[1†]
[1]APRIL Lab, Zhejiang University [2]Youtu Lab, Tencent
[3]RIKEN Center for Advanced Intelligence Project
{186368, 21832066, wenzhouchen}@zju.edu.cn, yongliu@iipc.zju.edu.cn
{swordli, caseywang, yingtai, jasoncjwang, feiyuehuang}@tencent.com
shuo.chen.ya@riken.jp

Abstract

Inspired by biological evolution, we explain the rationality of Vision Transformer by analogy with the proven practical Evolutionary Algorithm (EA) and derive that both of them have consistent mathematical representation. Analogous to the dynamic local population in EA, we improve the existing transformer structure and propose a more efficient EAT model, and design task-related heads to deal with different tasks more flexibly. Moreover, we introduce the spatial-filling curve into the current vision transformer to sequence image data into a uniform sequential format. Thus we can design a unified EAT framework to address multi-modal tasks, separating the network architecture from the data format adaptation. Our approach achieves state-of-the-art results on the ImageNet classification task compared with recent vision transformer works while having smaller parameters and greater throughput. We further conduct multi-modal tasks to demonstrate the superiority of the unified EAT, *e.g.*, Text-Based Image Retrieval, and our approach improves the rank-1 by +3.7 points over the baseline on the CSS dataset.[3]

1 Introduction

Since Vaswani *et al*. [69] introduce the Transformer that achieves outstanding success in the machine translation task, many improvements have been made to this method [15, 38, 21]. Recent works [64, 82, 43] led by ViT [23] have achieved great success in the field of many vision tasks by replacing CNN with transformer structure. In general, these works are experimentally conducted to verify the effectiveness of modules or improvements, but they may lack other forms of supporting evidence.

Inspired by biological population evolution, we explain the rationality of Vision Transformer by analogy with the proven effective, stable, and robust Evolutionary Algorithm (EA), which has been widely used in practical applications. Through analogical analysis, we observe that the training procedure of the transformer has similar attributes to the naive EA, as shown in Figure 1. Take the one-tier transformer (*abbr.*, TR) as an example. ***1)*** TR processes a sequence of patch embeddings while EA evolutes individuals, both of which have the same vector formats and necessary initialization. ***2)*** The Multi-head Self-Attention (MSA) among patch embeddings in TR is compared with that of (sparse) global individual crossover among all individuals in EA, in which local and dynamic population concepts are introduced to increase running speed and optimize results [63, 59]. ***3)*** Feed-Forward

[*]Work done during an intership at Tencent Youtu Lab.
[†]Corresponding author.
[3]Code is available at `https://github.com/TencentYoutuResearch/BaseArchitecture-EAT`.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

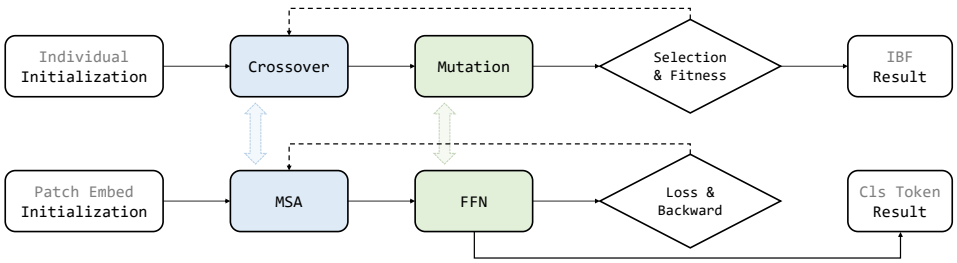

Figure 1: Analogy of EA (top) and Transformer (bottom) pipelines. For simplicity, only one layer of the Transformer structure is displayed here.

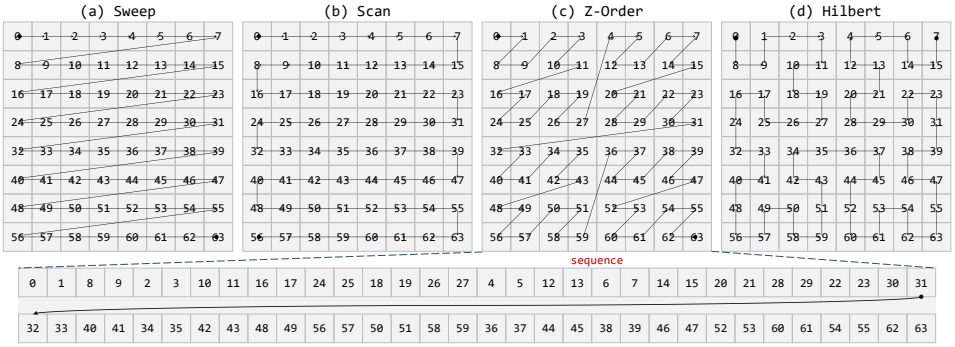

Figure 2: Different SFC indexing methods, taking 2D images with side length of 8 as an example.

Network (FFN) in TR enhances embedding features that is similar to the individual mutation in EA. ***4)*** During training, TR optimizes the network through backpropagation while EA optimizes individuals through selection and fitness calculation. ***5)*** TR chooses the enhanced Classification Token (Cls Token) as the target output, while EA chooses Individual with the Best Fitness (IBF). Meanwhile, we deduce the mathematical characterization of crossover and mutation operators in EA (*c.f.*, Equations 5,8) and find that they have the same mathematical representation as MSA and FFN in TR (*c.f.*, Equations 6,9), respectively. Inspired by the characteristics in the crossover step of EA, we propose a novel *EA-based Transformer* (EAT) that intuitively designs a local operator in parallel with global MSA operation, in which the local operator can be instantiated as 1D convolution, local MSA, *etc*. Subsequent experiments demonstrate the effectiveness of this design in that it could reduce parameters and improve the running speed and boost the network performance, which is consistent with the results of EAs in turn. Current TR-based models would initialize different tokens for different tasks, and they participate in every level of calculation that is somewhat incompatible with other tokens for internal operations. Therefore, we design task-related heads docked with transformer backbone to complete final information fusion, which is more flexible for different tasks learning and suitable for the transfer learning of downstream tasks.

Our designed local operator receives the same data format as the global MSA branch, *i.e.*, a sequence that is conform to NLP. Therefore, the generally used high dimension operations such as 2D reshape [79, 77, 80] are not required, which brings the possibility to standardize multi-modal data (*e.g.*, 1D sentence, 2D image, and 3D video) of input into consistent sequence data in *one* unified model. To accomplish this target, we introduce the space-filling curve (SFC) concept to standardize the multi-modal data format and design an SFC module. As shown in Figure 2, taking 2D image as an example, the top half represents four kinds of SFCs: Sweep, Scan, Z-Order, and Hilbert, while the bottom two lines represent the sequenced image by Z-Order SFC. Thus the image ($\in \mathbb{R}^{8\times8\times3}$) is specifically re-indexed and arranged ($\in \mathbb{R}^{64\times3}$) by the predefined SFC before feeding the network, realizing uniform sequence input. The difference between SFCs mainly reflects how the 1D sequence preserves the 2D spatial structure, and the SFC can be extended to 3D and higher dimensions.

Specifically, we make the following four contributions:

- In theory, we explain the rationality of Vision Transformer (TR) by analogy with Evolutionary Algorithm (EA) and derive that they have consistent mathematical representation.
- For the method, we improve a unified EAT model by analogy with dynamic local population concept in EA and design a Task-related Head to deal with various tasks more flexibly.

- On framework, we introduce Space-Filling Curve (SFC) module as a bridge between rasterized 2D image data and serialized 1D sequence data. This makes it possible to integrate the unified paradigm that uses a unified model to solve multi-modal tasks, keeping the network architecture and data structure independent of the data dimensionality. Note that only several 1D operators are required for our method, meaning that it is very friendly to the underlying optimization workload of different platforms.
- Massive experiments on classification and multi-modal tasks demonstrate the superiority and flexibility of our approach.

2 Related Work

2.1 Evolution Algorithms

Evolution algorithm is a subset of evolutionary computation in computational intelligence, and it belongs to modern heuristics [60, 70]. Inspired by biological evolution, general EA contains reproduction, crossover, mutation, and selection steps, and has been proven to be effective and stable in many application scenarios [29]. Moscato *et al*. [47] propose the Memetic Algorithm (MA) for the first time in 1989, which applies a local search process to refine solutions. Later Differential Evolution (DE) appears in 1995 [62] is arguably one of the most competitive improved variant [18, 49, 52, 24]. The core of DE is a differential mutation operator, which differentiates and scales two individuals in the same population and interacts with the third individual to generate a new individual. In contrast to the aforementioned global optimization, Local Search Procedures (LSPs) aim to find a solution that is as good as or better than all other solutions in its "neighborhood" [44, 14, 9, 41, 28]. LSP is more efficient than global search in that a solution can be verified as a local optimum quickly without associating the global search space. Also, some works [44, 14, 41] apply the Multi-Population Evolutionary Algorithm (MPEA) to solve the constrained function optimization problems relatively efficiently. Recently, Li *et al*. [41] point out that the ecological and evolutionary history of each population is unique, and the capacity of Xylella fastidiosa varies among subspecies and potentially among populations. Therefore, we argue that the local population as a complement to the global population can speed up the EA convergence and obtain better results. In this paper, we explain and improve the naive transformer structure by analogy with the validated evolutionary algorithms, where a parallel local path is designed inspired by the local population concept in EA.

2.2 Vision Transformers

Since Transformer structure is proposed for the machine translation task [69], many improved language models [51, 21, 53, 54, 7] follow it and achieve great results. Some later works [32, 74, 15, 38, 3] improve the basic transformer structure for better efficiency. Inspired by the high performance of transformer in NLP and benefitted from abundant computation, recent ViT [23] introduces the transformer to vision classification for the first time and sparks a new wave of excitement for many vision tasks, *e.g*., detection [8, 85], segmentation [81, 11, 66], generative adversarial network [36, 72], low-level tasks [10], video processing [4], general backbone [76, 75, 43], self-supervision [12, 2, 13, 48], neural architecture search [73, 40], *etc*. Many researchers have focused on improving the basic transformer structure [64, 82, 77, 26, 65, 79, 19, 80, 16, 68, 61, 17, 42], which is more challenging than other application-oriented works. Among these methods, DeiT [64] is undoubtedly a star job that makes the transformer performance better and training more data-efficient. Based on DeiT, this work focuses on improving the basic structure of the vision transformer, making it perform better considering both accuracy and efficiency. Besides, our approach supports multi-modal tasks using only one unified model while other transformer-based methods such as DeiT do not.

2.3 Space Filling Curve

A space-filling curve maps the multi-dimensional space into the 1D space, and we mainly deal with 2D SFCs in this paper. SFC acts as a thread that passes through all the points in the space while visiting each point only once, *i.e*., every point except the starting and ending points is connected to two adjacent line segments. There are numerous kinds of SFCs, and the difference among them is in their ways of mapping to the 1D space. Peano [50] first introduces a mapping from the unit interval to the unit square in 1890. Hilbert [33] generalizes the idea to a mapping of the whole space. G.M.Morton [46] proposes Z-order for file sequencing of a static two-dimensional geographical database. Subsequently, many SFCs are proposed [45, 67, 56], *e.g*., Sweep, Scan, Sierpiński, Lebesgue, Schoenberg, *etc*. Some researchers further apply SFC methods to practical applications [5, 83, 6], such as data mining and bandwidth reduction, but so far, almost none has been

applied to computer vision. By introducing SFC in our EAT, this paper aims to provide a systematic and scalable paradigm that uses a unified model to deal with multi-modal data, keeping the network architecture and data structure independent of the data dimensionality.

3 Preliminary Transformer

The Transformer structure in vision tasks usually refers to the encoder and mainly builds upon the MSA layer, along with FFN, Layer Normalization (LN), and Residual Connection (RC) operations.

MSA is equivalent to the fusion of several SA operations that jointly attend to information from different representation subspaces, formulated as:

$$\begin{aligned} \mathrm{MultiHead}(Q,K,V) &= \text{ Concat }\left(\text{ Head }_1,\text{ Head }_2,\ldots,\text{ Head }_{\mathrm{h}}\right) W^O, \\ \text{here Head }_{\mathrm{i}} &= \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \\ &= \mathrm{softmax}\left[\frac{QW_i^Q\left(KW_i^K\right)^T}{\sqrt{d_k}}\right] VW_i^V = AVW_i^V, \end{aligned} \tag{1}$$

where $W_i^Q \in \mathbb{R}^{d_m \times d_k}, W_i^K \in \mathbb{R}^{d_m \times d_k}, W_i^V \in \mathbb{R}^{d_m \times d_v}, W^O \in \mathbb{R}^{hd_v \times d_m}$ are parameter matrices; d_m is the input dimension, while d_k and d_v are hidden dimensions of each projection subspace; h is the head number; $A \in \mathbb{R}^{l \times l}$ is the attention matrix and l is the sequence length.

FFN consists of two cascaded linear transformations with a ReLU activation in between:

$$\mathrm{FFN}(x) = \max\left(0, xW_1 + b_1\right) W_2 + b_2, \tag{2}$$

where W_1 and W_2 are weights of two linear layers, while b_1 and b_2 are corresponding biases.

LN is applied before each layer of MSA and FFN, and the transformed $\hat{x}$ is calculated by:

$$\hat{x} = x + [\text{MSA} \mid \text{FFN}](\text{LN}(x)). \tag{3}$$

4 EA-based Transformer

In this section, we firstly expand the association between operators in EA and modules in Transformer, and derive a unified mathematical representation for each set of corresponding pairs, expecting an evolutionary explanation for *why Transformer architecture works*. Then we introduce the SFC module to sequence 2D image that conforms to standard NLP format. Thus we may only focus on designing one unified model to solve multi-modal data. Finally, an EA-based Transformer only containing 1D operators is proposed for vision tasks, and we argue that this property is more suitable for hardware optimization in different scenarios.

4.1 EA Interpretation of Transformer Structure

Analogous to EA, the transformer structure has conceptually similar sub-modules, as shown in Figure 3. For both of methods, we define the individual (patch embedding) as $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,d}]$, where d indicates sequence depth. Denoting l as the sequence length, the population (patch embeddings) can be defined as $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_l]^{\mathrm{T}}$.

Crossover Operator *vs.* **SA Module**. In EA, the crossover operator aims at creating new individuals by combining parts of other individuals. In detail, for an individual $\boldsymbol{x}_1 = [x_{1,1}, x_{1,2}, \ldots, x_{1,d}]$, the operator will randomly pick another individual $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,d}] (1 \leq i \leq l)$ in the global population, and then randomly replaces features of $\boldsymbol{x}_1$ with $\boldsymbol{x}_i$ to form the new individual $\hat{\boldsymbol{x}}_1$:

$$\hat{\boldsymbol{x}}_{1,j} = \begin{cases} \boldsymbol{x}_{i,j} & if\ randb(j) \leqslant CR \\ \boldsymbol{x}_{1,j} & otherwise \end{cases}, \quad s.t.\ j = 1, 2, \ldots, d, \tag{4}$$

where $randb(j)$ is the j-th evaluation of a uniform random number generator with outcome $\in [0, 1]$, CR is the crossover constant $\in [0, 1]$ that is determined by the user. We re-formulate this process as:

$$\begin{aligned} \hat{\boldsymbol{x}}_1 &= \boldsymbol{w}_1 \cdot \boldsymbol{x}_1 + \boldsymbol{w}_i \cdot \boldsymbol{x}_i \\ &= \boldsymbol{w}_1 \cdot \boldsymbol{x}_1 + 0 \cdot \boldsymbol{x}_2 + \cdots + \boldsymbol{w}_i \cdot \boldsymbol{x}_i + \cdots + 0 \cdot \boldsymbol{x}_l \\ &= W_1^{cr} X_1 + 0 X_2 + \cdots + W_i^{cr} X_i + \cdots + 0 X_l, \end{aligned} \tag{5}$$

where $\boldsymbol{w}_1$ and $\boldsymbol{w}_i$ are vectors filled with zeros or ones, representing the selection of different features of $\boldsymbol{x}_1$ and $\boldsymbol{x}_i$. W_1^{cr} and W_i^{cr} are corresponding diagonal matrix representations.

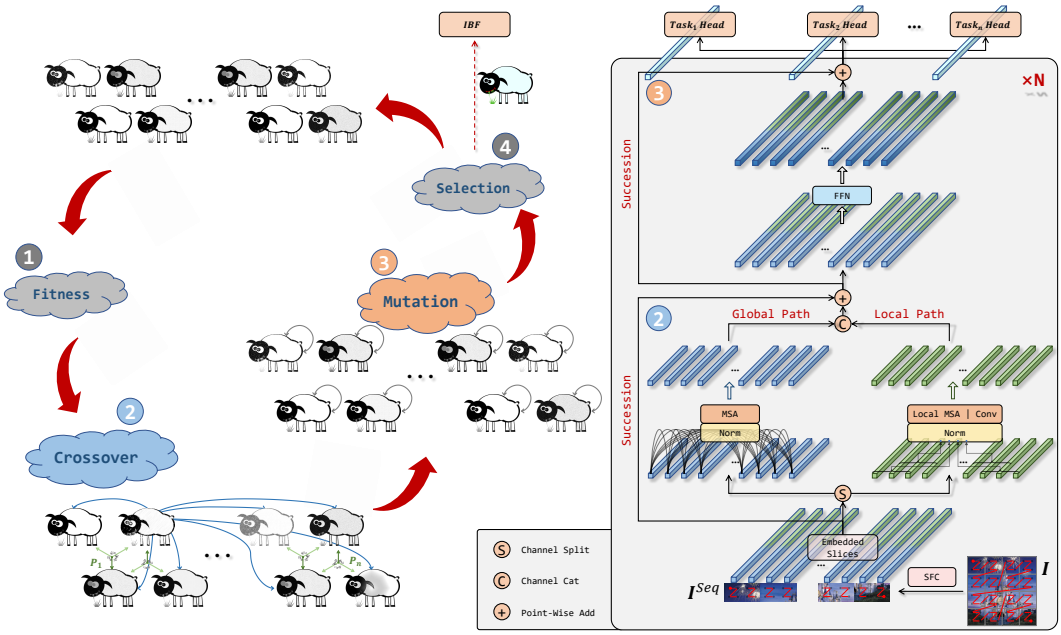

Figure 3: Structure of the proposed EAT. The right part illustrates the detailed procedure inspired by the left evolutionary algorithm. The blue and green lines in EA represent information interactions among individuals in global and local populations, which feeds back the design for global and local paths in our EAT. $P_1, \cdots, P_n$ represent biological evolution inside each local population. The proposed SFC module maps the 2D image into a 1D sequence so that only 1D operators are required.

For the SA module in transformer (MSA degenerates to SA when there is only one head), each patch embedding interacts with all embeddings. Without loss of generality, $\boldsymbol{x}_1 = [x_{1,1}, x_{1,2}, \ldots, x_{1,d}]$ interacts with the whole population, *i.e.*, X, as follows:

$$\begin{aligned} \hat{\boldsymbol{x}}_1 &= A_1 V_1 + A_2 V_2 + \cdots + A_l V_l \\ &= A_1 W^V X_1 + A_2 W^V X_2 + \cdots + A_l W^V X_l, \end{aligned} \tag{6}$$

where $A_i (i = 1, 2, \cdots, l)$ are attention weights of all patch embedding tokens that are calculated from queries and keys, W^V is the parameter metric for the value projection. By comparing Equations 5 with 6, we find that they share the same formula representation, and the crossover operation is a sparse global interaction while SA has more complex computing and modeling capabilities.

Mutation Operator *vs.* **FFN Module**. In EA, the mutation operator injects randomness into the population by stochastically changing specific features of individuals. For an individual $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,d}] (1 \leq i \leq l)$ in the global population, it goes through the mutation operation to form the new individual $\hat{\boldsymbol{x}}_i$, formulated as follows:

$$\hat{x}_{i,j} = \begin{cases} rand(v_j^L, v_j^H) x_{i,j} & if\ randb(j) \leqslant MU \\ x_{i,j} & otherwise \end{cases}, \quad s.t.\ j = 1, 2, \ldots, d, \tag{7}$$

where the MU is the mutation constant $\in [0, 1]$ that the user determines, v_j^L and v_j^H are lower and upper scale bounds of the j-th feature. Similarly, we re-formulate this process as:

$$\begin{aligned} \hat{\boldsymbol{x}}_i &= \boldsymbol{w}_i \cdot \boldsymbol{x}_i = [w_{i,1} x_{i,1}, w_{i,2} x_{i,2}, \cdots, w_{i,l} x_{i,l}] \\ &= W_i^{mu} X_i, \end{aligned} \tag{8}$$

where $\boldsymbol{w}_i$ is a randomly generated vector that represents weights on different characteristic depths, while W_i^{mu} is the corresponding diagonal matrix representation.

For the FFN module in transformer, each patch embedding carries on directional feature transformation through cascaded linear layers (*c.f.*, Equation 2). Take one-layer linear as an example:

$$\hat{\boldsymbol{x}}_i = W_1^{FFN} X_i, \tag{9}$$

where W_1^{FFN} is the weight of the first linear layer of the FFN module, and it is applied to each position separately and identically. By comparing Equations 8 and 9, we also find that they have the

same formula representation. FFN module is more expressive because it contains cascaded linear layers and non-linear ReLU activation layers in between, as depicted in Equation 2.

Population Succession *vs.* **Residual Connection**. In the evolution of the biological population, individuals at the previous stage have a certain probability of inheriting to the next stage. This phenomenon is expressed in the transformer in the form of residual connection, *i.e.*, patch embeddings of the previous layer are directly mapped to the next layer.

Best Individual *vs.* **Task-Related Token**. EAT chooses the enhanced task-related token (*e.g.*, classification token) associated with all patch embeddings as the target output, while EA chooses the individual with the best fitness score among the population as the object.

Necessity of Modules in Transformer. As described in the work [30], the absence of the Crossover operator or Mutation operator will significantly damage the model's performance. Similarly, Dong *et al*. [22] explore the effect of MLP in Transformer and find that MLP stops the output from degeneration. Furthermore, removing MSA in Transformer would also greatly damage the effectiveness of the model. Thus we can conclude that global information interaction and individual evolution are necessary for Transformer, just like the global Crossover and individual Mutation in EA.

4.2 Detailed Architecture of EAT

Local Path. Inspired by works [44, 14, 41, 63, 59] that introduce local and dynamic population concept to the evolutionary algorithm, we analogically introduce a local operator into the naive transformer structure in parallel with global MSA operation. As shown in Figure 3, the input features are divided into global features (marked blue) and local features (marked green) at the channel level with ratio p, which are then fed into global and local paths to conduct feature interaction, respectively. The outputs of the two paths recover the original data dimension by concatenation operation. Thus the improved module is very flexible and can be viewed as a plug-and-play module for the current transformer structure. This structure also implicitly introduces the design of multi-modal fusion, analogous to grouping convolution with the group equaling two. Specifically, the local operator can be 1D convolution, local MSA, 1D DCN [84], *etc*. In this paper, we adopt 1D convolution as the local operator that is more efficient and parameter friendly, and the improved transformer structure owns a constant number of sequentially executed operations and $O(1)$ maximum path length between any two positions, *c.f.*, Appendix A. Therefore, the proposed structure maintains the same parallelism and efficiency as the original vision transformer structure. And p is set to 0.5 for containing the minimal network parameters and FLOPs, proved by Proposition 1 and Proposition 2 in Appendix B.

Proposition 1. *The numbers of global and local branches in the $i-th$ mixed attention layer (MA^i) is d_1 and d_2. For the $i-th$ input sequence $F^i \in \mathbb{R}^{l\times d}$, the parameter of MA^i is minimum when $d_1 = d_2 = d/2$, where l and d are sequence length and dimension, respectively.*

Proposition 2. *The numbers of global and local branches in the $i-th$ mixed attention layer (MA^i) is d_1 and d_2. For the $i-th$ input sequence $F^i \in \mathbb{R}^{l\times d}$, the FLOPs of MA^i is minimum when $d_1 = d/2 + l/8, d_2 = d/2 - l/8$, where l and d are sequence length and dimension, respectively.*

Task-Related Head. Furthermore, we remove the Cls token that has been used since ViT [23], and propose a new paradigm to vision transformer that uses the task-related Head to obtain the corresponding task output through the final features. In this way, the transformer-based backbone can be separated from specific tasks. Each transformer layer does not need to interact with the Cls token, so the computation amount reduces slightly from $O((n+1)^2)$ to $O(n^2)$. Specifically, we use cross-attention to implement this head module. As shown in Figure 4, K and V are output features extracted by the transformer backbone, while Q is the task-related token that integrates information through cross-attention. M indicates that each Head contains M layers and is set to 2 in the paper. This design is flexible, with a negligible computation amount compared to the backbone.

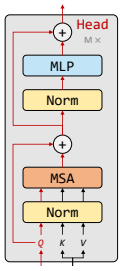

Figure 4: Structure of the proposed task-related Head.

Convergence Analysis. The convergence of Adam has been well studied in previous work [55]. It can be verified that global path and local path are both Lipschitz-smooth and gradient-bounded, as long as the transformer embedding is Lipschitz-smooth and gradient-bounded. In this case, the iteration sequence of the training algorithm converges to a stationary point of the learning objective with a convergence rate $O(1/sqrt(T))$, where T is the number of iterations.

4.3 Space-Filling Curve Module

In this subsection, we mainly deal with two-dimensional SFCs. The mathematical definition of SFC is a surjective function that continuously maps a closed unit interval:

$$\omega = \{z \mid 0 \leq z \leq 1\} = [0, 1], \text{ where } z \in \omega \tag{10}$$

as the domain of the function and is mapped onto the unit square:

$$\Omega = \{(x, y) \mid 0 \leq x \leq 1, \quad 0 \leq y \leq 1\}, \text{ where } (x, y) \in \Omega \tag{11}$$

which is our range of the function. One can easily define a surjective mapping from ω onto Ω, but it is not injective [67], *i.e.*, not bijective. Nevertheless, for a finite 1D sequence that maps 2D pixels, it contains bijective property.

Remark 1. *According to the definition of SFC, it must be surjective. For a finite 1D to 2D mapping, there is bound to be an intersection if the SFC is not injective, where there are at least two 1D points mapping to one same 2D point, i.e., $z_1 = SFC(x, y)$ and $z_2 = SFC(x, y)$. This is in contradiction to the finite SFC filling definition. Therefore, a finite SFC is bijective, and we can carry out two-way data lossless transformation through the pre-defined SFC.*

We propose an SFC module to map 2D images to the uniform 1D sequential format based on the above content. Therefore, we can address sequence inputs to handle multi-modal tasks in *one* unified model, *e.g.*, TIR and VLN. Taking image sequence as an example. As shown in the right bottom of Figure 3, the input image $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$ goes through the SFC module to be $\boldsymbol{I}^{Seq} \in \mathbb{R}^{HW \times 3}$, formulated as:

$$\boldsymbol{I}_k^{Seq} = \boldsymbol{I}_{i,j}, \text{ where } k = SFC(x, y), \tag{12}$$

where $1 \leq i \leq H$, $1 \leq j \leq W$, $1 \leq k \leq HW$, and SFC can be any manually specified space-filling curve. The difference among SFCs mainly reflects how the 1D sequence preserves the 2D spatial structure. As an extension, SFC can be defined to three or higher dimensional data space and can also serialize the intermediate network features, *e.g.*, the output feature maps of stage 3 in ResNet [31]. In this paper, we use *Embedded Slices* instead of *Embedded Patches* as the input of backbone that is consistent with NLP input format. As shown in Figure 5, we further propose a new SFC named SweepInSweep (SIS) that is equivalent to the input mode of the existing transformer-based methods, *e.g.*, ViT and DeiT, where the slice size equaling 16 is equivalent to that patch size equaling 4. The paper uses this mode for SFC by default. More visualizations of different SFCs are presented in Appendix C.

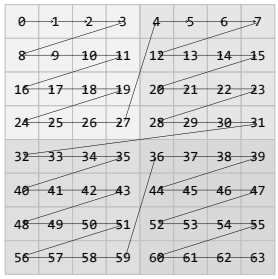

Figure 5: Indexing illustration of the proposed SweepInSweep SFC.

5 Experiments

5.1 Implementation Details.

We choose SOTA DeiT [64] as the baseline and follow the same experimental setting. By default, we train each model for 300 epoch from scratch without pre-training and distillation. The classification task is evaluated in ImageNet-1k dataset [20], and we conduct all experiments on a single node with 8 V100 GPUs. TIR is conducted on Fashion200k [27], MIT-States [34], and CSS [71] datasets, while VLN is performed on the R2R navigation dataset [1]. Detailed EAT variants can be viewed in Appendix D, and we supply the train and test codes in the supplementary material.

5.2 Comparison with SOTA Methods

The classification task is conducted on the ImageNet-1k dataset without external data. As noted in Table 1, we list four kinds of EAT of different magnitudes with *backbone + head* format, *i.e.*, Tiny, Small, Middle, and Base. They have comparable parameters with SOTA DeiT, but *a higher accuracy, a lower inference memory occupancy, and more throughput* in both GPU (tested with batch size equaling 256 in single V100) and CPU (tested with batch size equaling 16 in i7-8700K @3.70GHz). Our method is trained in 224 resolution for 300 epochs without distillation, so our result is slightly lower than large CNN-based EfficientNet in terms of Top-1. Moreover, we visualize comprehensive Top-1 results of different methods under various evaluation indexes in Figure 6 and present the results of a series of EAT models with different model sizes. ***1)*** Left two subgraphs show that our EAT outperforms the strong baseline DeiT, and even obtains better results than EfficientNet in both GPU and CPU throughput, *e.g.*, EAT-M. ***2)*** Besides, our approach is very competitive for

Table 1: Comparison with SOTA CNN-based and Transformer-based methods on ImageNet-1k dataset. Reported results are from corresponding papers. ⚗: Distillation; 1ke: Training 1000 epochs.

Network	Params. ↓ (M)	GFlops. ↓	Images/s ↑ GPU	Images/s ↑ CPU	Image Size	Top-1	Inference Memory
CNN-Based Nets							
ResNet-18	11.7M	1.8	4729	77.9	224^2	69.8	1728
ResNet-50	25.6M	4.1	1041	21.1	224^2	76.2	2424
ResNet-101	44.5M	7.8	620	13.2	224^2	77.4	2574
ResNet-152	60.2M	11.5	431	9.4	224^2	78.3	2694
RegNetY-4GF	20.6M	4.0	976	22.1	224^2	80.0	2828
RegNetY-8GF	39.2M	8.0	532	12.9	224^2	81.7	3134
RegNetY-16GF	83.6M	15.9	316	8.4	224^2	82.9	4240
EfficientNet-B0	5.3M	0.4	2456	49.4	224^2	77.1	2158
EfficientNet-B2	9.1M	1.0	1074	22.6	260^2	80.1	2522
EfficientNet-B4	19.3M	4.5	313	6.1	380^2	82.9	5280
EfficientNet-B5	30.4M	10.4	145	3.0	456^2	83.6	7082
EfficientNet-B7	66.3M	38.2	48	1.0	600^2	84.3	14650
NFNet-F0	71.5M	9.6	574	12.5	256^2	83.6	2967
Transformer-Based Nets							
ViT-B/16	86.6M	17.6	291	10.3	384^2	77.9	2760
ViT-L/16	304M	61.6	92	2.7	384^2	76.5	4618
DeiT-Ti	5.7M	1.3	2437	89.8	224^2	72.2	1478
DeiT-S	22.1M	4.6	927	31.4	224^2	79.8	1804
DeiT-B	86.6M	17.6	290	10.2	224^2	81.8	2760
EAT-Ti	5.7M (4.8M + 0.9M)	1.0	2442	95.4	224^2	72.7	1448
EAT-S	22.1M (18.5M + 3.6M)	3.8	1001	34.4	224^2	80.4	1708
EAT-M	49.0M (41.0M + 8.0M)	8.4	519	18.4	224^2	82.1	2114
EAT-B	86.6M (72.4M +14.2M)	14.8	329	11.7	224^2	82.0	2508
DeiT-Ti ⚗	5.9M	1.3	2406	87.1	224^2	74.5	1476
DeiT-Ti / 1ke ⚗	5.9M	1.3	2406	87.1	224^2	76.6	1476
EAT-Ti ⚗	5.7M (4.8M + 0.9M)	1.0	2442	95.4	224^2	74.8	1448
EAT-Ti / 1ke ⚗	5.7M (4.8M + 0.9M)	1.0	2442	95.4	224^2	77.0	1448

resource-restrained scenarios, *e.g.*, having a lower inference memory occupancy and parameters while obtaining considerable accuracy. ***3)*** Also, we find accuracy saturation in transformer structure that is also mentioned in recent works [65, 82], and we look forward to follow-up works to solve this problem. Furthermore, some techniques to help improve the accuracy of the model are applied, and our EAT model obtains a better accuracy.

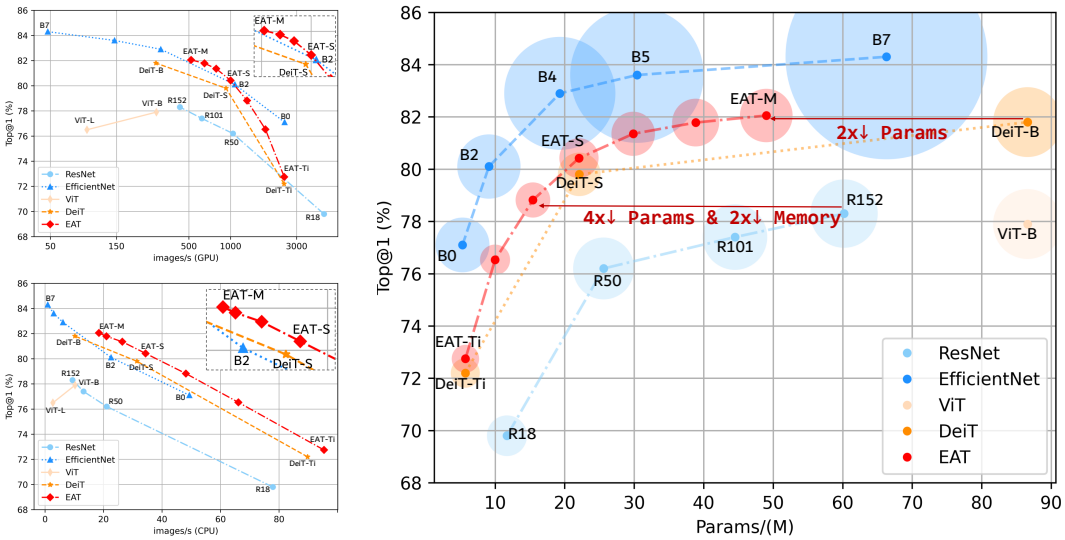

Figure 6: Comparison of different methods under different evaluation indexes. From left to right are GPU throughput, CPU throughput, and model parameters. The smaller the circle in the right sub-figure, the smaller the inference memory occupancy. Please zoom in for a better visual effect.

Table 2: Retrieval performance on three datasets in rank-1. The vision and language models are replaced by other nets, marked in the lower right footnote as V and L. *: Our reproduction; -: No corresponding result. The best number is in bold.

Method	CSS	MIT-States	Fashion200k
TIRG [71]	-	12.2	14.1
TIRG*	70.1	13.1	14.0
+ TR_L	70.7	13.2	14.3
+ EAT_L	71.0	13.4	14.9
+ R101_V	73.0	14.3	19.0
+ EAT_V	73.5	14.9	19.9
+ EAT_{V+L}	**73.8**	**15.0**	**20.1**

Table 3: VLN experiment on the R2R dataset. *: Our reproduction with suggested parameters by authors. NE: Navigation Error; SPL: Success weighted by (normalized inverse) Path Length; nDTW: normalized Dynamic Time Warping.

	Method	NE ↓	SPL ↑	nDTW ↑
Seen	PTA*	3.98	0.61	0.71
	+ EAT_L	**3.84**	**0.62**	**0.72**
	+ EAT_V	3.95	0.61	0.71
	+ EAT_{V+L}	3.95	0.61	0.71
Unseen	PTA*	6.61	0.28	0.48
	+ EAT_L	**6.41**	**0.33**	**0.51**
	+ EAT_V	6.63	0.29	0.49
	+ EAT_{V+L}	6.43	0.32	**0.51**

Table 4: Ablation study for several items on ImageNet-1k. *Default* represents the baseline method based on EAT-B. The gray font indicates that the corresponding parameter is not be modified.

Ablation Items	Head Layers	Local Ratio	FFN Ratio	Top-1	Ablation Items	Kernel Size	Local Operator	SFC Mode	Image Size	Top-1
Default	1	0.50	4			3	1D Conv	SIS	224^2	80.264
Head Layer	0	0.50	4	79.692	Kernel Size	1	1D Conv	SIS	224^2	79.128
	2	0.50	4	80.422		5	1D Conv	SIS	224^2	80.256
	3	0.50	4	80.435		7	1D Conv	SIS	224^2	80.052
	4	0.50	4	80.454	Local Operator	3	Local MSA	SIS	224^2	79.940
	5	0.50	4	80.446		3	DCN	SIS	224^2	68.070
Local Ratio	1	0.25	4	80.280	SFC Mode	3	1D Conv	Sweep	256^2	71.280
	1	0.75	4	79.518		3	1D Conv	Scan	256^2	74.164
FFN Ratio	1	0.50	2	77.422		3	1D Conv	Hilbert	256^2	79.842
	1	0.50	3	79.176		3	1D Conv	Z-Order / SIS	256^2	80.438

5.3 Multi-Modal Experiments

Text-based Image Retrieval. TIR is the task of searching for semantically matched images in a large image gallery according to the given search query, *i.e.*, an image and a text string. Table 2 illustrates the quantitative results. Naive transformer and EAT_L have the same number of layers, and EAT_V has similar parameters as R101_V. The results show that EAT brings positive benefits by replacing either original sub-network, even though vision and language models share one structure.
Vision Language Navigation. We further assess our approach by VLN experiments, where an agent needs to follow a language-specified path to reach a target destination with the step-by-step visual observation. We choose the PTA [39] as the base method, and Table 3 shows results under the same experimental setting. Our unified EAT obtains comparable results with PTA in the seen mode but achieves greater improvement in the unseen mode, meaning better robustness. When only replacing the vision EAT, the result changes very little. We analyze the reason that the used visual feature dimension of EAT-B is smaller than the original ResNet-152 (from 2048 to 384), and reinforcement learning is sensitive to this. Overall, our unified EAT still contributes to the PTA method.

5.4 Ablation Study

Table 4 shows ablation results for several items on ImageNet-1k. ***1) Head Layer.*** Within a certain range, the model performance increases with the increase of the head layer, and it equals 2 in the paper for trading off model effectiveness and parameters. ***2) Local Ratio.*** Large local ratio causes performance falling and increases the parameters (*c.f.*, Section 4.2), so p equaling 0.5 is a better choice. ***3) FFN Ratio.*** Similar to the conclusion of work [69], a lower FFN ratio results in a drop in accuracy. ***4) Kernel Size.*** Large kernel can lead to accuracy saturation, so we set it to 3 in the paper. ***5) Local Operator.*** We replace the local operator with local MSA and 1D DCN. The former slightly reduces the accuracy (a structure similar to the global path may result in learning redundancy), while the latter greatly reduces the accuracy (possibly hyper-parameters are not adapted). ***6) SFC Mode.*** Since some SFCs, *e.g.*, Hilbert and Z-Order, only deal with images with the power of side lengths, we set the image size to 256 here. Results show that the images serialized by Z-Order or SIS get better results because they can better preserve 2D spatial information that is important for 1D-convolution.

5.5 Model Interpretation by Attention Visualization

We average attention maps of each head layer and visualize them to illustrate which parts of the image the model is focusing on. As shown in Figure 7, Head1 pays more attention to subjects that are meaningful to the classification results, while the deeper Head2 integrates features of Head1 to form the final vector for classification that focuses on more comprehensive areas. Also, Grad-CAM [58] is applied to highlight concerning regions by our model and results consistently demonstrate the effectiveness of our proposed Head module. Note that the Cls Head contains two head layers, denoted as Head1 and Head2 in the figure. We also visualize attention maps of middle layers in Appendix F and find that the global path prefers to model long-distance over DeiT.

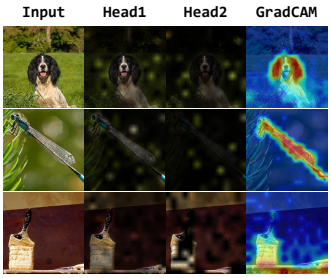

Figure 7: Visualization of attention maps for head layers and Grad-CAM results.

6 Conclusions

This paper explains the rationality of vision transformer by analogy with EA, which brings inspiration for the neural architecture design. Besides, the introduced SFC module serializes the multi-modal data into a sequential format, and we achieve the paradigm of using one unified model to address multi-modal tasks. Abundant experimental results demonstrate the effectiveness and flexibility of our approach. We hope this work may bring some enlightenment to network interpretability and thinking to design the unified model for multi-modal tasks. Recently, we observe that EA has been successfully used in other areas, *e.g.*, RL (PES [57], ERL [37]), NAS (SPOS [25], CARS [78]), Hyperparameter Adjustment (PBT [35]), *etc*. Nevertheless, how to combine heuristic algorithms such as EA with other fields is still a difficult problem, and we will further explore potential possibilities in the future.

Acknowledgment We thank all authors for their excellent contributions as well as anonymous reviewers and chairs for their constructive comments. This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61836015.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[2] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021.

[3] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. In *International Conference on Learning Representations*, 2021.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

[5] T. Bially. Space-filling curves: Their generation and their application to bandwidth reduction. *IEEE Transactions on Information Theory*, 15(6):658–664, 1969.

[6] Christian Böhm. Space-filling curves for high-performance data mining. *arXiv preprint arXiv:2008.01684*, 2020.

[7] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[9] Andreina I Castillo, Carlos Chacón-Díaz, Neysa Rodríguez-Murillo, Helvecio D Coletta-Filho, and Rodrigo PP Almeida. Impacts of local population history and ecology on the evolution of a globally dispersed pathogen. *BMC genomics*, 21:1–20, 2020.

[10] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *arXiv preprint arXiv:2012.00364*, 2020.

[11] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[12] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.

[13] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised visual transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[14] Ziyi Chen and Lishan Kang. Multi-population evolutionary algorithm for solving constrained optimization problems. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 381–395. Springer, 2005.

[15] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

[16] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *Arxiv preprint 2102.10882*, 2021.

[17] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021.

[18] Swagatam Das and Ponnuthurai Nagaratnam Suganthan. Differential evolution: A survey of the state-of-the-art. *IEEE transactions on evolutionary computation*, 15(1):4–31, 2010.

[19] Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. *arXiv preprint arXiv:2103.10697*, 2021.

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

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. *arXiv preprint arXiv:2103.03404*, 2021.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[24] Manolis Georgioudakis and Vagelis Plevris. A comparative study of differential evolution variants in constrained structural optimization. *Frontiers in Built Environment*, 6:102, 2020.

[25] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. In *European Conference on Computer Vision*, pages 544–560. Springer, 2020.

[26] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021.

[27] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1463–1471, 2017.

[28] Phan Thi Hong Hanh, Pham Dinh Thanh, and Huynh Thi Thanh Binh. Evolutionary algorithm and multifactorial evolutionary algorithm on clustered shortest-path tree problem. *Information Sciences*, 553:280–304, 2021.

[29] William Eugene Hart, Natalio Krasnogor, and James E Smith. Memetic evolutionary algorithms. In *Recent advances in memetic algorithms*, pages 3–27. Springer, 2005.

[30] Ahmad Hassanat, Khalid Almohammadi, Esra' Alkafaween, Eman Abunawas, Awni Hammouri, and VB Prasath. Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach. *Information*, 10(12):390, 2019.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] Ruining He, Anirudh Ravula, Bhargav Kanagal, and Joshua Ainslie. Realformer: Transformer likes residual attention. *arXiv e-prints*, pages arXiv–2012, 2020.

[33] David Hilbert. Über die stetige abbildung einer linie auf ein flächenstück. In *Dritter Band: Analysis· Grundlagen der Mathematik· Physik Verschiedenes*, pages 1–2. Springer, 1935.

[34] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015.

[35] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

[36] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 2021.

[37] Shauharda Khadka and Kagan Tumer. Evolution-guided policy gradient in reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1196–1208, 2018.

[38] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

[39] Federico Landi, Lorenzo Baraldi, Marcella Cornia, Massimiliano Corsini, and Rita Cucchiara. Perceive, transform, and act: Multi-modal attention networks for vision-and-language navigation. *arXiv preprint arXiv:1911.12377*, 2019.

[40] Changlin Li, Tao Tang, Guangrun Wang, Jiefeng Peng, Bing Wang, Xiaodan Liang, and Xiaojun Chang. Bossnas: Exploring hybrid cnn-transformers with block-wisely self-supervised neural architecture search. *arXiv preprint arXiv:2103.12424*, 2021.

[41] Xiaoyu Li, Lei Wang, Qiaoyong Jiang, and Ning Li. Differential evolution algorithm with multi-population cooperation and multi-strategy integration. *Neurocomputing*, 421:285–302, 2021.

[42] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021.

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[44] David E McCauley. Genetic consequences of local population extinction and recolonization. *Trends in Ecology & Evolution*, 6(1):5–8, 1991.

[45] Mohamed F Mokbel, Walid G Aref, and Ibrahim Kamel. Analysis of multi-dimensional space-filling curves. *GeoInformatica*, 7(3):179–209, 2003.

[46] Guy M Morton. A computer oriented geodetic data base and a new technique in file sequencing. *empty*, 1966.

[47] Pablo Moscato et al. On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. *Caltech concurrent computation program, C3P Report*, 826:1989, 1989.

[48] Kodai Nakashima, Hirokatsu Kataoka, Asato Matsumoto, Kenji Iwata, and Nakamasa Inoue. Can vision transformers learn without natural images? *arXiv preprint arXiv:2103.13023*, 2021.

[49] Nikhil Padhye, Pulkit Mittal, and Kalyanmoy Deb. Differential evolution: Performances and analyses. In *2013 IEEE Congress on Evolutionary Computation*, pages 1960–1967. IEEE, 2013.

[50] Giuseppe Peano. Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen*, 36(1):157–160, 1890.

[51] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[52] Adam P Piotrowski and Jarosław J Napiórkowski. The grouping differential evolution algorithm for multi-dimensional optimization problems. *Control and Cybernetics*, 39:527–550, 2010.

[53] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *empty*, 2018.

[54] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[55] S Reddi, Manzil Zaheer, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Proceeding of 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, 2018.

[56] Hans Sagan. *Space-filling curves*. Springer Science & Business Media, 2012.

[57] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[58] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[59] Anabela Simões and Ernesto Costa. The influence of population and memory sizes on the evolutionary algorithm's performance for dynamic environments. In *Workshops on Applications of Evolutionary Computation*, pages 705–714. Springer, 2009.

[60] Andrew N Sloss and Steven Gustafson. 2019 evolutionary algorithms review. *arXiv preprint arXiv:1906.08870*, 2019.

[61] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. *arXiv preprint arXiv:2101.11605*, 2021.

[62] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[63] Kay Chen Tan, Tong Heng Lee, and Eik Fun Khor. Evolutionary algorithms with dynamic population size and local exploration for multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 5(6):565–588, 2001.

[64] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[65] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021.

[66] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv preprint arXiv:2102.10662*, 2021.

[67] Levi Valgaerts. Space-filling curves an introduction. *Technical University Munich*, 2005.

[68] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. *arXiv preprint arXiv:2103.12731*, 2021.

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[70] Pradnya A Vikhar. Evolutionary algorithms: A critical review and its future prospects. In *2016 International conference on global trends in signal processing, information computing and communication (ICGTSPICC)*, pages 261–265. IEEE, 2016.

[71] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2019.

[72] Zhaoyi Wan, Haoran Chen, Jielei Zhang, Wentao Jiang, Cong Yao, and Jiebo Luo. Facial attribute transformers for precise and robust makeup transfer. *arXiv preprint arXiv:2104.02894*, 2021.

[73] Hanrui Wang, Zhanghao Wu, Zhijian Liu, Han Cai, Ligeng Zhu, Chuang Gan, and Song Han. Hat: Hardware-aware transformers for efficient natural language processing. *arXiv preprint arXiv:2005.14187*, 2020.

[74] Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[75] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[76] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

[77] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

[78] Zhaohui Yang, Yunhe Wang, Xinghao Chen, Boxin Shi, Chao Xu, Chunjing Xu, Qi Tian, and Chang Xu. Cars: Continuous evolution for efficient neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1829–1838, 2020.

[79] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. *arXiv preprint arXiv:2103.11816*, 2021.

[80] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[81] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020.

[82] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

[83] Liang Zhou, Chris R Johnson, and Daniel Weiskopf. Data-driven space-filling curves. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[84] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.

[85] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A Properties of Different Layers

Table 5: Parameters (Params), floating point operations (FLOPs), Sequential Operations, and Maximum Path Length for different layer types. Assume that the input data is a sequence of length l and depth d. k is the kernel size for 1D convolution and local self-attention, n is the query length, and inf stands for infinity.

Layer Type	Params	FLOPs	Sequential Operations	Maximum Path Length
Linear	$(d+1)d$	$2d^2l$	$O(1)$	$O(inf)$
1D Convolution	$(kd+1)d$	$2d^2lk$	$O(1)$	$O(n/k)$
Self-Attention	$4(d+1)d$	$8d^2l+4dl^2+3l^2$	$O(1)$	$O(1)$
Local Self-Attention	$4(d+1)d$	$8d^2l+4dlk+3ld$	$O(1)$	$O(n/k)$
Cross-Attention	$4(d+1)d$	$4d^2l+(4dl+2d^2+3l)n$	$O(1)$	$O(1)$

We adopt 1D convolution as the local operator that is more efficient and parameter friendly, and the improved transformer structure owns a constant number of sequentially executed operations and $O(1)$ maximum path length between any two positions.

B Analysis of the Local Ratio

Proposition 3. *The numbers of global and local branches in the $i-th$ mixed attention layer (MA^i) is d_1 and d_2. For the $i-th$ input sequence $F^i \in \mathbb{R}^{l\times d}$, the parameter of MA^i is minimum when $d_1 = d_2 = d/2$, where l and d are sequence length and dimension, respectively.*

Proof. Given the $F^i \in \mathbb{R}^{l\times d}$ and MA^i with d_1 and d_2, the overall parameters $Params^i = 4(d_1+1)d_1 + (d_2+1)d_2 + (kd_2+1)d_2$ according to Table 5 (k is the kernel size), and it is factorized as follows:

$$\begin{aligned} Params^i &= 4(d_1+1)d_1 + (d_2+1)d_2 + (kd_2+1)d_2 \\ &= 4{d_1}^2 + (k+1){d_2}^2 + 4d_1 + 2d_2 \end{aligned} \tag{13}$$

Based on $d_1 + d_2 = d$, we have

$$Params^i = (5+k){d_2}^2 - (8d+2)d_2 + 4d^2 + 4d \tag{14}$$

Applying the minimum value formula of a quadratic function to the equation 14, we can obtain the minimum value $2d^2 + 3d + 1/8$, where $d_1 = d/2 - 1/8$ and $d_2 = d/2 + 1/8$. Given that d_1 and d_2 are integers, we make $d_1 = d_2 = d/2$. Therefore, the minimum value of equation 14 becomes $2d^2 + 3d$ that is nearly half of the original self-attention layer, *i.e.*, $4d^2 + 4d$, according to Table 5. □

Proposition 4. *The numbers of global and local branches in the $i-th$ mixed attention layer (MA^i) is d_1 and d_2. For the $i-th$ input sequence $F^i \in \mathbb{R}^{l\times d}$, the FLOPs of MA^i is minimum when $d_1 = d/2 + l/8, d_2 = d/2 - l/8$, where l and d are sequence length and dimension, respectively.*

Proof. Given the $F^i \in \mathbb{R}^{l\times d}$ and MA^i with d_1 and d_2, the overall FLOPs $FLOPs^i = 8{d_1}^2l + 4d_1l^2 + 3l^2 + 2{d_2}^2l + 2{d_2}^2lk$ according to Table 5 (k is the kernel size), and it is factorized as follows:

$$\begin{aligned} FLOPs^i &= 8{d_1}^2l + 4d_1l^2 + 3l^2 + 2{d_2}^2l + 2{d_2}^2lk \\ &= (8{d_1}^2 + 4d_1l + 3l + 2{d_2}^2 + 2{d_2}^2k)l \end{aligned} \tag{15}$$

Based on $d_1 + d_2 = d$, we have

$$FLOPs^i = [(10+2k){d_2}^2 - (16d+4l)d_2 + 8d^2 + 4dl + 3l]l \tag{16}$$

Applying the minimum value formula of a quadratic function to the equation 16, we can obtain the minimum value $(4d^2 + 2ld - l^2/4 + 3l)l$, where $d_1 = d/2 - l/8$ and $d_2 = d/2 + l/8$. Given that d_1 and d_2 are integers and l is usually much smaller than d in practice, we make $d_1 = d_2 = d/2$ that is consistent with the above proposition 1. Therefore, the minimum value of equation 16 becomes $(4d^2 + 2ld + 3l)l$ that is nearly half of the original self-attention layer according to Table 5, *i.e.*, $(8d^2 + 4ld + 3l)l$. □

C Visualization of Hilbert SFC

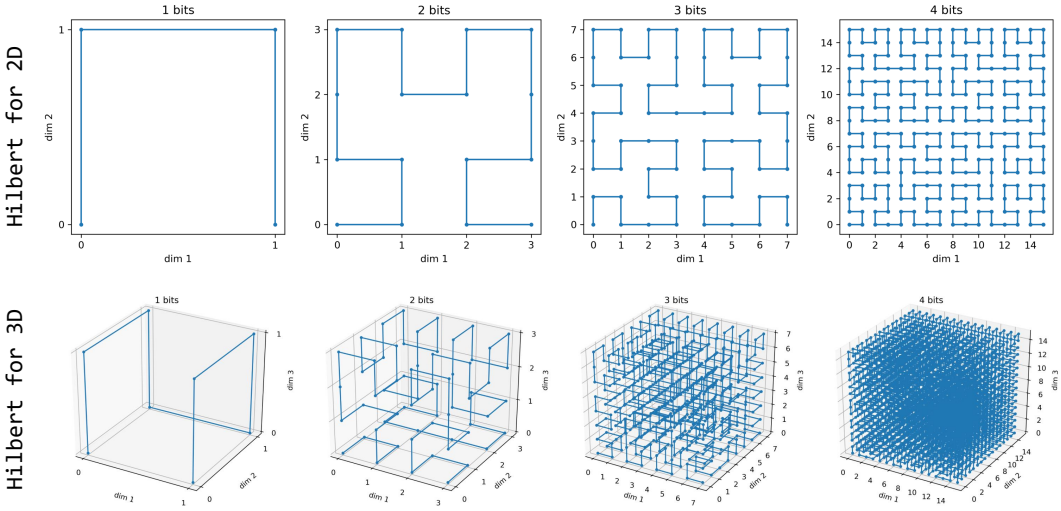

Figure 8: Visualization of Hilbert SFC in 2D and 3D format under different image bits.

We visualize the SFC of Hilbert for a more intuitive understanding in Figure 8.

D Variants of EAT

Table 6: Detailed settings of our EAT variants.

Model	Emb. Dim.	H. in MSA	Layers	Head Layers	FFN Ratio	Local Ope.	Local Ratio	Kernel Size	SFC Mode	Image Size	Params.
EAT-Ti	192	2	12	2	4	1D Conv	3	0.5	SIS	224^2	5.7M
EAT-S	384	3	12	2	4	1D Conv	3	0.5	SIS	224^2	22.1M
EAT-M	576	4	12	2	4	1D Conv	3	0.5	SIS	224^2	49.0M
EAT-B	768	6	12	2	4	1D Conv	3	0.5	SIS	224^2	86.6M

Table 6 shows detailed settings of our proposed four EAT variants.

E Visualization of Attention Map in Task-Related Head

Taking the classification Head as an example, we visualize the attention map in the Head to intuitively explain why the model works. Specifically, we choose EAT-S here, which contains two layers for the classification Head, and each Head contains eight heads in the inner cross-attention layer. Here, capital Head indicates task-related Head, while lowercase head represents multi-head in the cross-attention. As shown in Figure 9, we normalize values of attention maps to [0, 1] and draw them on the right side of the image. Results show that different heads focus on different regions in the image, and the deeper Head2 integrates features of Head1 to form the final vector for classification that focuses on a broader area.

Furthermore, we average eight attention maps of each head layer and use it as the global attention map that represents which parts of the image the corresponding head layer is focusing on, as shown in Figure 11. Also, Grad-CAM [58] is applied to produce a coarse localization map highlighting the crucial regions in the image. By analyzing results, we find that both visualization methods focus more on the subjects in the image, demonstrating the effectiveness of our proposed Head module.

F Visualization of Attention Map in Middle Layers

We also visualize some attention maps for middle layers, taking the heads in the fourth and sixth layers as an example. As shown in Figure 10, compared with DeiT without local modeling, our EAT pays more attention to global information fusion, where more significant values are found at off-diagonal locations. We analyze the reason for this phenomenon that the parallel local path takes responsibility for some of the local modelings that would have been the responsibility of the MSA.

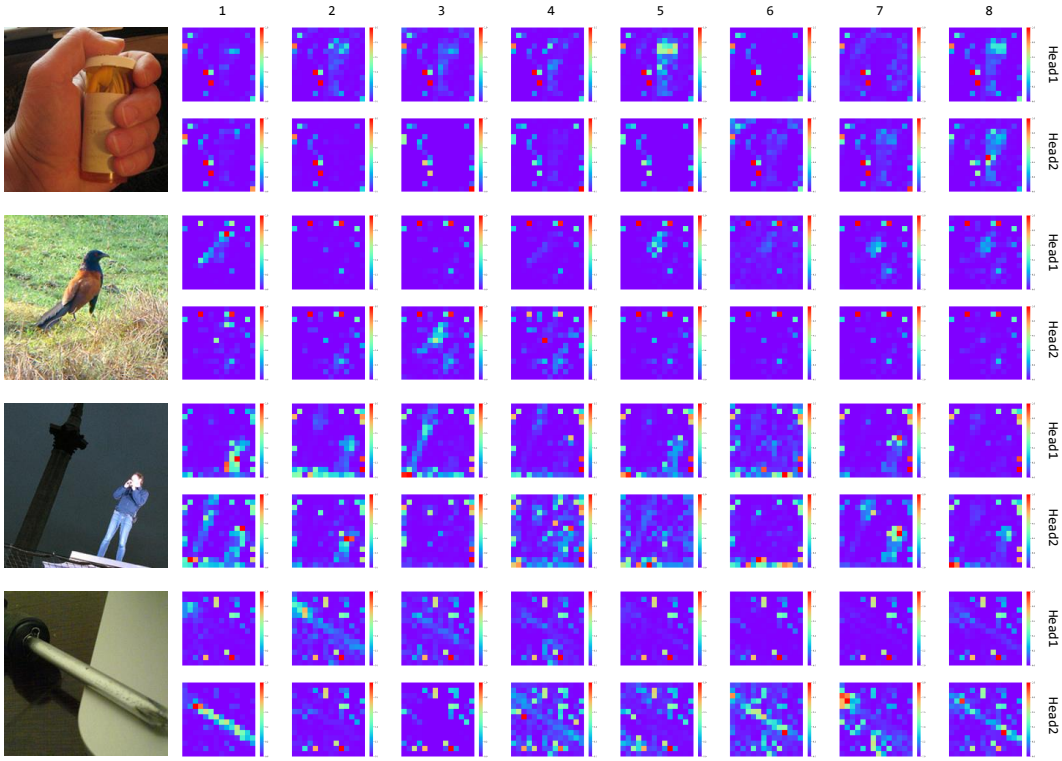

Figure 9: Visualization of Attention Maps in the classification head. We display two head layers for each image and eight attention maps for eight heads in each head layer.

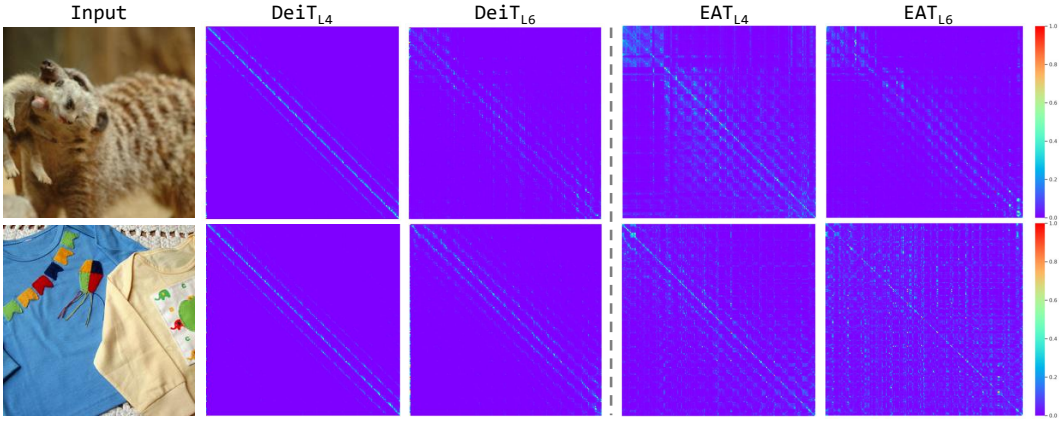

Figure 10: Visualization of attention maps for the fourth and sixth middle layers. The first column shows the input images; the second and third columns are visualized attention maps for DeiT, while the fourth and fifth columns for our EAT-S.

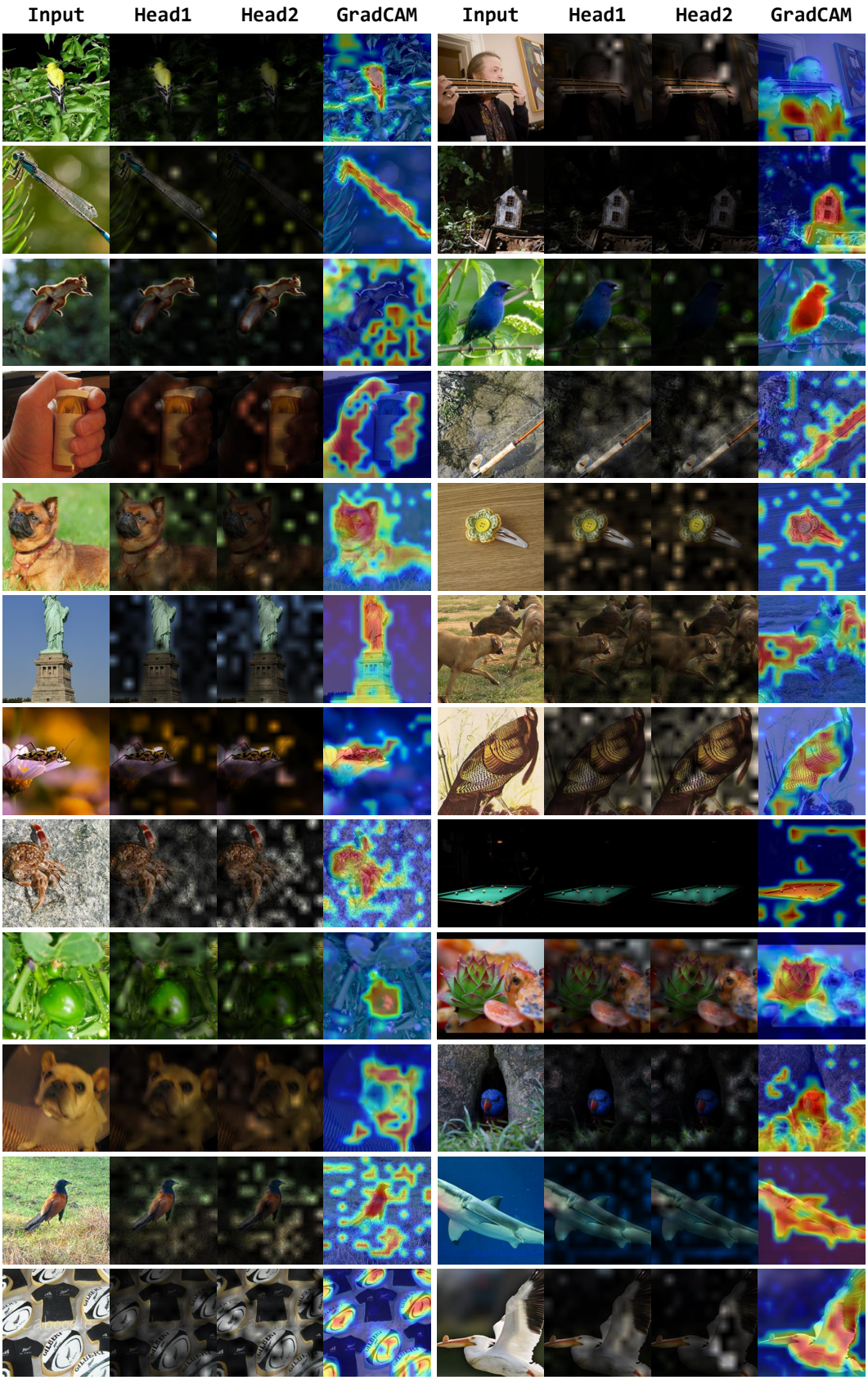

Figure 11: Visualization of overall attention for different head layers and Grad-CAM results.