EAT: An Enhancer for Aesthetics-Oriented Transformers

Shuai He
Beijing University of Posts and Telecommunications
Beijing, China
hs19951021@bupt.edu.cn

Anlong Ming*
Beijing University of Posts and Telecommunications
Beijing, China
mal@bupt.edu.cn

Shuntian Zheng
Beijing University of Posts and Telecommunications
Beijing, China
zhengshuntian@bupt.edu.cn

Haobin Zhong
Beijing University of Posts and Telecommunications
Beijing, China
zhonghaobin2023@bupt.edu.cn

Huadong Ma
Beijing University of Posts and Telecommunications
Beijing, China
mhd@bupt.edu.cn

ABSTRACT

Transformers have shown great potential in various vision tasks, but none of them have surpassed the best convolutional neural network (CNN) model on image aesthetics assessment (IAA) tasks. IAA is a challenging task in multimedia systems that requires attention to both foreground and background, as well as robustness to noisy and redundant labels. The global and dense attention mechanism of Transformers, designed for saliency-oriented tasks, may miss important aesthetic information in the background, increase the computational cost and slow down the convergence on IAA tasks. To address these issues, we propose an Enhancer for Aesthetics-Oriented Transformers (EAT). EAT uses a deformable, sparse and data-dependent attention mechanism that learns where to focus and how to refine attention by offsets. EAT also guides the offsets to balance the attention between foreground and background according to dedicated rules. We evaluate EAT-enhanced Transformers on four representative datasets: AVA, TAD66K, FLICKR-AES and AADB, and demonstrate they outperform the previous methods with fewer training epochs. Code is available.

CCS CONCEPTS

• **Computing methodologies** → **Computational photography**.

KEYWORDS

image aesthetics assessment; attention mechanism; neural networks

ACM Reference Format:
Shuai He, Anlong Ming, Shuntian Zheng, Haobin Zhong, and Huadong Ma. 2023. EAT: An Enhancer for Aesthetics-Oriented Transformers. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, Ottawa, ON, Canada, 10 pages. https://doi.org/10.1145/3581783.3611881

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611881

1 INTRODUCTION

As digital photography expands rapidly, image aesthetics assessment (IAA) has become one of the most important criteria to automatically assess whether the image meets users' aesthetic preferences [7, 33]. It is also an essential step in imaging measurements among manufacturers to evaluate the performance of smartphones and cameras [1, 22]. IAA methods are currently under development, usually guided by the understanding that the results of these methods should be statistically consistent with the perception of human observers.

Although CNN models have yielded many remarkable achievements, they are susceptible to the attention dispersion phenomenon in IAA tasks and may even fail to properly focus on the aesthetic information of the foreground, as proved by the work of [10]. Compared to CNN counterparts, Transformer-based models are driven by attention mechanisms, which have larger receptive fields and excel at modeling long-range dependencies, thus allowing models to avoid the above problem.

Nevertheless, applying Transformers for IAA is challenging because the attention mechanism of Transformers is a double-edged sword. Most Transformer models have been designed in a saliency-oriented manner to solve classic classification or segmentation problems; consequently, they can more easily focus on or even overemphasize salient objects or foreground areas, which also means that they are more prone to ignore background regions and result in ***attention bias*** problem. For aesthetic tasks, however, a lack of the attention to a background is inconsistent with the original intention of a photographic work, e.g., hierarchical compositions are usually formed with the deliberate consideration of background regions [14, 32]. Most Transformer-based models tend to generate large prediction errors for background-sensitive images (Fig. 2). Therefore, Transformer-based models ***have not comprehensively surpassed CNN models on IAA tasks yet***, to our knowledge. Moreover, the superfluous attention in Transformers usually leads to unnecessarily computational cost and slow convergence on IAA tasks and may even result in overfitting on small IAA datasets.

To guide the IAA model to locate more reasonable regions, we present an Enhancer for Aesthetics-Oriented Transformers (EAT) based on the deformable attention, which is able to learn where to locate interest points and how to refine attention by means of offsets for IAA. Specifically, our interest points are generated by an interest network instead of being manually specified. The candidate

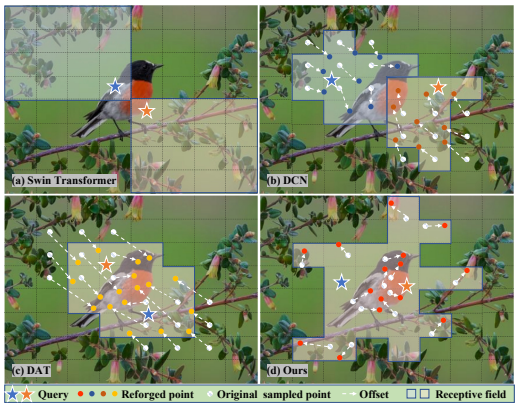

Figure 1: To compute self-attention in different/shared queries, (a) adopts shifted windows, (b) and (c) manually specify dense sample points instead of queries and learn foreground-oriented offsets, while (d) EAT obtains sparse sample points via a learning approach and learns aesthetics-oriented offsets.

Figure 2: On the images with "background-sensitive" content (e.g., those tagged with "sky" in AVA), Transformer-based models [8, 15, 21, 35, 37] perform 17% worse than average in terms of the SRCC metric, while our EAT can effectively perceive aesthetic information of these images to solve this issue.

keys/values generated from these interest points can be quickly transferred to regions with rich aesthetic information (Fig. 1(d)). This design makes the number of interest points controllable and reduces their redundancy in space, which endows our EAT with linear computational complexity. To solve attention bias issue, the model dynamically refines the offsets of interest points in the back-propagation process. During the early training process, the model is encouraged to explore aesthetic information in a larger space, especially the background areas away from salient objects. In the later stage of training, the model exploits explored information to calculate aesthetic relationships. This strategy prevents convergence to sub-optimal interest points, which results in missing of aesthetic information in the background.

Our contributions are concluded as follows:

- This is the ***first time***, to our knowledge, to reveal the attention bias problem of Transformer-based methods in IAA, which brings the necessity of rebalancing the attention between foreground and background to the forefront of the community.
- Our proposed EAT adopts a new deformable attention with dedicated offset strategies to solve the above issue, which explores the adaptation of transformers and can be integrated into different transformers for IAA tasks.
- EAT achieves state-of-the-art performance on four representative IAA datasets, with faster training speed and lower computational cost than most IAA methods.

2 RELATED WORK

IAA Models. General IAA encompasses three types of tasks: binary classification (aesthetically positive or negative) [5, 23], aesthetic score regression [26, 31] and score distribution prediction [2, 30, 34], while personalized IAA adapts a generic aesthetics model for individual user's preference.[24, 29]. CNN models have been widely applied to extract aesthetic information and map visual features to annotated labels. To equip models with some prior knowledge, a common practice is to pre-train CNNs on ImageNet dataset [6] for basic information. However, the retrieved basic information would usually be lost during the fine-tuning process on IAA datasets [10], which makes such models insufficiently understand aesthetics and cannot locate on effective salient regions. Nonetheless, these models' ability to evenly distribute attention to the background region enables them to acquire more contextual information, yielding superior performance compared to current Transformer-based approaches that rely solely on salient areas.

Recently, Transformers have been introduced into IAA tasks (e.g., MaxViT [35], MUSIQ [15], and ViT [8]). However, unlike the tasks of classification and recognition with attention only focusing on salient objects, IAA is inevitably dependent on background information. Undeniably, transformer-based models can focus on salient objects or semantically meaningful content, while these models respond slightly to background regions without significant features. Thus, most of the transformer-based methods require much longer training epochs to converge than the existing CNNs on IAA tasks.

In this paper, attention bias in IAA tasks is revealed for the first time, and we enhance the existing transformer attention mechanism to address this problem. Specifically, we balance the attention between foreground and background by guiding the direction and speed of attention updates, while utilizing a sparse attention mechanism instead of global attention to achieve lower training epochs and reduced computational costs.

Deformable Convolution and Attention. Although deformable transformers ***have not been applied to IAA tasks yet***, they are a promising solution to improve attention efficiency and facilitate flexible adaptation to target regions. In deformable convolution [4, 41], the grid-based sampling locations of standard convolution are offset to flexible spatial locations by means of displacements learned with respect to the preceding feature maps. Some works [3, 36, 37, 42] have combined deformable convolution with the attention mechanism of Transformers and have obtained impressive

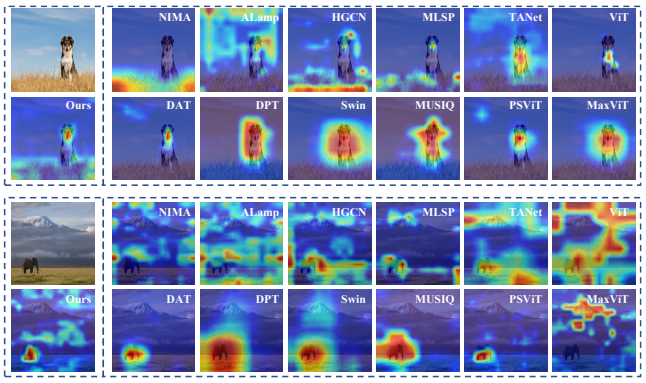

Figure 3: The saliency maps generated by 13 models for IAA tasks, while our EAT captures more reasonable and aesthetically relevant regions in both foreground and background by varying degrees of attention.

results on many challenging tasks. Such a deformable attention mechanism is able to learn offsets from queries and then combine reference points to shift keys and values in order to steer attention to target regions.

For IAA tasks, a naive implementation of the above idea leads to unreasonable and slow convergence, because the hand-crafted reference/sampling points generated from uniform grids are chosen in an aesthetics-agnostic (Fig. 1), and then adjusted in a saliency-oriented manner. However, these points may not be reasonable or necessary, and long training epochs are inevitable for the attention weights to be learned from hand-crafted points to meaningful locations, thus leading to slow training convergence and inferior performance. To address the issues of conventional deformable works, we dynamically assign interest points to control the computational complexity of attention. Moreover, our interest points offsets are differentiable, allowing them to be flexibly placed in appropriate locations during the back-propagation process, while offsets can assist in fixing attention bias in each training step.

3 PRELIMINARIES

Taking a feature map X as the input to obtain queries Q, keys K and values V, the classical multihead self-attention (MSA) can be calculated as:

$$\text{MSA} = \text{Softmax}\left(\mathbf{Q}^{(m)}\mathbf{K}^{(m)T}/\sqrt{D} + Bias\right)\mathbf{V}^{(m)T}, \tag{1}$$

where D is the query/key dimension, m denotes the m-th attention head, and $Bias$ is a bias matrix, as shown in [21].

The above calculation's complexity is dependent on the entire feature map, which significantly increases the amount of calculation. To address this issue, deformable attention was developed to sample feature points. Consider an input feature map $F \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the feature channel dimension, height, and width, respectively. The deformed points $P \in \mathbb{R}^{n\times n}$ can be sampled using a uniform grid size Δx and Δy:

$$(P_x, P_y) = F_{(s_x\Delta x, s_x\Delta y)}, s_x, s_y = 1, 2, \ldots, n. \tag{2}$$

Subsequently, the offsets (O_x, O_y) are introduced to the regular grid sampling locations in the standard attention, which can be learned during the back-propagation process. Following this, features X_s are sampled at the locations of deformed points as keys and values, calculated by projection matrices $K = X_sW_k, V = X_sW_v$.

However, the initial deformed points (P_x, P_y) are set manually, thus having a high degree of redundancy, and they are prone to fall into local optimum even after offset adjustment.

Besides, according to the work of Li *et al.* [18], the attention mechanism exhibits a local bias, whereby areas with higher attention weights lead to a greater increase in update speed, which in turn promotes even higher attention weights. This issue is also evident in the IAA tasks, we simplify this process as [18]:

$$\frac{\partial L}{\partial A_s} - \frac{\partial L}{\partial A_b} \approx (\sum_{j=1}^{N} A_j\beta_j - \|y\|)(\beta_s - \beta_b). \tag{3}$$

Here, A_s, A_b represent the attention weights of the salient and background areas, respectively, while β_s and β_b are the corresponding update speeds, y denotes the target. When $A_s > A_b$, this backward propagation properties would lead $\frac{\partial L}{\partial A_s} < \frac{\partial L}{\partial A_b}$.

IAA models are usually pre-trained on classification datasets to obtain better performance. However, during the fine-tuning process on IAA datasets, the update speed of β_s is faster than β_b and then lead $\beta_s > \beta_b$, $A_s > A_b$. Based on the above analysis, it can be concluded that the attention mechanism gives greater weight to salient areas and neglects crucial aesthetic information in the background, which results in an attention bias towards the foreground.

To solve the aforementioned issues, this paper introduces two enhancements: 1) Since ***the number of interest points*** can affect the computational complexity of the entire attention mechanism, we propose to learn adaptive, sparse, and data-dependent interest points instead of manually sampling them. This approach effectively reduces the redundancy and computational burden of the attention mechanism. 2) In deformable attention, the updates to the attention ***are controlled by offsets***. Therefore, we design a dedicated offset guide module to regulate the speed and direction of the weight updates. The design details of our method are as follows.

4 METHOD

In this section, we show how to enhance the existing attention mechanism in a Transformer to alleviate attention bias, while reducing computational loss and improving training speed. Specially, ***we specify the enhancement strategy into a deformable Transformer EAT***. While in implementation, there are three challenges versus standard deformable works: 1) fixing attention bias in each training step, 2) compressing training epochs then lowering FLOPs and 3) designing a module that can be seamlessly integrated into existing deformable methods ***without altering network structure***.

4.1 Enhanced Attention Module.

Interest Generate Module. In order to achieve differentiable interest points, and be able to flexibly replaced the manual reference points module in the existing deformable transformer without

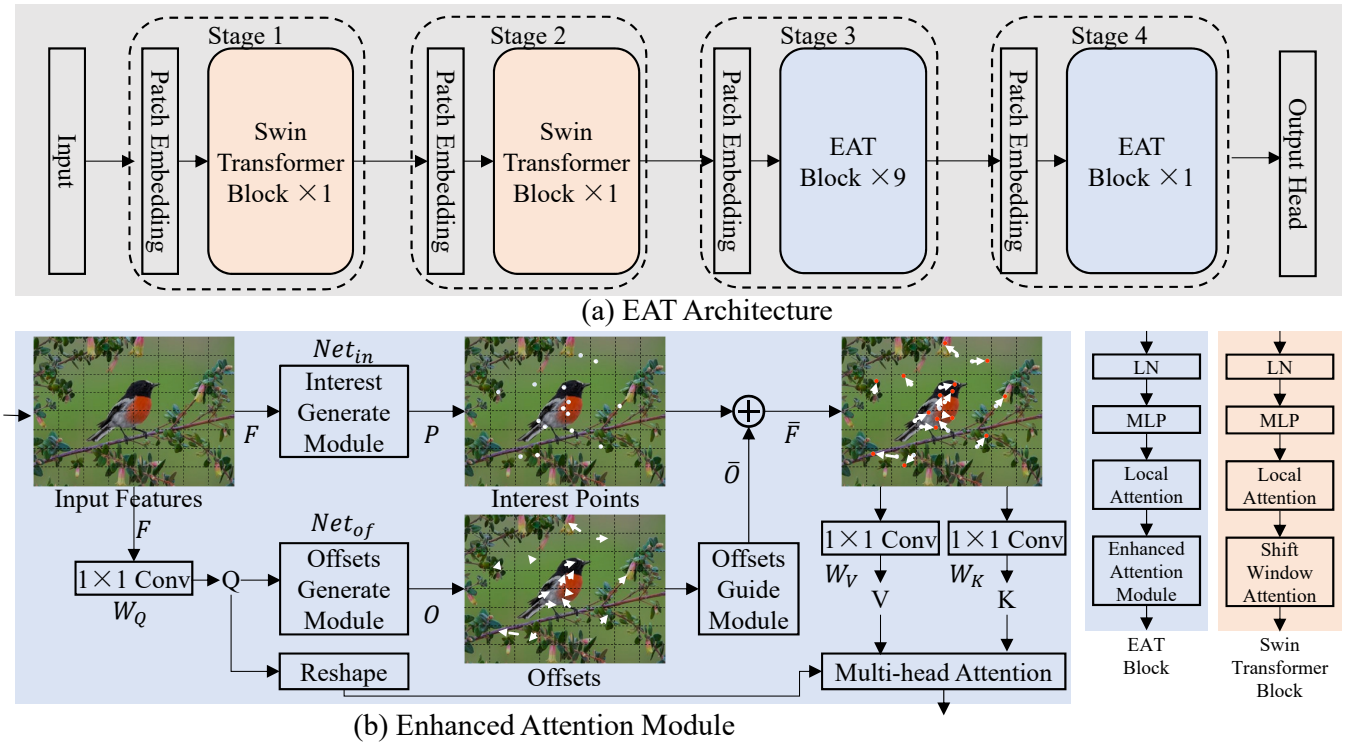

Figure 4: A typical structure, denoted as EAT, incorporates the proposed enhancement strategy. (a) Illustrates the overall architecture of EAT and (b) shows feature processing for the enhanced attention module (EAM). Initial interest points are learned from the input features by the Interest Generation Module, while original offsets are generated by the Offset Generation Module from queries and refined by the Offset Guidance Module. Finally, sampled features are obtained by bilinear interpolation for keys and values.

adding too much extra computation, we use a lightweight Interest Generation Module Net_{in} to generate interest points:

$$(P_x, P_y) = Tanh(Net_{in}(F)), \tag{4}$$

which includes a convolutional layer to map F to features $F_{in} \in \mathbb{R}^{2\times H_{in}\times W_{in}}$, where $2 \times H_{in} \times W_{in}$ is the output size, with the two channels representing the abscissa P_x and ordinate P_y of each point of interest ($P = (P_x, P_y)$). By changing the kernel size k_{in} of the convolutional layer to control the number of generated interest points ($H_{in} \times W_{in}$), we can flexibly adjust the complexity of the attention calculation while reducing the redundancy of the attention space. To facilitate subsequent processing and accelerate convergence, we normalize the coordinate values to the range (-1, +1), where (-1, -1) indicating the top left and (+1, +1) indicating the bottom right.

Offsets Generate and Guide Module. To obtain the offset for each interest point, the feature maps are linearly projected to obtain query tokens $Q = FW_Q$, which are then fed into another lightweight module Net_{of} to generate the offsets:

$$(O_x, O_y) = Tanh(Net_{of}(Q)). \tag{5}$$

To adapt the interest points, the output dimensions of Net_{of} are the same as those of Net_{in}, which are adjusted based on the kernel size $k_{of} = k_{in}$. Here, O_x and O_y represent the amplitudes of the abscissa and ordinate ($O = (O_x, O_y)$), respectively.

Since models for aesthetic tasks are usually pre-trained on ImageNet [6], such a network prefers to make most of the interest points closing to salient objects while scattering the rest in the background. By assigning each point an offset, the classical works have attempted to progressively guide all attention points to salient objects [36, 37]. In contrast, IAA prefers to preserve some interest points in the background to retrieve as much aesthetic information as possible. Thus, a guidance module Φ is designed to refine the offsets:

$$(\bar{O}_x, \bar{O}_y) = \Phi(O_x, O_y). \tag{6}$$

Firstly, in module Φ, to regulate the update speed of attention in both salient and background areas, we divide the values of the offsets and interest points into four quadrants, calculate whether each point and the direction of its offset are in the same quadrant, and multiply by different w values to adjust offset amplitude:

$$w = \begin{cases} 1, \text{ if } ReLU(O_xP_x)ReLU(O_yP_y) > 0 \\ \eta, \text{ otherwise} \end{cases} \tag{7}$$

where $0 < \eta < 1$. The purpose of this function is to weaken the ability of the offsets for pulling points that are not in the same quadrant to the regions of salient objects, which encourages those interest points that are away from salient objects (not in the same

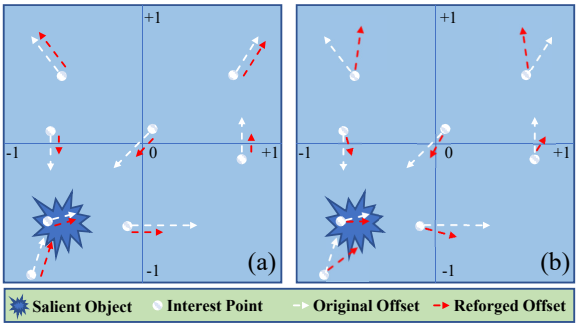

Figure 5: Offset Guidance Module. (a) Weaken the amplitude of offset if it is not in the same quadrant as interest points. (b) Randomly rotate the direction of offset.

quadrant) to retrieve aesthetic information from the background, as shown in Fig. 5(a).

Secondly, to avoid falling into local optimum, we regulate the update direction of attention by randomly rotating the offset direction in the early stage of training (Fig. 5(b)). The final offset $\bar{O} = (\bar{O}_x, \bar{O}_y)$ can be formulated as:

$$(\bar{O}_x, \bar{O}_y) = [w(O_x cos\alpha - O_y sin\alpha), w(O_y cos\alpha + O_x sin\alpha)] \quad (8)$$

where α is used to control the level of exploration and can be expressed as:

$$\alpha = (1 - \frac{e}{E}) \cdot Random(-\frac{\pi}{2}, \frac{\pi}{2}), \quad (9)$$

where e is the current training epoch, and E is the total number of training epochs. Specifically, formulation (8) drives the model to tend to retrieve more unknown/uninteresting aesthetic information in the early training stage, and to retrieve as much background information as possible, thereby enhancing the model's robustness to background information. In the later stage of training, the randomness is reduced, and the model optimizes its decisions by exploiting the explored information. Then, the sampled feature map $\bar{F}$ that is used to generate keys and values can be calculated as:

$$\bar{F} = B[F, Tanh((P_x, P_y) + (\bar{O}_x, \bar{O}_y))]. \quad (10)$$

We adopt bilinear interpolation, $B[\cdot, \cdot]$, to sample features from the feature maps, and the sampled features are fed into the key and value projections to obtain $K = \bar{F}W_K$ and $V = \bar{F}W_V$. Finally, the standard multi-head attention mechanism of equation (1) is applied to calculate the output. Moreover, the back-propagation process and computational complexity are provided in the **Appendix**.

4.2 Overall Architecture

The EAT architecture is presented in Fig. 4 (a). In our implementation, an input image is first embedded by a 4×4 convolution to obtain patch tokens. Then, two Swin Transformer blocks [21] make up stage 1 and stage 2, and the feature maps are firstly processed by shift window attention to locally aggregate aesthetic information. Subsequently, several proposed EAT blocks make up stage 3 and stage 4 to process these feature maps globally and model the long-range relations between the foreground and background. Finally, an output head is adopted to predict aesthetic score.

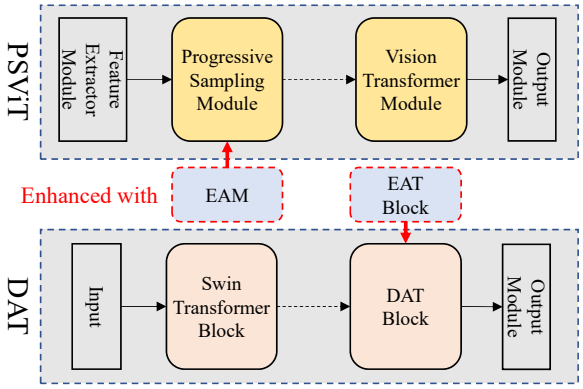

Figure 6: Examples of enhancing other deformable transformer architectures by EAT.

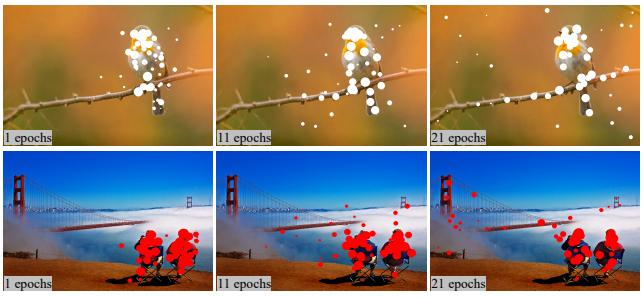

Figure 7: Interest points are visualized during training, with circle size indicating attention weights across multiple heads. Larger circles represent higher weights.

This stacking configuration is mainly considered from two factors. Firstly, our attention mechanism can reduce redundancy in the attention space by specifying interest points, but it may result in information loss if it is used in an early stage. Secondly, the main purpose of our structure is protecting important aesthetic information in the background while paying attention to the foreground information. However, at the early stage, the local features are difficult to distinguish foreground and background information.

5 EXPERIMENTS

5.1 Experimental Settings

Benchmark Datasets. We performed model evaluations on four representative datasets, AVA [28], TAD66K [10], FLICKR-AES [29] and AADB [17], which are the general, theme-oriented general, personalized aesthetic and aesthetic attribute datasets, respectively, for IAA tasks. The AVA dataset contains approximately 250,000 images, and each image is associated with a distribution of scores in a range of 1–10 rated by approximately 250 raters. For a fair comparison, we employ the most commonly train-test split way, e.g., in [2, 10, 26–28, 34] to generate 235,528 images for training and 20,000 images for test. The TAD66K dataset contains 66,000 images covering 47 popular themes, and each image has been densely annotated with scores from 1 to 10 by more than 1200 people based

Table 1: Comparing 16 models on AVA [28], we retrained Transformer-based models with publicly available codes marked by '*' for optimal performance. All models were pre-trained on ImageNet-1K and η of our model is set to 0.25. To ensure fair comparisons, we trained our EAT on three input sizes: 224×224, 512×512 (highlighted in yellow), and 640×640 (highlighted in green). Note: HGCN utilized a less commonly employed dataset splitting method.

	CNN-based models								Tranformer-based models									
Metric (AVA)	[34] NIMA 224	[26] ALamp 224	[2] AFDC 230	[30] HGCN 224	[11] MLSP 629	[10] TANet 224	[13] ResNext 512	[12] POC 640	[38] PSViT* B224	[3] DPT* L224	[21] Swin B224	[15] MUSIQ 512	[15] ViT B224	[37] DAT* B224	[35] MaxViT 512	Ours 224	Ours 512	Ours 640
FLOPs	-	-	-	-	-	-	-	-	264G	264G	240G	111G	880G	240G	120G	140G	200G	280G
Params	56M	99M	45M	44M	24M	40M	43M	1853M	21M	61M	87M	27M	88M	87M	31M	87M	87M	87M
Epoch	40+	50+	-	60+	40+	110+	-	-	70+	100+	70+	90+	80+	60+	300+	25+	**25+**	**30+**
PLCC↑	.636	.671	.671	.687	.757	.765	.781	.795	.645	.720	.737	.738	.739	.739	.745	**.770**	**.790**	**.814**
SRCC↑	.612	.666	.649	.665	.756	.758	.780	.794	.701	.694	.736	.726	.728	.738	.708	**.759**	**.786**	**.803**
Ratio↑	.751	.807	.779	.786	.925	**.940**	.940	.955	.847	.874	.901	.891	.913	.909	.877	.929	**.950**	**.961**

Table 2: Comparing 13 models on TAD66K [10] with the same FLOPs and Param values as in Table 1, all pre-trained on ImageNet-1K.

Metric (TAD66K)	CNN-based models						Transformer-based models							
	NIMA 224	ALamp 224	AFDC 230	HGCN 224	MLSP 629	TANet 224	ViT* B224	PSViT* B224	DPT* L224	Swin* B224	MUSIQ* 512	MaxViT* 512	DAT* B224	**Ours 224**
Epochs↓	80+	150+	-	80+	76+	110+	120+	160+	160+	100+	90+	100+	80+	**30+**
PLCC↑	.405	422	-	.493	.508	.531	.421	.431	.445	.454	.517	.518	.527	**.546**
SRCC↑	.390	.411	-	.486	.490	.513	.393	.408	70	.419	.489	.490	.499	**.517**
MSE↓	.021	.019	-	.020	.019	.016	.201	.022	.021	.201	.018	.018	.016	**.015**

on dedicated theme evaluation criteria. We used the official train-test split way [10], 52,248 for training and 14,079 for testing. The FLICKR-AES dataset consists of 40,000 images whose aesthetic scores range from 1 to 5 to reflect different levels of image aesthetics, with each image rated by 5 raters. We used the official train-test split way [29], 35,263 images rated by 173 users were used as the training set, and the rest of the 4,737 images rated by 37 users were used as the testing set. The AADB dataset collected 10,000 images evaluated by a total of 190 users. We use 22 workers and their rated images as the testing set, and the remaining 168 workers and the labeled images as the training set. The train-test split way is same with previous works [20, 39, 40].

Evaluation Metrics. We adopt two popular evaluation metrics, the spearman rank correlation coefficient (SRCC) and the pearson linear correlation coefficient (PLCC). In addition, we use the SRCC/accuracy ratio instead of the binary classification accuracy (aesthetically negative or positive) to better evaluate the general performance ([10, 11]). The metrics for the TAD66K dataset also include the mean squared error (MSE) loss. Since the ground truth in the AVA dataset consists of the score distribution, we use the earth mover's distance (EMD) loss to measure the distance between the ground-truth and predicted distributions.

Benchmark Models. We compared our EAT with 8 state-of-the-art (SOTA) general IAA models on the AVA and TAD66K datasets and 9 SOTA personalized IAA models on the FLICKR-AES dataset, and we selected 3 SOTA Transformer-based methods (MaxViT, MUSIQ, and ViT) with verified performance on the AVA dataset. Furthermore, we retrained 4 shift attention Transformer methods (PSViT [36], DPT [3], Swin [21], and DAT [37]) based on the official codes. To the best of our knowledge, there are currently no Transformer-based methods that have verified performance on the TAD66K dataset; therefore, we also retrained all 7 Transformer-based methods.

Training Details. Our method based on the PyTorch and optimized with an Adam [16] optimizer. We adopt nni (AutoML toolkit for hyper-parameter tuning) for the batch size and initial learning rate settings, and our learning rate is fixed with no decay rate strategy. For training the TAD66K, AADB, and FLICKR-AES datasets, we used mean squared error (MSE) loss. We opted for earth mover's distance (EMD) loss for training AVA.

5.2 Performance Evaluations

AVA Dataset. Table 1 lists the results of EAT and 13 other models on the AVA dataset. Compared with the SOTA CNN-based models, our EAT achieves the best performance in terms of the SRCC and PLCC, and it surpasses the previous best results of its Transformer-based counterparts by +3.4% in the PLCC, +2.8% in the SRCC and +1.8% in the ratio. Furthermore, EAT achieves significantly higher training speed, requiring -37.5% and -58.3% as many training epochs as the fastest CNN-based and Transformer-based models, respectively. The fewer training epochs and higher performance of EAT are mainly due to its sparse attention and flexibility in guiding the model to retrieve hierarchical aesthetic information.

TAD66K Dataset. Table 2 compares 13 models. Since the annotations of this dataset are more refined, with different annotation

Table 3: Our EAT and 9 baseline methods were evaluated on FLICKR-AES for personalized IAA, with performance results measured by SRCC.

Models (input size)	10 shot↑	100 shot↑
PAM [29] (224)	0.520±0.003	0.553±0.012
USAR [25] (224)	0.525±0.004	0.552±0.015
MT_IAA [19] (224)	0.523±0.004	0.582±0.014
PA_IAA [19] (224)	0.543±0.003	0.639±0.011
UG-PIAA [24] (224)	0.559±0.002	0.660±0.013
BLG-PIAA [39] (224)	0.561±0.005	0.669±0.013
TAPP-PIAA [20] (224)	0.591±0.007	0.685±0.012
PIAA-SOA [40] (?)	0.618±0.006	0.691±0.015
ResNext [13] (512)	0.612±?	0.706±?
Ours (224)	**0.593±0.005**	**0.693±0.013**
Ours (512)	**0.644±0.003**	**0.732±0.015**

evaluation criteria being adopted for each theme, it is more challenging to learn aesthetic features. Thanks to its flexible attention mechanism and data-dependent approach to specifying interest points, our EAT can extract the aesthetic information corresponding to different themes more effectively than all other methods. Our best model achieves +2.8% in the PLCC with -72.7% as many training epochs compared to the best CNN-based model and surpasses the best Transformer-based model by +3.6% in the PLCC, +3.5% in the SRCC and -62.5% as many training epochs.

FLICKR-AES Dataset. People with different personalities may observe images in different ways, which encourages models to select different levels of attention to model aesthetic information. We tested our model on the personalized aesthetic dataset FLICKR-AES. Table 3 shows that our model achieves the best SRCC of 0.689 (in the 100-shot case), surpassing the previous best SRCC results by +3.0%, which means that our model can use a smaller amount of data to learn more about individual preferences.

AADB Dataset. A comparison of our method with some PIAA approaches on the AADB dataset is presented in Table 6. As can be seen, EAT is able to achieve SOTA performance with improvements of 3.1% and 5.6%. This demonstrates that our EAT can be generalized to the PIAA tasks of real-world users and smaller datasets.

Cross-dataset Evaluations. To evaluate the generalization ability of our proposed EAT method, the cross-dataset evaluations are conducted as in previous work [20]. Table 7 shows that EAT outperforms TAPP-PIAA on all datasets, especially when trained on a small dataset with limited images.

5.3 Enhance for other Architectures

In order to demonstrate the efficacy of our EAT, and its seamless integration into existing deformable methods ***without altering entire network structure*** for mitigating attention bias, we applied EAT to two representative deformable transformers (Fig. 6), namely PSViT [38] and DAT [37]. Specifically, we enhanced these models with the EAT Block and the Enhanced Attention Module (EAM).

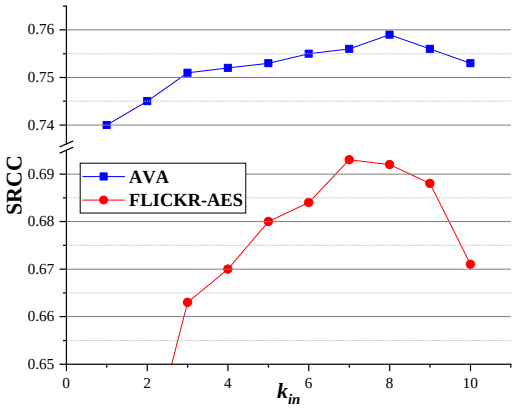

Figure 8: Ablation of different kernel sizes in the Interest Generation Module that larger k_{in} results in fewer interest points.

Table 4: Ablation of different attention mechanisms at different stages, denoted by *S, D,* and *E* for Swin, DAT, and EAT blocks respectively.

Stage 1	2	3	4	FLOPs	Params	SRCC↑
S	*S*	*S*	*S*	240G	87.0M	0.736
E	*S*	*S*	*S*	230G	86.9M	0.731
E	*E*	*S*	*S*	207G	86.9M	0.728
E	*E*	*E*	*S*	160G	86.6M	0.725
D	*D*	*D*	*D*	240G	87.0M	0.738
E	*D*	*D*	*D*	231G	86.9M	0.735
E	*E*	*D*	*D*	210G	86.8M	0.733
E	*E*	*E*	*D*	155G	86.7M	0.732
E	*E*	*E*	*E*	130G	86.7M	0.732
S	*E*	*E*	*E*	137G	86.8M	0.746
S	*S*	*E*	*E*	140G	87.0M	**0.759**
S	*S*	*S*	*E*	203G	87.0M	0.751

As shown in Table 8, the incorporation of our EAM module into PSViT results in a notable enhancement of performance on the AVA dataset, with an increase in SRCC by 0.723 (+3.1%) and an increase in PLCC by 0.707 (+9.6%), while simultaneously reducing the required training epochs by 43%. In addition, replacing the DAT block with our EAT block results in a 1.1% increase in SRCC and a 1.5% increase in PLCC, while reducing training epochs by 50%. Furthermore, our method incurs no additional computational costs. The reduction in training epochs is primarily due to the integration of our method, which effectively reduces redundancy in interest points and accelerates attention balance between background and salient regions.

Table 5: Ablation on the performance of different modules.

Models (AVA)	FLOPs	Params	SRCC↑
no Interest Generate	139.96G	86.9M	0.749
no Offsets Guide	139.96G	87.0M	0.747
no Offsets Generate+ no Offsets Guide	139.84G	86.9M	0.743
no Interest Generate+ no Offsets Generate+ no Offsets Guide	139.80G	86.8M	0.738

Table 6: Performance results for our EAT and 3 baseline methods on AADB, all pre-trained on FLICKR-AES.

Models (AADB)	SRCC↑	
	10-shot	100-shot
BA-PIAA [39]	0.450±0.001	0.513±0.005
BLG-PIAA [39]	0.497±0.003	0.545±0.007
PIAA-SOA [40]	0.509±0.003	0.557±0.007
TAPP-PIAA [20]	0.534±0.004	0.612±0.007
Ours	**0.551±0.004**	**0.646±0.005**

5.4 Ablation Study

Enhanced Attention Module. In Table 4, we replace the Enhanced Attention Module of our EAT with the shift attention mechanism of Swin transformer and deformable transformer in different stages due to their similar structures. The optimal choice is to place our attention module in the last stage, in this case, our model outperforms Swin by 3.1% in the SRCC with 41.7% FLOPs and outperforms DAT by 2.8% in the SRCC with 41.7% FLOPs.

Interest and Offset Modules. We evaluate the effectiveness of the interest module and two offset modules in Table 5. Without the assistance of the Interest Generation Module and the two offset modules (Generation and Guidance), the SRCC of our EAT is reduced by 1.3% and 2.1%, respectively. Removing the Offset Guidance Module alone will also cause an SRCC loss of 1.6%, further indicating the effectiveness of the proposed modules.

The Number of Interest Points. Here, we evaluate how the number of interest points affects the performance. When stride = 1, the number of interest points generated through convolution mainly depends on the kernel size k_{in} of the Interest Generation Module. Fig. 8 shows that insufficient or excessive points lead to a decline in model's performance; meanwhile, an increasing number of interest points will lead to an increase in computation and memory.

5.5 Visualization

Interest Points. The interest points are visualized in Fig. 7 by extracting their coordinates and corresponding attention weights as in previous works [4, 37]. The interest points are adaptively adjusted in accordance with the content, placing points around salient objects while also tending to pay more attention to background to seek hierarchical aesthetic information.

Table 7: Cross-dataset evaluations of EAT and TAPP-PIAA [20] for the 100-shot PIAA task.

Cross-dataset evaluations	Training Dataset	Testing Dataset	
		FLICKR-AES	AADB
TAPP-PIAA [20]	FLICKR-AES	0.685	0.540
	AADB	0.615	0.612
Ours	FLICKR-AES	**0.689**	**0.593**
	AADB	**0.632**	**0.646**

Table 8: Cross-architecture evaluations were conducted to enhance other deformable transformer methods, resulting in improved results on AVA.

Models	FLOPs	Params	Epochs	SRCC↑	LRCC↑
DAT [37]	264G	87M	60+	0.738	0.739
DAT+EAT Block	264G	87M	30+	**0.746**	**0.750**
PSViT [38]	264G	21M	70+	0.701	0.645
PSViT+EAM Module	264G	21M	40+	**0.723**	**0.707**

Saliency Maps. The GradCAM [9] is applied to visualize saliency maps in Fig. 3. Compared with other methods, the attention of our model not only shows a high response to the core areas of salient objects in an image but also shows some attention to the background area; thus, it is able to cover more highly aesthetics-correlated regions and is more in line with human perception.

Predictions for Images. The prediction examples are shown in the **Appendix** (Fig. 9). Similar to human cognition, EAT assigns higher scores to images that perform better in terms of important aesthetic attributes, such as composition, color, lighting, and depth of field. In contrast, in cases of incompatible color or unnatural boundaries between foreground and background, the corresponding predicted scores are usually lower. This further shows that EAT can enhance the model to capture vital information in a human manner.

6 CONCLUSIONS

This paper reveals the long-ignored influence of attention bias in IAA. To address this issue, we propose the Enhancer for Aesthetics-Oriented Transformer (EAT), which guides an IAA model to focus on more reasonable and aesthetically relevant regions. EAT achieves SOTA performance on four representative datasets with faster training speed and lower computational cost. Furthermore, we conduct a thorough analysis to examine the generalization capabilities of EAT. Comprehensive experiments demonstrate the effectiveness of our approach in enhancing Transformers for various IAA tasks. As future work, we will continue to explore the adaptation of Transformers for IAA and stimulate research with a broader perspective.

7 ACKNOWLEDGMENTS

This work was supported by the Funds for Creative Research Groups of China under Grant 61921003.

REFERENCES

[1] Emilie Baudin, François-Xavier Bucher, Laurent Chanas, and Frédéric Guichard. 2020. DXOMARK Objective Video Quality Measurements. *Electronic Imaging* 2020, 9 (2020), 166–1.

[2] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan. 2020. Adaptive Fractional Dilated Convolution Network for Image Aesthetics Assessment. In *CVPR*. 14114–14123.

[3] Zhiyang Chen, Yousong Zhu, Chaoyang Zhao, Guosheng Hu, Wei Zeng, Jinqiao Wang, and Ming Tang. 2021. Dpt: Deformable patch-based transformer for visual recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2899–2907.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 764–773.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *ECCV*. Springer, 288–301.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. Ieee, 248–255.

[7] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine* 34, 4 (2017), 80–106.

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

[9] Jacob Gildenblat and contributors. 2021. PyTorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam.

[10] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. 2022. Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks. *IJCAI* (2022).

[11] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. 2019. Effective aesthetics prediction with multi-level spatially pooled features. In *CVPR*.

[12] Jingwen Hou, Henghui Ding, Weisi Lin, Weide Liu, and Yuming Fang. 2022. Distilling knowledge from object classification to aesthetics assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 11 (2022), 7386–7402.

[13] Jingwen Hou, Weisi Lin, Guanghui Yue, Weide Liu, and Baoquan Zhao. 2022. Interaction-Matrix Based Personalized Image Aesthetics Assessment. *IEEE Transactions on Multimedia* (2022).

[14] Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nat. Rev. Neurol.* 2, 3 (2001), 194–203.

[15] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *ICCV*. 5148–5157.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*. Springer, 662–679.

[18] Bonan Li, Yinhan Hu, Xuecheng Nie, Congying Han, Xiangjian Jiang, Tiande Guo, and Luoqi Liu. 2023. DropKey. In *CVPR*.

[19] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. 2020. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE TIP* 29 (2020), 3898–3910.

[20] Yaohui Li, Yuzhe Yang, Huaxiong Li, Haoxing Chen, Liwu Xu, Leida Li, Yaqian Li, and Yandong Guo. 2022. Transductive aesthetic preference propagation for personalized image aesthetics assessment. In *Proceedings of the 30th ACM International Conference on Multimedia*. 896–904.

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

[22] Hao Lou, Heng Huang, Chaoen Xiao, and Xin Jin. 2021. Aesthetic Evaluation and Guidance for Mobile Photography. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2780–2782.

[23] Yiwen Luo and Xiaoou Tang. 2008. Photo and video quality evaluation: Focusing on the subject. In *Eur. Conf. Comput. Vis.* 386–399.

[24] Pei Lv, Jianqi Fan, Xixi Nie, Weiming Dong, Xiaoheng Jiang, Bing Zhou, Mingliang Xu, and Changsheng Xu. 2021. User-Guided Personalized Image Aesthetic Assessment based on Deep Reinforcement Learning. *arXiv:2106.07488* (2021).

[25] Pei Lv, Meng Wang, Yongbo Xu, Ze Peng, Junyi Sun, Shimei Su, Bing Zhou, and Mingliang Xu. 2018. USAR: An interactive user-specific aesthetic ranking framework for images. In *Proceedings of the 26th ACM international conference on Multimedia*. 1328–1336.

[26] Shuang Ma, Jing Liu, and Chang Wen Chen. 2017. A-Lamp: Adaptive Layout-Aware Multi-patch Deep Convolutional Neural Network for Photo Aesthetic Assessment. In *CVPR*. 722–731.

[27] Long Mai, Hailin Jin, and Feng Liu. 2016. Composition-Preserving Deep Photo Aesthetics Assessment. In *CVPR*. 497–506.

[28] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*. IEEE.

[29] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. 2017. Personalized image aesthetics. In *ICCV*. 638–647.

[30] Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. 2021. Hierarchical Layout-Aware Graph Convolutional Network for Unified Aesthetics Assessment. In *CVPR*. 8475–8484.

[31] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. 2018. Attention-based multi-patch aggregation for image aesthetic assessment. In *ACMMM*.

[32] C. Siagian and L. Itti. 2007. Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 2 (2007), 300–312.

[33] Wei-Tse Sun, Ting-Hsuan Chao, Yin-Hsi Kuo, and Winston H Hsu. 2017. Photo filter recommendation by category-aware aesthetic learning. *IEEE Transactions on Multimedia* 19, 8 (2017), 1870–1880.

[34] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural Image Assessment. *TIP* 27, 8 (2018), 3998–4011.

[35] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. 2022. MaxViT: Multi-Axis Vision Transformer. *ECCV* (2022).

[36] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 568–578.

[37] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. 2022. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4794–4803.

[38] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip HS Torr, Wayne Zhang, and Dahua Lin. 2021. Vision transformer with progressive sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 387–396.

[39] Hancheng Zhu, Leida Li, Jinjian Wu, Sicheng Zhao, Guiguang Ding, and Guangming Shi. 2020. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *TCYB* (2020).

[40] Hancheng Zhu, Yong Zhou, Leida Li, Yaqian Li, and Yandong Guo. 2021. Learning personalized image aesthetics from subjective and objective attributes. *IEEE Transactions on Multimedia* (2021).

[41] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. 2019. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9308–9316.

[42] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.

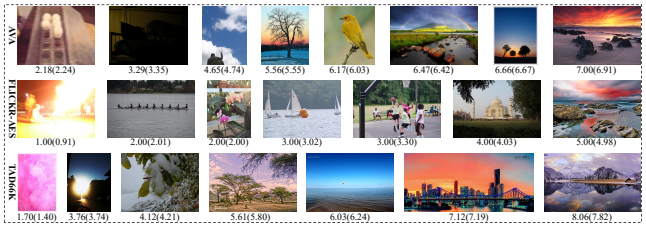

Figure 9: The images in AVA, FLICKR-AES and TAD66K datasets are visualized with their corresponding ground-truth and predicted scores displayed below.

A APPENDIX

Back-propagation Process. The differentiable components of interest points and offsets generate the sampled feature map $\bar{F}$ according to equation (10), and then $\bar{F}$ is employed to compute the output. To guarantee finding high-quality interest points, each offset i is dynamically refined during the back-propagation process, with the aim of minimizing the predicted loss at each training step t:

$$\frac{\partial(\bar{F}_t^i)}{\partial(\bar{O}_{t-1}^i)} = \frac{\partial B[F_{t-1}^i, \bar{P}_{t-1}^i]}{\partial(\bar{O}_{t-1}^i)} = \sum_q \frac{\partial B[q, \bar{P}_{t-1}^i]}{\partial(\bar{O}_{t-1}^i)} F(q), \tag{11}$$

where $\bar{P}_{t-1}^i = P_{t-1}^i + \bar{O}_{t-1}^i$, and q is non-zero on only the 4 integral points in F closest to $\bar{P}$ for bilinear interpolation. This process provides us with a differentiable offset mechanism that enables the loss gradients to flow back to the interest points.

Computational Complexity. The computational cost of our aesthetic multi-head self-attention (A-MSA) can be calculated as:

$$\Omega(A-MSA) = \underbrace{HWC^2 + 2(H_{in}W_{in})^2C + 3H_{in}W_{in}C^2}_{\text{self-attention operation}} + \underbrace{(k_{of}^2+3)H_{in}W_{in}C}_{\text{offset module}} + \underbrace{k_{in}^2 H_{in}W_{in}C}_{\text{interest module}}. \tag{12}$$

Although the offset and the interest modules cost additional linear complexity, most of the computational cost is controlled by H_{in} and W_{in}, rather than by H and W, while the former is much smaller than the latter. For example, when H_{in}=W_{in}=7, H=W=14, and k_{of}=k_{in}=8, A-MSA saves approximately 40% of the computational cost compared to the classical works [21, 37].

Failure Cases. In our study, we have observed two types of failure cases. Firstly, the EAT model shows lower prediction accuracy on certain extreme samples with ambiguous meaning. For instance, an image ("118003.jpg" in AVA) that is completely black is assigned a ground-truth score of 3.15, whereas our EAT model assigns a lower score of 1.3. Secondly, within the "art" theme category in the TAD66K dataset, our EAT model performs 20% worse than the average performance indicated by the SRCC metric (0.413 vs. 0.517).