# ELTA: An Enhancer against Long-Tail for Aesthetics-oriented Models

ICML24

Limin Liu*, Shuai He*, Anlong Ming*, Rui Xie, Huadong Ma
Beijing University of Posts and Telecommunications

## Introduction

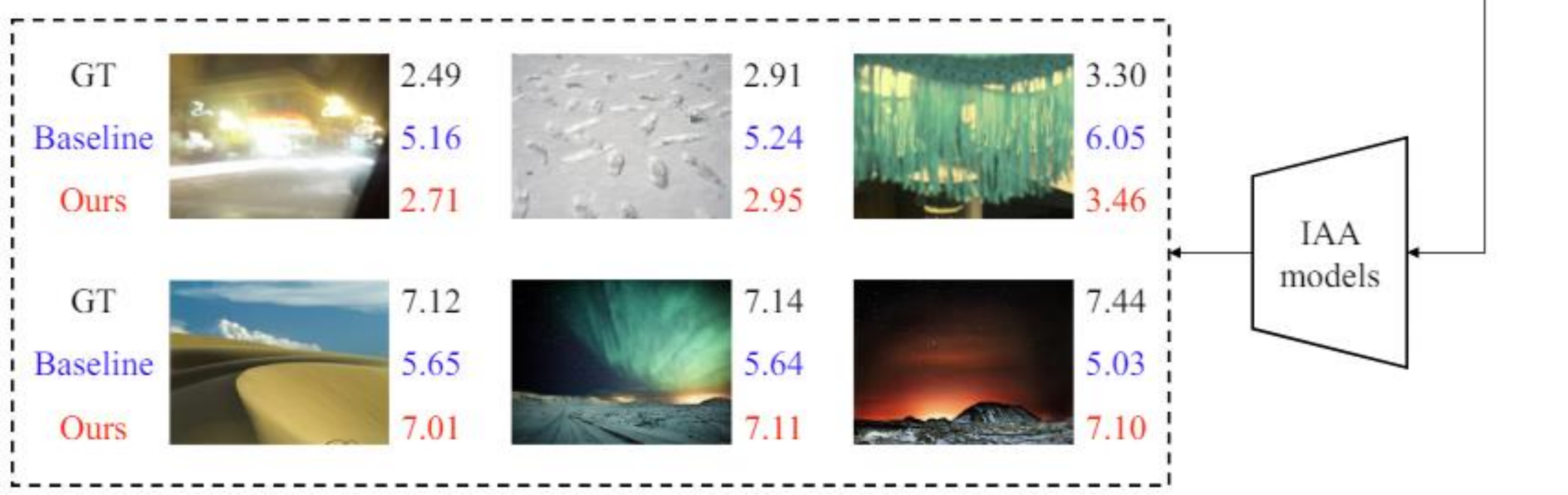
### ☞ *What is image aesthetics assessment?*

· Image aesthetics assessment (IAA) aims to assess image aesthetics based on human perception.

### ☞ *What is long-tailed IAA dataset?*

· Most of the images are located in the medium scores (e.g., 4-6), with few samples of high and low scores.

| Dataset | Ground-truth Range | | | | | | | | Distribution |
|---|---|---|---|---|---|---|---|---|---|
| | Minority | | | Majority | | | Minority | | |
| | [0, 2) | [2, 3) | [3, 4) | [4, 5) | [5, 6) | [6, 7) | [7, 8) | [8, 10] | |
| AVA | 6 | 491 | 7K | 60K | 116K | 43K | 3K | 46 | |
| AADB | 506 | 826 | 1K | 2K | 880 | 2K | 967 | 640 | |
| TAD66K | 180 | 2K | 7K | 11K | 12K | 11K | 7K | 1K | |
| PARA | 0 | 81 | 1K | 2K | 8K | 13K | 3K | 114 | |




### ☞ *Why the long-tail issue should be addressed?*

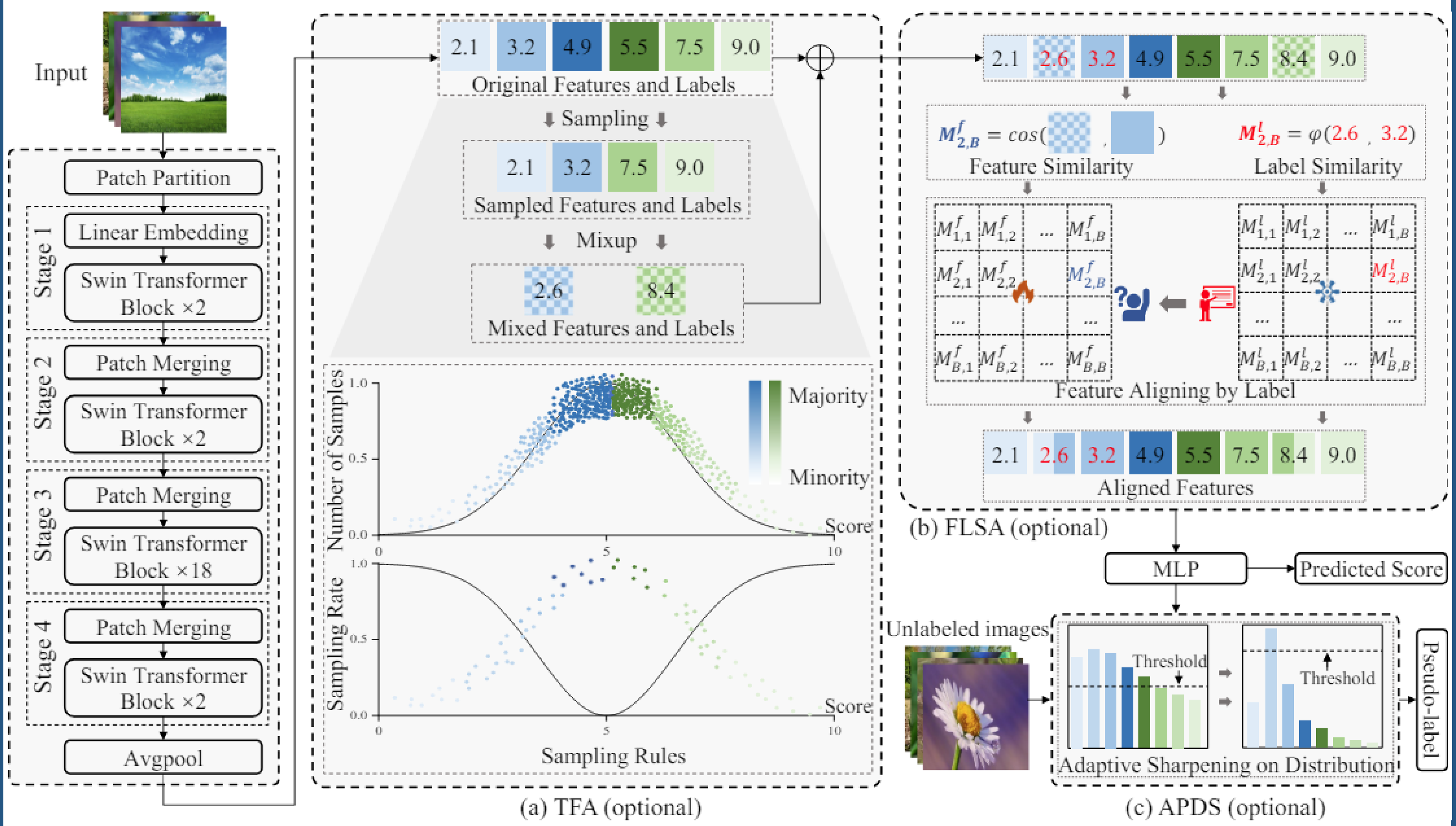
· Models trained on long-tailed data are biased towards majority and away from minority.

· The bias compromises the models' generalizability and fairness, resulting in low accuracy and insufficient differentiation in model scores.

### ☞ *Contributions*

- ***Long-tail in IAA***: This is the first solution proposed against long-tail for IAA models.
- ***ELTA***: Mitigates the data imbalance by augmenting minority features, aligning features to labels, and improving pseudo-labeling accuracy.

## ELTA




$$P(i) = \frac{\exp(|\bar{s} - s_i|/\tau_1)}{\sum_{k=1}^{B}\exp(|\bar{s} - s_k|/\tau_1)},$$

$$P(j|i) = \frac{\exp(\tau_2/|s_i - s_j|)}{\sum_{k=1,2,\cdots,i-1,i+1,\cdots,B}\exp(\tau_2/|s_i - s_k|)},$$

$$\lambda = \frac{P(i)}{P(i)+P(j)} = \frac{\exp(|\bar{s} - s_i|/\tau_1)}{\exp(|\bar{s} - s_i|/\tau_1) + \exp(|\bar{s} - s_j|/\tau_1)}$$

$$\tilde{\boldsymbol{z}}^k = \lambda \boldsymbol{z}_i^k + (1-\lambda)\boldsymbol{z}_j^k,$$
$$\tilde{\boldsymbol{y}} = \lambda \boldsymbol{y}_i + (1-\lambda)\boldsymbol{y}_j.$$

***Probabilistic Sampling Strategy for Mixup***

$$\tau_i(\beta) = e^{-\beta|\hat{s}_i - \bar{s}|},$$

$$\hat{\boldsymbol{y}}_i = \frac{e^{\boldsymbol{z}_i/\tau_i(\beta)}}{\sum_{j=1}^{B} e^{\boldsymbol{z}_j/\tau_j(\beta)}},$$

$$\beta^*, t^* = \arg\min_{\beta,t} \sum_{i=1}^{M} |\hat{s}_i(\beta,t) - s_i|$$

$$D' = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) | \max(\hat{\boldsymbol{y}}_i(\beta)) \geq t\}_{i=1}^{M}$$

***Adaptive Probability Distribution Sharpening***

## Experiment Results

| Dataset | Metric | CNN-based models | | | | Transformer-based models | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | NIMA | HGCN | BIAA | TANet | MaxViT | MUSIQ | EAT | Base. | Ours |
| AVA | $\mathcal{P}\uparrow$ | 0.636 | 0.687 | 0.668 | 0.765 | 0.745 | 0.738 | 0.770 | 0.743 | **0.777** |
| | $\mathcal{S}\uparrow$ | 0.612 | 0.665 | 0.651 | 0.758 | 0.708 | 0.726 | 0.759 | 0.735 | **0.764** |
| | $\mathcal{L}\downarrow$ | 0.655 | 0.675 | 0.653 | 0.630 | 0.600 | 0.647 | 0.490 | 0.616 | **0.438** |
| | $\mathcal{M}\downarrow$ | 0.322 | 0.321 | 0.382 | **0.237** | 0.317 | 0.305 | 0.313 | 0.295 | 0.302 |
| | $\mathcal{H}\downarrow$ | 0.648 | 0.660 | 0.568 | 0.729 | 0.531 | 0.628 | 0.433 | 0.513 | **0.426** |
| AADB | $\mathcal{P}\uparrow$ | 0.711 | 0.734 | 0.733 | 0.742 | 0.748 | 0.761 | 0.767 | 0.740 | **0.772** |
| | $\mathcal{S}\uparrow$ | 0.700 | 0.716 | 0.710 | 0.749 | 0.742 | 0.751 | 0.759 | 0.732 | **0.760** |
| | $\mathcal{L}\downarrow$ | 1.450 | 1.453 | 1.508 | 1.394 | 1.592 | 1.447 | 1.375 | 1.526 | **1.289** |
| | $\mathcal{M}\downarrow$ | 0.874 | 0.989 | 0.897 | 0.846 | **0.782** | 0.880 | 0.828 | 0.896 | 0.905 |
| | $\mathcal{H}\downarrow$ | 1.527 | 1.299 | 1.423 | 1.355 | 1.461 | 1.159 | 1.260 | 1.402 | **1.141** |
| TAD66K | $\mathcal{P}\uparrow$ | 0.405 | 0.493 | 0.431 | 0.531 | 0.513 | 0.517 | **0.546** | 0.507 | 0.539 |
| | $\mathcal{S}\uparrow$ | 0.390 | 0.486 | 0.417 | 0.513 | 0.484 | 0.489 | **0.517** | 0.478 | 0.496 |
| | $\mathcal{L}\downarrow$ | 1.851 | 1.808 | 1.734 | 1.598 | 1.570 | 1.627 | 1.591 | 1.621 | **1.457** |
| | $\mathcal{M}\downarrow$ | 0.812 | 0.780 | 0.876 | **0.682** | 0.746 | 0.728 | 0.782 | 0.793 | 0.812 |
| | $\mathcal{H}\downarrow$ | 1.690 | 1.370 | 1.669 | 1.651 | 1.402 | 1.460 | 1.175 | 1.354 | **1.162** |
| PARA | $\mathcal{P}\uparrow$ | 0.862 | 0.881 | 0.886 | 0.899 | 0.936 | 0.918 | 0.940 | 0.925 | **0.943** |
| | $\mathcal{S}\uparrow$ | 0.877 | 0.865 | 0.858 | 0.887 | 0.902 | 0.899 | 0.909 | 0.897 | **0.912** |
| | $\mathcal{L}\downarrow$ | 0.616 | 0.573 | 0.469 | 0.551 | 0.383 | 0.572 | 0.336 | 0.402 | **0.327** |
| | $\mathcal{M}\downarrow$ | 0.344 | 0.290 | 0.328 | 0.299 | 0.282 | 0.315 | 0.276 | 0.314 | **0.251** |
| | $\mathcal{H}\downarrow$ | 0.486 | 0.502 | 0.503 | 0.429 | 0.276 | 0.424 | **0.256** | 0.379 | 0.290 |

(a)

| Dataset | Metric | Model | | | |
|---|---|---|---|---|---|
| | | L&FDS | RankSim | Base. | Ours |
| AVA | $\mathcal{P}\uparrow$ | 0.752 | 0.759 | 0.743 | **0.777** |
| | $\mathcal{S}\uparrow$ | 0.740 | 0.753 | 0.735 | **0.764** |
| | $\mathcal{L}\downarrow$ | 0.588 | 0.542 | 0.616 | **0.438** |
| | $\mathcal{M}\downarrow$ | 0.303 | 0.297 | **0.295** | 0.302 |
| | $\mathcal{H}\downarrow$ | 0.509 | 0.485 | 0.513 | **0.426** |
| AADB | $\mathcal{P}\uparrow$ | 0.746 | 0.750 | 0.740 | **0.772** |
| | $\mathcal{S}\uparrow$ | 0.735 | 0.738 | 0.732 | **0.760** |
| | $\mathcal{L}\downarrow$ | 1.473 | 1.537 | 1.526 | **1.289** |
| | $\mathcal{M}\downarrow$ | 0.888 | **0.881** | 0.896 | 0.905 |
| | $\mathcal{H}\downarrow$ | 1.407 | 1.295 | 1.402 | **1.141** |
| TAD66K | $\mathcal{P}\uparrow$ | 0.509 | 0.514 | 0.507 | **0.539** |
| | $\mathcal{S}\uparrow$ | 0.483 | 0.487 | 0.478 | **0.496** |
| | $\mathcal{L}\downarrow$ | 1.611 | 1.560 | 1.621 | **1.457** |
| | $\mathcal{M}\downarrow$ | **0.787** | 0.796 | 0.793 | 0.812 |
| | $\mathcal{H}\downarrow$ | 1.369 | 1.343 | 1.354 | **1.162** |
| PARA | $\mathcal{P}\uparrow$ | 0.933 | 0.930 | 0.925 | **0.943** |
| | $\mathcal{S}\uparrow$ | 0.904 | 0.906 | 0.897 | **0.912** |
| | $\mathcal{L}\downarrow$ | 0.371 | 0.363 | 0.402 | **0.327** |
| | $\mathcal{M}\downarrow$ | 0.301 | 0.322 | 0.314 | **0.251** |
| | $\mathcal{H}\downarrow$ | 0.360 | 0.345 | 0.379 | **0.290** |

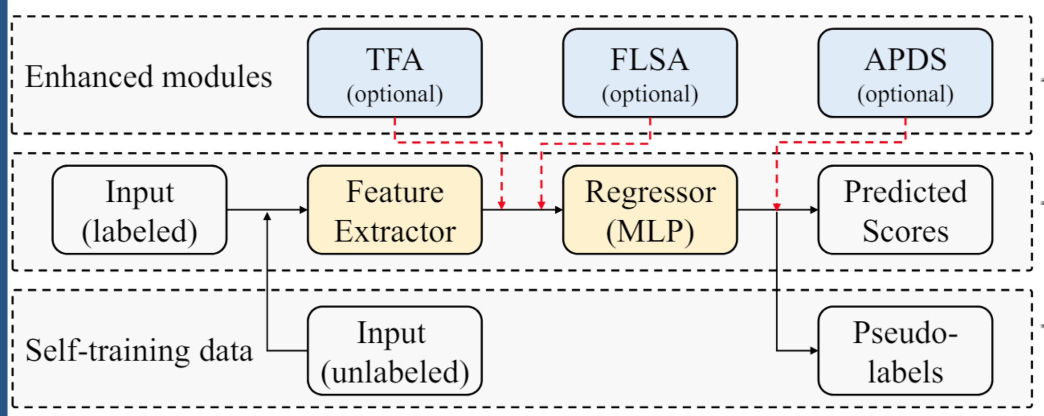
(b)

| Module | | | Metric | | | | |
|---|---|---|---|---|---|---|---|
| TFA | FLSA | APDS | $\mathcal{P}\uparrow$ | $\mathcal{S}\uparrow$ | $\mathcal{L}\downarrow$ | $\mathcal{M}\downarrow$ | $\mathcal{H}\downarrow$ |
| | | | 0.743 | 0.735 | 0.616 | 0.295 | 0.513 |
| ✓ | | | 0.749 | 0.740 | 0.518 | 0.323 | 0.474 |
| | ✓ | | 0.758 | 0.752 | 0.592 | 0.287 | 0.520 |
| | | ✓ | 0.760 | 0.749 | 0.575 | 0.286 | 0.480 |
| ✓ | ✓ | | 0.764 | 0.748 | 0.484 | 0.324 | 0.431 |
| ✓ | | ✓ | 0.762 | 0.756 | 0.461 | 0.314 | 0.445 |
| | ✓ | ✓ | 0.771 | 0.757 | 0.547 | **0.286** | 0.463 |
| ✓ | ✓ | ✓ | **0.777** | **0.764** | **0.438** | 0.302 | **0.426** |

(c)

(a) Comparing ELTA with 7 IAA methods on four datasets.
(b) Comparing ELTA with 2 Deep Imbalanced Regression (DIR) methods on four datasets.
(c) Ablation of different modules on the AVA dataset.

## Enhancing other Methods

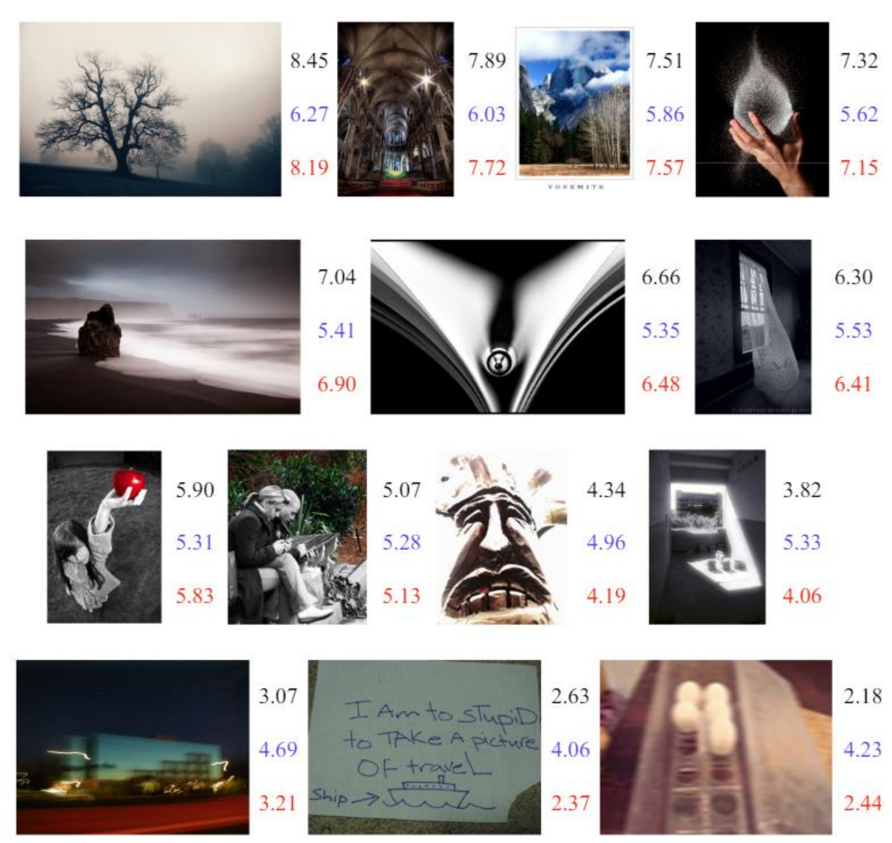



Above: Example of plugging our modules into other methods. The three modules proposed are all optional and can be easily integrated with other methods.

Right: Cross-architecture evaluations are conducted to enhance other IAA methods, resulting in improved results on AVA.

| Model | + Module | | | Metric | | | | |
|---|---|---|---|---|---|---|---|---|
| | TFA | FLSA | APDS | $\mathcal{P}\uparrow$ | $\mathcal{S}\uparrow$ | $\mathcal{L}\downarrow$ | $\mathcal{M}\downarrow$ | $\mathcal{H}\downarrow$ |
| NIMA | | | | 0.636 | 0.612 | 0.655 | **0.322** | 0.648 |
| | ✓ | ✓ | | 0.649 | 0.631 | 0.614 | 0.338 | 0.628 |
| | ✓ | ✓ | ✓ | **0.658** | **0.640** | **0.593** | 0.340 | **0.601** |
| HGCN | | | | 0.687 | 0.665 | 0.675 | **0.321** | 0.660 |
| | ✓ | ✓ | | 0.700 | 0.675 | 0.582 | 0.343 | 0.593 |
| | ✓ | ✓ | ✓ | **0.714** | **0.693** | **0.564** | 0.329 | **0.575** |
| BIAA | | | | 0.668 | 0.651 | 0.653 | 0.382 | 0.568 |
| | ✓ | ✓ | | 0.682 | 0.675 | 0.596 | 0.390 | 0.514 |
| | ✓ | ✓ | ✓ | **0.699** | **0.687** | **0.544** | **0.381** | **0.497** |
| MUSIQ | | | | 0.738 | 0.726 | 0.647 | **0.305** | 0.628 |
| | ✓ | ✓ | | 0.748 | 0.734 | 0.603 | 0.311 | 0.562 |
| | ✓ | ✓ | ✓ | **0.761** | **0.745** | **0.588** | 0.327 | **0.482** |
| MaxViT | | | | 0.745 | 0.708 | 0.600 | **0.317** | 0.531 |
| | ✓ | ✓ | | 0.750 | 0.728 | 0.557 | 0.340 | **0.426** |
| | ✓ | ✓ | ✓ | **0.759** | **0.742** | **0.532** | 0.333 | 0.428 |
| TANet | | | | 0.765 | 0.758 | 0.630 | **0.237** | 0.729 |
| | ✓ | ✓ | | 0.772 | 0.767 | 0.591 | 0.244 | 0.633 |
| | ✓ | ✓ | ✓ | **0.779** | **0.771** | **0.562** | 0.251 | **0.585** |
| EAT | | | | 0.770 | 0.759 | 0.490 | 0.313 | 0.433 |
| | ✓ | ✓ | | 0.777 | 0.765 | **0.450** | 0.310 | 0.426 |
| | ✓ | ✓ | ✓ | **0.780** | **0.768** | **0.450** | **0.307** | **0.413** |

## Evaluation Samples




The Baseline model, marked in blue, struggles to effectively differentiate the aesthetic quality of images. Typically, its evaluation results cluster within a 4-6.5 point range, which means significant evaluation errors for images with high or low aesthetic values. In contrast, our proposed ELTA model, indicated in red, demonstrates improved performance. The result is closer to ground truth.