Aesthetic Visual Question Answering of Photographs

AIART submission 133

Supplementary Material

A Distribution in VQA Dataset

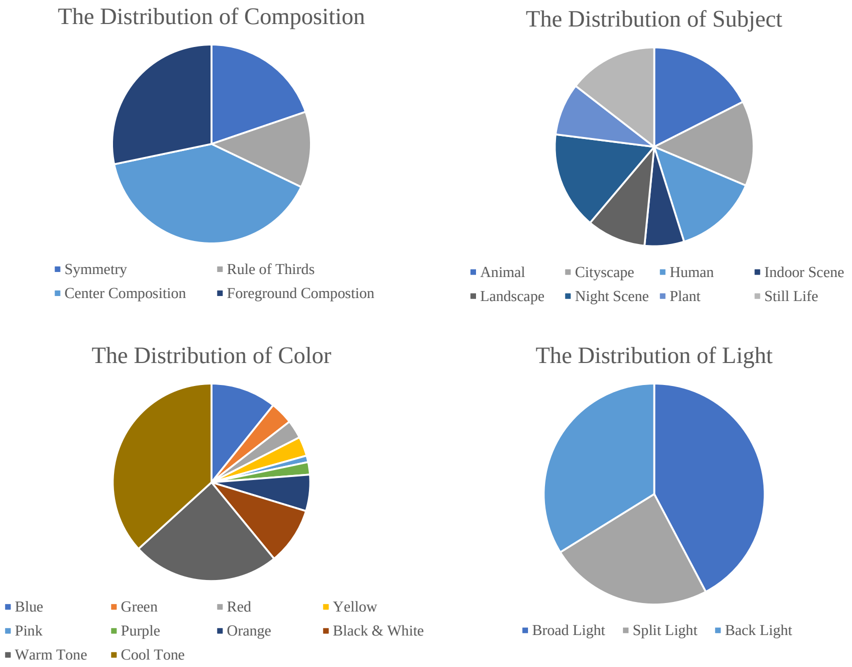


Fig. 1. The distribution of the basic aesthetic labels in the AesVQA dataset.

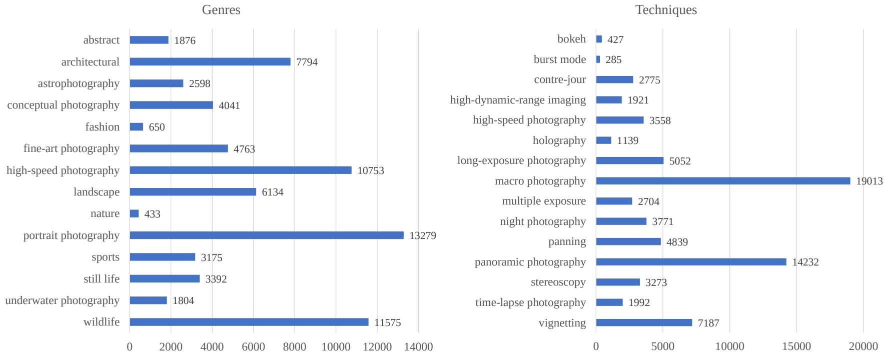
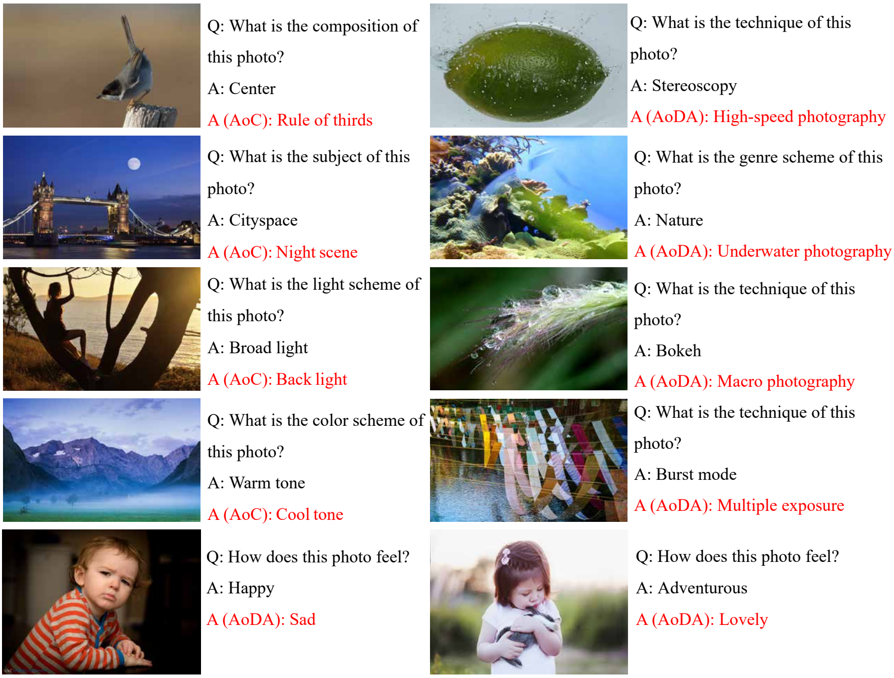


Fig. 2. The distribution of the genres labels in the AesVQA dataset.

B Subjective Labels and Photography Component



Q: What is the composition of this photo?
A: Center
A (AoC): Rule of thirds

Q: What is the technique of this photo?
A: Stereoscopy
A (AoDA): High-speed photography

Q: What is the subject of this photo?
A: Cityspace
A (AoC): Night scene

Q: What is the genre scheme of this photo?
A: Nature
A (AoDA): Underwater photography

Q: What is the light scheme of this photo?
A: Broad light
A (AoC): Back light

Q: What is the technique of this photo?
A: Bokeh
A (AoDA): Macro photography

Q: What is the color scheme of this photo?
A: Warm tone
A (AoC): Cool tone

Q: What is the technique of this photo?
A: Burst mode
A (AoDA): Multiple exposure

Q: How does this photo feel?
A: Happy
A (AoDA): Sad

Q: How does this photo feel?
A: Adventurous
A (AoDA): Lovely

Fig. 3. Question answers which before and after the adjustment of image's confidence and the adjustment of the distribution of answers.The **AoC** means the adjustment of image's classify confidence. The **AoDA** means the adjustment of distribution of answers. These two tricks will help the model achieving a higher accuracy.

C Adjustment of Confidence

Table 1. By randomly extracting 1000 pictures from each category, the confidence adjustment operation needed to be judged. The following table describes the pre-adjusted and post-adjusted range of each sub-category in the categories of genres and techniques:

Genres	Before Adjustment	After Adjustment
Abstract	0.536	0.536
Architectural	0.912	0.615
Astrophotography	0.990	0.653
Conceptual photography	0.760	0.760
Fashion	0.450	0.450
Fine-art photography	0.784	0.635
High-speed photography	0.856	0.654
Landscape	0.750	0.653
Nature	0.803	0.664
Portrait photography	0.869	0.685
Sports	0.427	0.427
Still life	0.752	0.645
Underwater photography	0.416	0.416
Wildlife	0.947	0.703
Techniques	Before Adjustment	After Adjustment
Bokeh	0.514	0.514
Burst mode	0.272	0.452
Contre-jour	0.634	0.634
High-dynamic-range imaging	0.595	0.595
Holography	0.402	0.402
Long-exposure photography	0.308	0.502
Macro photography	0.869	0.652
Multiple exposure	0.894	0.675
Night photography	0.613	0.613
Panning	0.597	0.597
Panoramic photography	0.749	0.678
Stereoscopy	0.875	0.658
Time-l apse photography	0.605	0.605
Vignetting	0.634	0.634

D Adjustment of Distribution of Answers

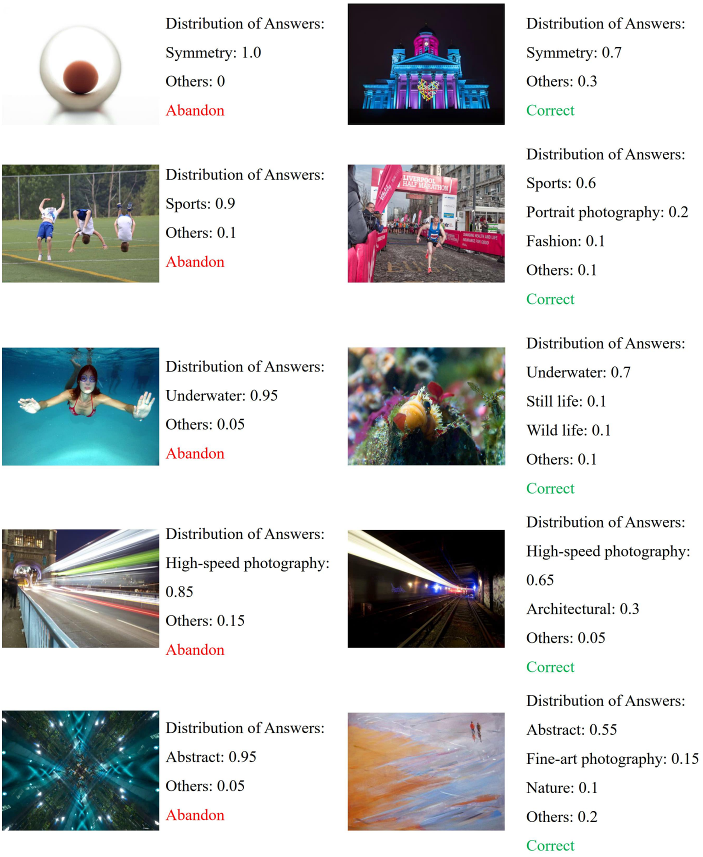


Fig. 4. The pictures on the left show that these pictures have more biased attributes, which can easily cause the model to fall into overfitting; the pictures on the right have appropriate "controversial" labels, which can allow the model to obtain better training results.

E Experiments

Table 2. The **AoC** in the table means the data with adjustment of image's class confidences, and the **AoDA** in the table means the adjustment of the distribution of answers for each question. "M1" means method 1, it is the LXMERT, the baseline model. "M2" means method 2, it is the Visual BERT. "M3" means method 3, it is the UNITER, the state-of-the art model in VQA. The training set in the AesVQA database was 58,168 images, and the test set and validation set were both 7,000 images, which were randomly mixed and assigned.

	M1	M1&AoC	M1&AoDA	M2	M2&AoC	M2&AoDA	M3	M3&AoC	M3&AoDA
Composition	53.4%	54.6%	54.8%	55.6%	56.3%	56.5%	57.6%	59.2%	**60.3%**
Color	55.3%	57.6%	58.8%	57.2%	59.5%	60.2%	58.2%	58.5%	**60.5%**
Light	70.6%	73.5%	67.3%	72.5%	**75.8%**	69.0%	71.5%	72.5%	67.8%
Subject	45.9%	50.2%	55.5%	48.3%	53.2%	57.1%	49.5%	53.9%	**58.0%**
Genres	48.5%	56.6%	54.7%	50.3%	58.5%	56.5%	52.2%	**61.2%**	60.7%
Techniques	42.4%	49.9%	50.3%	45.9%	51.2%	53.0%	47.2%	53.6%	**54.9%**
Subject	51.8%	52.6%	53.5%	53.9%	54.4%	55.8%	55.6%	57.8%	**59.4%**
All	57.7%	58.6%	59.3%	59.8%	60.7%	62.2%	60.3%	61.4%	**61.9%**