

# Image Composition Assessment with Saliency-augmented Multi-pattern Pooling

## Supplementary Material

Bo Zhang  
bo-zhang@sjtu.edu.cn

Li Niu\*  
ustcnewly@sjtu.edu.cn

Liqing Zhang  
zhang-lq@cs.sjtu.edu.cn

MoE Key Lab of Artificial Intelligence  
Shanghai Jiao Tong University  
Shanghai, China

In this document, we provide additional materials to supplement our main submission. We first present more details about constructing our Composition Assessment DataBase (CADB) in Section 1. Then, we describe the detailed consistency analysis of the collected composition scores in Section 2, which verifies that our composition quality annotations are reliable for scientific research. Next, we use some examples to illustrate the content bias in the CADB dataset in Section 3. Meanwhile, in Section 4, we describe the proposed weighted EMD loss and study the effect of using weighted EMD loss to mitigate the content bias. Besides, more implementation details of the proposed method are provided in Section 5. In Section 6, Section 7, Section 8, Section 9, and Section 10, experiments on the hyper-parameter, training set size, backbone, each composition pattern, and using more composition patterns further prove the effectiveness of our method. Then we compare the performance of our method and human raters in Section 11. Finally, in Section 12, we provide additional visualization results on images inside/outside our CADB dataset.

## 1 Our CADB Dataset

### 1.1 Data Collection

Recently, many large-scale aesthetic assessment datasets have been created to facilitate research on image aesthetic evaluation, like Aesthetic Visual Analysis database (AVA) [1], Aesthetics and Attributes DataBase (AADB) [2], Photo Critique Captioning Dataset (PCCD) [3], AVA-Comments [4], AVA-Reviews [5], FLICKER-AES [6], and DPC-Captions [8]. Therefore, we can build the CADB dataset upon those existing datasets. Table 1 provides a summary comparison of CADB to other related aesthetic datasets. Here we select real photos to construct our dataset, because we target at the real-world application of composition assessment. To the best of our knowledge, among them, only the images in AADB and PCCD datasets are all real photos, while the images in other datasets (*e.g.*, AVA dataset) may be heavily edited or synthetic. Besides, PCCD dataset contains 4,235 images downloaded from a professional photo critique website, most of which are taken by professional

\*Corresponding author.

Dataset	Images	All Real Photo	Composition Score	Raters
AVA [13]	255,530	N	N	-
PCCD [14]	4,235	Y	Y	1
AADB [15]	10,000	Y	N	-
CADB(Ours)	9,497	Y	Y	5

Table 1: Comparison with existing aesthetic assessment datasets. Our CADB dataset is built upon the AADB dataset, taking into account its all real-world images, and more balanced distribution of professional and amateurish photos [15].

photographers and have relatively high composition quality. Differently, AADB dataset provides 10,000 images and contains a much more unbiased distribution of professional photos and amateurish photos. So we choose to construct our CADB dataset based on the AADB dataset.

## 1.2 Guidelines for Image Composition Evaluation

In this section, we show the annotation guidelines for evaluating the quality of image composition, aiming to make the annotation consistent across five individual raters who specialize in fine art. 1) We provide a composition rating scale from 1 to 5, where a larger score indicates better composition. 2) To help raters quickly learn the rules of image composition rating, we provide them with 100 images with high composition quality and 100 images with low composition quality selected from PCCD dataset [14] to serve as examples. 3) When assessing image composition quality, the photographic rules that should be considered are including but not limited to: *rule of thirds*, *centred composition*, *symmetry*, *repetition*, *shallow depth of field*, *diagonals*, *triangles*, *golden ratio*, *frame within the frame*, *leading lines*, *fill the frame*, *isolate the subject*, *vanishing point*, *juxtaposition*, *balancing elements*, and *object emphasis*. In addition, we draw a  $3 \times 3$  dotted grid on each image as auxiliary lines that divide the image into nine equal rectangles, which is displayed for the raters together with the original image. 4) The raters are requested to complete the composition rating independently, and the rating procedure for each single image should not be shorter than 20 seconds.

Besides, the same five raters annotate all images to mitigate the inconsistency across different raters. Following [15, 16], we average the composition scores of the five raters as the ground-truth composition mean score for each image.

## 1.3 Annotation Examples and Statistics

In Figure 1, we present some examples in our CADB dataset with five composition scores and composition mean score that is obtained by averaging those composition scores for each image. For better visualization, we divide these examples into three groups according to the composition mean score: images with high, low, medium scores. From Figure 1, we can roughly verify the validness of the composition annotations.

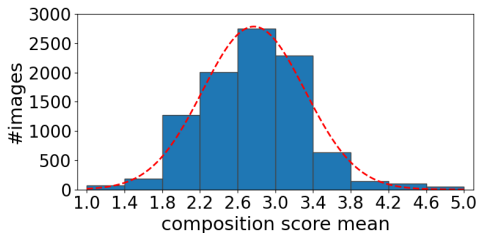
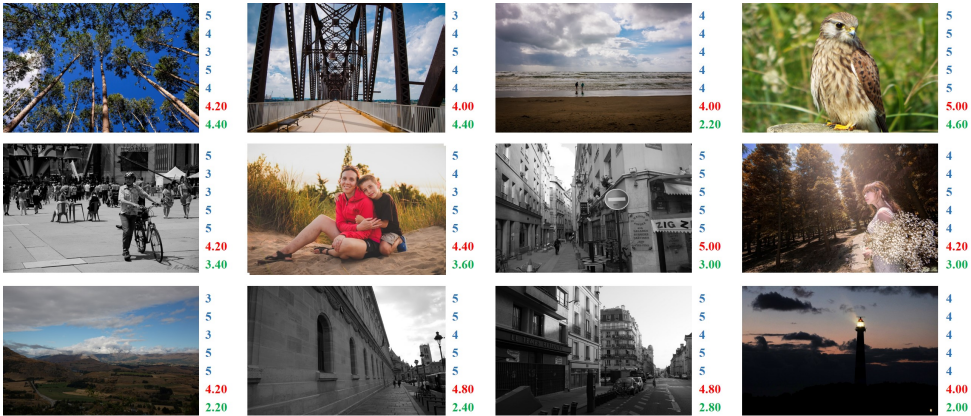


Figure 2: The distribution of composition mean score in our CADB dataset. The dashed line indicates the fitted Gaussian distribution.



(a) Examples with high composition mean scores



(b) Examples with low composition mean scores



(c) Examples with medium composition mean scores

Figure 1: Some example images in our CADB dataset with high/low/medium composition mean scores. We show five composition scores ranging from 1 to 5 provided by five raters in blue and the calculated composition mean score in red. We also show the aesthetic scores annotated by AADB dataset [10] on a scale from 1 to 5 in green.

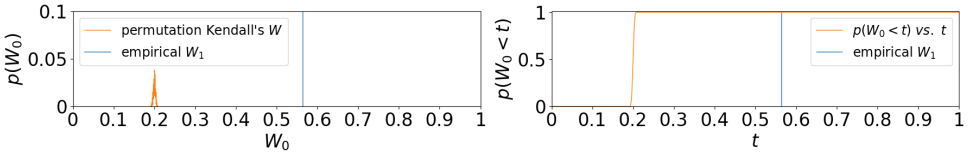


Figure 3: Permutation test on Kendall’s  $W$ . Left:  $p(W)$  vs.  $W$ . Right:  $p(W < t)$  vs.  $t$ .

In Figure 1, we also present the aesthetic scores annotated by AADB dataset [16], which is also rated by multiple individual human raters on a scale from 1 to 5 for the overall aesthetic quality, a larger score indicates higher aesthetic quality. We can see that, for some images, a high (*resp.*, low) composition score does not mean a high (*resp.*, low) aesthetic score. This is because that the composition assessment focuses on analyzing the placement of visual elements in the image, while aesthetic evaluation quantifies the aesthetic quality of the image in a comprehensive manner by taking not only image composition but also other visual factors (*e.g.*, interesting content, good lighting, color harmony, vivid color, motion blur, and shallow depth of field) into consideration. The essential difference between the above tasks further sheds light on the significance of specially developing methods for evaluating overall composition quality.

Furthermore, we calculate the distribution of composition mean score in Figure 2, where indicates that the scores are well fit by Gaussian distribution similar to the observation in AADB dataset [16] and AVA dataset [13]. As shown in Figure 2, the average and variance of composition mean score is 2.70 and 0.35, respectively.

## 2 Consistency Analysis of Annotations

Considering the subjective nature of human aesthetic activity [6, 16, 19], we carry out consistency analysis on the composition scores provided by multiple raters to verify that our CADB dataset is qualified for scientific evaluation. Following [16], we employ both Kendall’s concordance coefficient (also known as Kendall’s  $W$ ) and Spearman’s rank correlation coefficient (also known as Spearman’s  $\rho$ ) in the experiments. Kendall’s  $W$  indicates the agreement among multiple raters and accounts for tied ranks, the value of which varies from 0 (no agreement) to 1 (complete agreement). Spearman’s  $\rho$  is computed between the predicted and ground-truth composition score distribution to measure their closeness.

Since five raters annotated a collection of 10,200 images (including 240 sanity check images), we calculate an average Kendall’s  $W$  of 0.5734 over all images, which demonstrates significant consistency among different raters. Then, following [16], we conduct a permutation test over global Kendall’s  $W$  to obtain the distribution of  $W$  under the null hypothesis, the curves of which  $p(W)$  vs.  $W$  and  $p(W < t)$  vs.  $t$  are illustrated in Figure 3. We can observe that the empirical Kendall’s  $W$  on our CADB dataset is statistically significant from both curves.

Then, similar to [16], we investigate the consistency of composition scores at batch level and randomly split all annotated samples into multiple batches with each batch containing 100 images. For each batch, we calculate Kendall’s  $W$  to evaluate the consistency of annotations provided by different raters and confirm its statistical significance by using Benjamini-Hochberg procedure [1] controlling the false discovery rate (FDR) for multiple comparisons. At FDR level  $Q = 0.05$ , 100.0% batches have significant agreement, which means that all



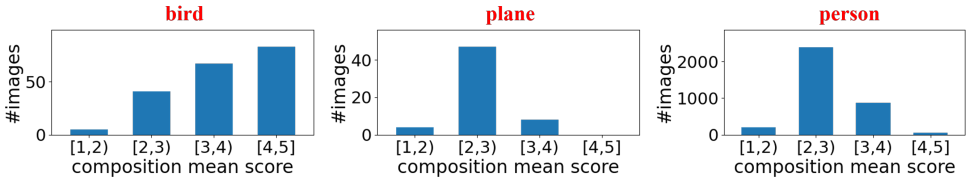


Figure 4: Illustration of biased categories in the CADB dataset. For each category (colored red), given images containing this object category, we count the occurrence of composition mean score in four score bins.

batches of our annotations have consistent composition scores, and our dataset is qualified for scientific evaluation.

Moreover, we also adopt Spearman’s  $\rho$  to measure the rank correlation between the composition scores of pairwise raters and test its statistical significance at batch level via Benjamini-Hochberg procedure. The  $p$ -value for each batch is computed by averaging the pairwise  $p$ -values in the current batch following [14]. At FDR level  $Q = 0.05$ , 98.04% batches have significant agreement, which further confirms the reliability of the composition quality annotations in our CADB dataset.

### 3 Content Bias

In Section 3 of the main text, we briefly mention the content bias issue in our CADB dataset. Here we provide a detailed description of this concept. Intuitively, photos of any object category have chances to be of high or low composition quality, which means that composition mean score  $\bar{y}$  should be approximately evenly distributed within each category. However, in our CADB dataset, as illustrated in Figure 4, we observe that there are some categories whose score distributions are concentrated in a very narrow interval, and we refer to these categories as biased categories. For example, as shown in in Figure 4, most bird photos are rated with high scores, probably because bird photos are more likely to be taken by professional photographers rather than amateurs. In this case, the network may find a shortcut to simply rate images based on their contents, which is dubbed as content bias in this paper.

To identify the biased categories, we first leverage Faster R-CNN [15] trained on Visual Genome [16] to detect objects for all images. We divide the range of composition mean score  $\bar{y}$  into  $M$  bins ( $M = 4$ ) with the bin size equal to 1 (e.g.,  $[1, 2)$ ). For the images containing each object category, we count the occurrence of  $\bar{y}$  in each bin and derive the score distribution over  $M$  bins. To measure the degree of bias, we compute the entropy of the score distribution for all categories. The category with an entropy below 0.1 is treated as a highly biased category whose associated images will be removed from our dataset. After this step, there are 9,497 images left. Then, we calculate the ratio of the maximum occurrence to the minimum non-zero occurrence in the  $M$  bins as  $r_c$  for the  $c$ -th category. A category is defined as an unbiased category if  $r_c \leq 1.5$  and otherwise a biased category. Furthermore, an image is defined as an unbiased image if its involving categories are all unbiased categories and otherwise a biased image.

For the test images in real-world applications, photos of any object category have chances to be of high or low composition quality. For better evaluation, we select 950 unbiased images to form the test set, which is closer to the test set in practice, and use the remaining 8,547 images (including both unbiased and biased ones) for training.

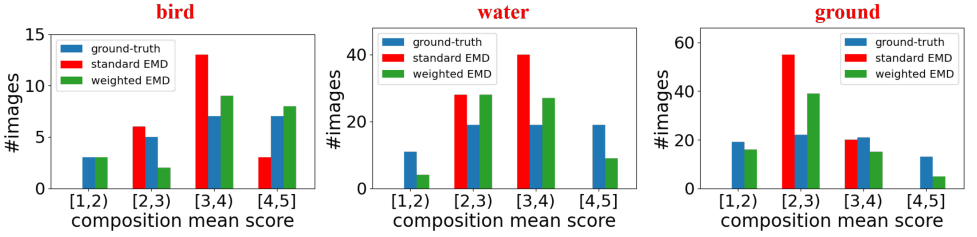


Figure 5: Illustration of using weighted EMD loss to eliminate content bias. For each category (colored red), given images containing this object category, we count three type of occurrences of composition mean scores in four score bins: ground-truth, predicted by the model trained with EMD loss, predicted by the model trained with our weighted EMD loss.

## 4 Weighted EMD Loss

In our CADB dataset, each image is rated by five raters, so both composition mean score and composition score distribution can be computed. Considering the intrinsic orderliness of our composition rating scale (see Section 1.2), we train our model to predict composition score distribution and adopt the normalized EMD loss [4], which has been widely used in aesthetic assessment [4, 20]. We assume that the ground-truth and predicted composition score distribution are  $\mathbf{y}$  and  $\hat{\mathbf{y}}$ , respectively. Then, the normalized EMD loss can be calculated by

$$\mathcal{L}_{EMD}(\mathbf{y}, \hat{\mathbf{y}}) = \left( \frac{1}{S} \sum_{s=1}^S |\text{CDF}_{\mathbf{y}}(s) - \text{CDF}_{\hat{\mathbf{y}}}(s)|^r \right)^{1/r}, \quad (1)$$

where  $S = 5$  is the scale of composition score in our dataset and  $r$  is a hyper-parameter.  $\text{CDF}_{\mathbf{y}}(s) = \sum_{i=1}^s y_i$  denotes the cumulative distribution function. We set  $r = 2$  following [4, 20]. The predicted composition mean score can be calculated as the expectation of the score distribution  $\sum_{i=1}^S i \cdot \hat{y}_i$ . As discussed in Section 3, we observe content bias in our dataset, that is, the images with certain object categories are more likely to have high or low composition scores. Training on such data, the network may find a shortcut to simply rate images based on their contents, leading to weak generalization ability to real-world photos. To eliminate the effect of content bias and prevent the model from learning a shortcut, we propose a strategy that assigns different weights to different samples when calculating EMD Loss. Specifically, as mentioned in Section 3, the range of composition mean score  $\bar{y}$  is divided into  $M$  bins. We use  $T_{m,c}$  to present the occurrence that the  $c$ -th category appears in  $m$ -th bin and calculate weights for each category via the strategy proposed in [4]:  $\alpha_{m,c} = \frac{\sum_{m=1}^M T_{m,c}}{M \cdot T_{m,c}}$ , which is inversely proportional to  $T_{m,c}$ . Given an image that contains  $C$  object categories and has a  $\bar{y}$  falling in the  $m$ -th bin, we take the minimum weight across all categories as its weight  $\beta = \min\{\alpha_{m,1}, \alpha_{m,2}, \dots, \alpha_{m,C}\}$ . Instead of *minimum*, we have also tried several other options (e.g., *maximum*, *median*, and *mean*), but *minimum* gives the best result. The weight  $\beta$  is different for different training samples. We precompute  $\beta$  for all training samples beforehand and assign sample-specific weight  $\beta$  to EMD loss (1) during training.

In the ablation study in Section 5.2 of the main text, we have confirmed that using weighted EMD loss can benefit model performance by eliminating content bias. To take a closer look at the advantage of weighted EMD loss in eliminating content bias, for each category, we analyze the distribution of ground-truth/predicted composition mean scores of

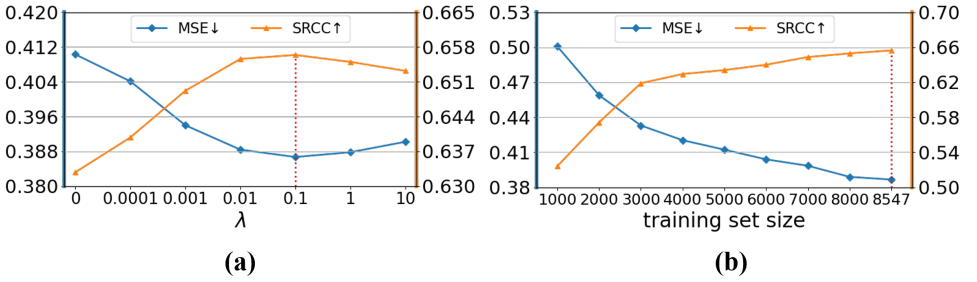


Figure 6: Analysis of hyper-parameters and training set size on our CADB dataset. (a) Performance variation of our model with different hyper-parameter  $\lambda$ . (b) Performance variation of our model with different training set size. The dashed vertical line denotes the default value used in our paper.

Backbone	MSE↓	EMD↓	SRCC↑	LCC↑
ResNet-18	0.3867	0.1798	0.6564	0.6709
ResNet-34	<b>0.3776</b>	<b>0.1794</b>	<b>0.6736</b>	<b>0.6808</b>
ResNet-50	0.399	0.1819	0.6539	0.6595
ResNet-101	0.4059	0.1824	0.6463	0.6563

Table 2: Performance of our method with different backbone networks.

images containing this object category.

Specifically, we first employ the ResNet18 [14] backbone trained with EMD loss (*resp.*, weighted EMD loss) to estimate composition mean scores for images in the testing set. Then, for each category, we collect the ground-truth/predicted composition mean scores of images containing this object category. After that, we visualize the distribution of composition mean score for each category in a similar way to Section 3. As illustrated in Figure 5, for each example category, we show three types of composition mean scores: ground-truth, predicted by the model trained with EMD loss, predicted by the model trained with proposed weighted EMD loss. Comparing the results of EMD loss and weighted EMD loss, we can see that training with weighted loss produces a much more unbiased distribution of composition mean score, which also looks closer to the ground-truth distribution. For example, for the results on water images in Figure 5, the ground-truth composition mean scores are approximately evenly distributed across four bins, while the composition mean score distribution of using EMD loss is concentrated in the intervals of [2,3) and [3,4). Differently, for the model trained with weighted EMD loss, the predicted composition mean score distribution is relatively balanced on all four bins, from which we can validate the advantages of the proposed weighted EMD loss on eliminating content bias qualitatively.

## 5 Implementation Details

We implement our model and conduct all experiments using Pytorch [15]. During the training stage, the backbone weights are pretrained on ImageNet [16] and other layers are randomly initialized. We adopt the Adam optimizer [17] and set the batch size as 16. Then, the initial learning rate of the layers in the backbone and the layers in the additional modules (*e.g.*, SAMP, AAFF, and prediction head) are set as  $1e^{-6}$  and  $1e^{-4}$ , respectively. This is because we noticed that using a small learning rate on the backbone results in easier and faster

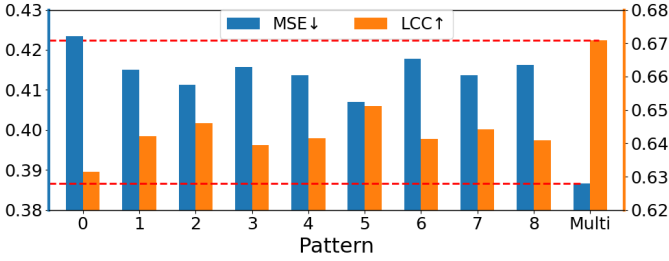


Figure 7: Results of our model using each individual composition pattern (pattern 1~8). Pattern 0 means the simplest pattern with only one partition. *Multi* means using all 8 patterns.

optimization in our experiments. Moreover, the learning rate of all layers is annealed by 0.1 every time the training loss plateaus. To prevent overfitting, a dropout rate of 0.5 is applied on each fully-connected layer of the additional modules, and we set weight decay as  $5e^{-5}$  for all layers in our network.

## 6 Hyper-parameter Analysis

There is a trade-off parameter  $\lambda$  before the attribute loss in Eq.(1) of the main text. We set the hyper-parameter according to cross-validation by splitting 20% training samples as validation set. We vary  $\lambda$  from 0 to 10 and present the results in Figure 6(a), in which we report Mean Squared Error (MSE) and Spearman’s Rank Correlation Coefficient (SRCC). Comparing the result without attribute loss ( $\lambda = 0$ ) and the result with  $\lambda = 0.1$ , we can see a clear gap between their performance. Therefore, we set  $\lambda = 0.1$  by default for all experiments. Moreover, the experimental results demonstrate that our method is robust when setting  $\lambda$  in the range of [0.01,10].

## 7 Different Training Set Size

As mentioned in Section 3 of the main text, we split the CADB dataset into training (8,547) and test (950) sets. To study the correlation between the test performance of our model and the training set size, we randomly select a certain amount of samples from training set to train the model and evaluate on the same test set. We vary the number of training samples from 1,000 to 8,000 with the step length of 1,000 and report the results in Figure 6(b) using MSE and SRCC. When the training set size increases, the model performance improves significantly, yet the performance growth slows down. When the training size gets larger than 8,000, the performance gain becomes negligible, demonstrating that our model capacity is compatible with the CADB dataset.

## 8 Different Backbone Network

We evaluate our method with different backbones on the CADB dataset and report results in Table 2. It can be seen that our method achieves the best result using ResNet-34 and the performance drop using ResNet-50 or ResNet-101 might be caused by overfitting. Given that ResNet-18 is efficient and can already receive good results, we adopt ResNet-18 as the default backbone in our method.



Patterns	MSE↓	EMD↓	SRCC↑	LCC↑
Patterns 1~ 8	<b>0.3867</b>	<b>0.1798</b>	<b>0.6564</b>	<b>0.6709</b>
Patterns 1~ 11	0.3876	0.1800	0.6558	0.6701

Table 3: Results of using more composition patterns. Based on the existing eight composition patterns, we add three more composition patterns (see Figure 10) in our model.

Human Rater	MSE↓	EMD↓	SRCC↑	LCC↑
1	0.1951	0.1403	0.8925	0.8944
2	0.4286	0.2089	0.7694	0.7811
3	0.5345	0.2328	0.7688	0.7705
4	<b>0.1606</b>	<b>0.1369</b>	<b>0.8990</b>	<b>0.9043</b>
5	0.1814	0.1430	0.8874	0.8934
SAMP-Net	0.3867	0.1798	0.6564	0.6709

Table 4: Comparison with human raters on the CADB dataset.

## 9 Effectiveness of Composition Pattern

Recall that we design eight composition patterns (see Figure 3(a) of the main text) for composition evaluation from different perspectives. To study the effectiveness of each pattern, we conduct experiments on our SAMP-Net with only a single pattern in SAMP. Moreover, we compare with the simplest pattern with only one partition (*i.e.*, global pooling), which is referred to as pattern 0. The experimental results are summarized in Figure 7, where we report MSE and Linear Correlation Coefficient (LCC). It can be seen that all the models with the designed patterns, including single pattern and multi-pattern, perform better than pattern 0, which indicates that our designed patterns are meaningful and helpful for composition assessment. Among the results using a single pattern, we find that pattern 5 performs best in terms of MSE, which might because visually important objects are often placed at the center of images. Furthermore, the multi-pattern model beats all single-pattern models, which again demonstrates the effectiveness of our SAMP module.

## 10 Using More Composition Patterns

Apart from existing eight patterns (see Figure 3(a) of the main text), to evaluate the effect of learning more diverse rules, we design three additional composition patterns in Figure 10. Pattern 9 is inspired by golden ratio [12]. Pattern 10 and pattern 11 concern more complex composition patterns. The results in Table 3 implies that using more composition patterns cannot achieve further improvement. We would like to explore other more composition patterns in the future.

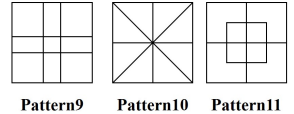


Figure 10: Three additional composition patterns.

## 11 Comparison with Human Ratings

We have shown that the proposed method outperforms existing methods in the Section 5.3 of the main text. To further analyze the capability of our method, we evaluate the performance

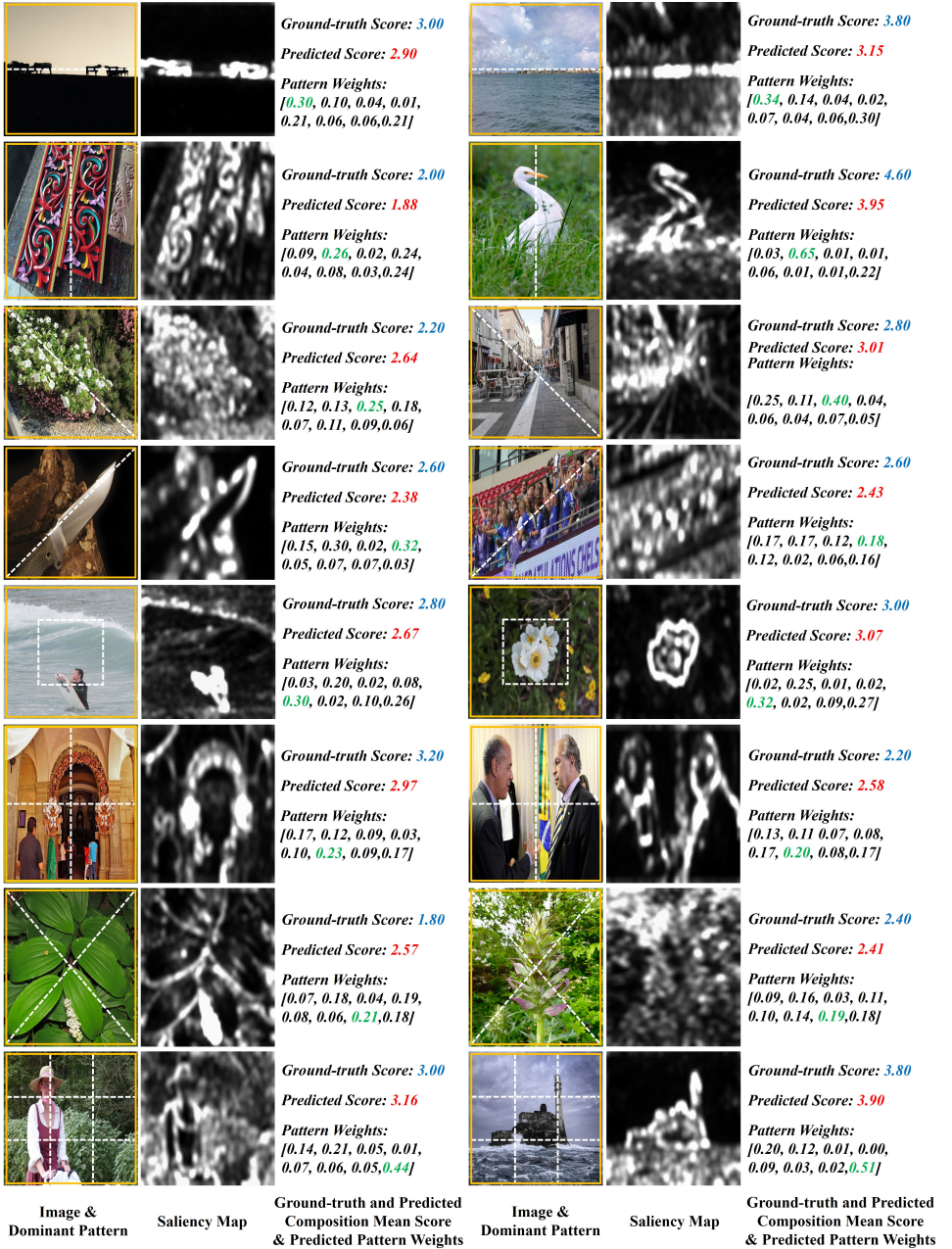


Figure 8: Visualization results of the proposed method on our CADB dataset. We show the estimated pattern weights and the largest weight is colored green. We also show the ground-truth/predicted composition mean score in blue/red.

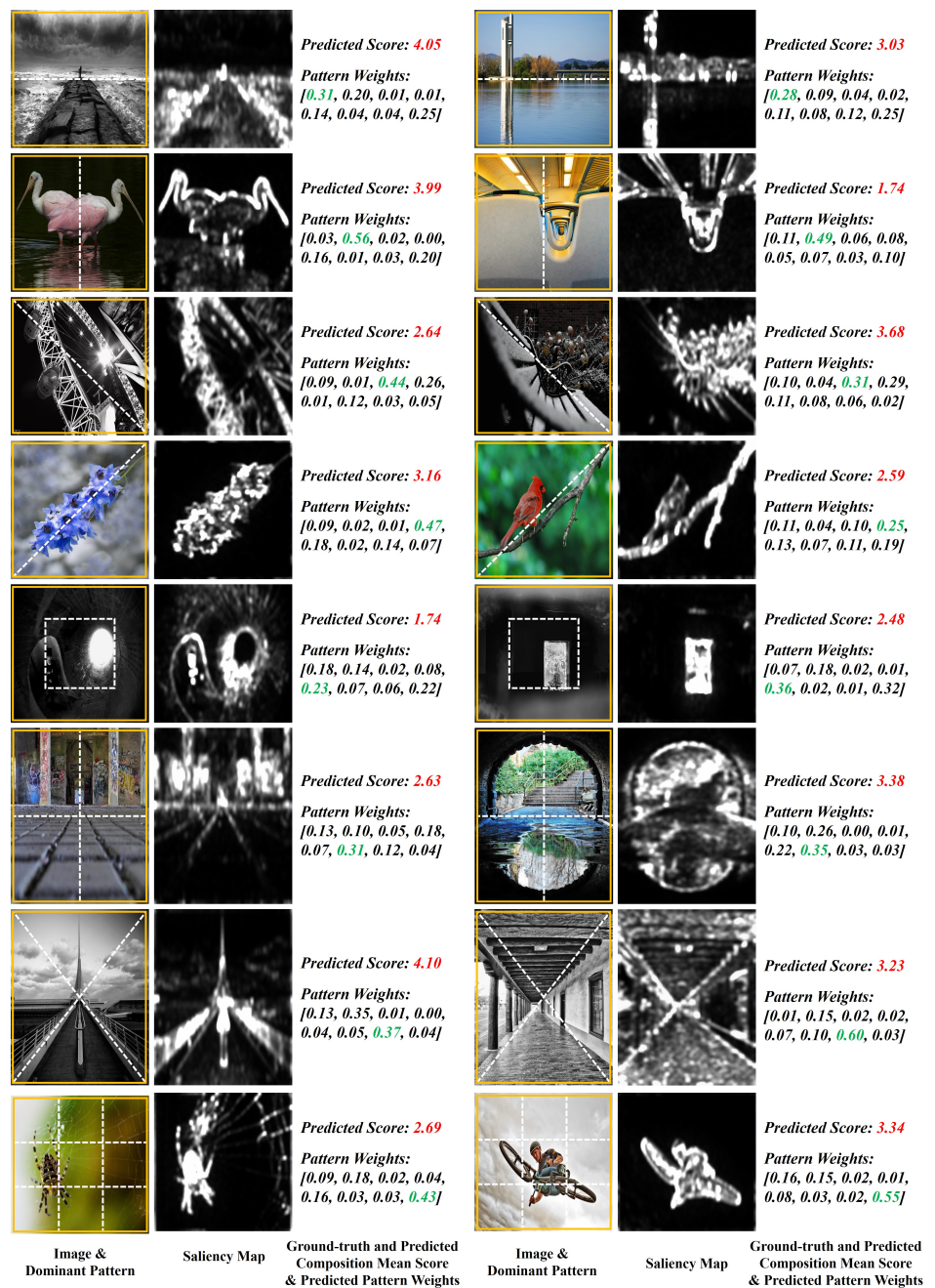


Figure 9: Visualization results of the proposed method on the PCCD dataset [14]. We show the estimated pattern weights and the largest weight is colored green. We also show the predicted composition mean score in red.

of each individual raters by comparing with the ground-truth in the same way. Unlike our model which predicts a composition score distribution, each rater only has one score for each test image, resulting in an one-hot score distribution. We summarize the results in Table 4. Interestingly, our method can outperform two of five human raters in terms of MSE and EMD. This may be due to the fact that our model is trained using the ratings of all raters and the prediction is close to the average distribution. However, considering SRCC and LCC, which indicate the ability to correctly rank different images according to their composition quality, we see that there is still a clear gap between our model and human raters.

## 12 Additional Visualization Results

Our SAMP-Net can facilitate composition assessment by integrating the information from multiple patterns and provide constructive suggestions for improving the composition quality. So we present additional examples in Figure 8, in which we show the input image, its saliency map, its ground-truth/predicted composition mean score, and its pattern weights. We refer to the composition pattern with the largest weight as the dominant pattern of the input image. For each pattern, we present two example images with this pattern as dominant pattern and draw this pattern on the image.

As discussed in Section 5.4 of the main text, the dominant pattern unveils from which perspective the input image is given a high or low score. For example, in Figure 8, in the right column of the second row, the vertical line of pattern 2 is parallel to the bird of the image, which looks more visually assuring to viewers. In the left column of the fourth row, pattern 4 implies that the knife is organised based on the diagonal line in the image. Since such images create a sense of visual balance and stability for viewer, the model estimates a relatively high score for them. On the contrary, in the left column of the second row in Figure 8, the carvings slightly deviate from their symmetrical axis under pattern 2. So the low score implies that maintaining horizontal symmetry may help to improve the composition quality. In the left column of the fifth row, per the relatively low score under pattern 5, the surfer is suggested to be moved towards the center. Those examples further validate the utility of our model for providing interpretable composition guidance.

Furthermore, we also test our model on some images outside the CADB dataset to show the generalization ability. Specifically, we test our model on some images collected from PCCD dataset [4] and show the results in Figure 9. Although the PCCD dataset contains the overall composition score, they only present one reviewer’s composition rating for each image and this reviewer (an anonymous website visitor) may be unprofessional, rendering the composition annotations of PCCD very noisy. Thus, we only report the composition score estimated by our model in Figure 9. We can see that our model can reasonably predict composition mean scores.

## References

- [1] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.
- [2] K. Chang, KH. Lu, and CS. Chen. Aesthetic critiques generation for photos. In *ICCV*, 2017.



- [3] Q. Chen, W. Zhang, N. Zhou, P. Lei, Y. Xu, Y. Zheng, and J. Fan. Adaptive fractional dilated convolution network for image aesthetics assessment. In *CVPR*, 2020.
- [4] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [5] M. Freeman. *The photographer's eye: Composition and design for better digital photos*. CRC Press, 2007.
- [6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [7] L. Hou, C.P. Yu, and D. Samaras. Squared earth mover's distance-based loss for training deep neural networks. *ArXiv*, abs/1611.05916, 2016.
- [8] X. Jin, L. Wu, G. Zhao, X. Li, X. Zhang, S. Ge, D. Zou, B. Zhou, and X. Zhou. Aesthetic attributes assessment of images. In *ACM-Multimedia*, 2019.
- [9] G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2): 137–163, 2001.
- [10] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [11] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016.
- [12] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
- [13] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.
- [14] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. The role of image composition in image aesthetics. In *ICIP*, 2010.
- [15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.
- [16] D. Präkel. *The fundamentals of creative photography*. Bloomsbury Publishing, 2010.
- [17] J. Ren, X. Shen, Z. Lin, R. Mech, and D. Foran. Personalized image aesthetics. In *ICCV*, 2017.
- [18] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2015.
- [19] A. Savakis, S. Etz, and A. Loui. Evaluation of image appeal in consumer photography. In *Human vision and electronic imaging V*, 2000.

- [20] H. Talebi and P. Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.
- [21] W. Wang, S. Yang, W. Zhang, and J. Zhang. Neural aesthetic image reviewer. *IET Computer Vision*, 13(8):749–758, 2019.
- [22] Y. Zhou, X. Lu, J. Zhang, and J.Z. Wang. Joint image and text representation for aesthetics analysis. In *ACM-Multimedia*, 2016.