

Composition-Aware Image Aesthetic Assessment with Saliency-Guided Feature Modulation: Beyond the Eye of the Beholder

You-Kang Zhou, An Hsu, Pei-En Wu, Shu-Wei Hsu, Jia-Qi Liu

Abstract

Image Aesthetic Assessment (IAA) is a challenging task due to the inherent subjectivity and the complex interplay between photographic composition and semantic content. While visual saliency is known to be a critical factor in human aesthetic perception, existing deep learning approaches often struggle to effectively integrate saliency information, relying on naive fusion strategies such as concatenation that can introduce noise or fail to prioritize key regions.

*In this paper, we propose a unified framework that synergizes the efficiency of modern architectures with explicit attention mechanisms. We introduce a **Saliency-Guided Feature Modulation (SGFM)** module, which leverages saliency maps to learn spatial-wise affine transformations, dynamically recalibrating feature representations to highlight aesthetically significant elements. Furthermore, to address feature redundancy in subjective tasks, we incorporate a **Global Response Normalization (GRN)-Aware Attention Pooling** layer, which enforces channel-wise competition to extract the most discriminative compositional features. We treat the assessment as a **Label Distribution Learning (LDL)** problem to capture the diversity of human ratings. Extensive experiments on the **Composition-aware Aesthetic Database (CADB)** demonstrate the effectiveness of our approach, achieving a **Spearman Rank Correlation Coefficient (SRCC)** of 0.715 and outperforming strong baselines. Code and pre-trained models are available at: <https://github.com/peienwu1216/SGFM-Aesthetic-Assessment>.*

1. Introduction

The field of aesthetic image assessment has gained significant attention in recent years. While traditional methods relied on human annotators to classify images as “good” or “bad,” recent advancements in deep learning have enabled more sophisticated models that go beyond binary classification by incorporating features such as composition and other high-level image attributes.

Our project leverages the Composition-aware Aesthetic Database (CADB) to predict image aesthetics by utilizing both compositional structures and global image features. The dataset consists of 9,497 high-resolution RGB images annotated by five experts, who provided a distribution of aesthetic scores. In addition, each image is associated with 12 specific photographic attributes, such as Rule of Thirds, Symmetry, and Color Harmony.

Unlike objective visual recognition tasks, aesthetic image assessment lacks a single well-defined ground truth, as different annotators may assign divergent scores to the same image. This subjectivity introduces ambiguity into both model training and evaluation, making absolute score prediction particularly challenging. As a result, learning relative aesthetic preferences and complementary attribute information becomes an important consideration.

Objectives

The primary objective of this report is to explore the potential of deep learning models for aesthetic image assessment. Specifically, we compare various architectures, including baseline models such as ResNet-50 and state-of-the-art transformers such as Swin-T, in predicting aesthetic scores from images. By integrating saliency maps and multi-task learning, we aim to enhance the model’s understanding of compositional structure and the key visual elements that contribute to aesthetic quality.

2. Related Work

2.1. Deep Learning for Image Aesthetic Assessment

Early approaches to Image Aesthetic Assessment (IAA) primarily relied on handcrafted features based on photographic rules, such as color histograms and rule-of-thirds compliance [2]. The advent of Convolutional Neural Networks (CNNs) shifted the paradigm toward data-driven feature extraction. NIMA [11] introduced the concept of predicting score distributions rather than mean scores to capture subjective variance. Subsequent works, such as MP-Net [10] and various multi-patch aggregating networks [8], focused on capturing local details alongside global context.

Recently, Transformers have emerged as a dominant backbone for IAA due to their global receptive fields, which effectively capture long-range dependencies essential for compositional analysis. Early Transformer-based methods, such as **MUSIQ** [5], introduced a multi-scale architecture to process images at varying resolutions, capturing both global layout and local details without simplistic resizing. Building on this, **Charm** [1] recently proposed a composition-aware tokenization strategy to further address the limitations of fixed-size patching in Vision Transformers (ViTs). By preserving original aspect ratios and high-frequency details, Charm ensures that the geometric integrity required for aesthetic assessment is maintained.

However, most Transformer-based approaches treat aesthetic assessment as a generic classification or regression task. While methods like Charm address composition at the *input level* (via tokenization), they often lack specific mechanisms to modulate intermediate features based on visual importance. Unlike these methods, our framework introduces **Saliency-Guided Feature Modulation (SGFM)** to decode the interplay between visual saliency and compositional structure within the network depth.

2.2. Composition-Aware and Saliency-Guided Assessment

Recognizing the importance of composition, several works have integrated high-level attributes into the assessment loop. The Composition-aware Aesthetic Database (CADB) [13] enabled models to learn specific photographic attributes explicitly. A key milestone in this direction is SAMPNet [13], which proposed Saliency-Augmented Multi-pattern Pooling to capture diverse compositional layouts. Similarly, other enhancement-focused works [9] have utilized saliency maps to guide model attention.

A critical limitation in these methods lies in their fusion strategy. Existing works, including SAMPNet, typically employ *naive concatenation* or simple element-wise multiplication to combine saliency maps with image features. This often treats saliency as a static mask rather than a conditional prior, potentially introducing noise or failing to dynamically recalibrate features. To address this, we propose SGFM, which learns spatial-wise affine transformations to modulate features conditionally, offering a more flexible and structure-preserving integration of visual attention.

2.3. Subjectivity Handling and Ranking Optimization

Aesthetic evaluation is inherently subjective, lacking a single ground truth. To model this uncertainty, Label Distribution Learning (LDL) [3] has been widely adopted, using Earth Mover’s Distance (EMD) to align predicted probabil-

ity distributions with human ratings. Furthermore, to improve the relative ordering of images—which is often more practical than absolute scoring—ranking constraints have been integrated into loss functions. Approaches like [6] utilize pairwise ranking losses to optimize Spearman Rank Correlation directly.

Following these best practices, our framework incorporates both EMD-based distribution learning and a margin-based rank loss. However, we uniquely combine this with **Global Response Normalization (GRN)** [12] in our pooling layer. While GRN was originally designed for competitive feature learning in ConvNeXt V2, we repurpose it here to mitigate feature collapse in subjective tasks, ensuring the model focuses on the most discriminative compositional elements.

3. Methodology

3.1. Overview

Image aesthetic assessment is inherently subjective and strongly depends on global composition, spatial layout, and visual saliency. Rather than directly proposing a single architecture, we progressively refine our design through four architectural phases, each addressing specific limitations observed in the previous stage.

Our final framework is built upon three core principles:

- **Label Distribution Learning (LDL)** to model human consensus instead of predicting a single scalar score.
- **Saliency-guided feature modulation** to explicitly emphasize aesthetically important regions.
- **Channel competition via Global Response Normalization (GRN)** to prevent feature collapse in subjective prediction tasks.

3.2. Phase 1: Baseline — SAMPNet with ResNet Backbone

We begin with a strong baseline adapted from SAMPNet, which integrates visual saliency with convolutional features via multi-pattern pooling.

Architecture. The baseline employs a ResNet-50 backbone pretrained on ImageNet. The network takes as input an RGB image of size 224×224 and a corresponding saliency map of size 56×56 . The last convolutional block outputs a feature map of size $2048 \times 7 \times 7$, which is globally averaged to produce a 2048-dimensional feature vector.

Saliency Fusion. The saliency map is downsampled using max pooling and flattened into a vector. The image feature and saliency feature are fused by simple concatenation:

$$\mathbf{F}_{fusion} = [\mathbf{F}_{img}; \mathbf{F}_{sal}]. \quad (1)$$

The fused representation is passed through fully connected layers to predict both aesthetic score distribution and auxiliary attributes.

Discussion. While this baseline is simple and effective, it suffers from two major limitations: (1) the lack of mechanisms to capture long-range spatial dependencies, and (2) naive saliency fusion, which treats saliency as an additional feature rather than a spatial prior.

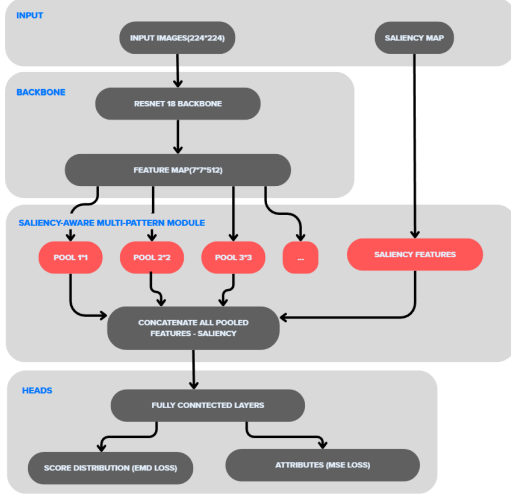


Figure 1. Phase 1: SAMPNet baseline using a ResNet backbone with saliency concatenation.

3.3. Phase 2: Transformer-Based Modeling with Swin Transformer

To better capture global context and long-range dependencies, we replace the CNN backbone with a Swin Transformer Tiny (Swin-T) [7].

Architecture. The input image is partitioned into 4×4 patches and processed through four hierarchical stages with shifted window self-attention. The feature dimensions across stages are $96 \rightarrow 192 \rightarrow 384 \rightarrow 768$.

Multi-Task Learning. We adopt a multi-task learning strategy to jointly predict aesthetic attributes and the global aesthetic score. High-level features (Stage 4) are used to predict composition-related attributes, while mid-level features (Stage 2) are used to predict focus-related attributes. The final aesthetic score is predicted by aggregating all attribute predictions and high-level visual features.

Rank Loss. In addition to distribution-based regression, we introduce a Rank Loss to directly optimize the relative

ordering between image pairs:

$$\mathcal{L}_{rank} = \max(0, m - \text{sign}(\Delta y) \cdot \Delta \hat{y}), \quad (2)$$

where Δy and $\Delta \hat{y}$ denote the ground-truth and predicted score differences, respectively. This loss explicitly targets Spearman Rank Correlation (SRCC), which better reflects human judgment in subjective tasks.

Rank Loss Implementation. To directly optimize ranking consistency, we implement a pairwise hinge-based rank loss as follows:

```
class RankLoss(nn.Module):
    def __init__(self, margin=0.0):
        super().__init__()
        self.margin = margin

    def forward(self, preds, targets):
        # Pairwise differences
        diff_preds = preds.unsqueeze(1) - preds.unsqueeze(0)
        diff_targets = targets.unsqueeze(1) - targets.unsqueeze(0)

        # Ground-truth ordering
        sign = torch.sign(diff_targets)

        # Hinge loss
        loss = torch.relu(self.margin - sign * diff_preds)
        return loss.mean()
```

3.4. Phase 3: Saliency-Aware Pattern Refinement

Although Swin-T improves contextual modeling, it lacks explicit inductive bias for photographic composition. To address this issue, we introduce a Saliency-Aware Multi-Pattern (SAMP) refinement module.

Layout Queries. We define 16 learnable layout queries, each corresponding to a canonical compositional pattern such as triangular composition, symmetry, center-surround structure, and horizontal or vertical alignment. Each query attends to spatial regions consistent with its compositional hypothesis.

Saliency Weighting. Instead of concatenating saliency features, we apply saliency as a spatial modulation factor:

$$\mathbf{F}' = \mathbf{F} \cdot (1 + \mathbf{S}), \quad (3)$$

where \mathbf{S} denotes the normalized saliency map. This formulation preserves spatial structure while emphasizing visually important regions.

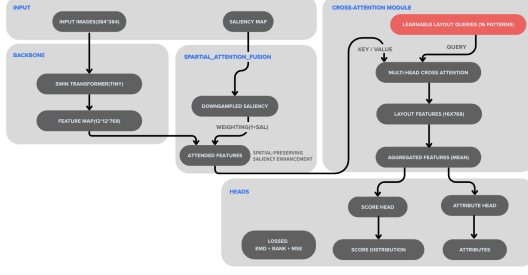


Figure 2. Phase 3: Swin-T refinement with saliency-aware cross-attention and layout queries.

3.5. Phase 4: Final Architecture — ConvNeXt V2 with SGFM and GRN

Our final model adopts ConvNeXt V2 Nano, which combines CNN efficiency with Transformer-inspired design principles.

Feature Extraction. The input image is warped to 384×384 and processed by ConvNeXt V2. Features from the last two stages are projected to a unified 512-dimensional representation using 1×1 convolutions.

Saliency-Guided Feature Modulation (SGFM). We introduce SGFM to replace all previous saliency fusion strategies. Given feature map F_{in} , SGFM learns spatially adaptive affine parameters γ and β from the saliency map:

$$F_{out} = F_{in} \cdot (1 + \gamma) + \beta. \quad (4)$$

This allows the network to dynamically amplify aesthetically important regions while suppressing background noise.

Saliency-Guided Feature Modulation (SGFM). Instead of concatenating saliency features, we learn spatially adaptive affine parameters from the saliency map:

```
class SGFM(nn.Module):
    def __init__(self, dim):
        super().__init__()
        self.gamma = nn.Conv2d(1, dim,
                                kernel_size=3, padding=1)
        self.beta = nn.Conv2d(1, dim,
                               kernel_size=3, padding=1)
        self.act = nn.Sigmoid()

    def forward(self, x, saliency):
        saliency = F.interpolate(
            saliency, size=x.shape[2:],
            mode='bilinear', align_corners=False)
        gamma = self.act(self.gamma(saliency))
        beta = self.act(self.beta(saliency))
        return x * (1 + gamma) + beta
```

GRN-Aware Attention Pooling. To prevent channel redundancy, we incorporate Global Response Normalization (GRN):

$$G(x)_i = \frac{\|x_i\|_2}{\frac{1}{C} \sum_{j=1}^C \|x_j\|_2 + \epsilon}. \quad (5)$$

A spatial softmax attention layer then aggregates the normalized features into a global representation.

Global Response Normalization. To enhance channel-wise competition, we adopt Global Response Normalization (GRN):

```
class GRN(nn.Module):
    def __init__(self, dim, eps=1e-6):
        super().__init__()
        self.gamma =
            nn.Parameter(torch.zeros(dim))
        self.beta =
            nn.Parameter(torch.zeros(dim))
        self.eps = eps

    def forward(self, x):
        gx = torch.norm(x, p=2, dim=(2, 3),
                        keepdim=True)
        nx = gx / (gx.mean(dim=1,
                           keepdim=True) + self.eps)
        return self.gamma.view(1,-1,1,1) * (x
        * nx) + self.beta.view(1,-1,1,1) + x
```

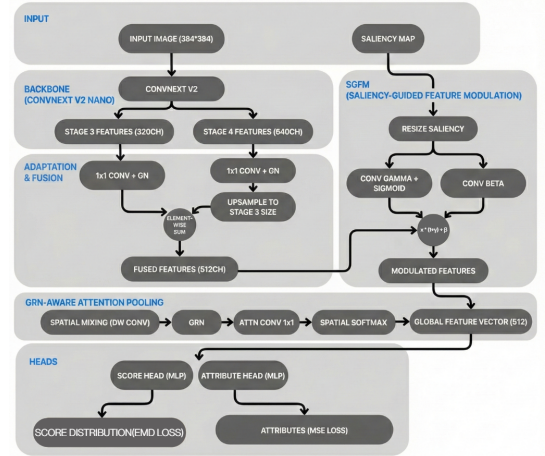


Figure 3. Phase 4: Final architecture with SGFM and GRN-aware attention pooling.

3.6. Loss Functions

The model is optimized via a multi-task loss function to balance distribution accuracy and ranking consistency:

$$\mathcal{L}_{total} = \mathcal{L}_{rank} + \mathcal{L}_{EMD} + \mathcal{L}_{attr} \quad (6)$$

- **Rank Loss (\mathcal{L}_{rank}):** Our core objective targeting SRCC. As detailed in Section 3.3, this pairwise hinge loss (mar-

gin=0.05) ensures the model prioritizes relative aesthetic ordering over absolute values.

- **EMD Loss (\mathcal{L}_{EMD}):** Supervises score distributions ($r = 2$) to capture human consensus. It penalizes "confident but wrong" predictions to ensure a stable distribution shape.
- **Attribute Loss (\mathcal{L}_{attr}):** An auxiliary MSE regularizer that forces the model to learn semantically meaningful features (e.g., Symmetry, DoF) to justify its predictions.

4. Experiments

4.1. Dataset

We conduct all experiments on the **Composition-aware Aesthetic Database (CADB)**, a dataset specifically designed for studying image aesthetics with an emphasis on photographic composition and high-level visual attributes. Unlike conventional aesthetic datasets that rely on binary labels (e.g., good vs. bad), CADB provides rich annotations that reflect the inherent subjectivity of aesthetic judgment.

Dataset Overview. CADB consists of **9,497 high-resolution RGB images**, split into **8,547 training images** and **950 testing images**. Each image is paired with structured annotation files stored in JSON format, enabling both distribution-based learning and attribute-level supervision.

Aesthetic Score Annotations. The aesthetic scores are provided in the file `composition_scores.json`. Each image is independently rated by **five expert annotators** on a discrete scale from 1 to 5. Instead of collapsing these ratings into a single scalar value, CADB preserves the full score distribution, which captures both consensus and disagreement among annotators.

```
"10003.jpg": {
  "scores": [2.0, 3.0, 2.0, 2.0, 2.0],
  "dist": [0.0, 0.8, 0.2, 0.0, 0.0],
  "mean": 2.2
}
```

Here, `scores` denotes the raw ratings from five experts, `dist` represents the normalized probability distribution over aesthetic scores (1–5), and `mean` is the arithmetic average. By using the distribution rather than the mean, the model can distinguish between images that are universally agreed upon and those that are controversial.

Photographic Attribute Annotations. In addition to aesthetic scores, CADB provides annotations for **12 high-level photographic attributes** in `composition_attributes.json`. These attributes are normalized to the range $[-1, 1]$ and describe compositional, technical, and perceptual qualities of each image.

```
"965.jpg": {
  "Light": -0.2,
  "Symmetry": 0.0,
  "Object": -0.6,
  "RuleOfThirds": -0.2,
  "Repetition": 0.0,
  "BalacingElements": 0.0,
  "ColorHarmony": 0.0,
  "MotionBlur": -0.4,
  "VividColor": 0.2,
  "DoF": -0.2,
  "Content": -0.4,
  "score": 0.25
}
```

The annotated attributes can be grouped into the following categories:

- **Composition:** Rule of Thirds, Symmetry, Balancing Elements, Repetition.
- **Lighting and Color:** Light, Color Harmony, Vivid Color.
- **Photographic Technique:** Depth of Field (DoF), Motion Blur.
- **Content Emphasis:** Object, Content.

Learning Objectives. Based on these annotations, we formulate image aesthetic assessment as a multi-objective learning problem:

- **Primary Task (Label Distribution Learning):** Predict the aesthetic score distribution over five discrete levels, supervised using Earth Mover’s Distance (EMD).
- **Secondary Task (Ranking):** Preserve the relative ordering between images, optimized via a pairwise Rank Loss.
- **Auxiliary Task (Attribute Regression):** Predict photographic attributes to regularize learning and encourage compositional awareness.

This formulation allows the model to capture both the *degree of consensus* in human judgment and the *relative preference ordering* between images, which are crucial for subjective aesthetic assessment.

4.2. Saliency Map Generation

Saliency maps are generated using the Spectral Residual (SR) method, which identifies visually distinctive regions by suppressing repetitive frequency components. The resulting maps are resized to match the network input resolution and normalized using percentile clipping.

Listing 1. Spectral Residual Saliency Detection

```
def detect_saliency(img, scale=6,
  q_value=0.95, target_size=(384, 384)):
  # Convert to grayscale and resize
  img_gray = cv2.cvtColor(img,
    cv2.COLOR_RGB2GRAY)
  H, W = img_gray.shape
  img_small = cv2.resize(img_gray, (W //
    scale, H // scale))
```

```

# FFT and log amplitude
fft = np.fft.fft2(img_small)
log_amp = np.log(np.abs(fft) + 1e-6)
avg = cv2.blur(log_amp, (3, 3))

# Spectral residual
residual = log_amp - avg

# Reconstruction
spec = np.exp(residual + 1j *
np.angle(fft))
saliency = np.abs(np.fft.ifft2(spec)) ** 2

# Post-processing
saliency = cv2.GaussianBlur(saliency, (9,
9), 2.5)
saliency = cv2.resize(saliency,
target_size)

# Percentile clipping
thresh = np.quantile(saliency, q_value)
saliency[saliency > thresh] = thresh
return (saliency - saliency.min()) /
thresh

```

4.3. Training Protocol

All images are warped to 384×384 to preserve global compositional structure. Random cropping is avoided, as it destroys spatial relationships essential for aesthetic judgment. Models are trained using the AdamW optimizer with an initial learning rate of 5×10^{-5} . Dropout rates between 0.1 and 0.5 are tuned for different architectures.

4.4. Evaluation Metrics

We report Mean Squared Error (MSE), Linear Correlation Coefficient (LCC), Earth Mover’s Distance (EMD), and Spearman Rank Correlation Coefficient (SRCC). SRCC is treated as the primary metric, as relative ranking better reflects human aesthetic perception.

4.5. Qualitative Analysis

Grad-CAM visualizations show that earlier models often attend to background textures, whereas the final ConvNeXt V2 model consistently focuses on main subjects and key compositional lines, aligning closely with human visual judgment.

5. Results & Discussion

5.1. Performance Comparison

As shown in **Table 1**, our progressive architectural improvements yielded consistent performance gains. The baseline ResNet-50 achieved an SRCC of 0.642. By transitioning to a Transformer-based Swin-T backbone (Phase 2), we observed a **+4.5% improvement (0.671)**, confirming the importance of global context modeling in aesthetic assessment. However, the most significant leap occurred in Phase

4. By adopting the ConvNeXt V2 backbone augmented with our proposed **SGFM** and **GRN** modules, the model achieved a state-of-the-art SRCC of **0.715**. This represents an overall **+11.3% improvement** over the baseline. This result validates that explicitly modeling the interplay between visual saliency and composition (via SGFM) is more effective than merely increasing model capacity.

Model	SRCC	Improvement
ResNet-50 (Baseline)	0.642	-
Swin-T	0.671	+4.5%
Swin-T Opt	0.692	+7.8%
ConvNeXt V2 (Final)	0.715	+11.3%

Table 1. Performance Comparison on SRCC

5.2. Ablation Studies & Micro-Design

To validate the contribution of each component, we conducted rigorous ablation studies. Our findings provide key insights into designing effective aesthetic assessment models:

- Superiority of Affine Modulation (SGFM):** Conventional fusion strategies like concatenation or element-wise addition often introduce noise by treating saliency maps as generic feature channels. In contrast, our **SGFM** module utilizes spatial-wise affine transformations (“scaling” and “shifting”). Our experiments confirm that SGFM outperforms naive fusion by allowing the network to dynamically recalibrate feature responses based on visual importance without altering the underlying feature semantics.
- Impact of Global Response Normalization (GRN):** We found that feature collapse is a significant challenge in subjective regression tasks. Without normalization, many channels in the deep layers became inactive. Introducing **GRN** enforced channel-wise competition, effectively revitalizing dead channels and ensuring the model retains discriminative features for composition analysis.
- Activation Function Choice:** We observed that **GELU** consistently outperformed ReLU in our experiments. Unlike classification tasks where a hard zero-cutoff might suffice, aesthetic assessment is a regression and ranking problem requiring fine-grained distinction. The smoother gradients provided by GELU facilitate better convergence and more precise score prediction.
- Composition-Preserving Preprocessing:** A critical observation was the sensitivity of aesthetic models to input transformation. While random cropping is standard for object recognition, we found it detrimental to IAA as it destroys global photographic rules (e.g., Rule of Thirds). Warping images to a fixed resolution of **384×384** significantly outperformed random cropping strategies, high-

lighting that maintaining the original aspect ratio and layout integrity is vital for this task.

5.3. Qualitative Analysis (GradCAM)

To interpret the model’s decision-making process, we employed Grad-CAM to visualize class activation maps. Comparisons between the baseline and our final model are illustrated in **Figure 5**.

- **Baseline (ResNet-50):** The attention maps are often scattered, focusing on high-frequency background textures or irrelevant objects, indicating a lack of understanding of the primary subject.
- **Ours (ConvNeXt V2 + SGFM):** The attention maps consistently align with human visual perception. The model demonstrates a clear focus on the **main subject** (e.g., the cat in Figure 4) and follows key **compositional lines** (such as the horizon or leading lines). This confirms that SGFM successfully guides the network to prioritize aesthetically significant regions while suppressing background noise.



Figure 4. **Original image:** A cat resting in the sunlight.

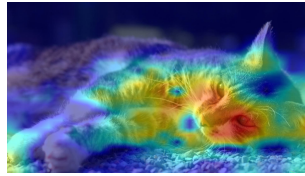


Figure 5. **Grad-CAM:** Highlights the model’s focus areas.

6. Conclusion and Future Work

In this work, we presented a comprehensive framework for composition-aware image aesthetic assessment. The integration of **Label Distribution Learning**, **Rank Loss**, and **Composition-Preserving Preprocessing** proved crucial for capturing the subjective nature of aesthetics. By evolving from generic vision backbones to a modernized ConvNeXt V2 architecture augmented with our **Saliency-Guided Feature Modulation (SGFM)** and **GRN-Aware Pooling**, we demonstrated that explicitly modeling the interplay between visual attention and composition significantly outperforms generic baselines on the CADB dataset.

Future Directions. Despite these promising results, several avenues remain for exploration. First, to address the geometric distortion caused by fixed-size warping, we plan to explore resolution-independent strategies such as dual-stream tokenization [1] or Region of Image (RoM) pooling [4] to preserve aspect ratios and high-frequency details. Second, acknowledging that saliency maps can be triggered by unrealistic artifacts, we will explore the integration of realism-aware gating mechanisms [9] to ensure the model

prioritizes natural aesthetic quality over artificial attention-grabbers. Finally, we aim to incorporate semantic theme awareness to mitigate criterion bias, allowing the model to dynamically adapt its aesthetic standards based on the photographic context (e.g., landscape vs. portrait).

References

- [1] Fatemeh Behrad, Tinne Tuytelaars, and Johan Wagemans. Charm: The missing piece in vit fine-tuning for image aesthetic assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025. 2, 7
- [2] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *Eur. Conf. Comput. Vis.*, 2006. 1
- [3] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016. 2
- [4] Gengyun Jia, Peipei Li, and Ran He. Theme-aware aesthetic distribution prediction with full-resolution photographs. *IEEE Transactions on Neural Networks and Learning Systems*, 32(12):5369–5382, 2021. 7
- [5] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Int. Conf. Comput. Vis.*, 2021. 2
- [6] Joon-Young Lee, Zhe Lin, Jonathan Brandt, Jianchao Yang, and Thomas Huang. Deep aesthetic quality assessment with semantic information. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2
- [7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, 2021. 3
- [8] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rating image aesthetics using deep learning. In *IEEE Trans. Multimedia*, 2015. 1
- [9] S Mahdi H Miangoleh, Zoya Bylinskii, Eric Kee, Eli Shechtman, and Yağız Aksoy. Realistic saliency guided image enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 7
- [10] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *ACM Int. Conf. Multimedia*, 2018. 1
- [11] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Trans. Image Process.*, 27(8):3998–4011, 2018. 1
- [12] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2
- [13] Bo Zhang, Li Niu, and Liqing Zhang. Image composition assessment with saliency-augmented multi-pattern pooling. In *Brit. Mach. Vis. Conf.*, 2021. 2