2025

Introduction to Machine Learning

Aesthetic  Image Rating

Presented by

GROUP 17

Date

16 December 2025

Project Github

github.com/peienwu1216/intro-ml-nycu-2025

Demo Website

aes.slasho.tw

Project Description

Project Description

01.

We built a model which can score images

Aesthetic quality is subjective and abstract.

Our goal is to build a Machine Learning model that can **quantify beauty and composition** just like a professional photographer, distinguishing high-quality shots from poor ones with **SOTA accuracy.**

4.4/5 🥰👍

High Aesthetics
Good Composition

1.8/5 🤮👎

Poor Lighting
Low Quality

Project Description

Xu, K., Deng, Y., Ding, M., Cheng, Z., & Lu, H. (2021). **Image composition assessment with saliency-augmented multi-patch network**. In Proceedings of the British Machine Vision Conference (BMVC 2021).

Method	MSE↓	EMD↓	SRCC↑	LCC↑
ResNet18	0.4534	0.1943	0.6025	0.6148
AADB [22]	0.4234	0.1923	0.6236	0.6415
MNA-CNN [32]	0.4260	0.1944	0.6108	0.6375
A-Lamp [31]	0.4230	0.1898	0.6270	0.6456
VP-Net [52]	0.4304	0.1948	0.6169	0.6285
RG-Net [28]	0.4398	0.1915	0.6026	0.6218
AFDC-Net [4]	0.4245	0.1910	0.6154	0.6388
SAMP-Net (Ours)	**0.3867**	**0.1798**	**0.6564**	**0.6709**

Our Model 0.3627 0.1515 0.7128 0.7124

02.

We Beat 上海交通大学‘s SOTA

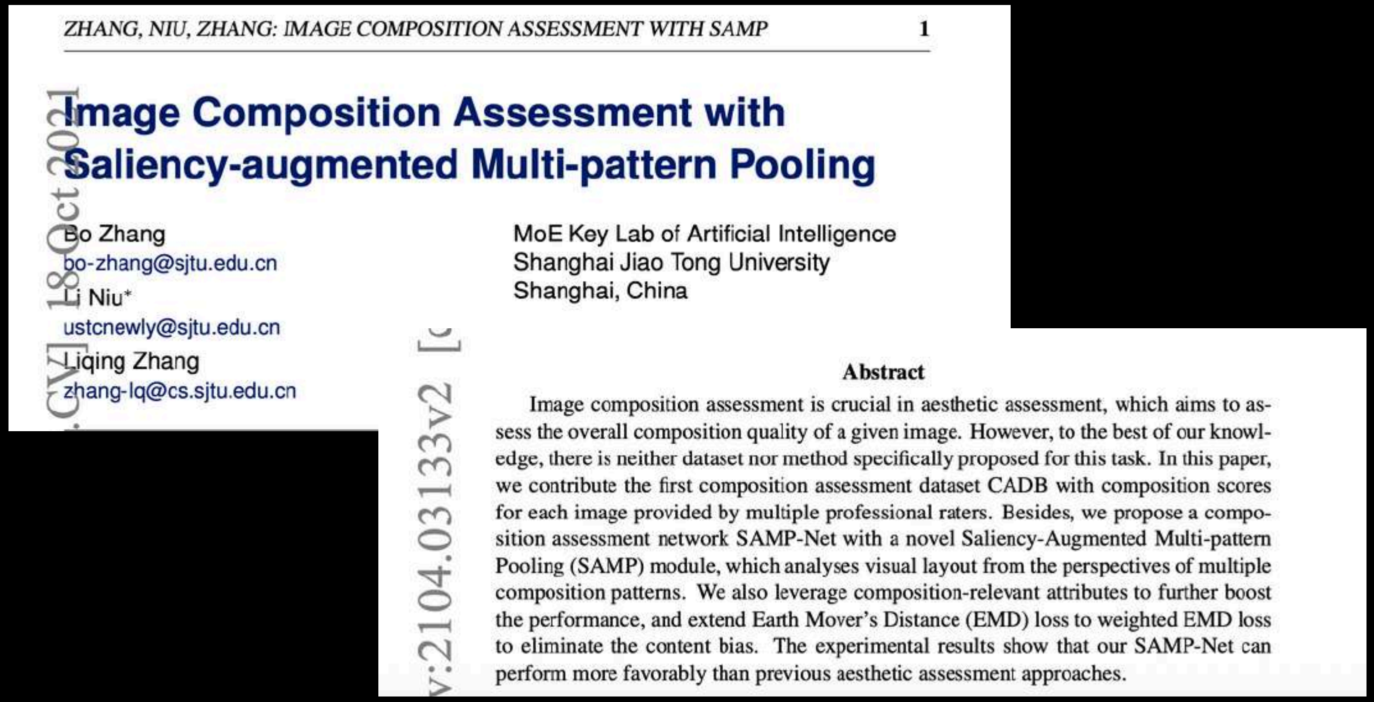

ZHANG, NIU, ZHANG: IMAGE COMPOSITION ASSESSMENT WITH SAMP 1

Image Composition Assessment with Saliency-augmented Multi-pattern Pooling

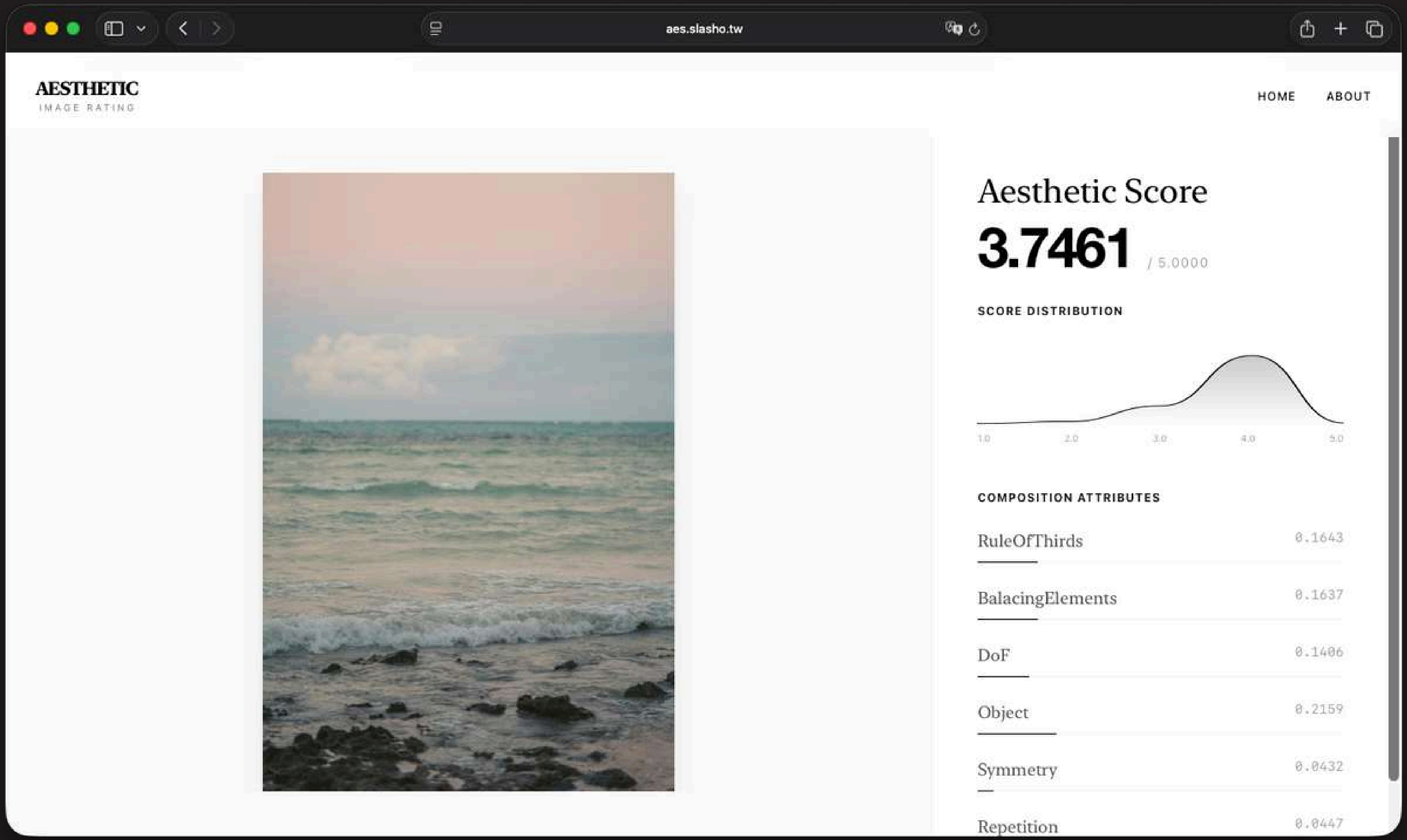
Bo Zhang
bo-zhang@sjtu.edu.cn
Li Niu*
ustcnewly@sjtu.edu.cn
Liqing Zhang
zhang-lq@cs.sjtu.edu.cn

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

Abstract

Image composition assessment is crucial in aesthetic assessment, which aims to assess the overall composition quality of a given image. However, to the best of our knowledge, there is neither dataset nor method specifically proposed for this task. In this paper, we contribute the first composition assessment dataset CADB with composition scores for each image provided by multiple professional raters. Besides, we propose a composition assessment network SAMP-Net with a novel Saliency-Augmented Multi-pattern Pooling (SAMP) module, which analyses visual layout from the perspectives of multiple composition patterns. We also leverage composition-relevant attributes to further boost the performance, and extend Earth Mover's Distance (EMD) loss to weighted EMD loss to eliminate the content bias. The experimental results show that our SAMP-Net can perform more favorably than previous aesthetic assessment approaches.

Demo


aes.slasho.tw
AESTHETIC
IMAGE RATING
HOME
ABOUT
Aesthetic Score
3.7461
/ 5.0000
SCORE DISTRIBUTION
COMPOSITION ATTRIBUTES
RuleOfThirds
0.1643
BalacingElements
0.1637
DoF
0.1406
Object
0.2159
Symmetry
0.0432
Repetition
0.0447

Define the problem

Our "Vision" The "Ground Truth" Paradox

Traditional ML

"Is this a cat?" → Objective Truth (Yes/No)

Cat ✅

Cat ❌

Aesthetic ML

Rate this beautiful ?

3.6/5 😐

4.5/5 😍

Our "Vision" From Plato to PyTorch

Beauty is an objective property of the universse.
– Theory of Forms

Plato

“Beauty is a "subjective universality" — a feeling of pleasure without a concept.”

Immanuel Kant

The Challenge:
We are not training a model to find "Truth", but to model **Human Consensus.**

A score of "4.5" isn't a physical measurement;
it's an aggregation of subjective opinions.
Difficulty: High variance, cultural bias, and lack of clear "right" answers.

Dataset

Composition Assessment DataBase

CADB Dataset

9,497 images

Train: 8,547 / Test: 950

The "5 Experts" (Ground Truth)

[3, 4, 3, 5, 4]

We use the distribution of these 5 votes to capture **consensus** vs **controversy** (e.g., a polarizing image vs a universally accepted one).

Attributes

1. RuleOfThirds (三分法)
2. BalacingElements (平衡元素)
3. Symmetry (對稱性)
4. Repetition (重複性)
5. Object (主體)
6. DoF (景深)
7. MotionBlur (動態模糊)
8. Light (光線)
9. ColorHarmony (色彩和諧)
10. VividColor (鮮豔色彩)
11. Content (內容)
12. Score (構圖分數)

Our Validation Set

Good 👍

Bad 👎

Unified Standard

We manually selected 10 pairs of “good” and “ bad” samples

With each member contributing 2 carefully selected pairs to define clear compositional and aesthetic criteria.

Rank (Precise) vs Score (Dubious)

Why "Accuracy" is a Trap in Aesthetics

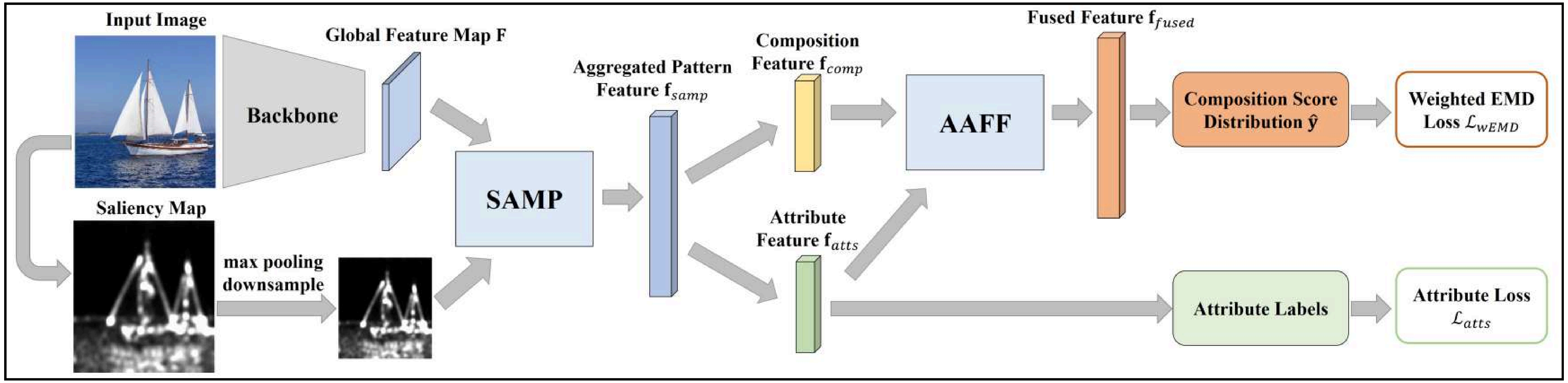
- The Threshold Problem: If we set "Good" > 3.5:
  - Prediction: 3.49 (Bad) | Truth: 3.51 (Good) → Fail.
  - This binary view ignores the nuance of aesthetic continuum.

SRCC (Spearman Rank Correlation Coefficient)

- Humans are bad at absolute scoring ("Is this a 72 or 75?"), but excellent at Ranking ("Is A better than B?").
- Conclusion: In subjective tasks, Relative Order > Absolute Value.
- SRCC: Measures Monotonicity.
  - If the model ranks Image A > Image B, and humans rank Image A > Image B, the model succeeds,
  - even if the absolute scores differ.

What We Have Been Through

Baseline: SNAP-Net



Zhang et al.,
"Image Composition Assessment with Saliency-augmented Multi-pattern Pooling",
arXiv 2021.

- Saliency-aware multi-pattern framework
- ResNet backbone for feature extraction
- Multi-pattern spatial layout modeling
- Joint composition score & attribute prediction

Transformer (Swin-T)

- Attempted to introduce Attention mechanisms to capture long-range dependencies.
- **Result: SRCC improved to 0.67.**
- Proved the potential of Transformer architecture, but training was less stable.

Final Form (ConvNeXt V2)

- Combines CNN efficiency with Transformer design philosophy
- GRN (Enhances channel competition to prevent Feature Collapse
- achieving the best balance of performance and efficiency..
- Result: **SRCC broke through 0.71**

Baseline (SNAP Net) → **Optimized Baseline Backbone** → **Optimized Best Backbone's Architecture** → **New Backbone with optimized Architecture**

Baseline (SAMPNet/ResNet)

- Original architecture using ResNet backbone.
- Limitation: Older feature extraction capabilities, insufficient global context capture.
- **SRCC: ~0.64**

Architecture Optimization (Swin-T Optimized)

- Added Spatial Attention on top of Swin-T to enhance composition understanding.
- Optimization: Fine-tuned the balance of Loss weights (EMD + Attribute + Rank).
- **Result: SRCC improved to 0.69.**

Evaluation Metrics

EMD

Earth Mover's Distance

- Measures the distance between predicted and true distributions. Lower is better.

MSE

Mean Squared Error

- Standard regression metric. Measures squared error between predicted and true mean scores.

LCC

Linear Correlation Coefficient

- Measures linear correlation between predicted and true scores.

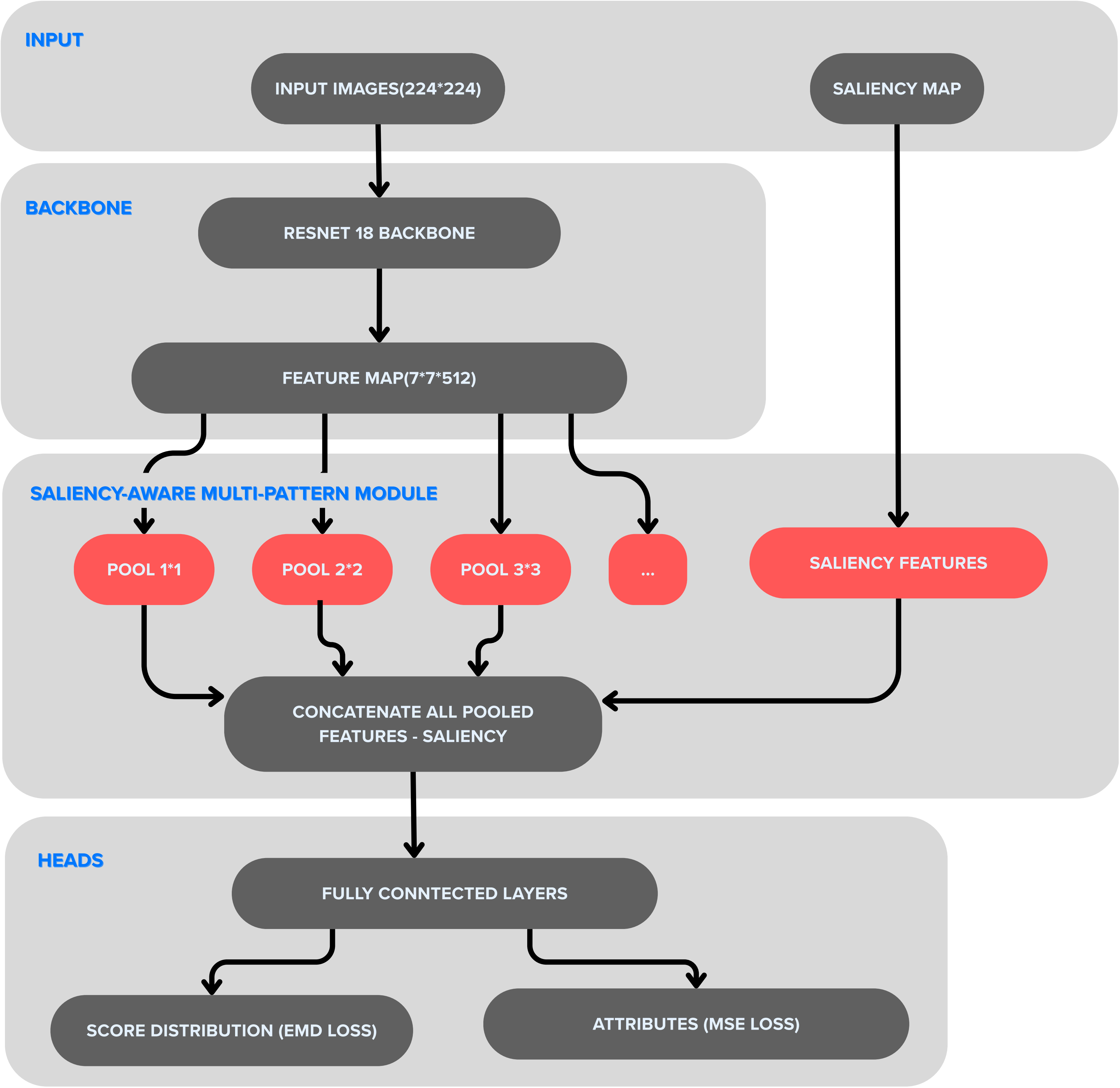
SRCC

Spearman Rank Correlation Coefficient

- Measures the accuracy of ranking (Monotonicity).
- Why? Aesthetics is relative. We care more about "A is better than B" than "A is 4.5".

Phase 1:

Baseline (SAMPNet/ResNet)



- **Architecture: SAMPNet (ResNet-18)**
- **Visual Focus: P1, P2, P3 (Multi-scale Pooling).**
- **Explanation:**
  - **We initially used multi-scale pooling to mimic "Spatial Pyramids".**
  - **Used simple Concatenation to fuse saliency maps.**

SRCC 0.65

Phase 2:

SNAP Net (transformer)

- Architecture: SNAP Net (Swin Transformer)
- Visual Focus: Window Attention.
- Explanation:
- Uses Self-Attention to capture Long-range dependencies.
- First introduction of Rank Loss to solve the ranking problem.

SRCC improved to 0.67

Attribute Loss (Auxiliary)

- Multi-task learning. Forces the model to understand Composition (Symmetry, DoF) to justify its score.
- Acts as a regularizer.

EMD Loss (Main)

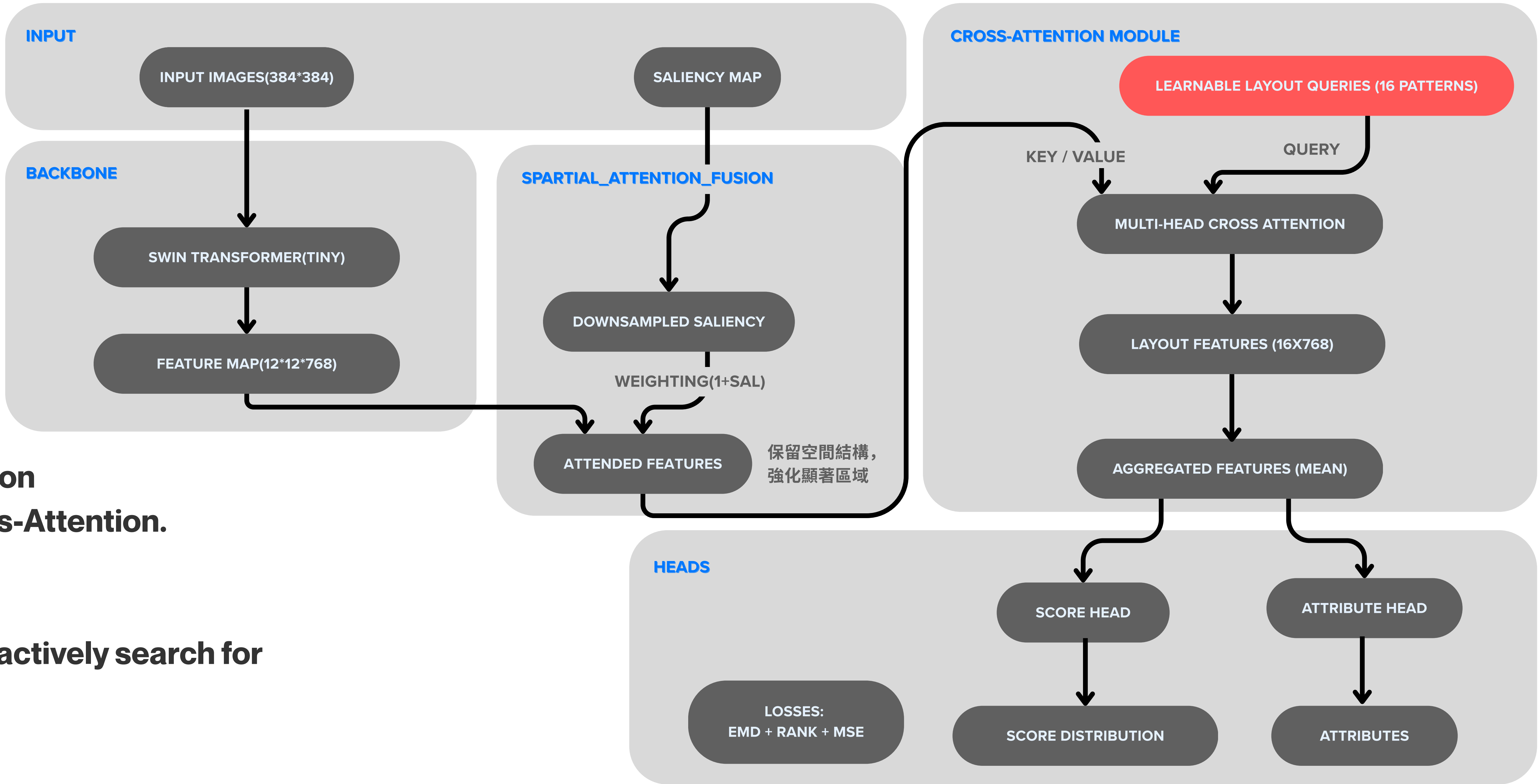
- Learns the shape of human opinion (the distribution).
- Penalizes "confident but wrong" predictions more than "uncertain" ones.

Rank Loss (The Secret Sauce)

秘

- Explicitly trains the model on pairs of images.
- Loss = max(0, -sign(True_Diff) * (Pred_Diff) + margin)
- Directly optimizes for SRCC.

Phase 3:



- **Architecture: Swin-T + Cross-Attention**
- **Visual Focus: Layout Queries & Cross-Attention.**
- **Explanation:**
- **This is the essence of Transformer.**
- **We designed 16 "Layout Queries" to actively search for composition patterns.**
- **Used (1+Sal) for saliency weighting.**

SRCC improved to 0.69

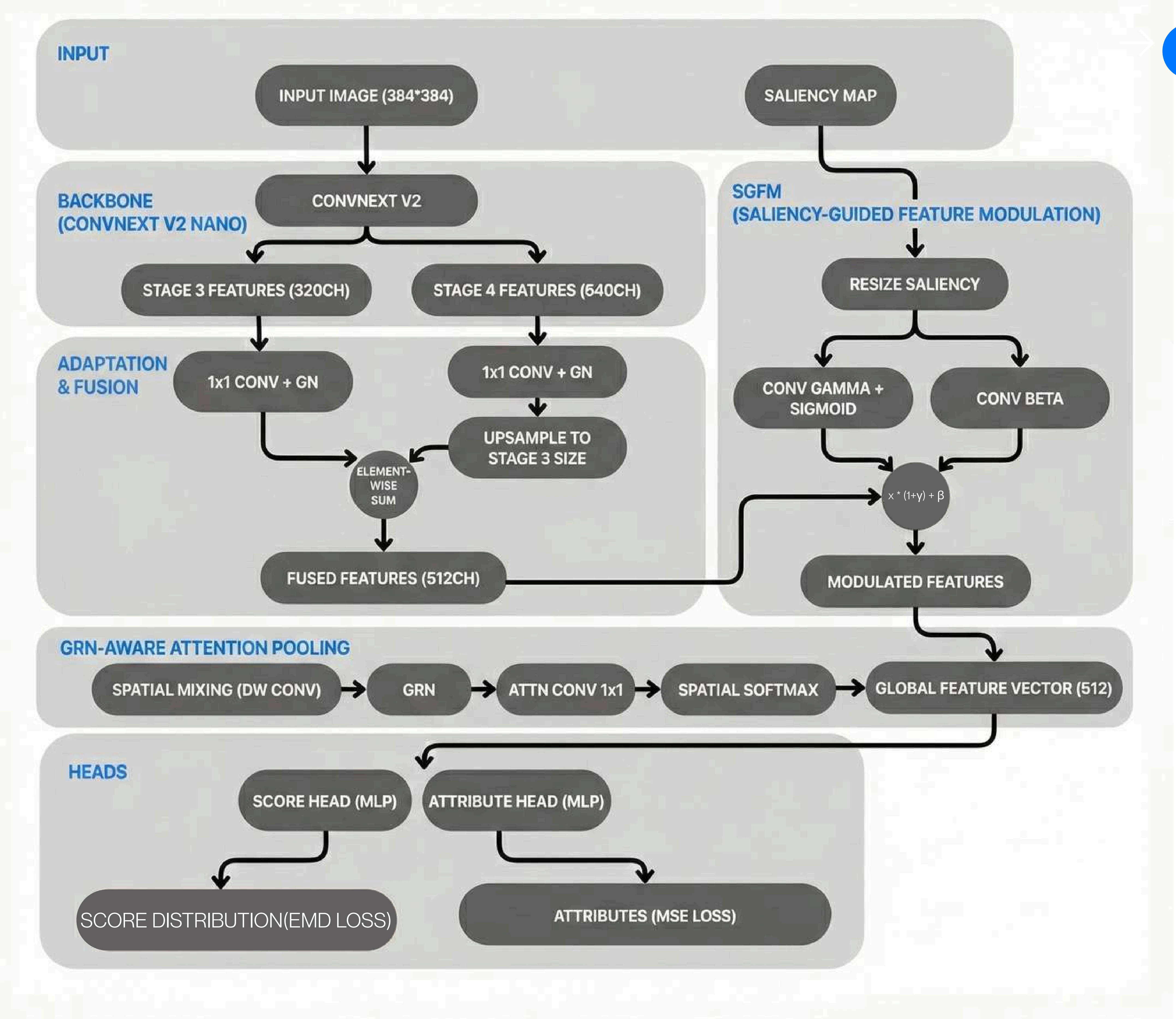
Group 17

Phase 4:

- Architecture: ConvNeXt V2 Nano
- Visual Focus: SGFM & GRN-Aware Pooling.
- Explanation:
- Used powerful SGFM (Affine Modulation) to recalibrate features.
- Added GRN in the pooling layer to filter invalid information, ensuring the model focuses on the most important aesthetic features.

SRCC improved to 0.71

baseline : 0.65



Ablation Studies

Xu, K., Deng, Y., Ding, M., Cheng, Z., & Lu, H. (2021). **Image composition assessment with saliency-augmented multi-patch network**. In Proceedings of the British Machine Vision Conference (BMVC 2021).

Method	MSE↓	EMD↓	SRCC↑	LCC↑
ResNet18	0.4534	0.1943	0.6025	0.6148
AADB [22]	0.4234	0.1923	0.6236	0.6415
MNA-CNN [32]	0.4260	0.1944	0.6108	0.6375
A-Lamp [31]	0.4230	0.1898	0.6270	0.6456
VP-Net [52]	0.4304	0.1948	0.6169	0.6285
RG-Net [28]	0.4398	0.1915	0.6026	0.6218
AFDC-Net [4]	0.4245	0.1910	0.6154	0.6388
SAMP-Net (Ours)	**0.3867**	**0.1798**	**0.6564**	**0.6709**
Our Model	**0.3627**	**0.1515**	**0.7128**	**0.7124**

Fusion Strategy

Concat vs Add vs Affine (SGFM)

Normalization Choice

BatchNorm vs LayerNorm vs GRN

Activation Function

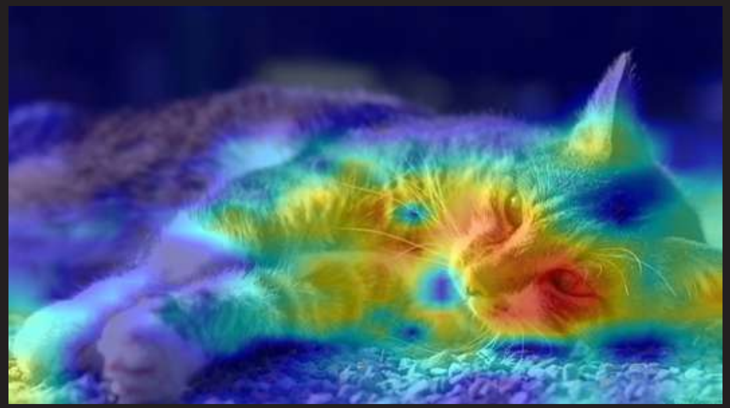
ReLU vs GELU

Loss Weight Sensitivity

Varying Rank Loss weight 0.2 (0.1~1)

GradCAM Sanity Check

origin image



GradCAM

Result

What our model can do?

Score Distribution (1~5)

- Our model outputs a full probability distribution (1-5), not just a mean score.
- Benefit: Distinguishes between "Universally Good" (sharp peak at 5) and "Controversial" (peaks at 1 and 5).

Granular Attribute Prediction:

- Predicts 6+ specific composition attributes (Rule of Thirds, Symmetry, etc.).
- Benefit: Provides **actionable feedback**. Instead of just saying "Bad photo", it says "Low Symmetry score".

Real-world Applications

AI Photography Assistant

Real-time composition suggestions in camera apps.

Aesthetic Quality Ranking

Analyzes and ranks burst sequences to identify the single best shot based on composition quality.

Composition-Aware Re-framing

Crop photos to any target aspect ratio while preserving well-composed regions.

Future Directions: Transfer Learning & Expansion

AADB (Attributes for Aesthetics Database)

The Data Gap

Contains rich attribute data but differs slightly from CADB.

Transfer Learning Strategy:

1. Pre-training on AADB: Learn fundamental aesthetic features (Color, Lighting) from AADB's larger scale.
2. Fine-tuning on CADB: Transfer this knowledge to CADB to learn specific composition rules (Rule of Thirds, Balancing Elements).

Leveraging More Fields

1. CADB offers unused metadata (e.g., object categories).
2. Future models can incorporate Semantic Awareness (e.g., "A portrait requires different composition than a landscape").

Group 17

Thanks