CVPR Nashville JUNE 11-15, 2025

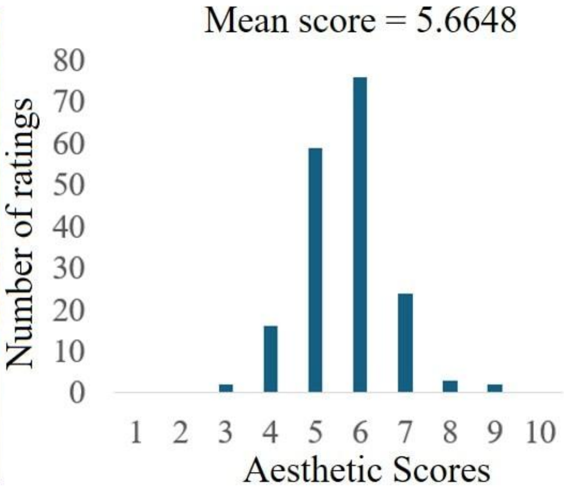
Charm: The Missing Piece in ViT Fine-Tuning for Image Aesthetic Assessment



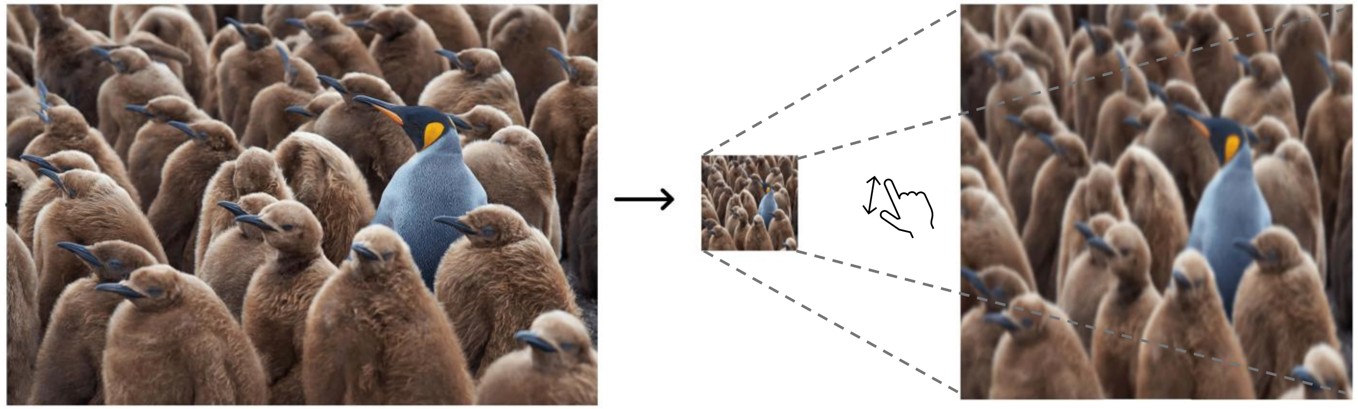
Fatemeh Behrad
Tinne Tuytelaars
Johan Wagemans

fwo erc KU LEUVEN

Image aesthetic assessment



Challenge



Original size

downscaled/
cropped input

Loss of
aesthetic
information

Current solutions

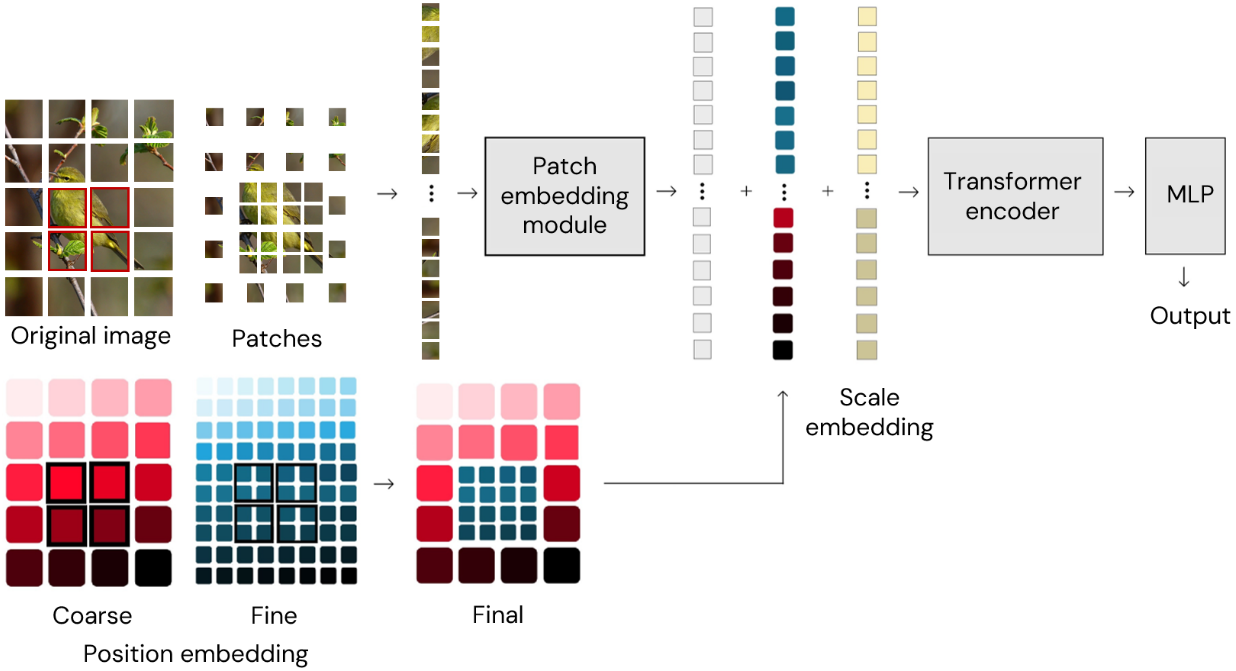
- Slow convergence using a batch size of one (CVPR 2016)
- Computationally expensive (CVPR 2020, ICPR 2022)
- Introduced for image classification and perform poorly on image aesthetic assessment (ICCV 2021, CVPR 2023, ICCV 2023, NeurIPS 2024)

Long Mai et al., Composition-preserving deep photo aesthetics assessment. CVPR 2016.
Qiuyu Chen et al., Adaptive fractional dilated convolution network for image aesthetics assessment. CVPR 2020.
Hossein Talebi and Peyman Milanfar. Learning to resize images for computer vision tasks. ICCV 2021.
Koustav Ghosal and Aljosa Smolic. Image aesthetics assessment using a graph attention network. ICPR 2022.
Jakob Drachmann Havtorn et al., Msvit: Dynamic mixed-scale tokenization for vision transformers. ICCV 2023.
Tomer Ronen et al., Vision transformers with mixed-resolution tokenization. CVPR 2023.
Mostafa Dehghani et al., Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. NeurIPS 2024.

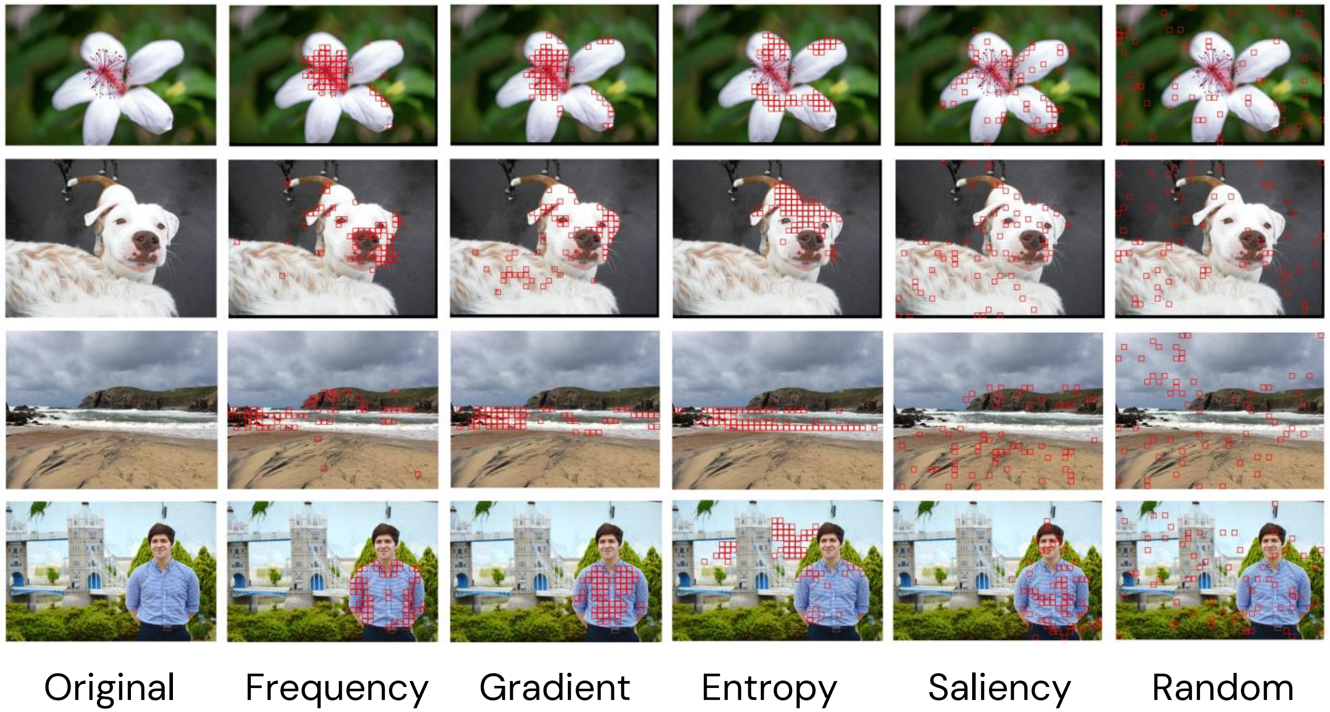
Method summary

Charm:

a novel tokenization approach that preserves **C**omposition, **H**igh-resolution, **A**spect **R**atio, and **M**ulti-scale information simultaneously.



Patch selection



Performance improvement over different *datasets*

Image aesthetic assessment

Dataset	Charm	PLCC	SRCC	ACC
AVA	-	0.734	0.732	0.808
	✓	0.779 (↑ 4.5%)	0.777 (↑ 4.5%)	0.826 (↑ 1.8%)
AADB	-	0.695	0.682	0.754
	✓	0.767 (↑ 7.2%)	0.754 (↑ 7.2%)	0.767 (↑ 1.3%)
TAD66k	-	0.429	0.401	0.646
	✓	0.488 (↑ 5.9%)	0.458 (↑ 5.7%)	0.794 (↑ 14.8%)
PARA	-	0.904	0.855	0.863
	✓	0.938 (↑ 3.4%)	0.905 (↑ 5%)	0.892 (↑ 2.9%)
BAID	-	0.428	0.342	0.750
	✓	0.439 (↑ 1.1%)	0.368 (↑ 2.6%)	0.763 (↑ 1.3%)

Image quality assessment

Dataset	Charm	PLCC	SRCC	ACC
SPAQ	-	0.911	0.907	0.907
	✓	0.919 (↑ 0.8%)	0.915 (↑ 0.8%)	0.917 (↑ 1.0%)
KonIQ10k	-	0.896	0.868	0.938
	✓	0.944 (↑ 4.8%)	0.93 (↑ 6.2%)	0.954 (↑ 1.6%)

Backbone: Dinov2-small

Performance improvement over different *backbones*

Model	Charm	PLCC	SRCC	ACC
Dinov2 -small	-	0.710	0.706	0.802
	✓	0.779 (↑ 6.9%)	0.777 (↑ 7.1%)	0.826 (↑ 2.4%)
ViT-small	-	0.687	0.679	0.794
	✓	0.762 (↑ 7.5%)	0.760 (↑ 8.1%)	0.827 (↑ 3.3%)
Dinov2 -large	-	0.734	0.732	0.808
	✓	0.783 (↑ 4.9%)	0.781 (↑ 4.9%)	0.828 (↑ 2%)

Dataset: AVA

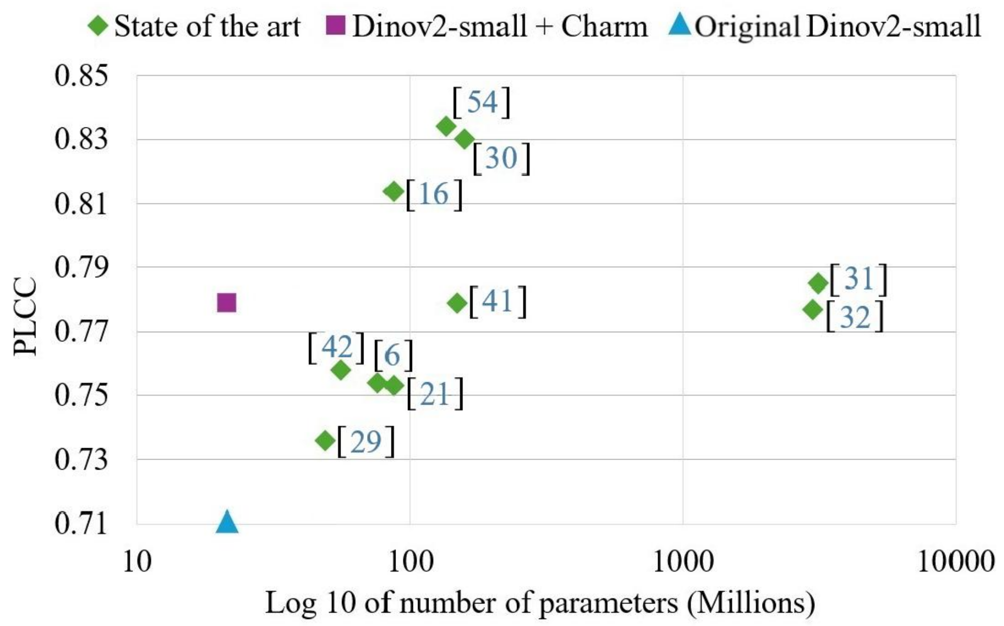
Comparison with existing methods

Model	AR	HR	MS	PLCC	SRCC	ACC
Ghosal et al.[12]	✓	-	-	0.764	0.762	-
Chen et al. [7]	✓	-	-	0.671	0.649	**0.832**
Dinov2-Small (+ Padding) *	✓	✓	-	0.709	0.703	0.801
MUSIQ [22]	✓	-	✓	0.738	0.726	0.815
FlexiViT [2]*	-	-	✓	0.737	0.735	0.812
Swin [33]*	-	-	✓	0.748	0.751	0.816
Dinov2-Small (+ Muller [46]) *	-	✓	-	0.682	0.675	0.794
ViT-small (+ Charm) *	✓	✓	✓	0.762	0.760	0.827
Dinov2-small (+ Charm) *	✓	✓	✓	<u>0.779</u>	<u>0.777</u>	0.826
Dinov2-large (+ Charm) *	✓	✓	✓	**0.783**	**0.781**	<u>0.828</u>

Our approach outperforms existing methods <u>in terms of preserving high resolution, aspect ratio, and multiscale information.</u>

State-of-the-arts in image aesthetic assessment

1. Multimodal models (text/attributes)



State-of-the-art models' performance on the AVA dataset

Charm achieves comparable performance using *only* image features and *fewer* parameters.

State-of-the-arts in image aesthetic assessment

2. Focus on other challenges in image aesthetic datasets:

   Example: long tails of rating distributions in image aesthetic datasets (ELTA)

Charm	ELTA	PLCC	SRCC	ACC
-	-	0.734	0.732	0.808
-	✓	0.742	0.742	0.811
✓	-	0.783	0.781	0.828
✓	✓	**0.787**	**0.786**	**0.829**

Dinov2-large performance on the AVA dataset

Charm can be *integrated* with state-of-the-art approaches for further performance improvement.

Computational analysis of Charm

Model	Input size	Charm	#tokens	ms	GMACs	MB
Dinov2 -small	224 x 224	-	256	**5.7**	**6.11**	**202.9**
	640 x 640	-	2070	32.8	84.01	2091.8
		✓	2-scale:512	7.3 (↓ 77.7%)	13.46 (↓ 84%)	346.0 (↓ 83.5%)
		✓	3-scale:700	9.3 (↓ 71.6%)	19.60 (↓ 76.7%)	494.3 (↓ 76.4%)

Dinov2-small inference cost breakdown for processing one single image.

Conclusion

Charm balances computational cost with the preservation of crucial aesthetic information for achieving optimal performance in image aesthetic assessment.

Any questions?

fatemeh.behrad@kuleuven.be

https://arxiv.org/abs/2504.02522