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

Fatemeh Behrad, Tinne Tuytelaars, Johan Wagemans

KU Leuven University, Belgium

KU LEUVEN
fwo erc GitHub

CVPR Nashville JUNE 11-15, 2025

Motivation

Original size

Fixed-size input (downscaled/cropped) suitable for batch processing and efficient computation
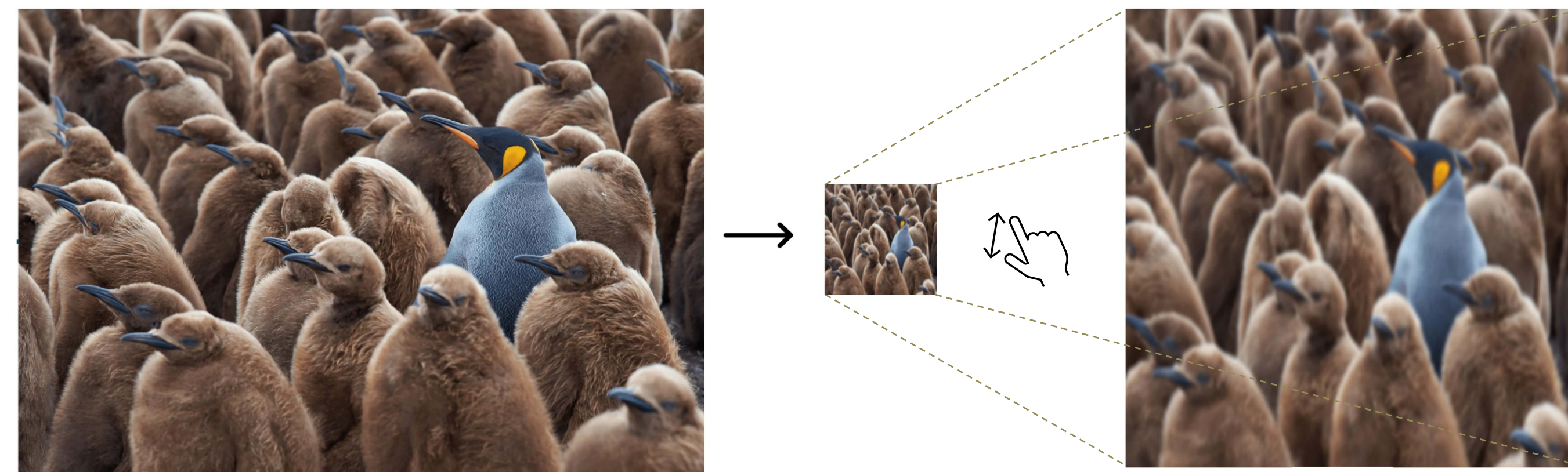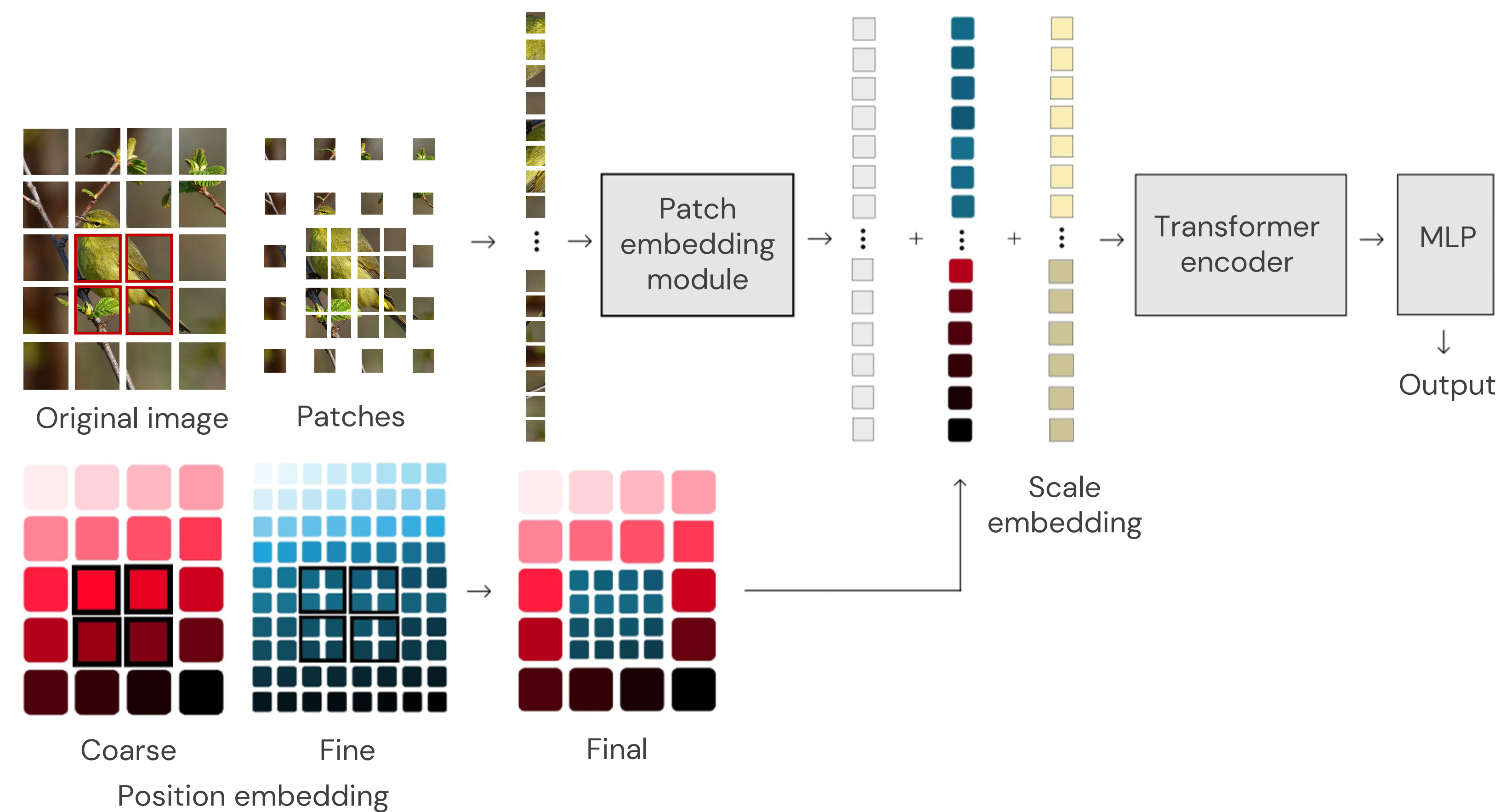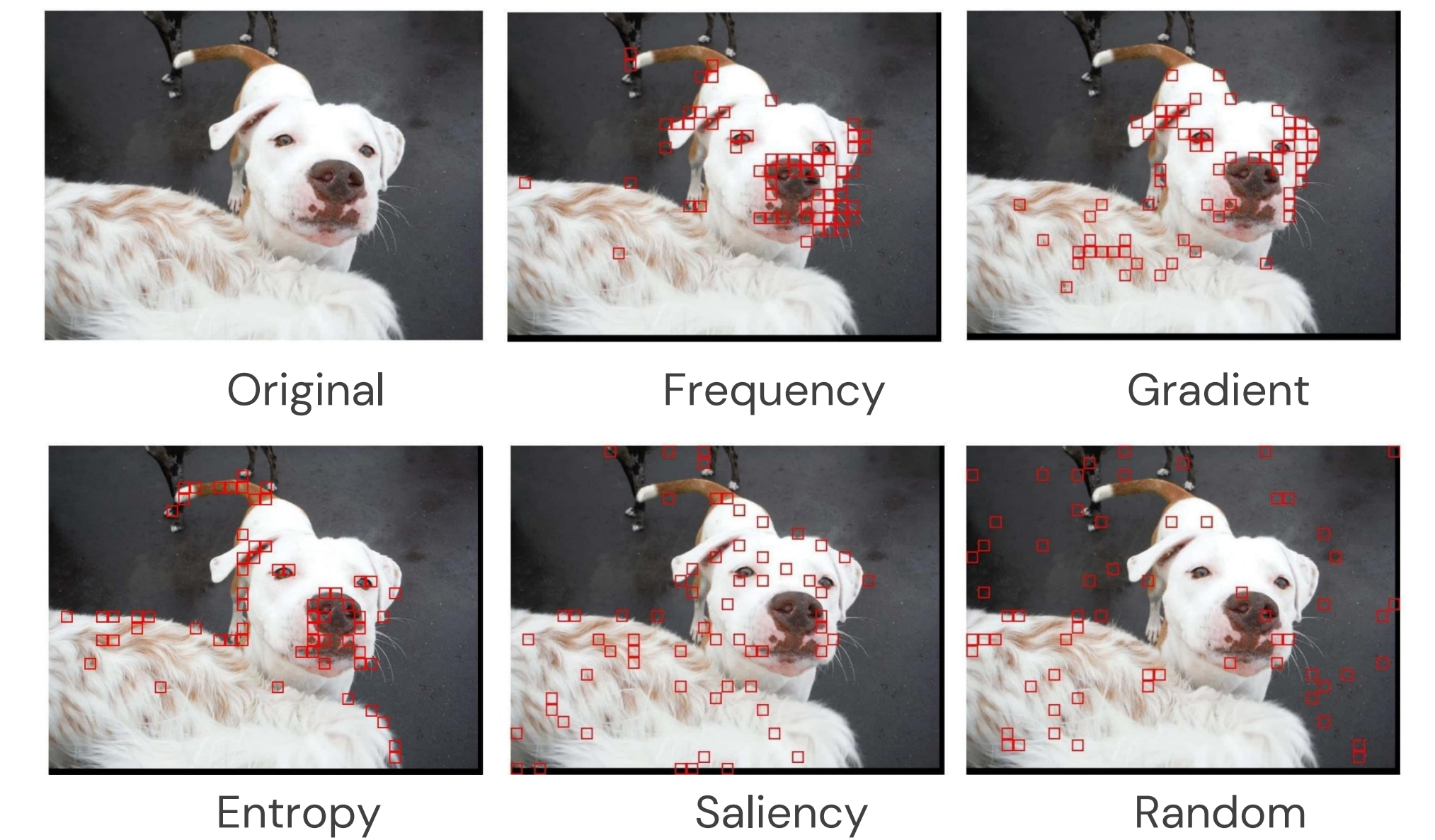
🚫 Loss of aesthetic information

- Downscaling and cropping can **distort aesthetic properties** by altering composition, changing aspect ratios, and removing high-resolution details, resulting in **inaccurate** image aesthetic assessment (IAA).
- Existing solutions often **slow convergence**, **degrade IAA model performance**, and **increase computational costs considerably**.

Proposed approach

Original image — Patches

Coarse — Fine — Final
Position embedding

Patch embedding module → + Scale embedding + → Transformer encoder → MLP → Output

- **Main idea**: Keep important regions in high resolution while downscaling the rest.

Original — Frequency — Gradient
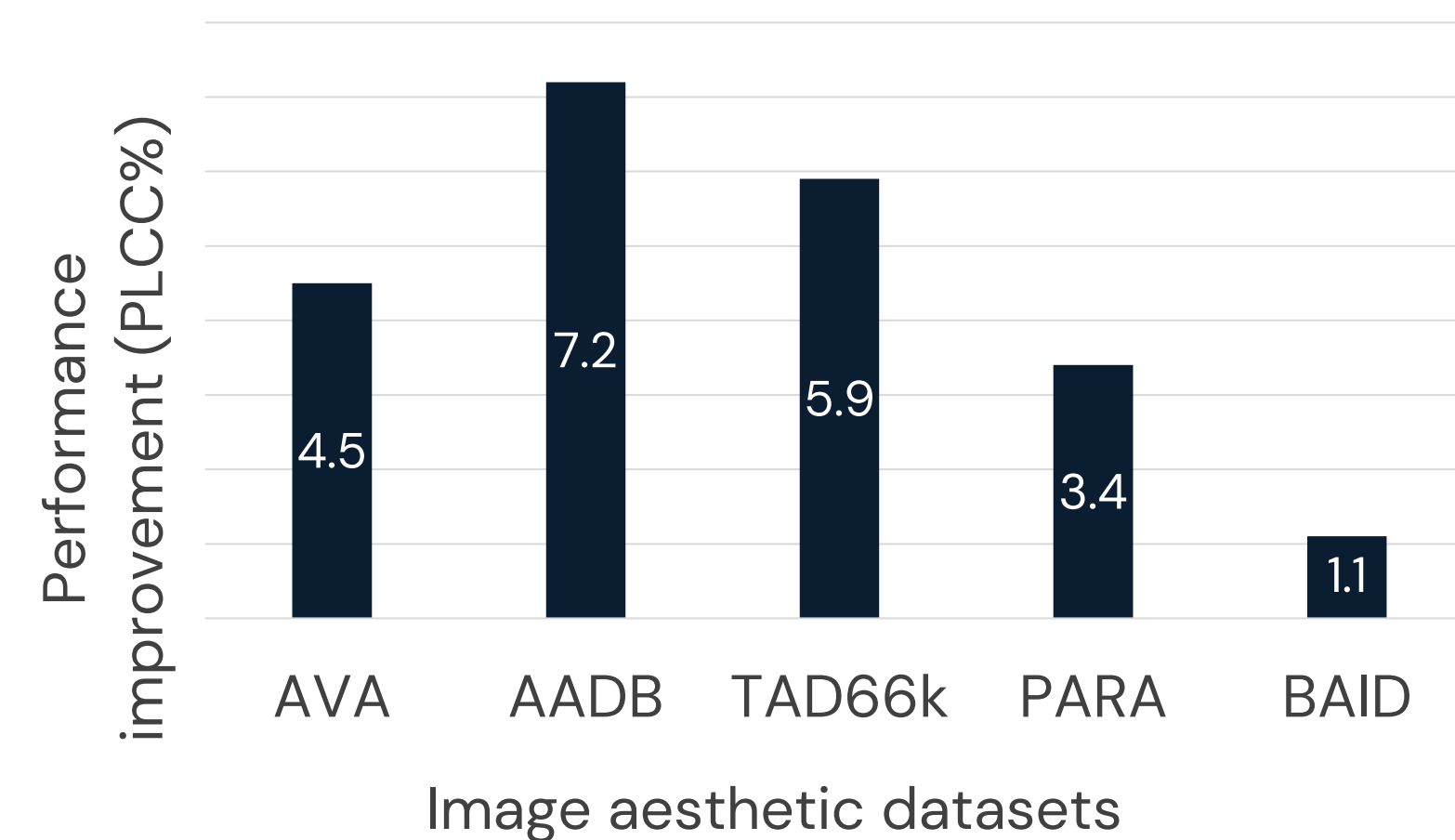Entropy — Saliency — Random
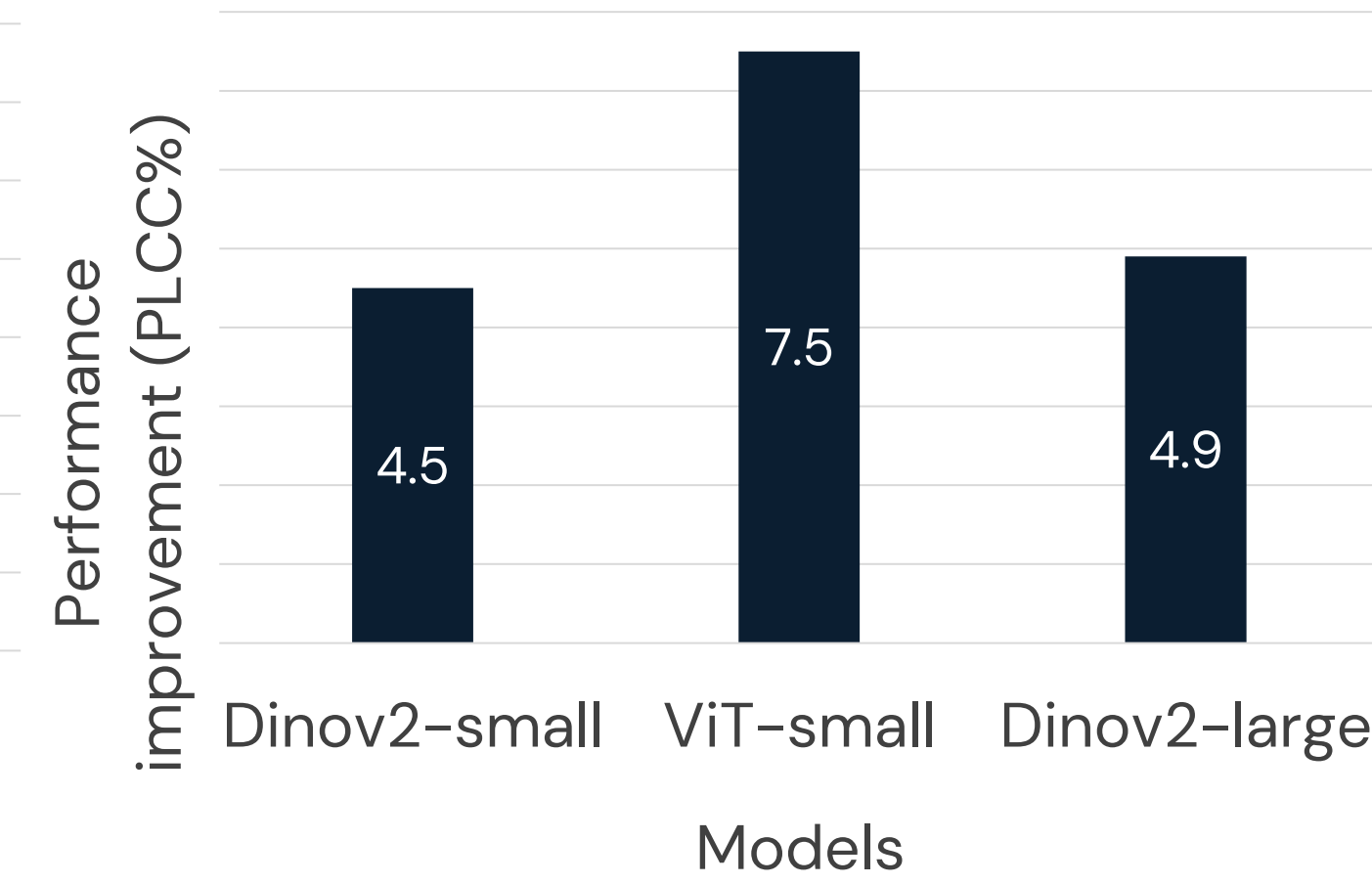
Patch selection strategies

Best approach:
- Training: Random
- Inference: Frequency

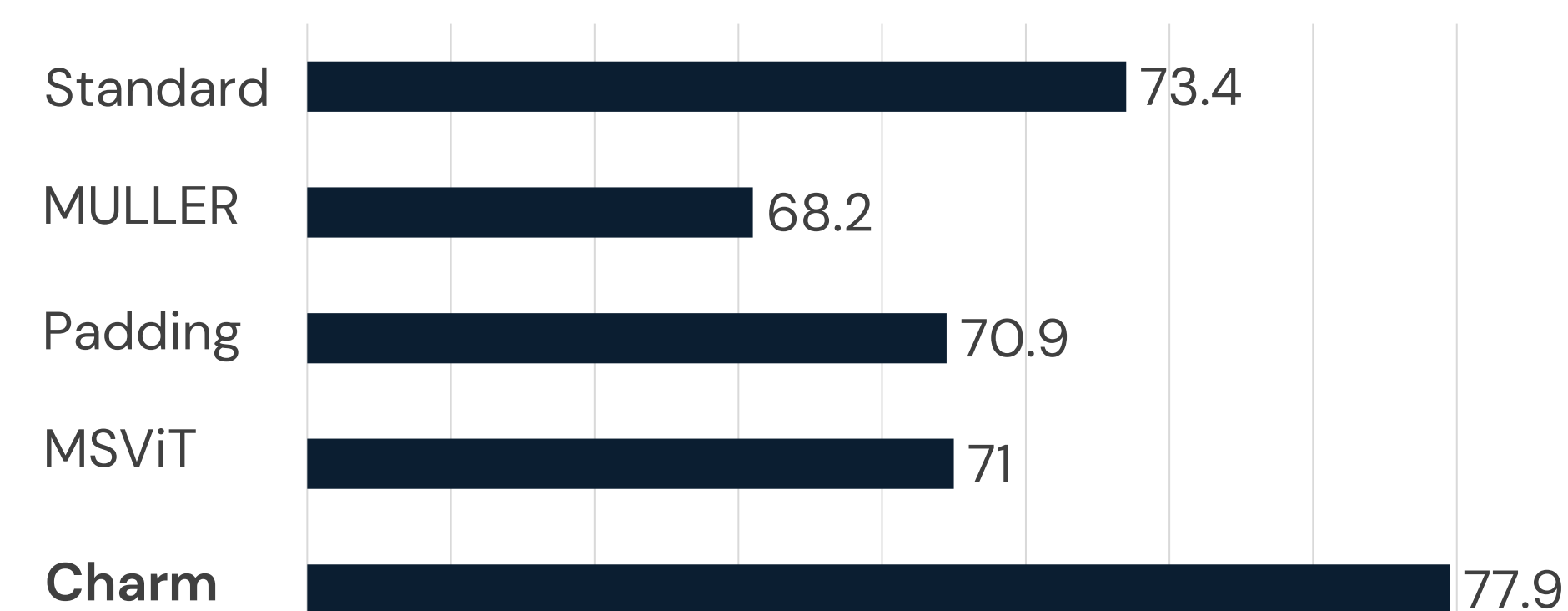## Evaluation

Performance improvement of Dinov2-small with Charm

Performance improvement of different models with Charm on the AVA dataset

	PLCC%
Standard	73.4
MULLER	68.2
Padding	70.9
MSViT	71
Charm	77.9

Dinov2-small performance (PLCC%) using different image preprocessing approaches on the AVA dataset

Model	Input size	Charm	#tokens	ms	GMACs	MB
Dinov2 -small	224 x 224	-	256	**5.7**	**6.11**	**202.9**
		-	2070	32.8	84.01	2091.8
	640 x 640	✓	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 a single image.

- Charm **significantly improves the performance** of various ViTs on different IAA datasets while remaining **computationally efficient**.

Conclusion

- **Charm** boosts IAA model performance by efficiently preserving **C**omposition, **H**igh resolution, **A**spect **R**atio, and **M**ulti-scale information without overwhelming computational resources.