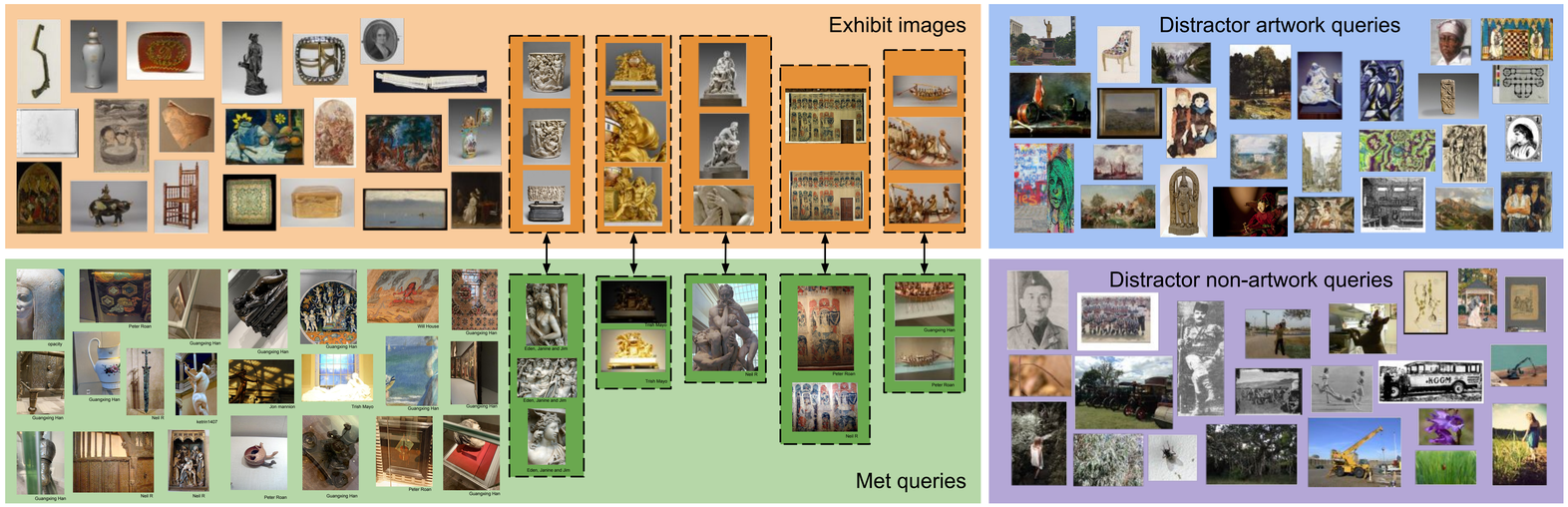
The Met Dataset: Instance-level Recognition for Artworks

Nikolaos-Antonios Ypsilantis[1], Noa Garcia[2], Guangxing Han[3], Sarah Ibrahimi[4], Nanne Van Noord[4], Giorgos Tolias[1]

[1]Czech Technical University in Prague, [2]Osaka University, [3]Columbia University, [4]University of Amsterdam

NEURAL INFORMATION PROCESSING SYSTEMS



Summary

New dataset for **large-scale Instance-level Recognition** (ILR) for the domain of artworks

Dataset collection

- **Exhibit images**: training set
  - Studio condition images from the Met museum in New York, depicting multiple views of artworks
  - Each artwork: one class
- **Met queries**: test/val set
  - Images of exhibits taken by Met visitors
  - Collected from Flickr + our own photos
- **Distractor queries**: test/val set
  - Artwork (non-Met) and non-artwork categories
  - Collected from Wikimedia Commons

Set	Type	#Images			#Classes
		Met	Distr. art.	Distr. non-art.	
Train	Exhibit	397,121	-	-	224,408
Val	Query	129	1,168	868	111 + 1
Test	Query	1,003	10,352	7,964	734 + 1

+1 class denotes distractors

Proposed benchmark

- **Task**: ILR for artworks
  - Predict class and classification confidence for query images
- **Evaluation metrics**
  - Accuracy (ACC): on Met queries only
  - Global Average Precision (GAP): on all queries

Dataset challenges

- **High inter-class similarity**

- **Domain shift between training and query set**

- **Robustness to (out-of-distribution) distractor queries**
  - GAP penalizes confident predictions for distractors
- **Long-tailed**
  - ∼60% of classes are represented by a single image

Experiments and results

- **Arc-face loss is superior to Cross-entropy**
  - DNet: representation + cosine classifier

Training method (ResNet18 backbone)	GAP	ACC
DNet + Cross-entropy	9.6	30.6
DNet + Arc-face [1]	16.9	36.6

- **Parametric classifier performs worse than kNN**

Training method (ResNet18 backbone)	GAP	ACC
DNet + Arc-face	16.9	36.6
DNet + Arc-face + kNN	23.7	47.4

- **Best performing model: self-supervised + supervised representation learning + kNN**

Training method (ResNet18 backbone)	GAP	ACC
ImageNet pre-training only	15.9	42.3
SimSiam [2] (self-supervised)	26.8	45.6
Con-Syn (contrastive)	30.4	49.4
Con-Syn+Real (contrastive)	29.8	48.8
Con-Syn+Real-closest (contrastive)	**32.5**	**50.0**

  - Pairs used by training methods

- **Pretraining comparisons**
  - SfM landmarks [3]: similar task, different domain
  - SemArt dataset [4]: different task, similar domain
  - Semi-weakly supervised (1B images) [5]: large-scale pretraining

Pretraining method (ResNet50 backbone)	GAP	ACC
ImageNet	22.2	46.4
SfM landmarks	26.6	48.6
SemArt (author attribute)	1.8	18.0
SemArt (type attribute)	7.9	31.9
Semi-weakly supervised (1B images)	**30.4**	**56.3**

References: **[1]**: Deng et al.: Arcface: Additive angular margin loss for deep face recognition, **[2]**: Chen & He: Exploring Simple Siamese Representation Learning, **[3]**: Radenovic et al.: Fine-tuning CNN Image Retrieval with No Human Annotation, **[4]**: Garcia et al.: Context-aware embeddings for automatic art analysis, **[5]**: Yalniz et al.: Billion-scale semi-supervised learning for image classification