SynGallery: A Synthetic Gallery of Real Paintings for Instance-Level Artwork Recognition

Patryk Bartkowiak · Jakub Markil · Bartosz Kotrys · Dominik L. Michels · Sören Pirk · Wojtek Pałubicki

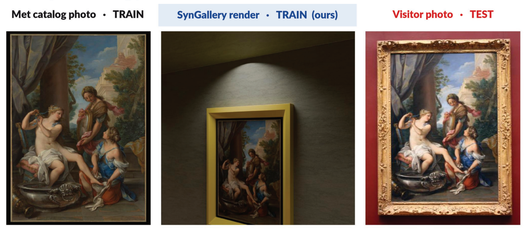
ECCV 2026 | MALMÖ | SEPT 8–12

Train on synthetic galleries, test on visitor photos

Museums already own clean catalog scans of their art. We place each scan in a procedurally generated 3D gallery and render it from five viewpoints. At an equal number of training images, a model trained on the renders beats one trained on the museum's own studio photographs.

- **SynGallery:** 24,490 renders of 4,898 real Met paintings, five viewpoints each. The catalog image *is* the canvas texture, so instance identity is preserved exactly.
- **Stronger signal than real photos:** at a matched 12,403-image budget, synthetic-only raises closed-world GAP⁻ from 67.18 to **73.47**. Added to the full Met set it beats the published benchmark model, 36.1 → **38.48** GAP.
- **Variation, not realism:** viewpoint coverage accounts for most of the gain. Simulating phone-camera artifacts (blur, noise, JPEG) reduces performance.

One painting, three domains



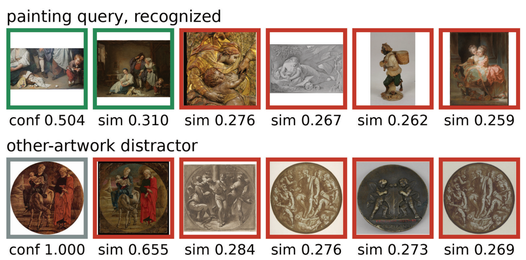
The domain gap. The Met benchmark trains on frontal studio scans and tests on handheld visitor photos. A SynGallery render keeps the scan's exact pixels as the canvas and adds the conditions of a real visit: viewing angle, gallery lighting, a frame, sometimes glass.

The Met benchmark setup

397k catalog training images (224,408 exhibit classes)	**60.8%** of classes have exactly one catalog photo
1,003 real visitor query photos taken on site	**18,316** distractor queries that must be rejected, not matched

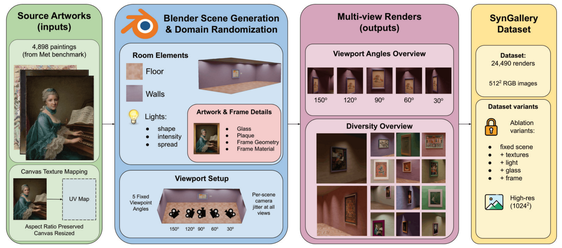
- Train and test look nothing alike: clean frontal scans vs. handheld shots at an angle, under gallery light, often through glass.
- Collecting real visitor photos per artwork does not scale - but every museum already has a digitized catalog.

The open-set retrieval task



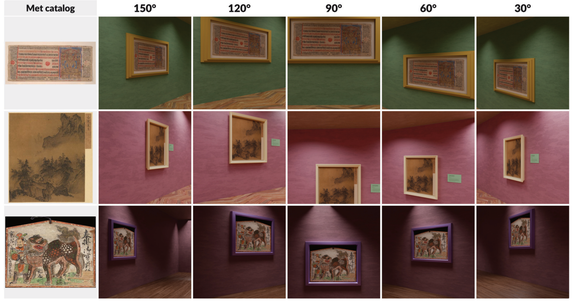
Open-set retrieval with our best model. A query (left, with its kNN-softmax confidence) and its five nearest database images (cosine similarity); green frames mark the query's own class. **Top:** a real visitor photo, correctly recognised. **Bottom:** a distractor with no correct match returns visually similar Met exhibits at confidence 1.000 - exactly what GAP penalises.

Procedural 3D scene generation



Identity-preserving rendering. The catalog image becomes the canvas texture, scaled to its own aspect ratio - never cropped, retouched or replaced. Everything *around* it is resampled per painting: wall/ceiling/floor materials, Poisson-disk ceiling lights, one of 27 frame geometries, plaque (p=0.5), glass (p=0.25) and per-view camera jitter. Blender 5.1 / Cycles, 512² RGB (1024² variant).

Same artwork, five viewpoints



Five fixed cameras sit on a horizontal arc at 30°–150° around the canvas, all aimed at the painting. Oblique views foreshorten and clip it exactly as a visitor's photo does; the distortion comes from the camera pose - the painting itself is never warped or edited.

What the generator produces

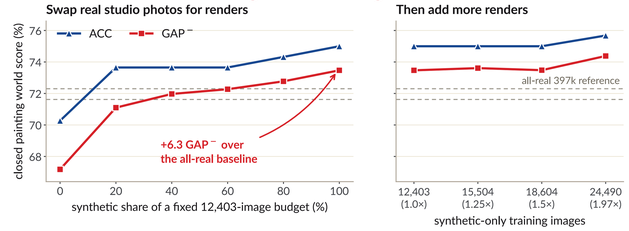
Uncurated random sample. The spread of rooms, floors, wall colours, frames, plaques, lighting and viewpoints across the 24,490 renders. No render is hand-picked or hand-tuned.

Full Met benchmark

model (training data)	GAP	GAP⁻	ACC
best published single model	36.1	52.4	55.0
our reproduction (397k real)	35.97	52.14	54.64
from scratch, 397k real + **SynGallery**	38.48	55.04	**57.63**
fine-tuned, + **SynGallery** only	**38.99**	**55.15**	57.23
fine-tuned, + **SynGallery** + real	38.66	53.73	55.53

1,003 Met queries + 18,316 distractors against the 397k database. Identical architecture (ResNet-18 SWSL, GeM pooling, whitening FC, contrastive loss) - the *only* change is adding synthetic gallery views. No extra real data, no retrieval-method change.

Real-to-synthetic mixing and scaling

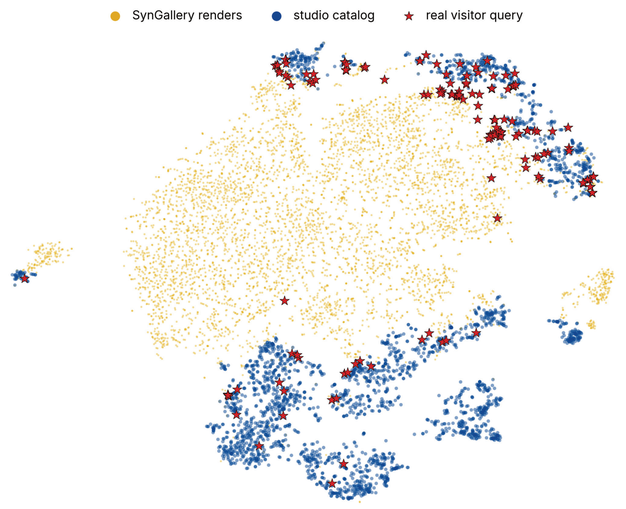


Closed painting world, 148 real visitor queries. GAP⁻ rises at every mixing step from all-real to all-synthetic; scaling synthetic-only to all 24,490 renders reaches GAP⁻ **74.38** / ACC **75.68** - above the 397k all-real reference (71.62 / 72.30) with roughly **16× fewer training images**.

Evaluation protocol

- **Full benchmark:** 1,003 Met queries + 18,316 distractors (10,352 other-artwork, 7,964 non-artwork) against all 397,121 catalog images.
- **Closed painting world:** 148 real visitor queries of the 4,898 SynGallery painting classes against their 12,403 catalog photos. No distractors, so it isolates recognition from rejection.
- **Scoring:** kNN on cosine similarity of multi-scale descriptors, class confidence temperature-softmaxed; GAP ranks every query by confidence. GAP⁻ drops the distractors, GAP keeps them - so high GAP needs calibration, not just accuracy.

DINOv3 embedding analysis



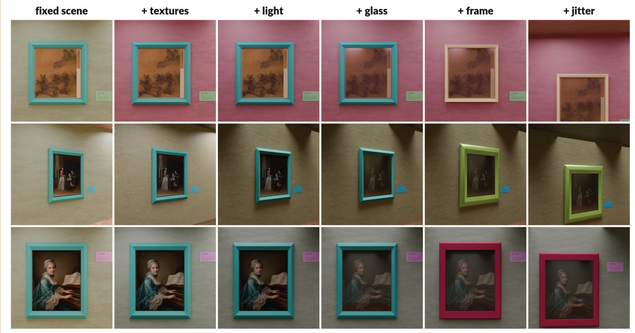
DINOv3 ViT-L embeddings, t-SNE. Studio catalog photos form compact clusters and the real visitor queries sit on top of them. The renders form a cluster of their own, *no closer* to the visitor photos than the catalog scans are. So their advantage as training data cannot come from resembling the test domain - it comes from the view variation they add.

Dataset ablation

training renders (synthetic-only)	GAP⁻	ACC
fixed scene (no randomization)	73.78	75.00
+ textures	73.94	75.00
+ light	74.42	75.68
+ glass (p=0.25)	75.32	**76.35**
+ frame	72.61	73.65
+ camera jitter *(default)*	74.38	75.68
default at 1024²	**75.41**	**76.35**
default, 3 viewpoints (14,694)	74.31	75.68
default, frontal view only (4,898)	**68.55**	**70.27**
all-real reference (12,403 photos)	67.18	70.27

Closed painting world: 148 real visitor queries against the 12,403-image painting database, k/τ by 2-fold cross-validation. Scene randomization has a mixed, modest effect: all factors combined add 0.6 GAP⁻ over a fixed scene. **Collapsing five viewpoints to one frontal view costs 5.8 GAP⁻** - nearly the entire advantage over real photos.

The randomization ladder



Each column adds one factor to the same scene; the last is the default SynGallery configuration. The renders look very different at each step, yet the full ladder moves GAP⁻ by 0.6 (73.78 → 74.38). Scene detail is not what the model learns from.

Effect of simulated capture artifacts

degradation injected during training (p=0.5)	Δ GAP⁻
JPEG compression	−0.2
sensor noise + resolution loss	−1.7
motion blur	−2.9
all three families combined	−3.8

Real queries *do* contain these artifacts, so simulating them might be expected to help. Instead it destroys the fine evidence instance recognition depends on - brushwork, texture, small compositional detail - and weakens the embedding more than it prepares the model for degraded queries. Realism does not improve either: under DINOv3, every artifact family moves the renders *further* from real visitor photos.

Takeaways

- Synthetic **scene-level view variation** is a stronger training signal for instance-level artwork retrieval than the museum's own studio photographs.
- **Viewpoint coverage is the single biggest factor.** Photorealism is not required, and simulated capture artifacts reduce performance.
- Any digitized collection can be turned into gallery training data this way - no new photography.
- **Next:** synthetic distractor classes (statues, furniture) to close the open-set gap; depth, masks and camera parameters; stronger occlusion and reflections.

Project syngallery.github.io
Code github.com/SynGallery/syngallery-bench
Data huggingface.co/datasets/patryk-bartkowiak/SynGallery
Paper arxiv.org/abs/2607.18907

Supported by the European Union through EOSC-ARENA (GA No. 101292597) and by the PIAST-AI Factory project, co-financed by the EuroHPC Joint Undertaking through the Digital Europe Programme and by Poland (GA No. 101250730). Conducted in collaboration with ArtiCollect.