Predicting aesthetic appreciation of images

Naila Murray
NAVER LABS Europe
2nd April 2019

Data-driven approaches to aesthetic prediction

Most prediction problems in computer vision involve largely objective properties:

- Object recognition
- Semantic segmentation
- Local feature matching
- Etc.

... and are successfully tackled using data-driven, machine learning methods.

Can we learn models for visual aesthetics using data-driven methods?

Data-driven approaches to aesthetic prediction

Requirements for data-driven aesthetic prediction:

- Annotated datasets
- Effective image representations

Global outline

AVA: A large-scale dataset for aesthetic visual analysis
Collaborators:
Luca Marchesotti
Florent Perronnin

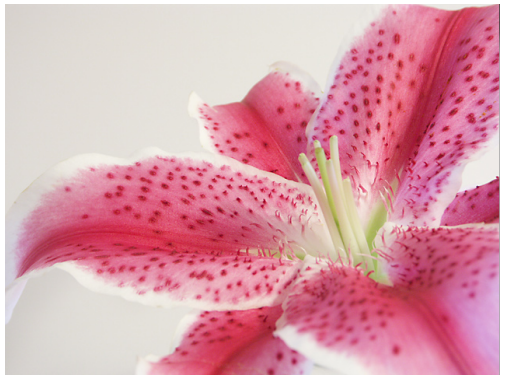
A deep architecture for unified aesthetic prediction
Collaborators:
Albert Gordo

Global outline

AVA: A large-scale dataset for aesthetic visual analysis
Collaborators:
Luca Marchesotti
Florent Perronnin

A deep architecture for unified aesthetic prediction
Collaborators:
Albert Gordo

Aesthetic prediction

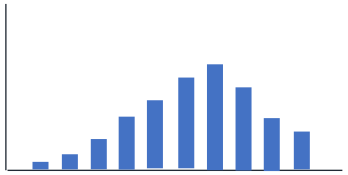
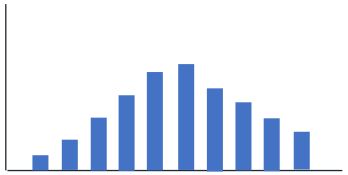
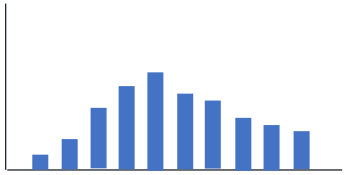
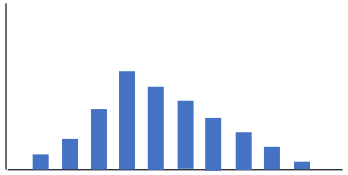



Binary classification


Mean score prediction

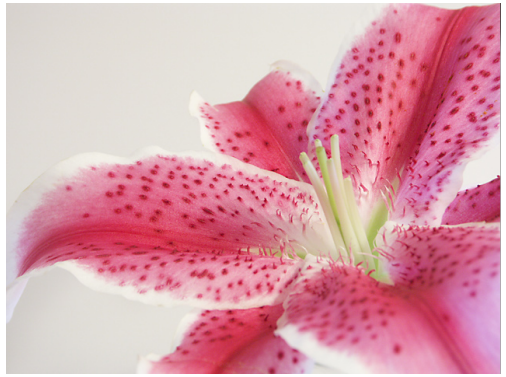
7.65 **6.32** … **4.81** **4.15**

Score distribution prediction

NAVER LABS

Machine learning-based approaches

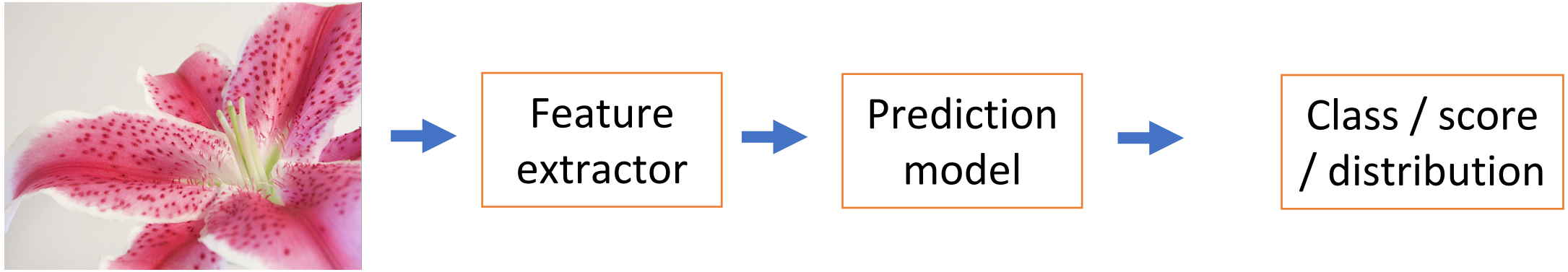


Feature extractor

Prediction model

Class / score / distribution

Machine learning-based approaches

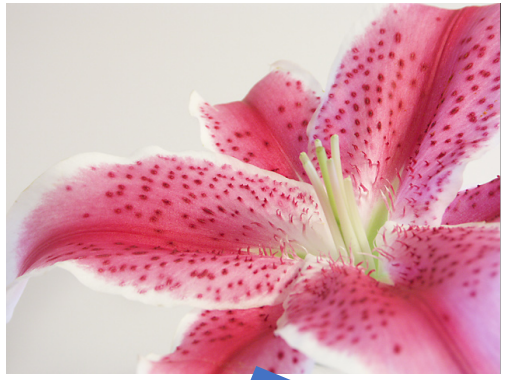


Features:

- Handcrafted for aesthetics: e.g. color histograms, saturation, composition [1]
- Generic: e.g. SIFT features [2]

[1] Datta et al. Studying aesthetics in photographic images using a computational approach. ECCV 2006.
[2] Marchesotti et al. Assessing the aesthetic quality of photographs using generic image descriptors. ICCV 2011.

Machine learning-based approaches



Feature extractor

Prediction model

Class / score / distribution

Deep model
(feature extractor / predictor)

Data for training aesthetic prediction models

Aesthetic appreciation is ***subjective***, i.e. there is no “right” answer.

Most annotations capture a consensus opinion or statistic, *i.e.* “How would most people score or label this image?”

Because of subjectivity, training requires large-scale data:

- Large number of images
- Large number of annotations per image

Data for training aesthetic prediction models

Public datasets collect data from 3 main sources:

Flickr:
- Image-sharing website
- Users can rate images from 1-5 stars

Photo.net:
- Online community for photographers
- Images are scored from 1-7

Dpchallenge:
- Online community for photographers
- Photography challenges are central
- Submissions are scored from 1-10

DPChallenge.com

TITLE: Skyscape

Description:

Make the sky the subject of your photo this week.

Stats

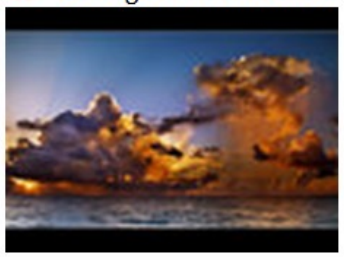
Voting Dates:
13/07/2010 - 19/07/2010

Numbers & Statistics:
Submissions: 136
Disqualifications: 1
Votes: 16,009
Comments: 595

Average Score: 5.64014

1st place with an average vote of **7.4831**

2nd place with an average vote of **7.0328**



3rd place with an average vote of **6.9333**

4th place with an average vote of **6.8547**

5th place with an average vote of **6.7073**

6th place with an average vote of **6.6667**

7th..

8th..

9th..

NAVER LABS

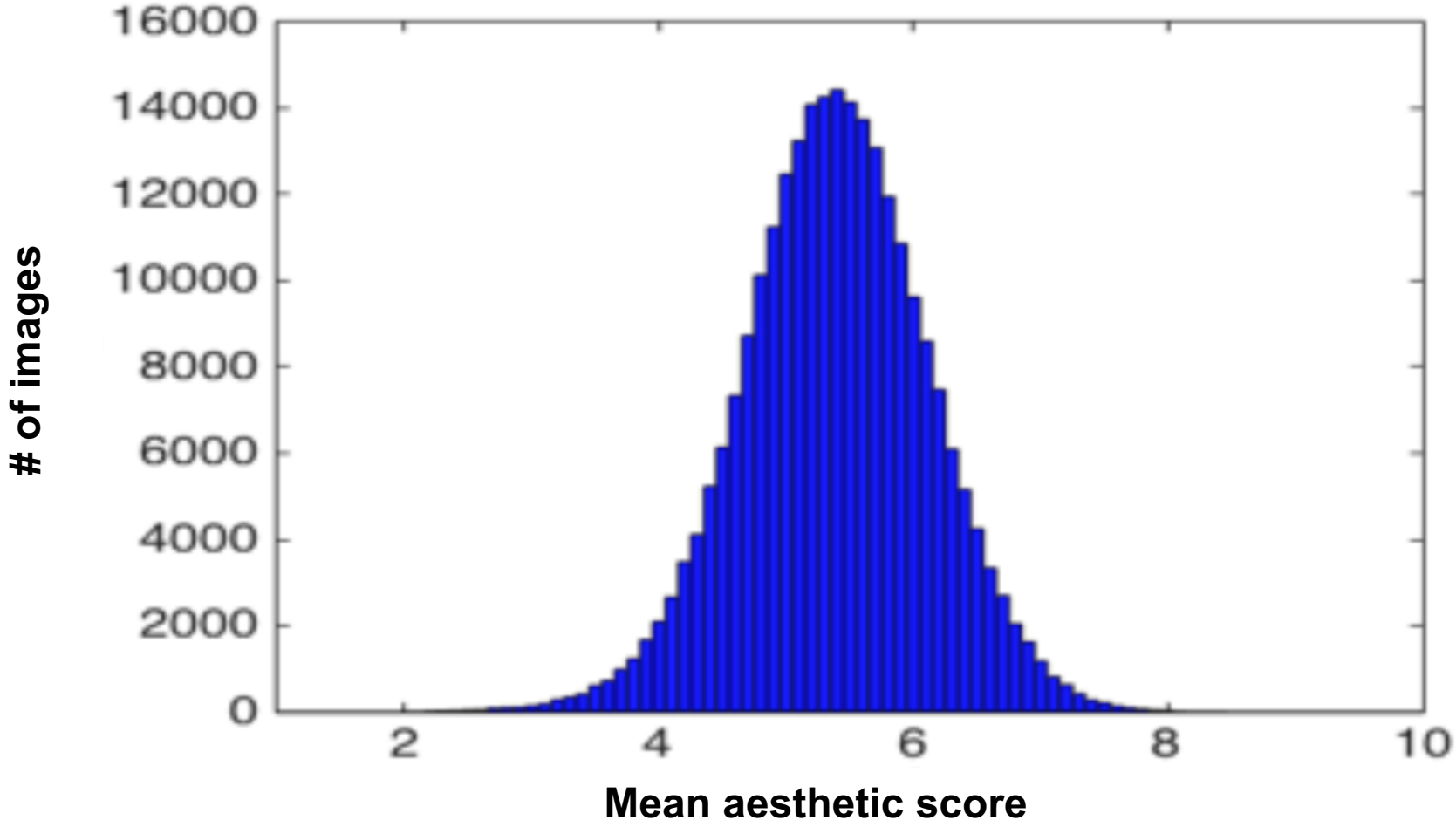
AVA: A large-scale dataset for aesthetic visual analysis

AVA [1,2] at a glance:

- # of images: approx. 255K
- # of associated challenges: 963
- # of votes per image: 78 – 549 (210 on average)
- Semantic annotations: approx. 200K images with 1 tag; 66 tags in total
- Weak style annotations: 14K images annotated with 14 style categories
- Weak textual annotations: 2.9M comments

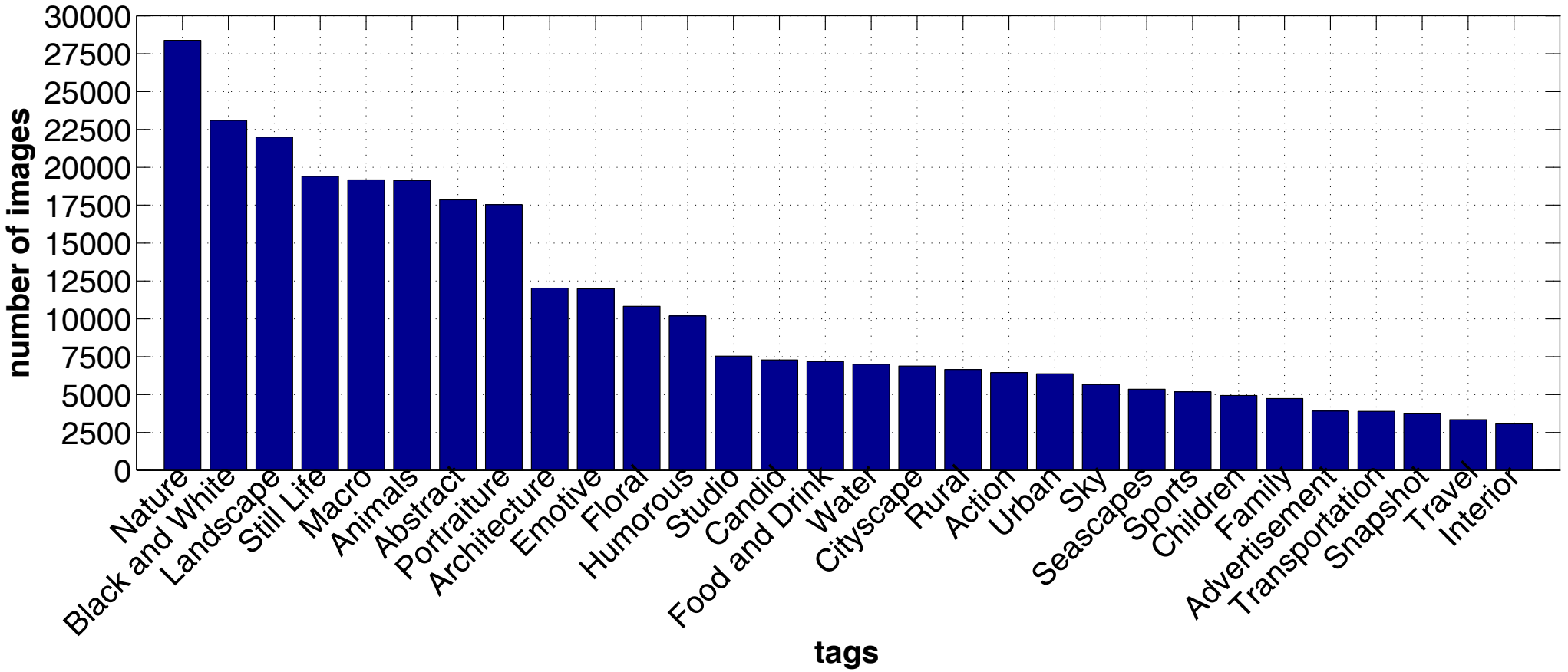
[1] Murray et al. AVA: A Large-Scale Database for Aesthetic Visual Analysis. CVPR 2012.
[2] Marchesotti et al. Discovering beautiful attributes for aesthetic image analysis . IJCV 2015.

AVA: mean score distribution



NAVER LABS

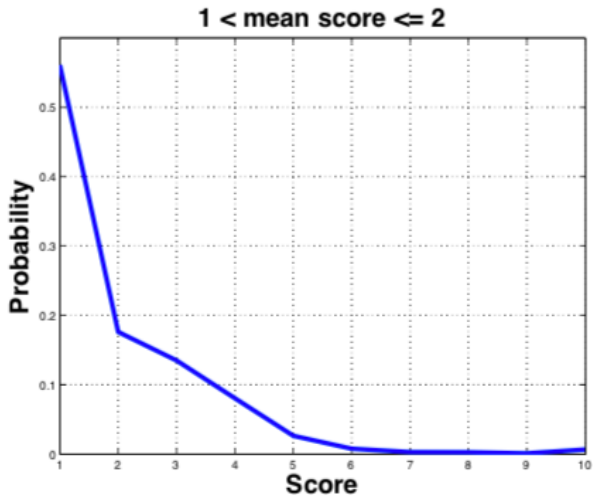
AVA: distribution of semantic labels



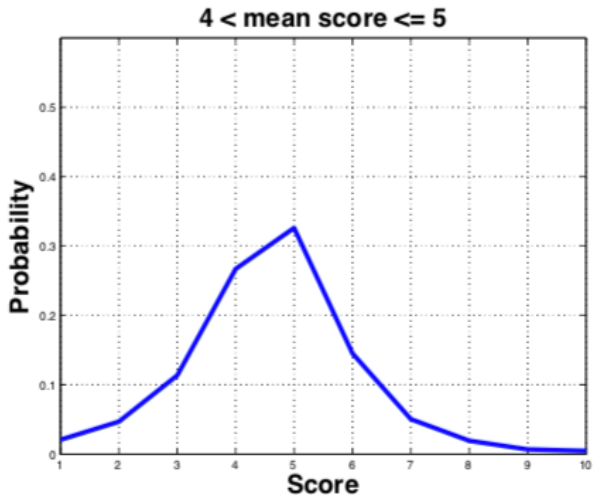
Note: Max 2 tags / image

NAVER LABS

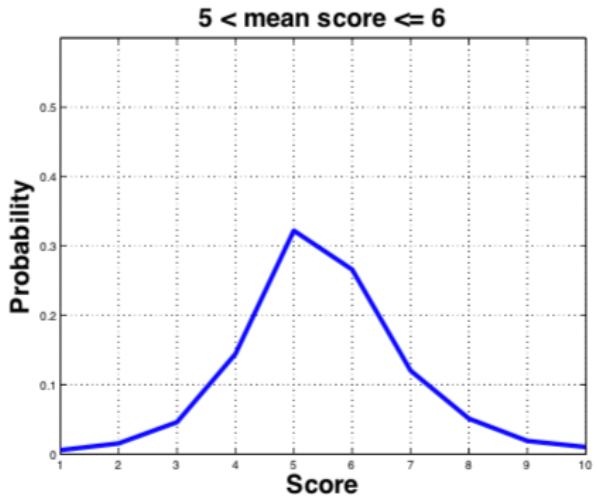
Averaged distributions for images with different mean scores



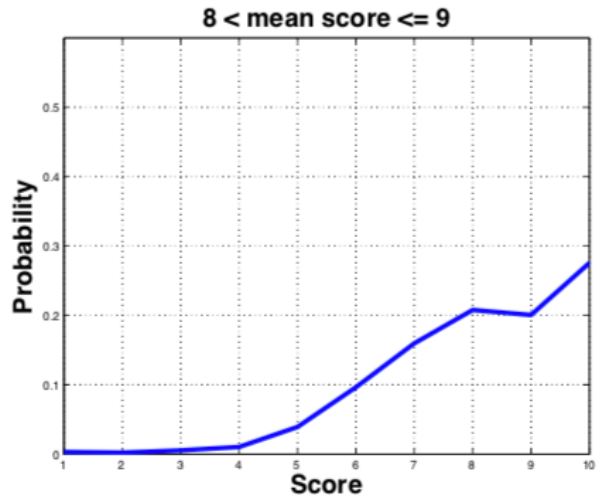
(a)



(b)



(c)



(d)

NAVER LABS

Modelling score distributions

We fit several probability distribution models to our distributions:

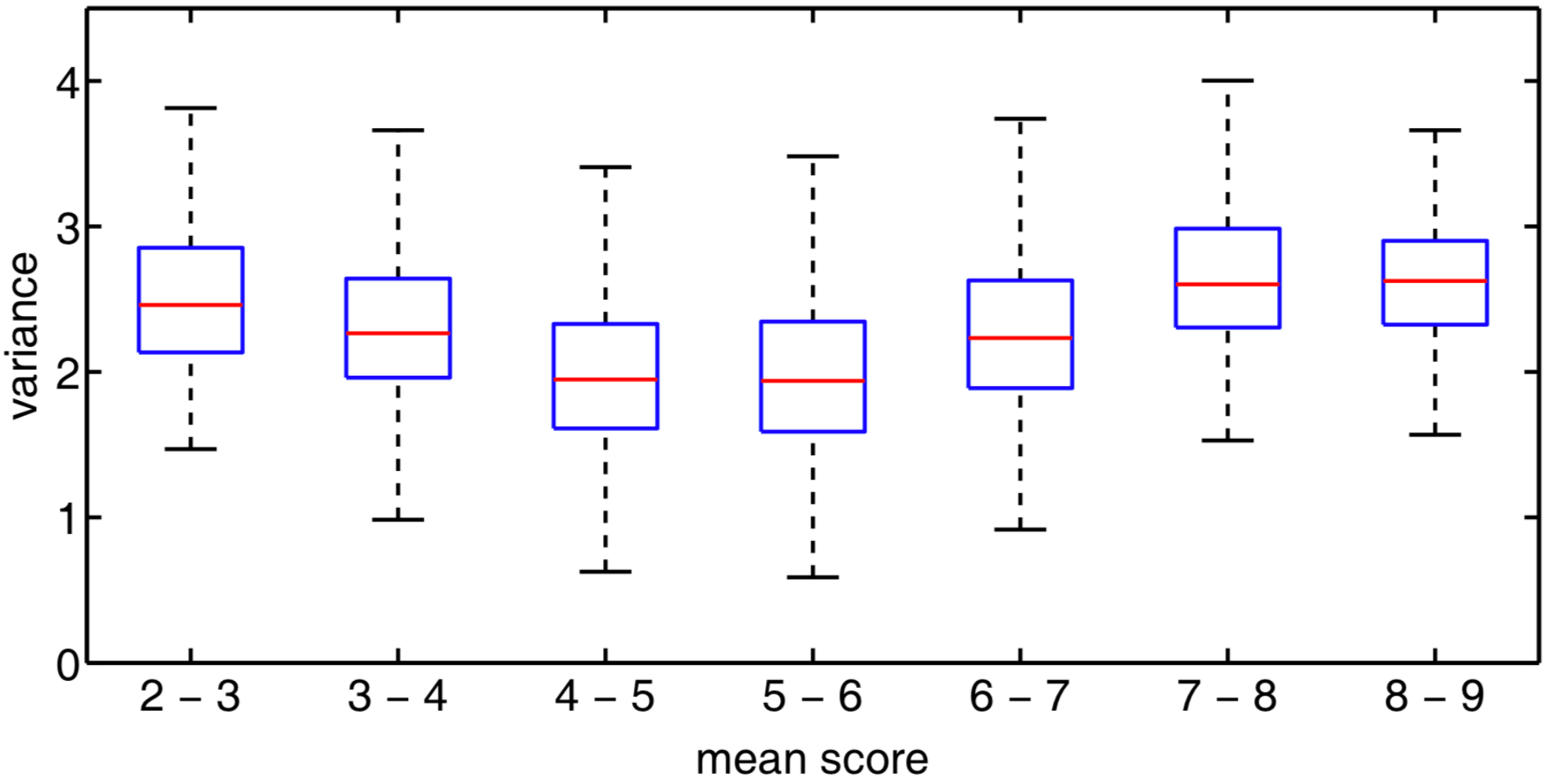
- Gaussian, Beta, Weibull and Generalized Extreme Value
- Gaussian functions perform adequately for most images
- Floor / ceiling effect at extremes of rating scale [1]
  - Better modelled with a Gamma distribution (Γ)

[1] Cramer & Howitt. The SAGE dictionary of statistics, 1st edn. 2004. p. 21 (entry “ceiling effect”), p. 67 (entry “floor effect”)

Modelling score distributions

Mean score	Average RMSE		
	Gaussian	Γ	Γ'
1-2	0.1138	**0.0717**	0.1249
2-3	0.0579	**0.0460**	0.0633
3-4	**0.0279**	0.0444	0.0325
4-5	**0.0291**	0.0412	0.0389
5-6	**0.0288**	0.0321	0.0445
6-7	0.0260	**0.0250**	0.0455
7-8	**0.0268**	0.0273	0.0424
8-9	0.0532	0.0591	**0.0403**
Average RMSE	**0.0284**	0.0335	0.0429

Variance of score distributions



NAVER LABS

Global outline

AVA: A large-scale dataset for aesthetic visual analysis
Collaborators:
Luca Marchesotti
Florent Perronnin

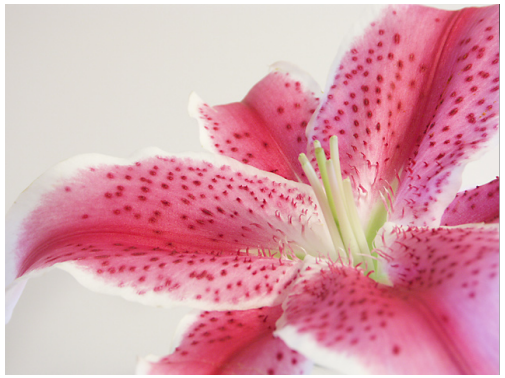
A deep architecture for unified aesthetic prediction
Collaborators:
Albert Gordo

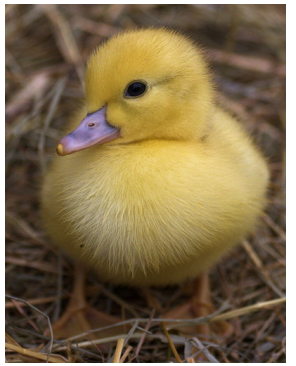
Global outline

AVA: A large-scale dataset for aesthetic visual analysis
Collaborators:
Luca Marchesotti
Florent Perronnin

A deep architecture for unified aesthetic prediction
Collaborators:
Albert Gordo

Aesthetic visual analysis



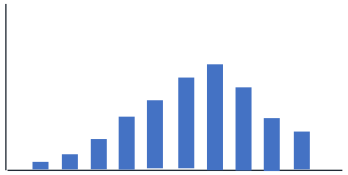
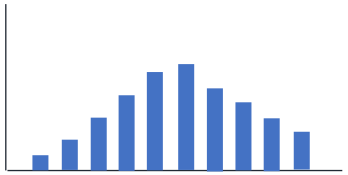
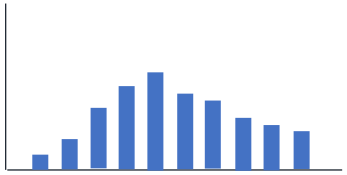
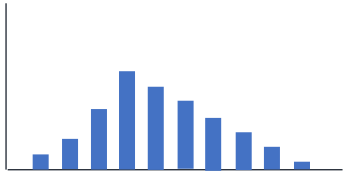



Binary classification


Mean score prediction

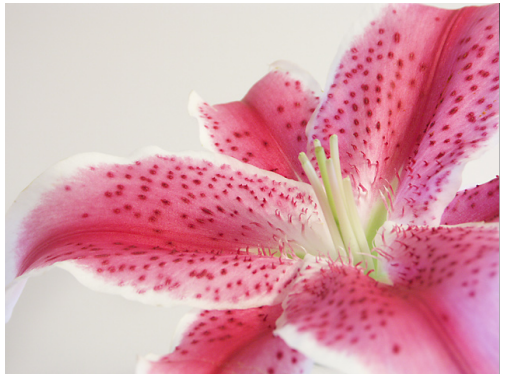
7.65 **6.32** … **4.81** **4.15**

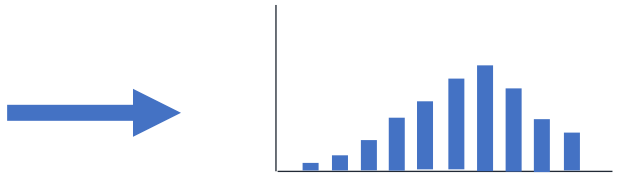
Score distribution prediction

NAVER LABS

Obtaining annotations for AVA





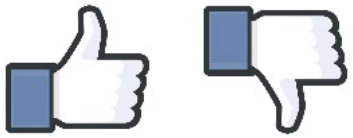
Score distribution

Calculate mean

7.65

Mean score

Apply threshold



Binary class

NAVER LABS

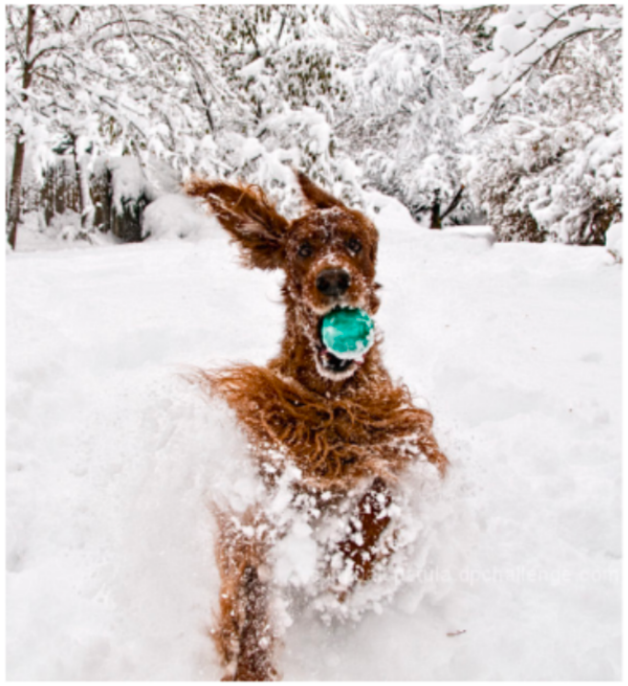
Obtaining annotations for AVA

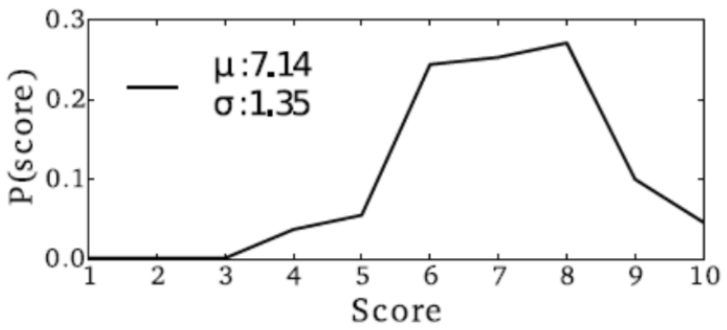
Most works perform binary prediction

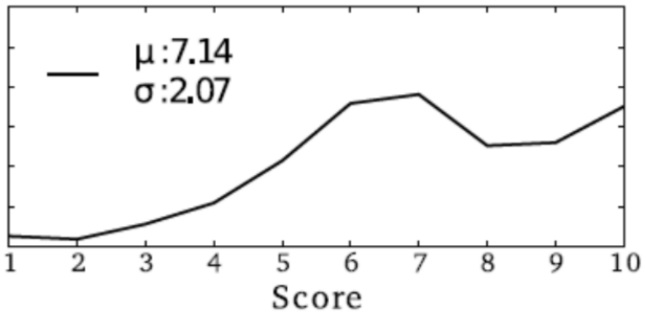
But …

Score distributions are

more informative







NAVER LABS

Predicting aesthetic score distributions

Previous work:

- Wu et al. [1]:
  - Used traditional features
  - Focused on mitigating unreliable scores from annotators using reweighting
- Jin et al. [2]:
  - Used deep model and χ^2error
  - Focused on reweighting samples to mitigate imbalance in score distribution of training set

Our approach [3]:

- Uses deep architecture
- Explores how to leverage semantic information

[1] Wu et al. Learning to predict the perceived visual quality of photos. ICCV 2011.
[2] Jin et al. Image aesthetic predictors based on weighted cnns. ICIP 2016.
[3] Murray & Gordo. A deep architecture for unified aesthetic prediction. arXiv 2017.

Leveraging semantic information

Aesthetic properties are dependent on image content

NAVER LABS

Leveraging semantic information for deep learning

No large-scale aesthetics datasets exist that also have semantic annotations

- AVA is only partially-annotated (max 2 tags from a predefined list)

Recent works incorporate semantic information by using:

- Weak labels [1]:
  - Cluster images and use cluster IDs as semantic label
  - Train network for auxiliary task of predicting weak semantic labels
- Auxiliary networks [2]:
  - Use output from scene categorizer as additional features

[1] Mai et al. Composition-preserving deep photo aesthetics assessment. CVPR 2016.
[2] Kong et al. Photo aesthetics ranking network with attributes and content adaptation. ECCV 2016.

Leveraging semantic information

Our approach - Inspired by “model distillation” [1]:

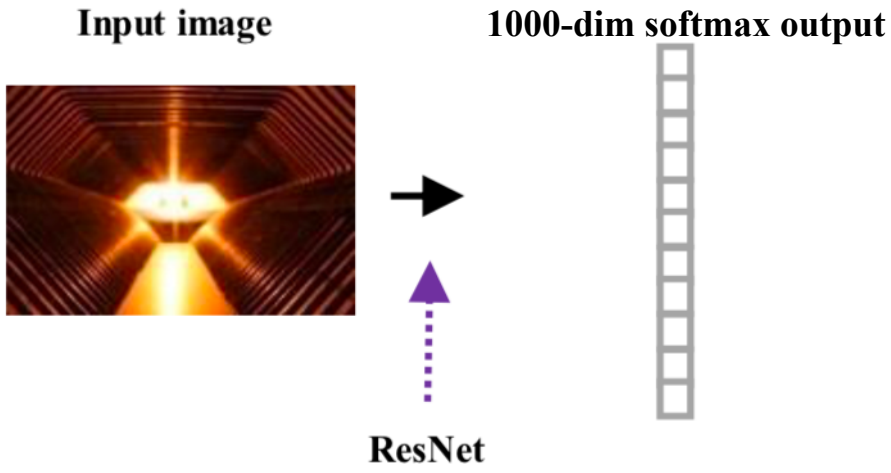
- In a nutshell, use a trained “teacher network” to supervise a “student” network:
  - The softmax output of the teacher network is used as the target for training the student
  - Conceived for network compression, *i.e.* student network is usually more compact than original teacher network
- We use a teacher network, trained for ImageNet classification, to:
  - Provide semantic pseudo-ground-truth in the form of softmax probabilites.
  - Pre-train the weights of our aesthetic prediction model, i.e. provide a better initialization.

[1] Hinton et al. Distilling the knowledge in a neural network. NIPS Deep Learning and Representation Learning Workshop, 2014.

Leveraging semantic information

Our approach in detail:

1. Extract softmax result from original ResNet model trained on ImageNet

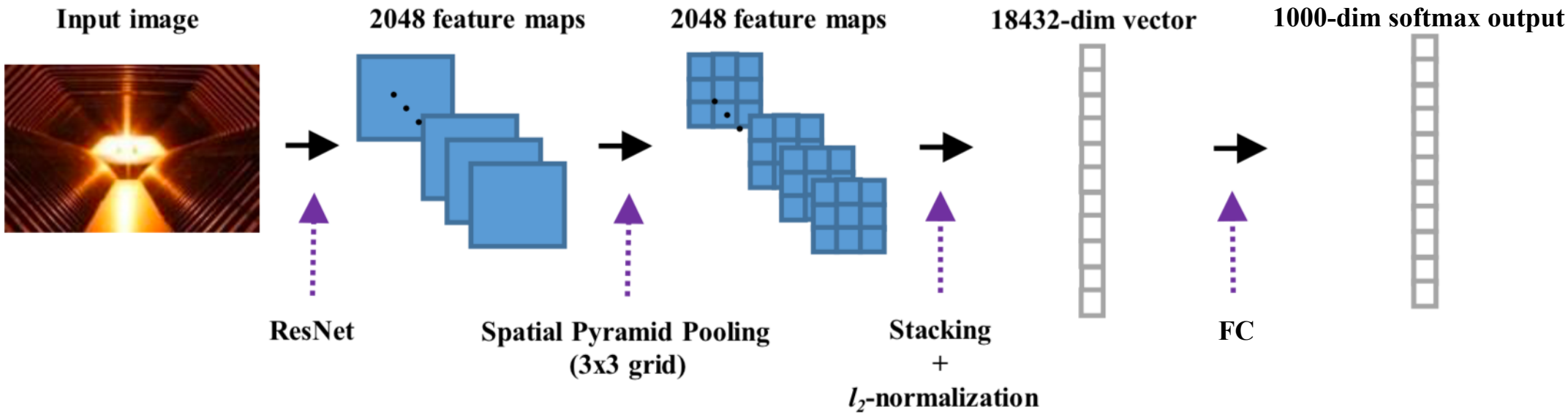


NAVER LABS

Leveraging semantic information

Our approach in detail:

2. Train modified architecture using softmax result as target

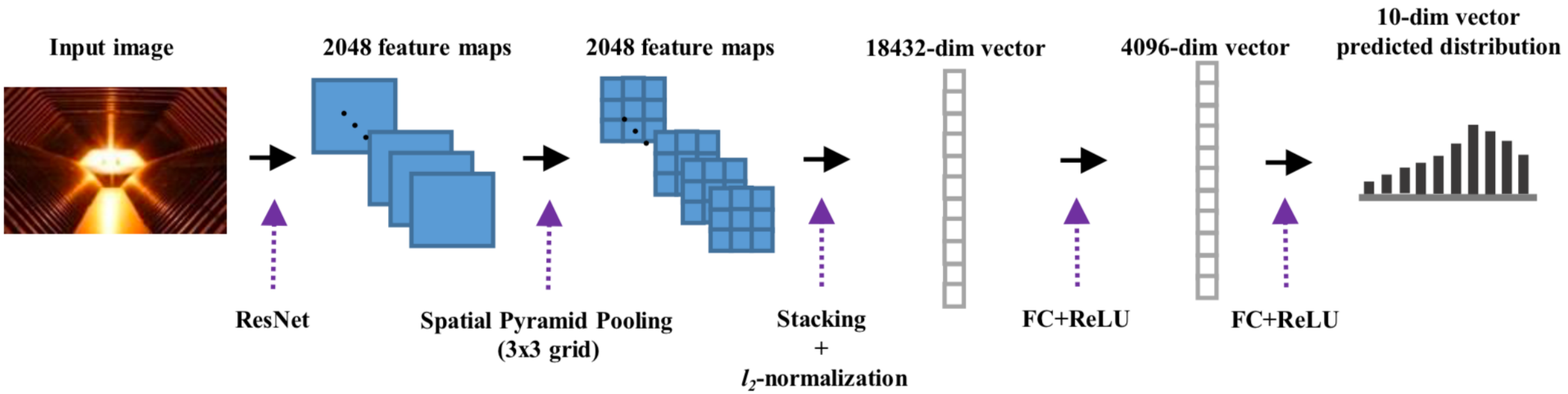


NAVER LABS

Leveraging semantic information

Our approach in detail:

3. Fine-tune previous model for score distribution prediction



NAVER LABS

Experiments

We evaluated our learned model on 3 tasks:

- Aesthetic quality (binary) classification
- Aesthetic mean score prediction
- Aesthetic score distribution prediction

Experiments

Comparing different variants of the proposed approach:

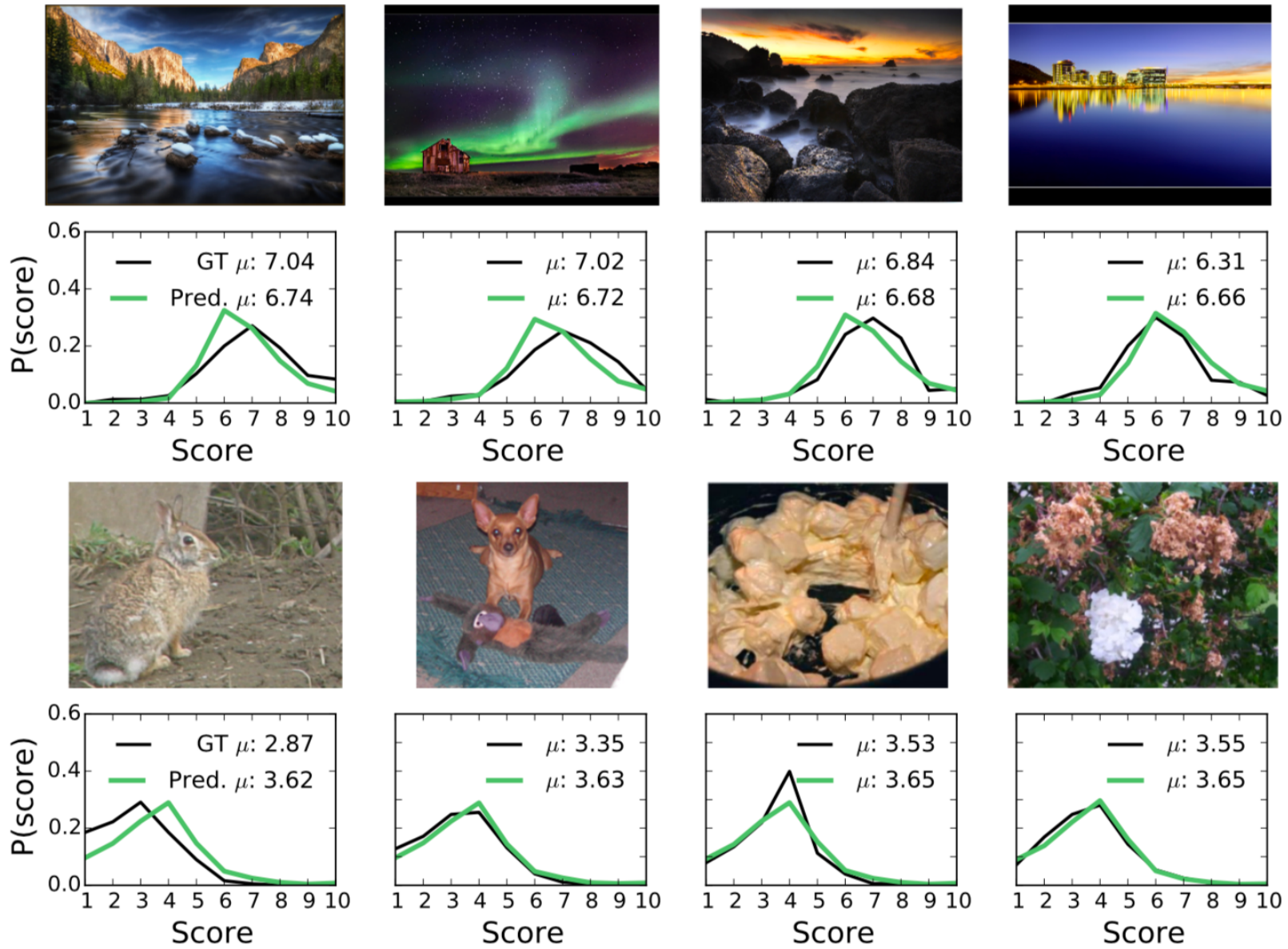
Method	KLDIV ↓	CDLoss ↓	MSE ↓	ρ ↑	Acc.(%) ↑
VGG-ImN-mean	-	-	0.314	0.659	78.4
VGG-ImN-dist-Eucl.	0.134	0.0821	0.383	0.560	76.3
VGG-ImN-dist	0.128	0.0673	0.311	0.665	78.7
ResNet-ImN-dist	0.105	0.0618	0.285	0.702	79.7
ResNet-AVA-dist	0.104	0.0609	0.280	0.707	79.8
ResNet-AVA-dist-m	**0.103**	**0.0607**	**0.279**	**0.709**	**80.3**

Experiments

Comparing to state-of-the-art methods:

Method	CDLoss ↓	MSE ↓	ρ ↑	Acc.(%) ↑
Wu *et al.* [44]†	0.1061	-	-	-
Jin *et al.* [16]-HP †	-	0.636	-	-
Jin *et al.* [16]-Reg †	-	0.337	-	-
Mai *et al.* [30]	-	-	-	77.4
Kong *et al.* [21]	-	-	0.558	77.8
Zhou *et al.* [46]	-	-	-	78.2
`VGG-ImN-dist`	0.0673	0.311	0.665	78.7
APM	**0.0607**	**0.279**	**0.709**	**80.3**

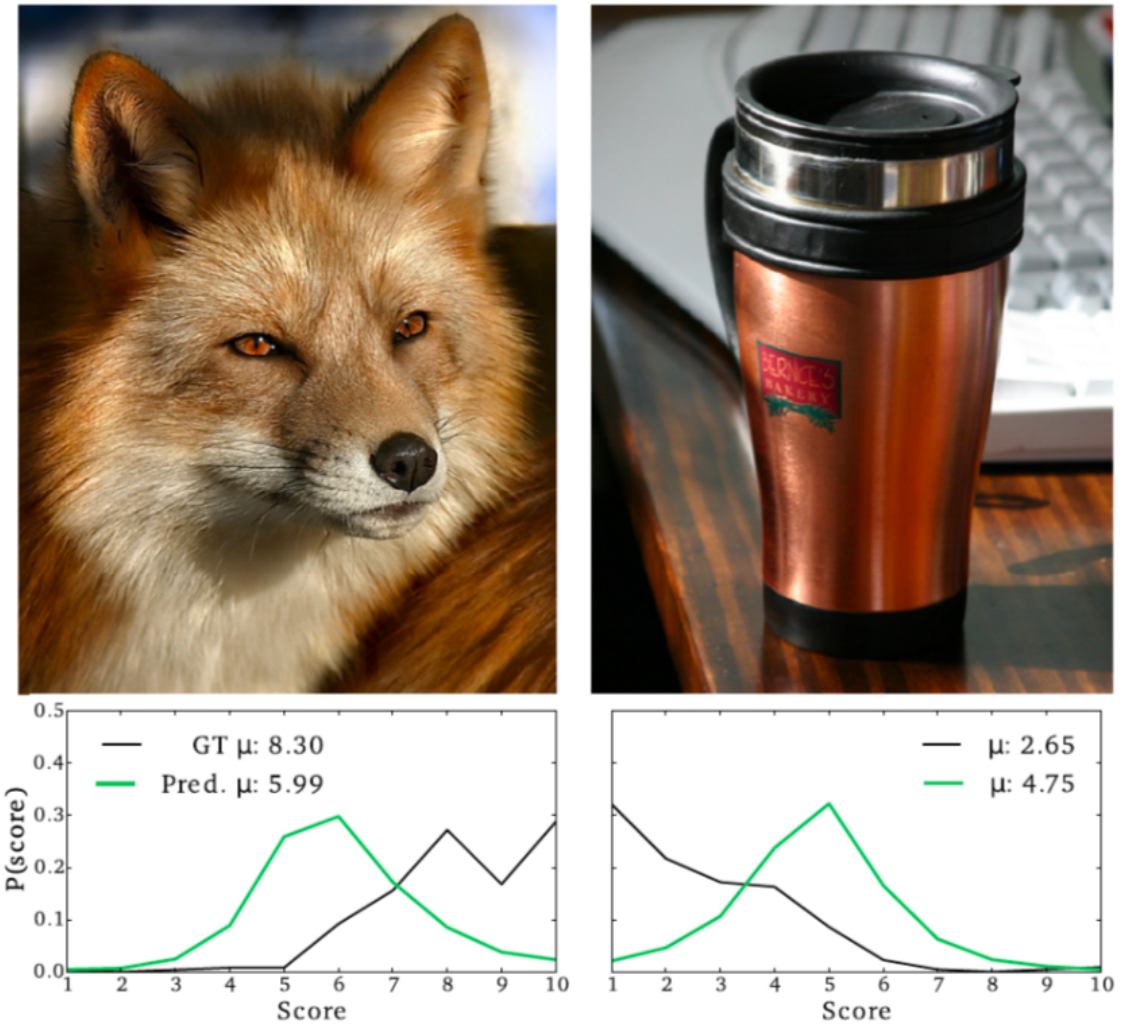
Qualitative results



NAVER LABS

Failure cases

The model fails for images with extremely good/bad scores:



NAVER LABS

A step towards interpretability

We want to know not just how aesthetically pleasing an image is but ***why.***

Some works visualize the learned convolutional weights in early layers

- Too low-level for aesthetics

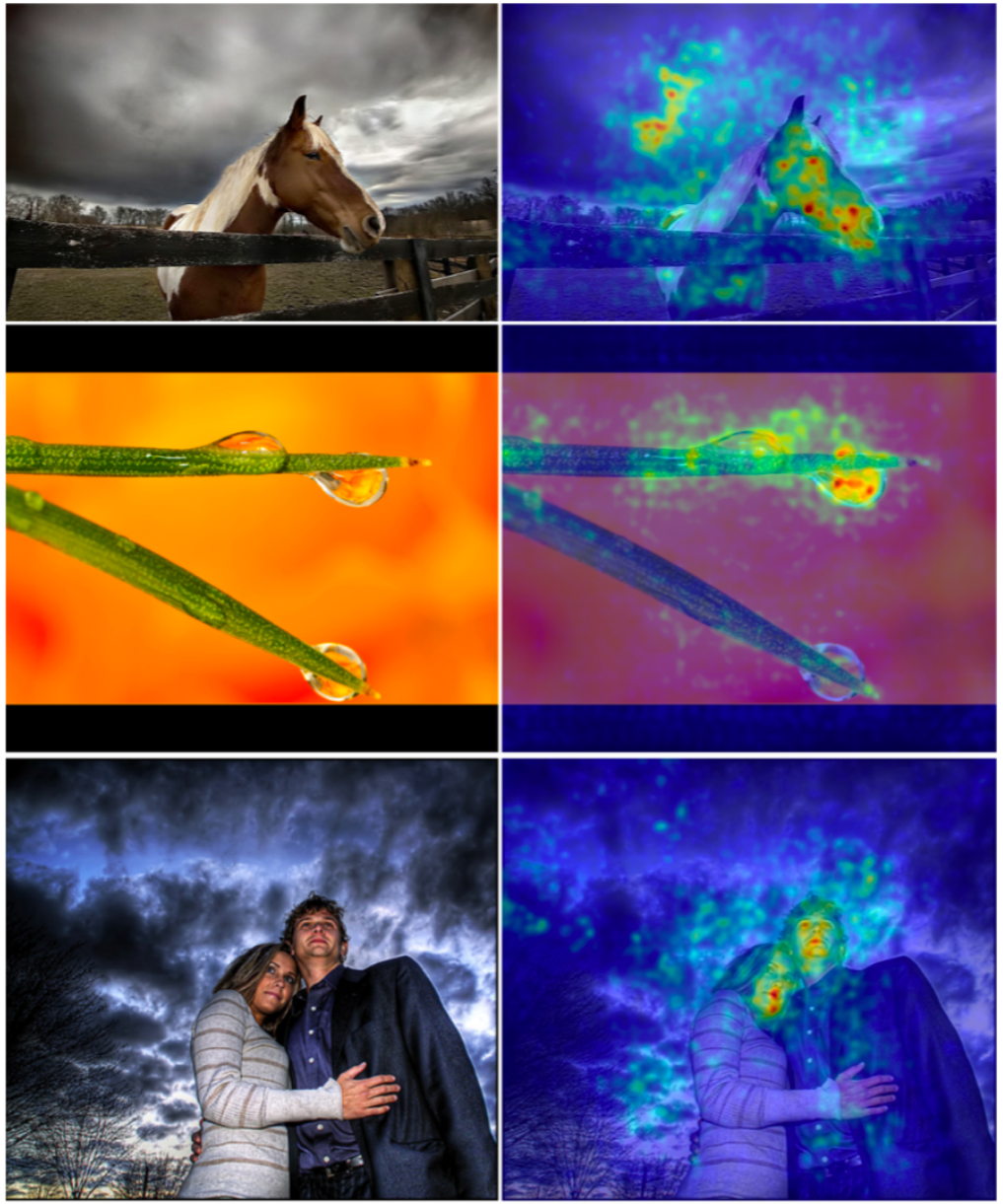
We propose an adversarial approach:

- Modify an image by gradient descent such that it has a lower/higher predicted score
- Visualize the regions that were most modified

Adversarial images & their heatmaps

Observations:

- Much easier to worsen an aesthetic score than improve it
- Most modified regions tend to contain salient objects



NAVER LABS

From predicting to generating

Conditioning generative image models on aesthetic properties

Quite a few recent works on conditioning generative image models:

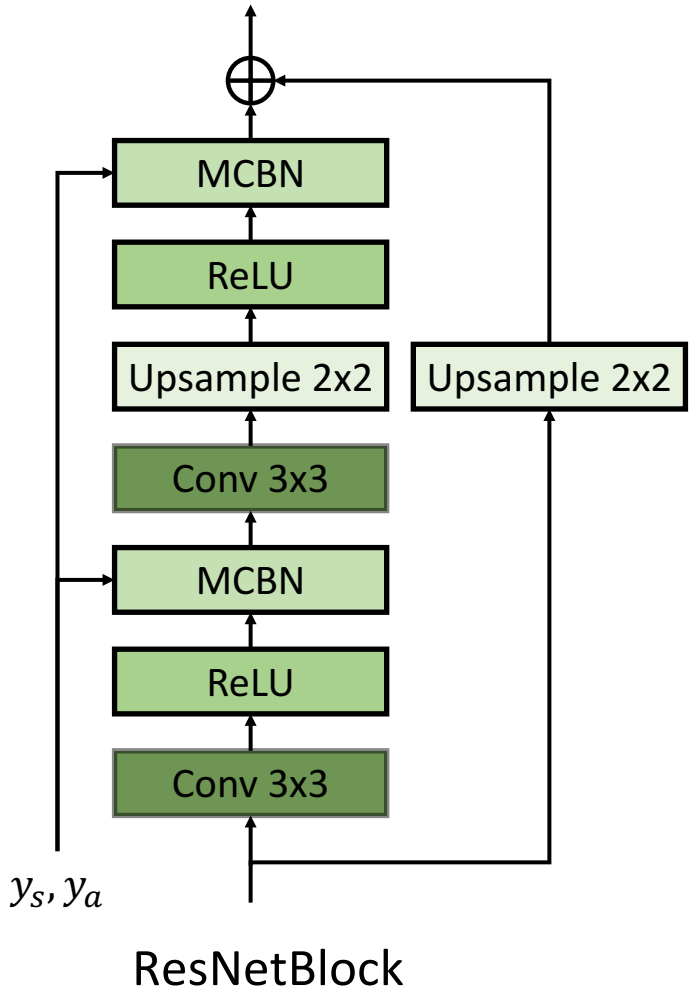
- On categorical variables [1]
- On human pose [2]
- On semantic layout [3]

We propose a GAN for photographic fine art (PFAGAN):

- The discriminator and generator are conditioned on semantics *and* aesthetics.
- We use normalized score histograms to condition on aesthetics.

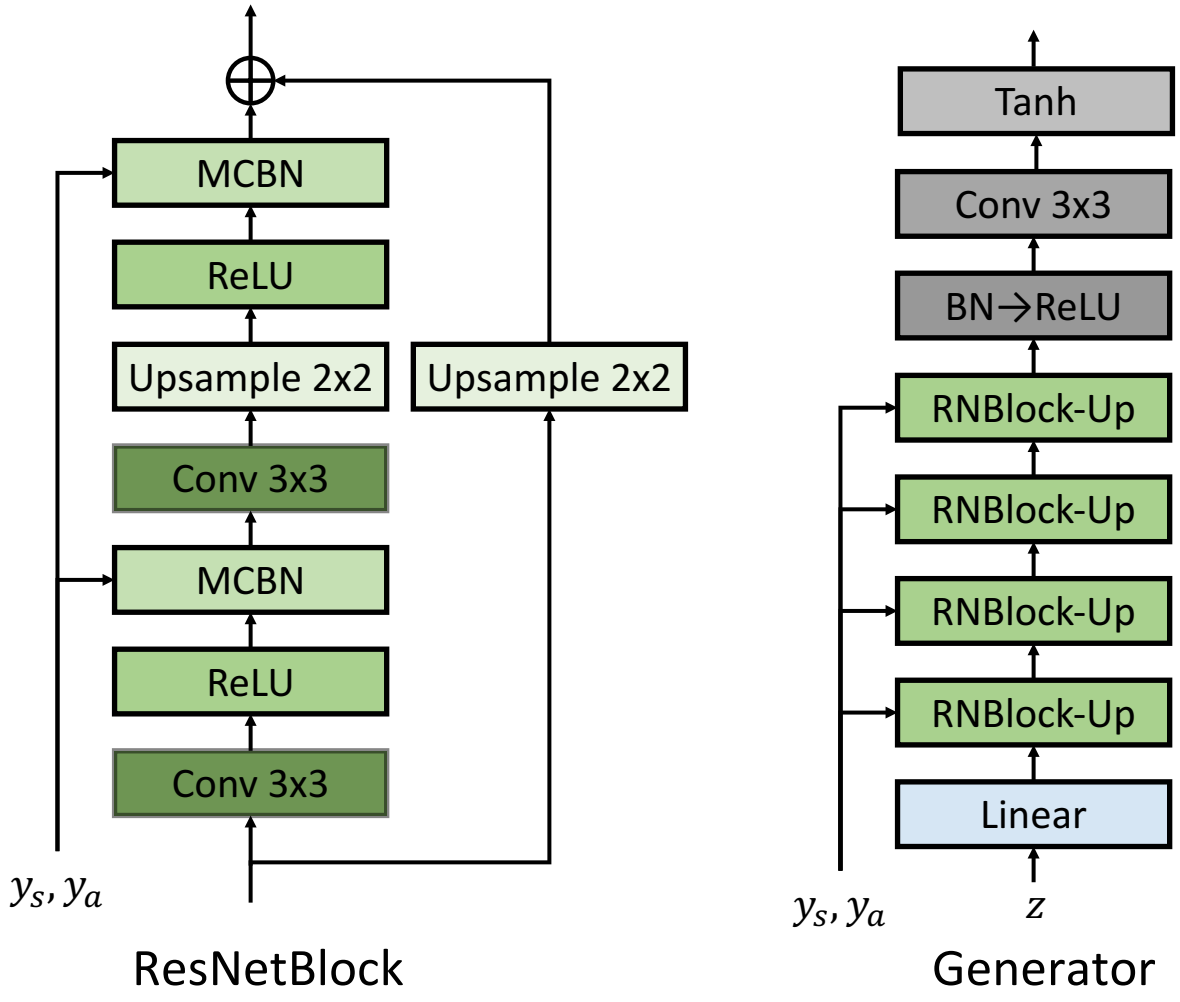
[1] Mirza & Osindero. Conditional generative adversarial nets. *arXiv*, 2014.
[2] Dong et al. Soft-Gated Warping-GAN for Pose-Guided Person Image Synthesis. NeurIPS 2018.
[3] Chen & Koltun. Photographic Image Synthesis with Cascaded Refinement Networks. ICCV 2017.

PFAGAN architecture



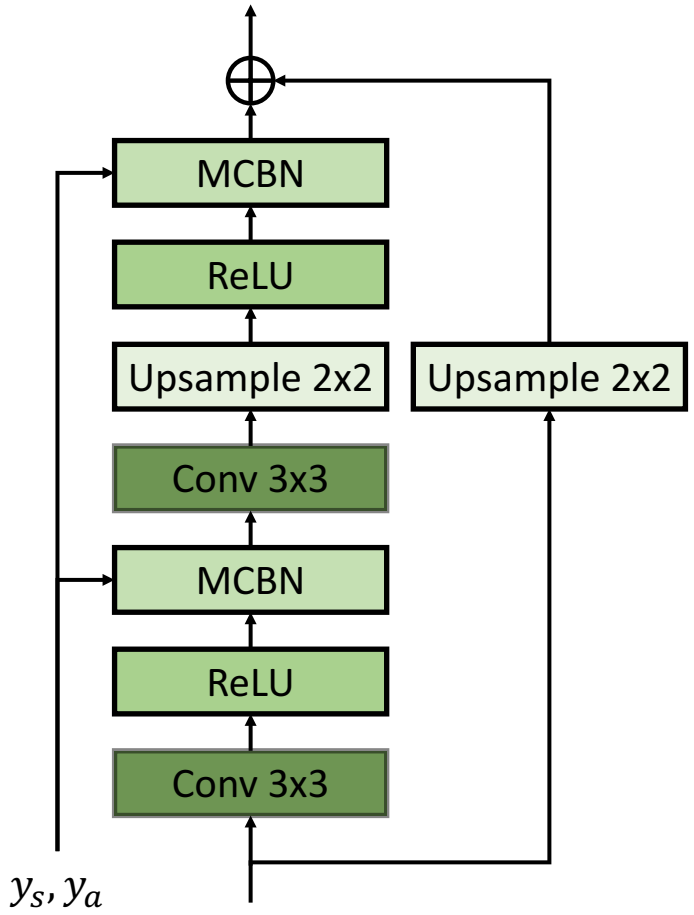
Murray. PFAGAN: An aesthetics-conditional GAN for generating photographic fine art. Under review.

PFAGAN architecture

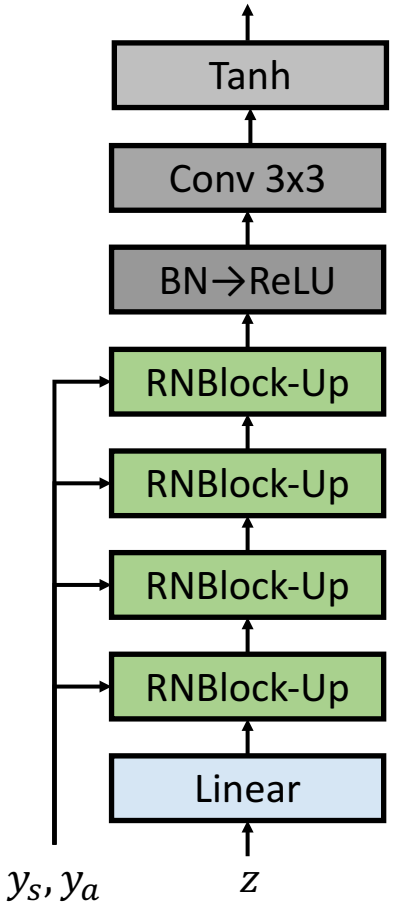


Murray. PFAGAN: An aesthetics-conditional GAN for generating photographic fine art. Under review.

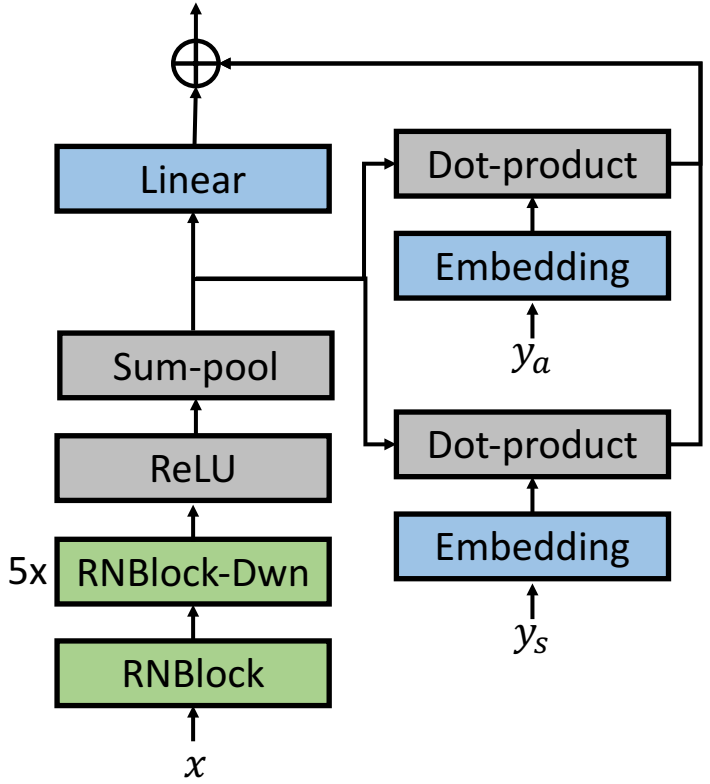
PFAGAN architecture



ResNetBlock



Generator

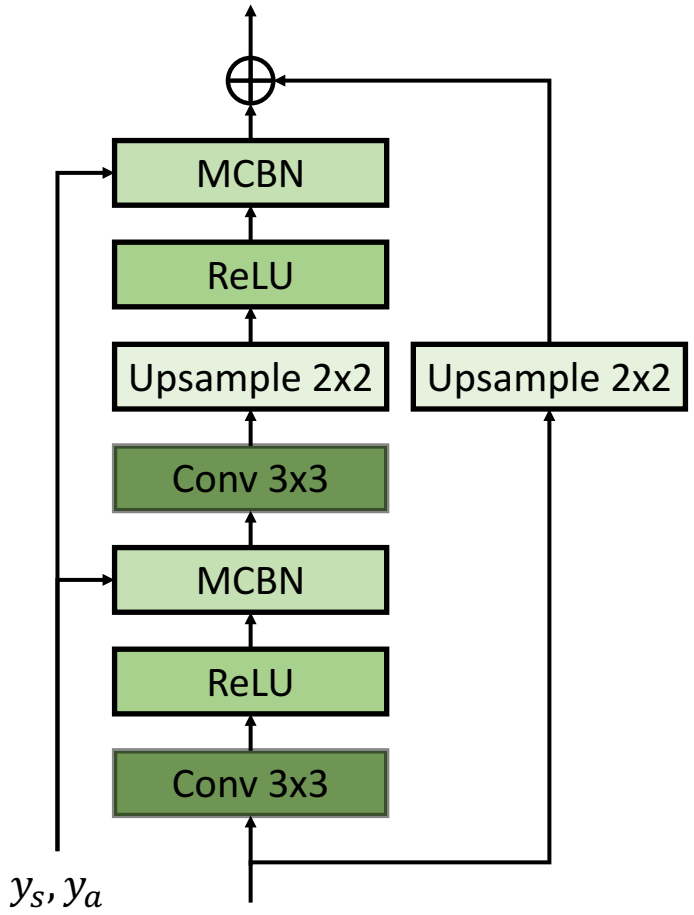


Discriminator

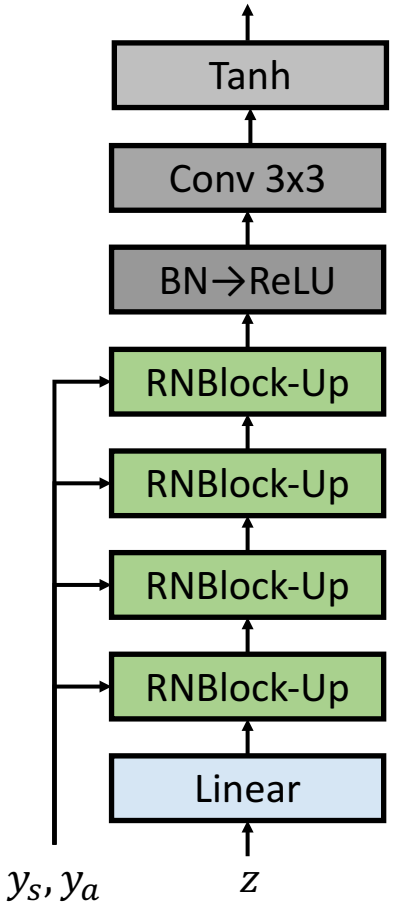
Murray. PFAGAN: An aesthetics-conditional GAN for generating photographic fine art. Under review.

NAVER LABS

PFAGAN architecture

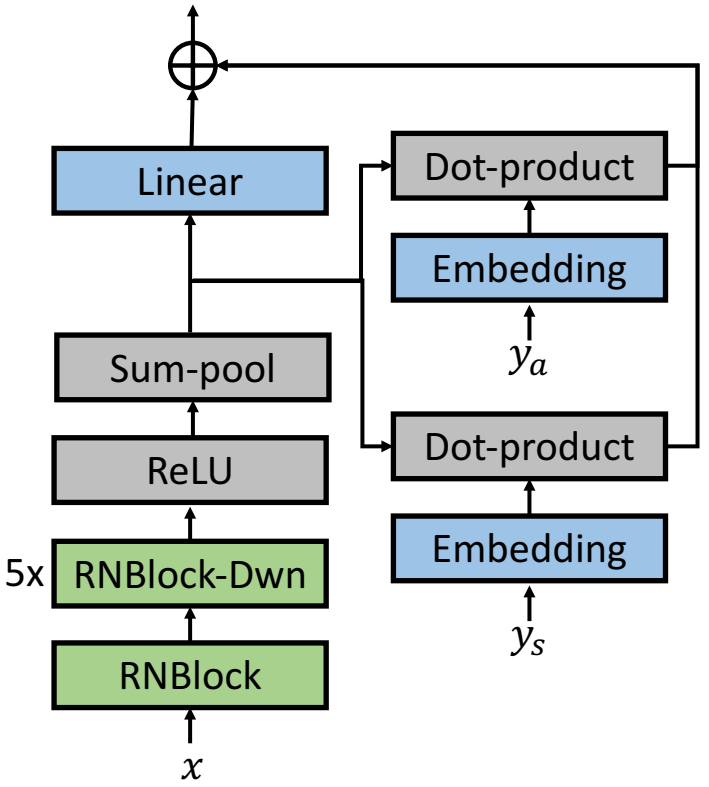


ResNetBlock



Generator

Similar architecture to Miyato & Koyama*



Discriminator

*Miyato and Koyama. cgans with projection discriminator. *ICLR*, 2018.

NAVER LABS

Conditioning the generator

To condition the generator we propose a “mixed-conditional” BN layer:

$$o_{i,c} = \lambda_{s,a,c}\left(\frac{h_{i,c} - \mu_c}{\sigma_c}\right) + \beta_{s,a,c}$$

Where $\boldsymbol{\lambda}_{s,a}$ and $\boldsymbol{\beta}_{s,a}$ are computed using affine transformations:

$$\boldsymbol{\lambda}_{s,a} = \boldsymbol{y}'_a W_s^{\lambda} + \boldsymbol{b}_s^{\lambda}; \qquad \boldsymbol{\beta}_{s,a} = \boldsymbol{y}'_a W_s^{\beta} + \boldsymbol{b}_s^{\beta}$$

Each semantic category s is associated with two affine transforms.

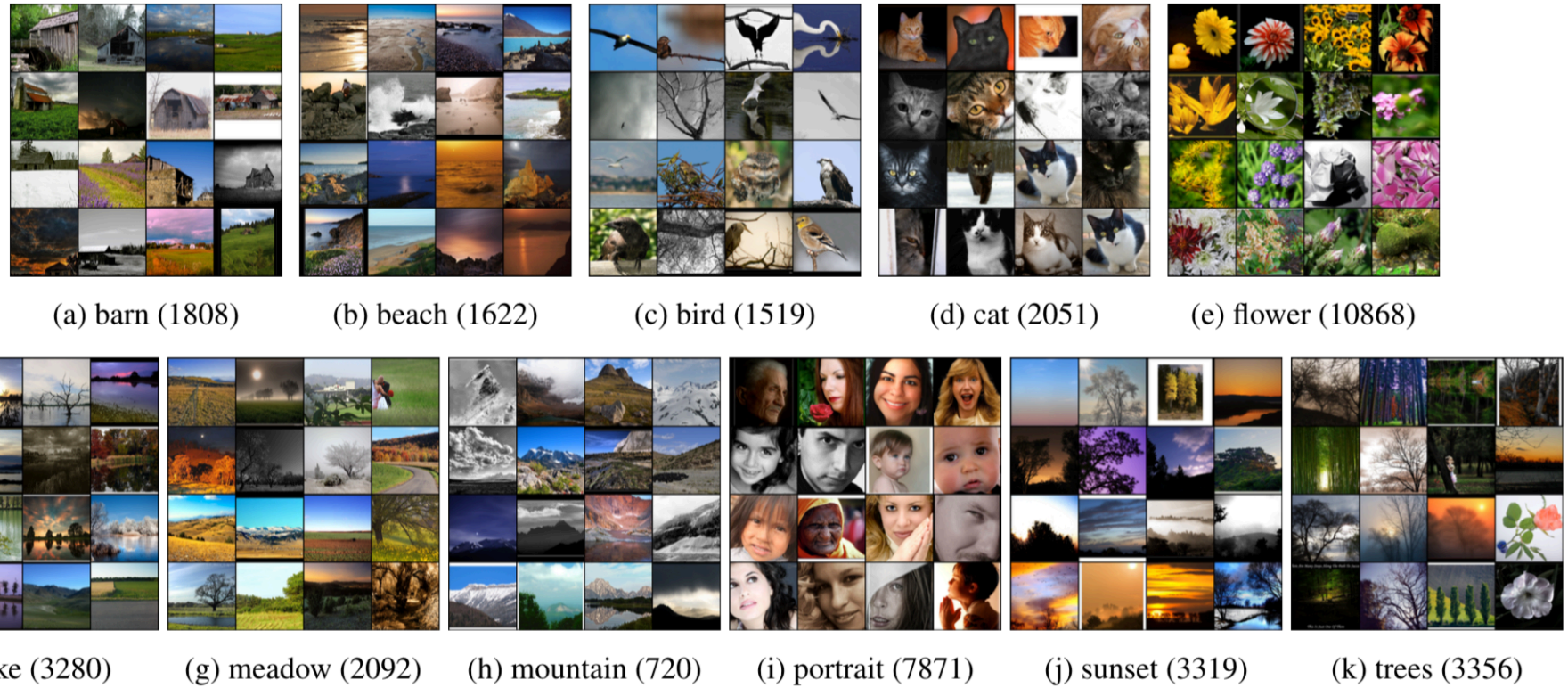
Related to both conditional batch normalization [1] and conditional instance normalization [2].

[1] De Vries et al. Modulating early visual processing by language. *NIPS*, 2017.
[2] Dumoulin and Kudlur. A learned represen tation for artistic style. *ICLR*, 2017.

Training PFAGAN with AVA

We use normalized score histograms to condition on aesthetics.

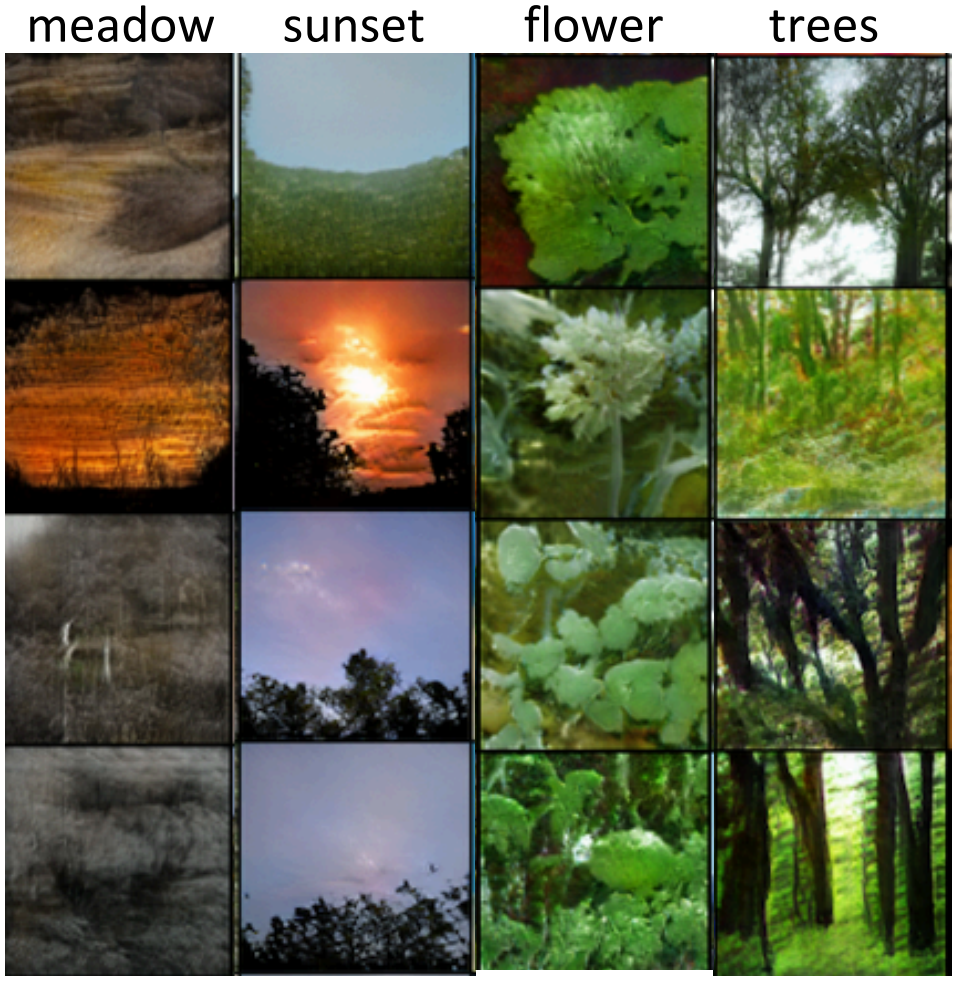
We cluster the images in AVA to obtain the following semantic categories:



Murray. PFAGAN: An aesthetics-conditional GAN for generating photographic fine art. Under review.

NAVER LABS

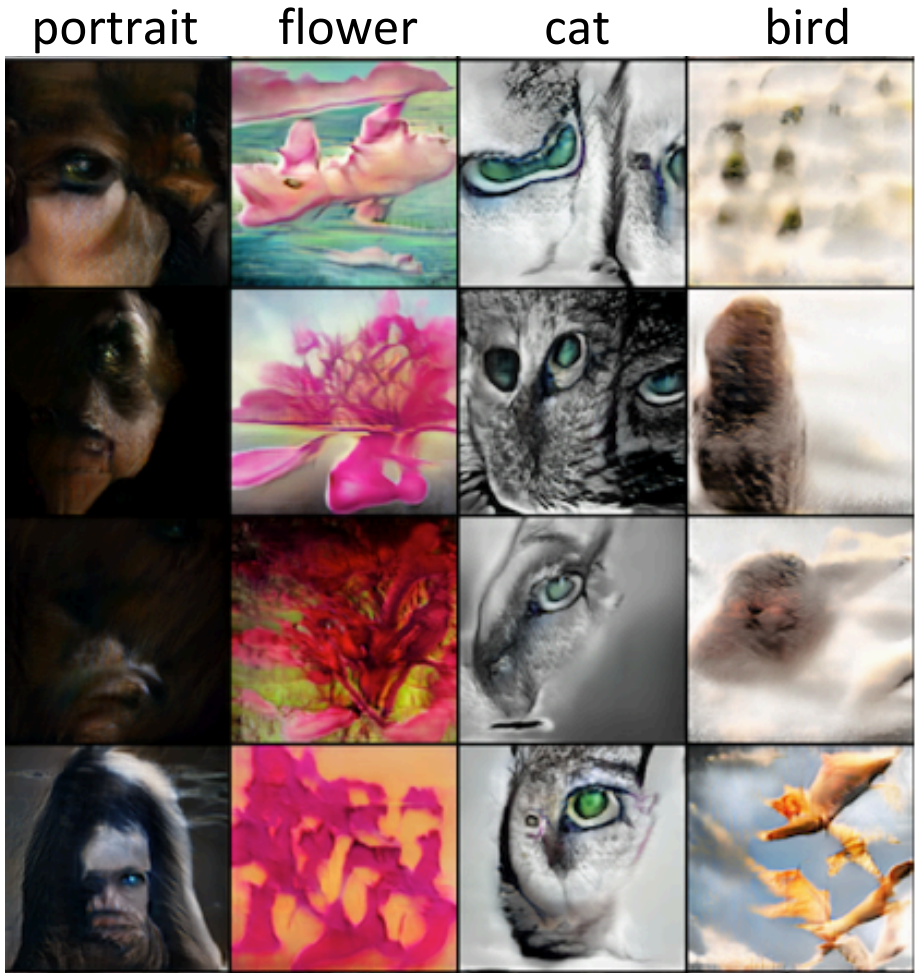
Generated samples: successful classes

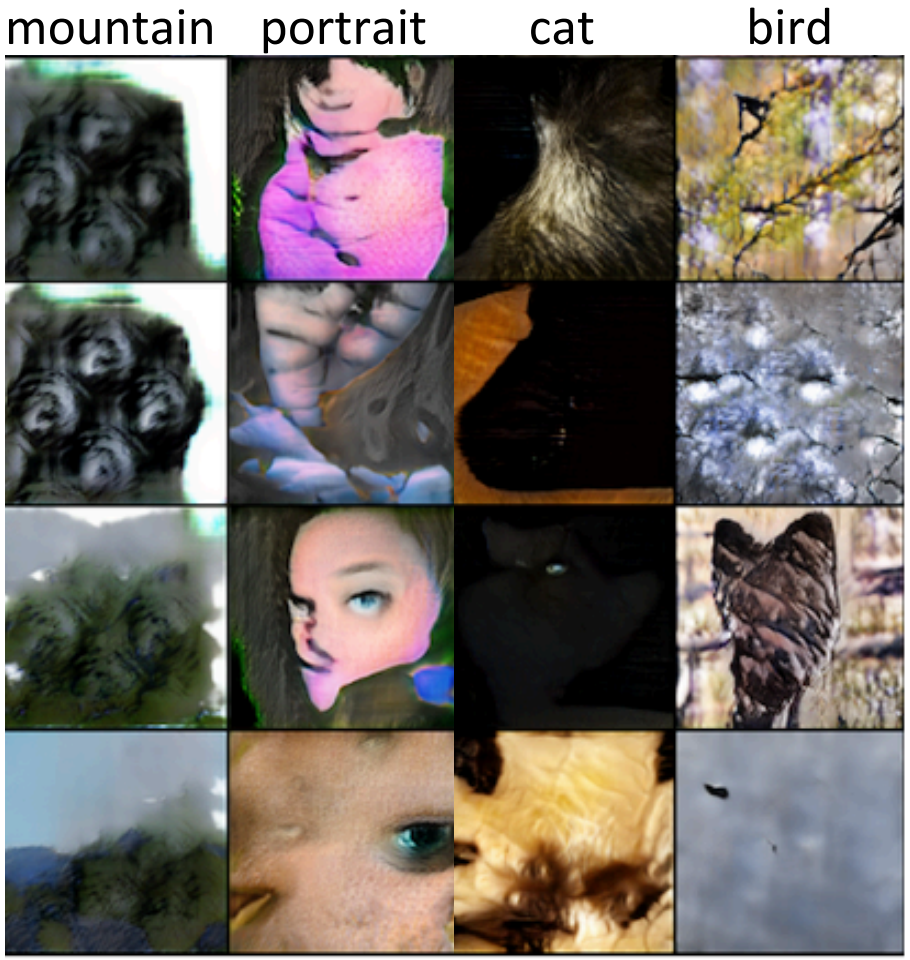


Murray. PFAGAN: An aesthetics-conditional GAN for generating photographic fine art. Under review.

NAVER LABS

Generated samples: failure cases





Murray. PFAGAN: An aesthetics-conditional GAN for generating photographic fine art. Under review.

NAVER LABS

Global summary

Supervised learning of image aesthetics models is effective given:

- Large-scale datasets.
- Powerful deep image representations.

Major challenge is dataset bias:

- Most images are “average”.
- Content is biased towards a few categories (landscapes, portraits, etc.).
- Raters are self-selected and participation is unbalanced.

Thank you!