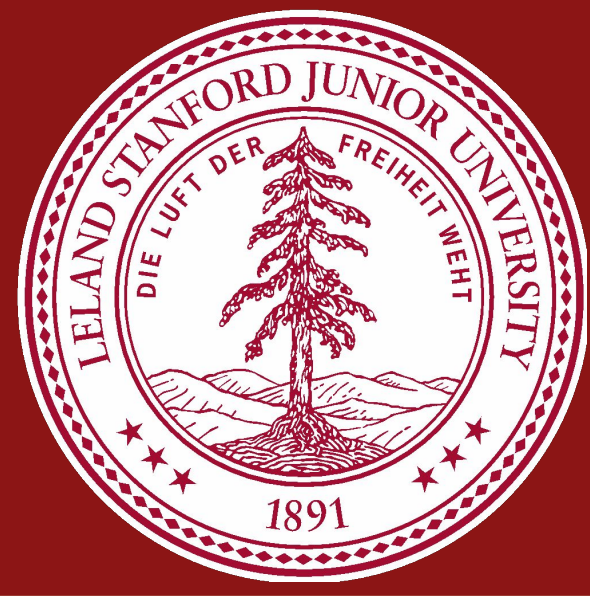




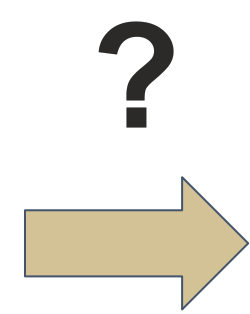
# Building a Neural Historical Art Classifier with Discriminative Features and Attention Feature Extraction

Michael Cao<sup>1</sup>, Eish Maheshwari<sup>1</sup>

<sup>1</sup>Department of Computer Science, Stanford University

## Problem

- Essential task: extract information to analyze and classify artwork
- Evaluating historical art is expensive and time-consuming
  - contextualized objective info + subjective opinion from trained appraisers
- We aim to implement an effective deep learning model to predict art style/genre, given digital art image



Impressionism: 89%  
Baroque: 5%  
Medieval: 2%  
Renaissance: 2%  
...

## Background

A labeled Historic Art Dataset w/ 45.6k images of European artworks from the 3rd to 19th century along with simple baselines on the classification task

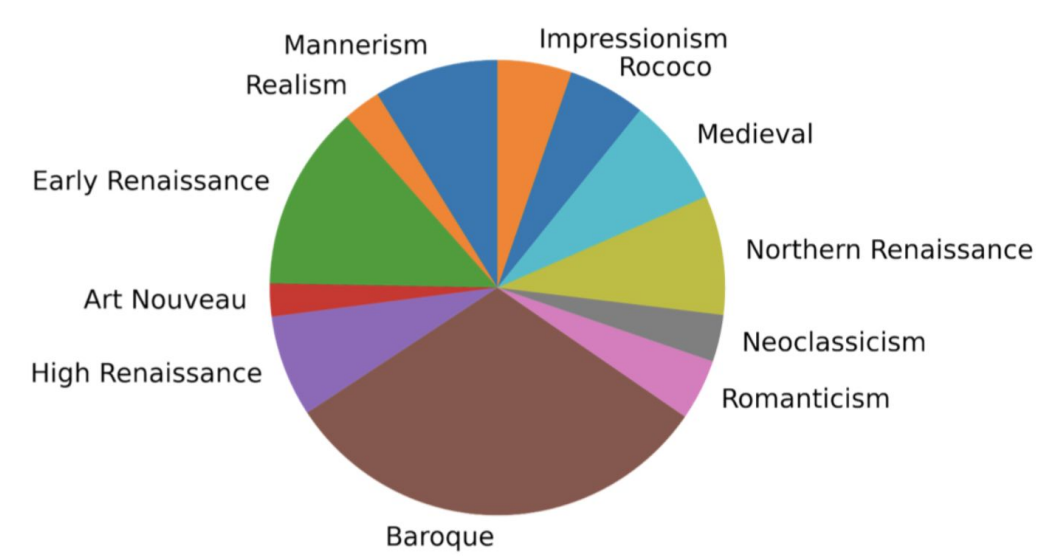


**Fig 1: a)** Class distribution across the dataset showing heavy class imbalances across different art styles. This suggests the difficulty of the problem as there is limited data available for certain time periods **b)** Performance of current best baselines on the same task.

| Model              | Accuracy |
|--------------------|----------|
| Random Forest      | 38%      |
| Fully-Connected NN | 41%      |
| VGG11              | 48%      |
| ResNet + Data Aug  | 51%      |

The main issues we found with existing modeling approaches are: **class imbalances, large intra-class variation, and small inter-class differences.**

Our goal is to address these issues with a three-pronged approach:

- 1) Data augmentation
- 2) Custom texture-based GLCM features
- 3) Attention maps with Grad-CAM for DDRL

We propose modifying a pre-trained ImageNet-based CNN with new texture features, attention mapping, and discriminative feature extraction to address artwork classification.

## Methods

### Data Augmentation + Preprocessing

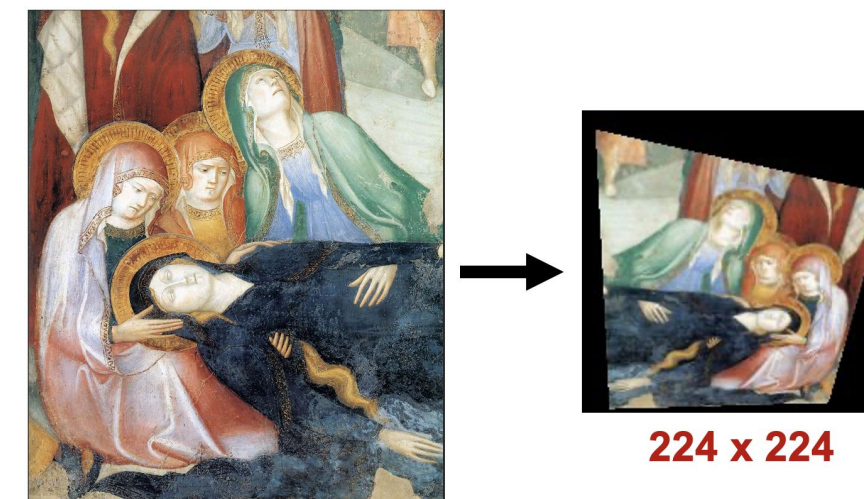
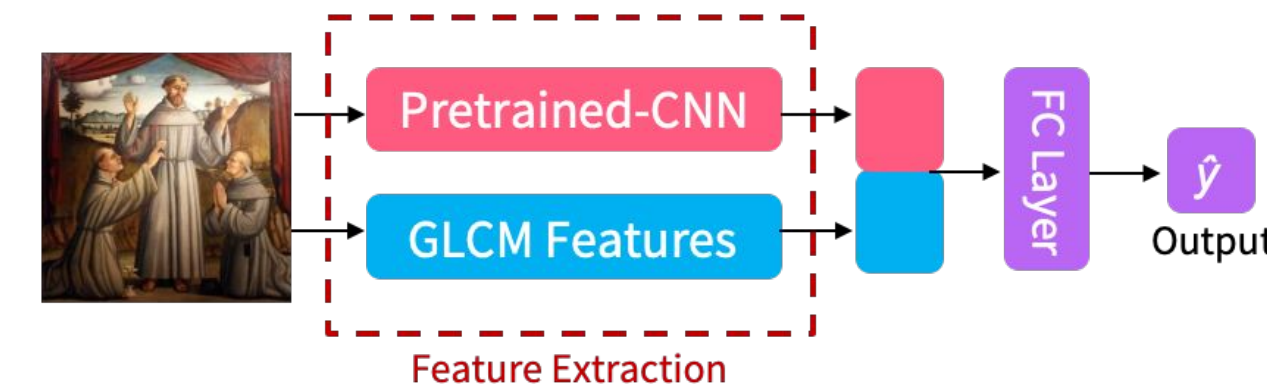


**Fig 2:** Geometric augmentation strategies applied on the training data space including implemented horizontal/vertical flips, perspective shifts, random cropping, rotations. No color-based augmentation was used. All images resized to 224x224, normalized pixels to be in range [0, 1], and converted to RGB.

### Gray-Level Co-Occurrence Matrices (GLCM)



GLCM matrix -  $(i, j)$  is the number of occasions pixel with intensity  $i$  is adjacent to pixel with intensity  $j$

**Fig 3:** Architecture of GLCM-Net. Images are fed through pre-trained CNN and GLCM feature extractor. Outputs of both are flattened, concatenated, and linearly projected into a lower feature space to create a hybrid feature before being run through a 2-layer feedforward network with dropout and ReLU activation.

### Deep Discriminative Representation Learning with Attention Mapping (DDRL-AM)

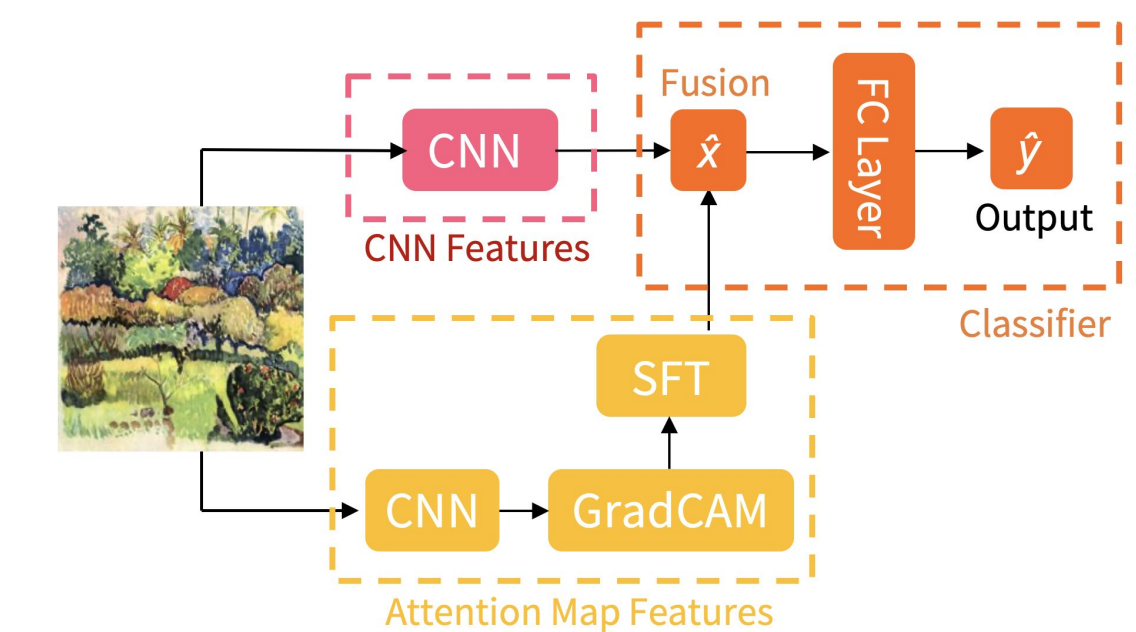


**Fig 4:** Architecture of DDRL-AM model. Grad-CAM attention maps are generated from a fine-tuned ResNet model. A spatial feature transformer extracts features from the RGB stream and attention maps, which are then fused with CNN-extracted features, and fed through a linear classification layer

## Experiments

### Baseline Model Selection

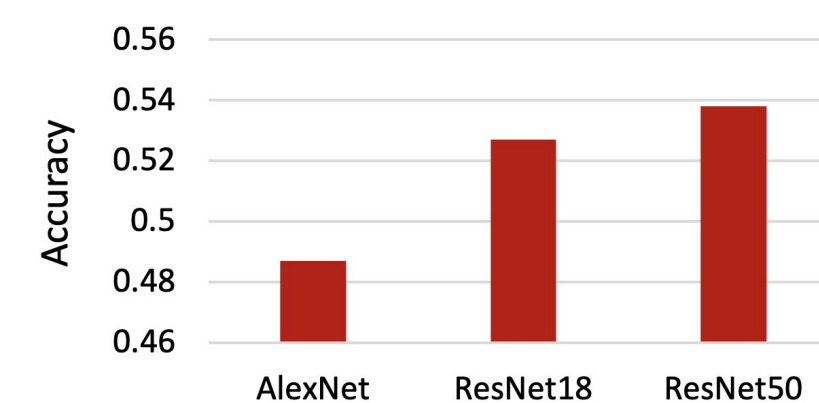


**Fig 5:** Performance of various baseline pretrained models (Best: ResNet50). Deeper models perform better.

### Data Augmentation Strategies

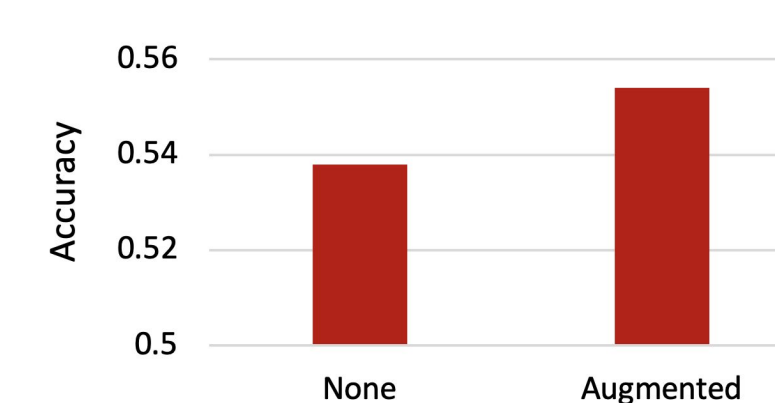


**Fig 6:** Performance of geometric augmentation on the data-space (Augment = 2.97% increase)

### GLCM Net

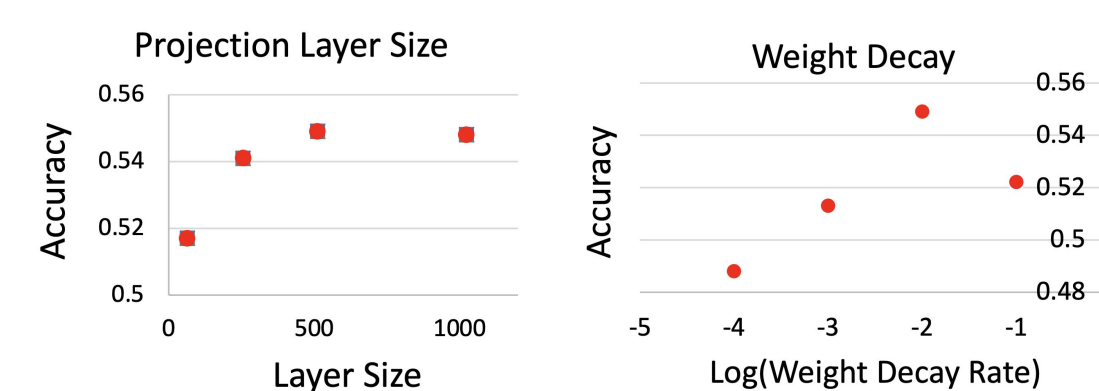


**Fig 7:** Hyperparameter tuning of GLCM-Net. Ideal performance was with projection layer size of 512 and weight decay of 1e-2 with accuracy = 54.9%. Overall, there was little improvement with the GLCM-Net.

### DDRL-AM

| Weight Decay | Dev Accuracy |
|--------------|--------------|
| 1e-1         | 0.395        |
| <b>1e-3</b>  | <b>0.402</b> |
| 5e-4         | 0.386        |

**Fig 8:** Hyperparameter tuning for DDRL-AM model. We test with multiple weight decay parameters to test different regularization strengths. Best model at 40.2% accuracy

### Grad-CAM Attention Analysis

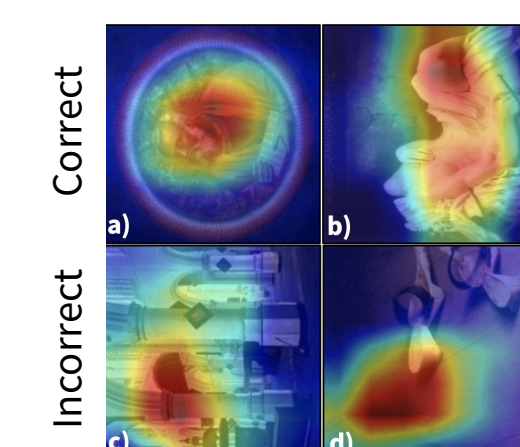


**Fig 10:** GradCAM visualizations for fine-tuned ResNet50 w/ augmentation.

**a)** Medieval  
**b)** Early Renaissance,  
**c)** Predicted: High Renaissance,  
Actual: Baroque,  
**d)** Predicted: Neoclassicism,  
Actual: Northern Renaissance,

DDRL Correct, Data Aug Incorrect



Data Aug Correct, DDRL Incorrect

**Fig 11:** DDRL-AM performs best on images with clear, defined objects since attention mapping can focus on region of interest. In complex images, attention struggles to focus, so simpler architectures perform better

## Results and Discussion

### Best Model: ResNet50 + Geometric Data Augmentation

- Overall accuracy of 55.4%
- Geometric data augmentation improved accuracy by 2.97% compared to our baseline and 6.6% compared to best previous published baseline
- Helped address class imbalances in our dataset and reduce overfitting

### GLCM texture features are ineffective in combination with CNNs

- 0.9% reduction in accuracy from augmented dataset
- Minimal effect on performance → don't add predictive power to CNN-features
  - Classification layer weights GLCM 4.1% as much as CNN-features
- Small dimension results in CNN-features overpowering GLCM features

### Localized attention is suboptimal for artwork genre classification

- Maximum accuracy of 40.2%, observed significant overfitting
- Potentially due to mediocre performance of baseline model, which was used to generate attention maps for DDRL-AM
- Attention might not help as artwork emphasizes textures and general features; fails on general patterns, complex multi-object images

## Future Work and References

- Search for pre-trained baseline better suited towards feature extraction rather than image/object recognition
- Try more advanced data augmentation and regularization strategies
- Improve classification layer architecture for GLCM and DDRL-AM models

Blake Howell. Using convolutional neural networks to predict completion year of fine art paintings. 2017.  
Jun Li, Daoyu Lin, Yang Wang, Guangluan Xu, Yunyan Zhang, and Chibiao Ding. Deep discriminative representation learning with attention map for scene classification. 2020.  
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015.  
Robert Haralick, Kumarasamy Shanmugam, and Itshak Dinstein. Textural features for image classification. *IEEE Trans. on Systems, Man and Cybernetics*, SMC-3(6):610–621, 1973.

**Fig 9:** Confusion Matrix for the best performing model. Apart from diagonals, rows represent false negatives & columns represent false positives. Artwork from similar periods were often confused with each other, suggesting a need for more expressive power.