How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval

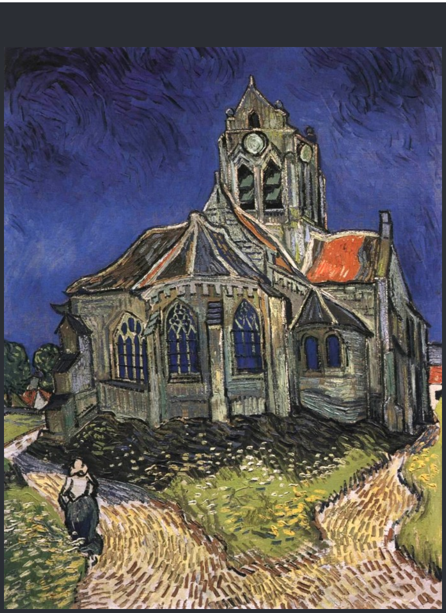
Noa Garcia & George Vogiatzis

4th Workshop on Computer Vision for Art Analysis

Motivation

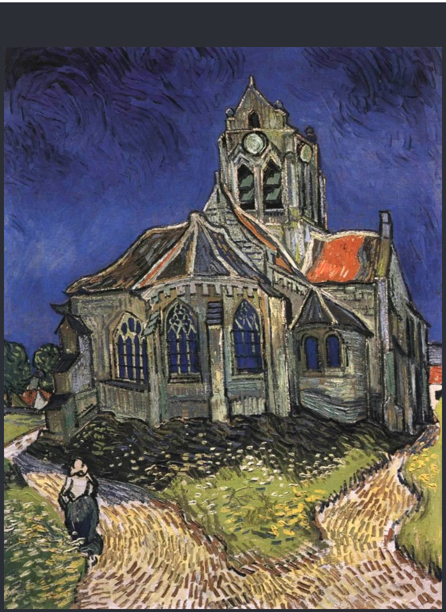
Semantic Art Understanding

In this painting the church in Auvers has been transformed by the artist into a vision using form and colour. Painted in portrait format, the church towers up before the onlooker like a fortification. The path leading to it forks in the foreground into two narrow paths passing the church on either side. On the path to the left, her back turned toward us, a peasant woman is walking into the distance. The path is bathed in light, while the church is viewed against the backdrop of a dark blue sky that merges with the black-blue of the night sky at the edges of the picture. The brushwork is restless and full of movement, and the forms of the church are distorted in the Expressionist manner.



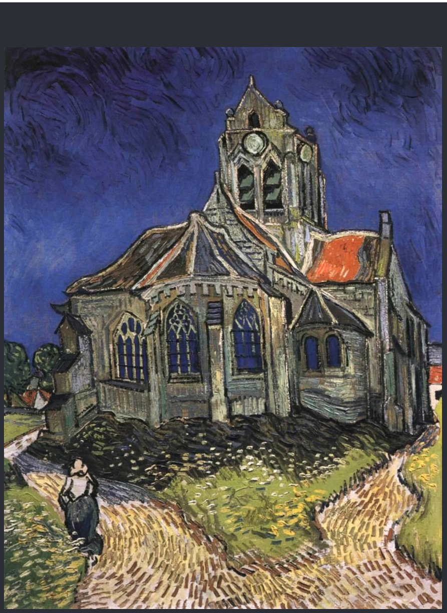
Semantic Art Understanding

In this painting the church in Auvers has been transformed by the artist into a vision using form and colour. Painted in portrait format, the church towers up before the onlooker like a fortification. The path leading to it forks in the foreground into two narrow paths passing the church on either side. On the path to the left, her back turned toward us, a peasant woman is walking into the distance. The path is bathed in light, while the church is viewed against the backdrop of a dark blue sky that merges with the black-blue of the night sky at the edges of the picture. The brushwork is restless and full of movement, and the forms of the church are distorted in the Expressionist manner.



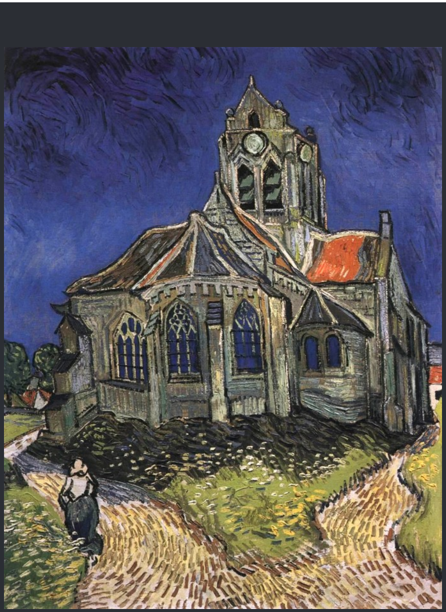
Semantic Art Understanding

In this painting the church in Auvers has been transformed by the artist into a vision using form and colour. Painted in portrait format, the church towers up before the onlooker like a fortification. The path leading to it forks in the foreground into two narrow paths passing the church on either side. On the path to the left, her back turned toward us, a peasant woman is walking into the distance. The path is bathed in light, while the church is viewed against the backdrop of a dark blue sky that merges with the black-blue of the night sky at the edges of the picture. The brushwork is restless and full of movement, and the forms of the church are distorted in the Expressionist manner.



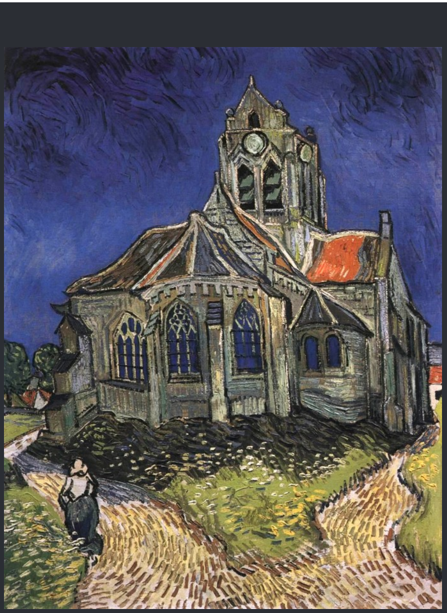
Semantic Art Understanding

In this painting the church in Auvers has been transformed by the artist into a vision using form and colour. Painted in portrait format, the church towers up before the onlooker like a fortification. The path leading to it forks in the foreground into two narrow paths passing the church on either side. On the path to the left, her back turned toward us, a peasant woman is walking into the distance. The path is bathed in light, while the church is viewed against the backdrop of a dark blue sky that merges with the black-blue of the night sky at the edges of the picture. The brushwork is restless and full of movement, and the forms of the church are distorted in the Expressionist manner.



Semantic Art Understanding

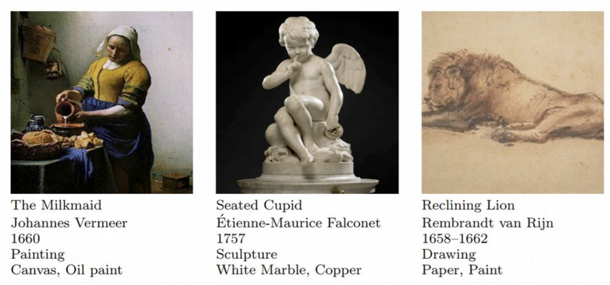
In this painting the church in Auvers has been transformed by the artist into a vision using form and colour. Painted in portrait format, the church towers up before the onlooker like a fortification. The path leading to it forks in the foreground into two narrow paths passing the church on either side. On the path to the left, her back turned toward us, a peasant woman is walking into the distance. The path is bathed in light, while the church is viewed against the backdrop of a dark blue sky that merges with the black-blue of the night sky at the edges of the picture. The brushwork is restless and full of movement, and the forms of the church are distorted in the Expressionist manner.



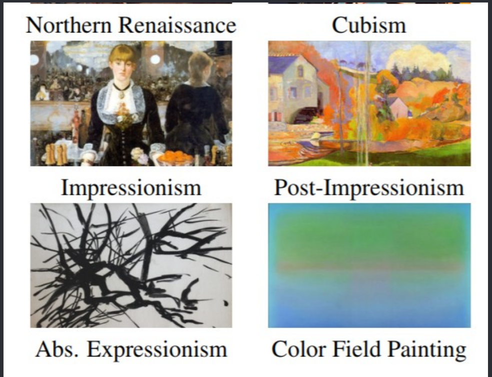
Related Work

PRINTART, 2012

Painting-91, 2014

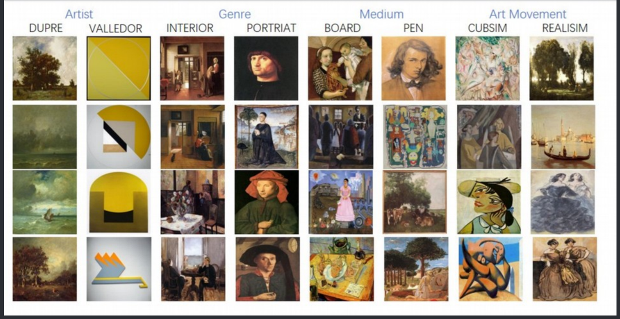


Rijksmuseum, 2014



Wikipaintings, 2014

Paintings Database, 2014



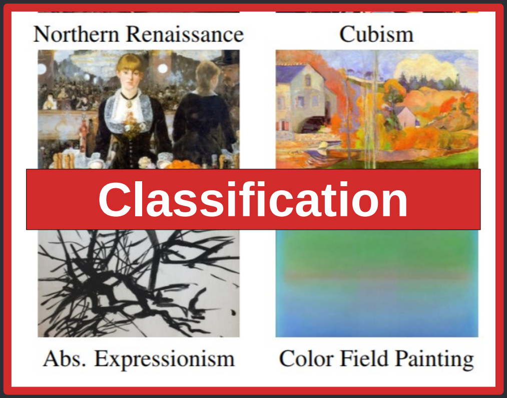
Art500k, 2016

Related Work

PRINTART, 2012

Painting-91, 2014

Rijksmuseum, 2014



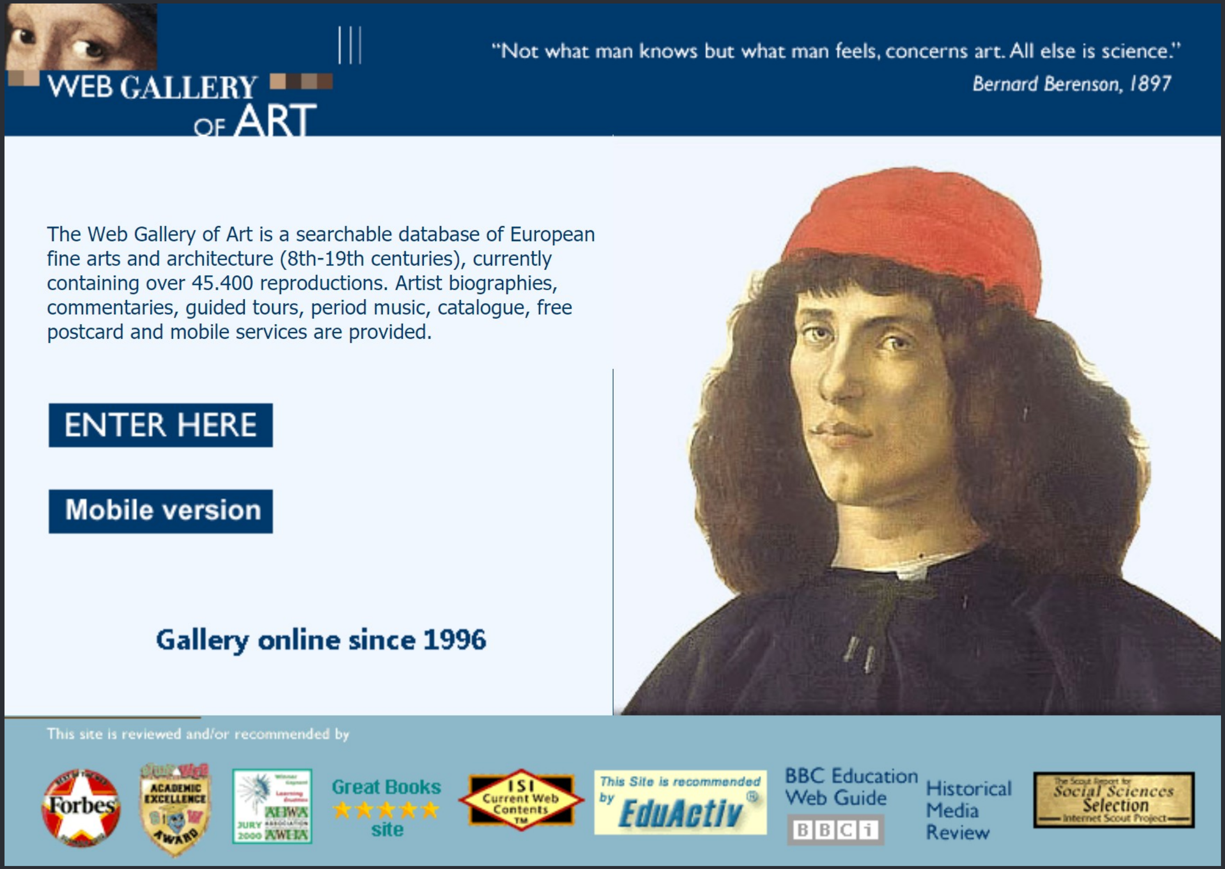
Wikipaintings, 2014

Paintings Database, 2014

Art500k, 2016

SemArt Dataset

Data collected from the Web Gallery of Art



"Not what man knows but what man feels, concerns art. All else is science."
Bernard Berenson, 1897

The Web Gallery of Art is a searchable database of European fine arts and architecture (8th-19th centuries), currently containing over 45.400 reproductions. Artist biographies, commentaries, guided tours, period music, catalogue, free postcard and mobile services are provided.

Gallery online since 1996

This site is reviewed and/or recommended by

Forbes
ACADEMIC EXCELLENCE AWARD
AEWA
JURY 2000 AWHA
Great Books ★★★★★ site
ISI Current Web Contents
This Site is recommended by EduActiv
BBC Education Web Guide
BBCi
Historical Media Review
The Scout Report for Social Sciences Selection Internet Scout Project

https://www.wga.hu/

SemArt Dataset

Each sample in the dataset is a triplet

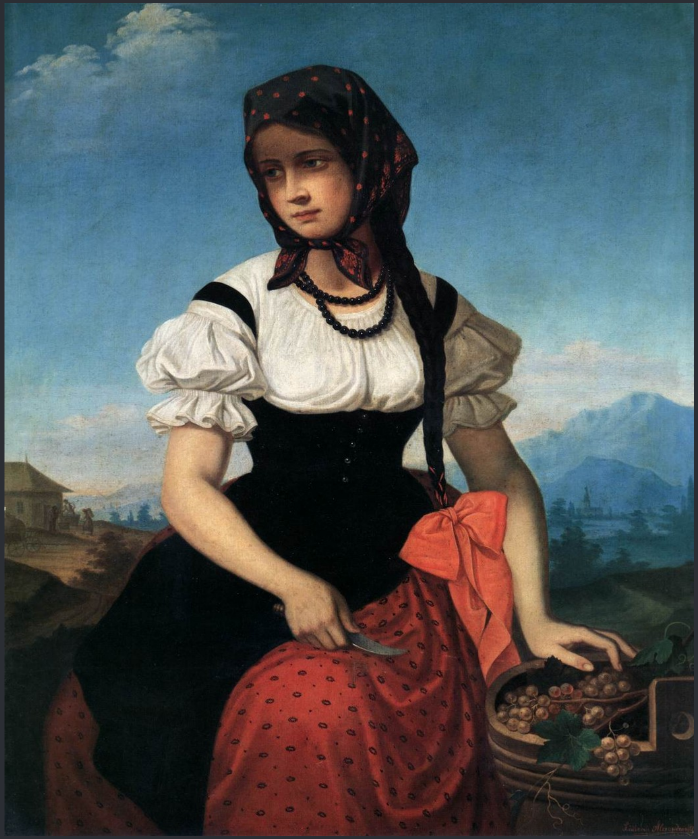
Title: Grape Harvest Girl
Author: Ljubomir Aleksandrova
Type: Genre
School: Other
Timeframe: 1851-1900

In Croatia, Bosnia and Herzegovina, and in northern Serbia, depending on the kind of harvest, people celebrate harvest season by dressing themselves with fruits of the harvest.

image, attributes and comments

SemArt Dataset

Each sample in the dataset is a triplet



Title: Grape Harvest Girl
Author: Ljubomir Aleksandrova
Type: Genre
School: Other
Timeframe: 1851-1900

In Croatia, Bosnia and Herzegovina, and in northern Serbia, depending on the kind of harvest, people celebrate harvest season by dressing themselves with fruits of the harvest.

image, attributes and comments

SemArt Dataset

Each sample in the dataset is a triplet

Title: Grape Harvest Girl
Author: Ljubomir Aleksandrova
Type: Genre
School: Other
Timeframe: 1851-1900

In Croatia, Bosnia and Herzegovina, and in northern Serbia, depending on the kind of harvest, people celebrate harvest season by dressing themselves with fruits of the harvest.

image, attributes and comments

SemArt Dataset

Each sample in the dataset is a triplet

Title: Grape Harvest Girl
Author: Ljubomir Aleksandrova
Type: Genre
School: Other
Timeframe: 1851-1900

In Croatia, Bosnia and Herzegovina, and in northern Serbia, depending on the kind of harvest, people celebrate harvest season by dressing themselves with fruits of the harvest.

image, attributes and comments

SemArt Dataset

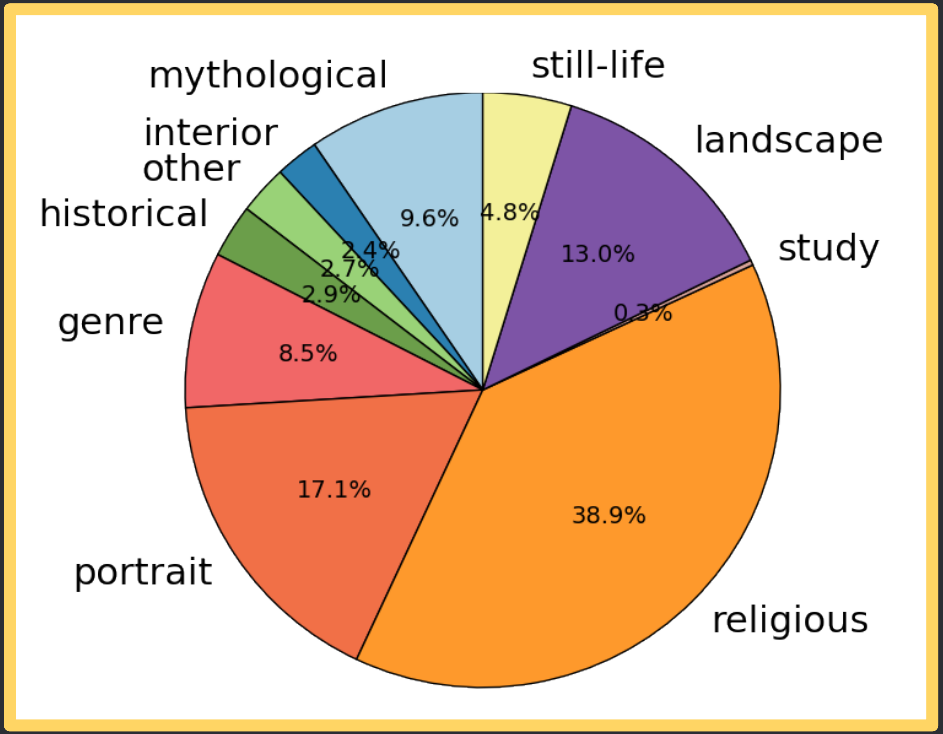
Attributes

Author, Title, Date, Technique, Type, School, Timeframe

SemArt Dataset

Attributes

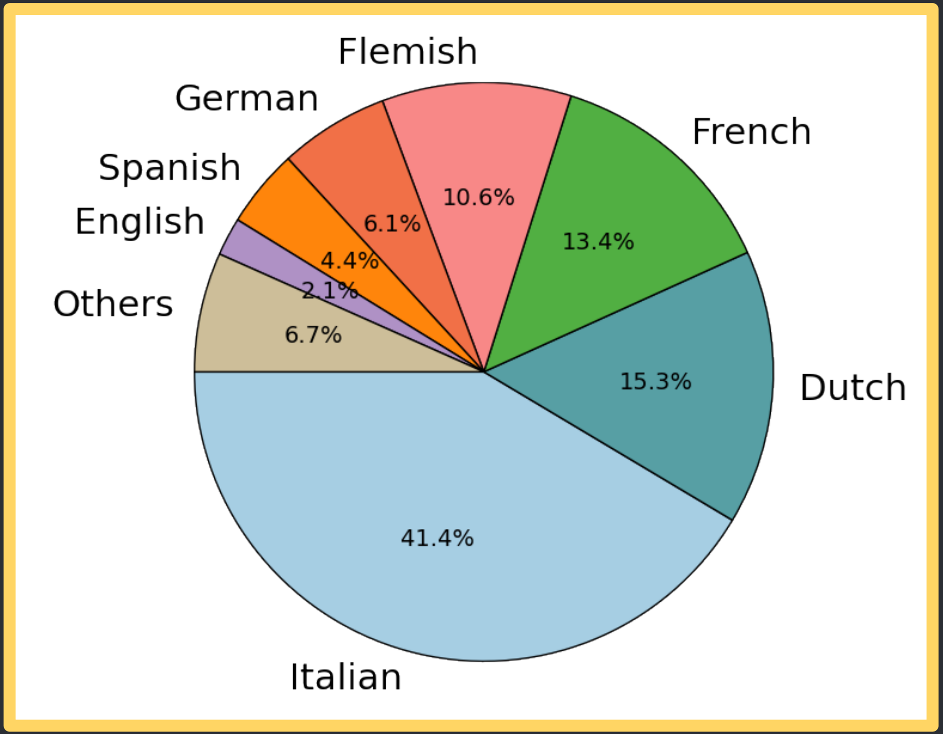
Author, Title, Date, Technique, **Type**, School, Timeframe



SemArt Dataset

Attributes

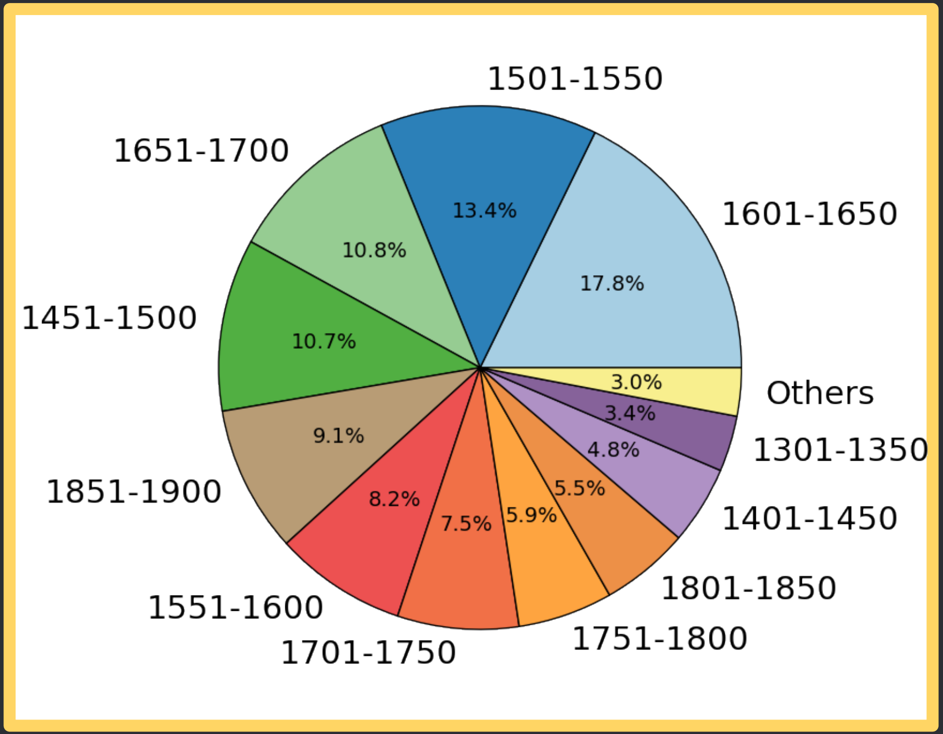
Author, Title, Date, Technique, Type, School, Timeframe



SemArt Dataset

Attributes

Author, Title, Date, Technique, Type, School, Timeframe



SemArt Dataset

Comments

70% with 100 words or less

The painting depicts a still-life with roses, tulips and other flowers resting on a ledge. It demonstrates the elegance, refinement, and technical brilliance cultivated during the painter's formative years in Italy.

In Croatia, Bosnia and Herzegovina, and in northern Serbia, depending on the kind of harvest, people celebrate harvest season by dressing themselves with fruits of the harvest.

This landscape depicts ships moored off a rocky coastline with fishermen unloading their catch.

This view of Florence is one of a number of views by Lear based upon on the spot sketches he produced in 1861

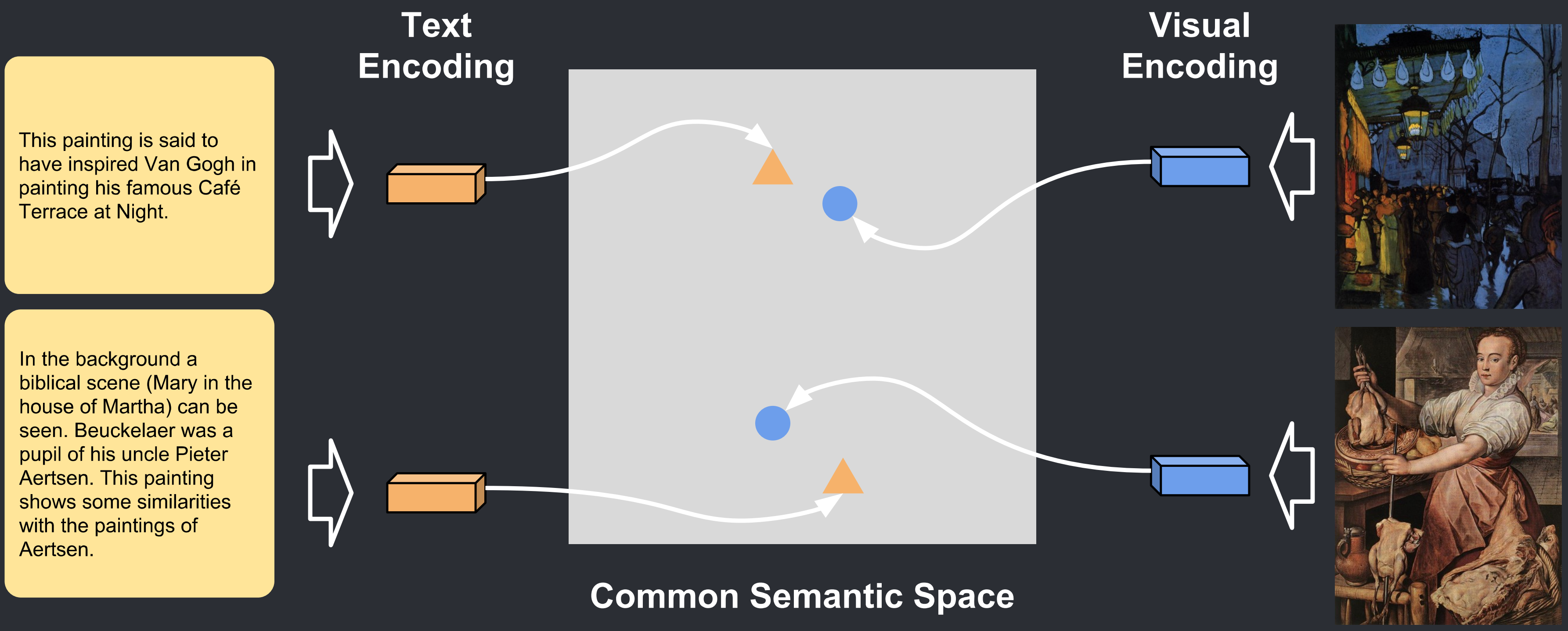
SemArt Dataset

Data splits

Partition	Num. Triplets	%
Training	19,244	90
Validation	1,069	5
Test	1,069	5
Total	21,383	100

Text2Art Challenge

Multi-modal retrieval



This painting is said to have inspired Van Gogh in painting his famous Café Terrace at Night.

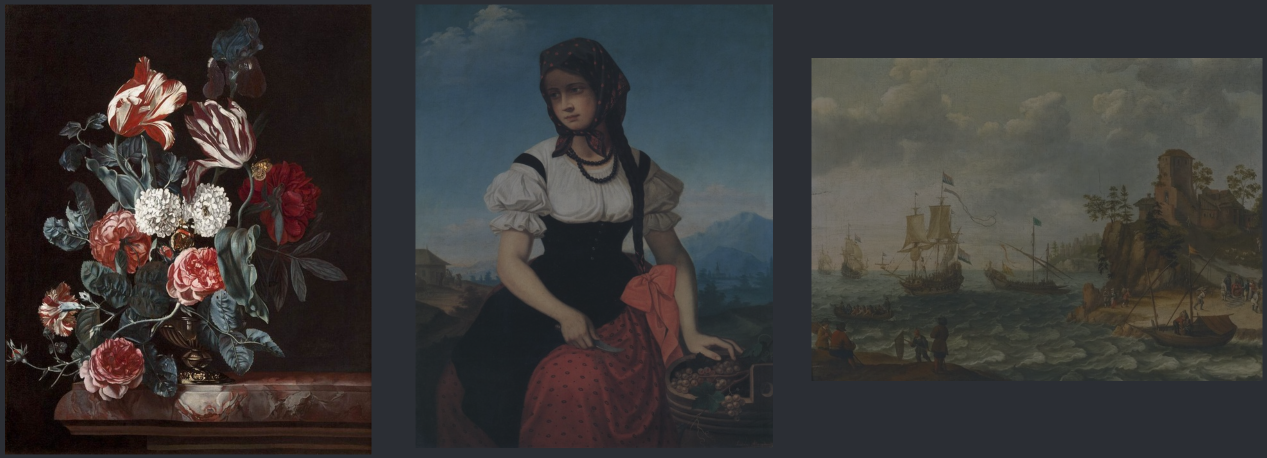
In the background a biblical scene (Mary in the house of Martha) can be seen. Beuckelaer was a pupil of his uncle Pieter Aertsen. This painting shows some similarities with the paintings of Aertsen.

Text2Art Challenge

Text-to-Image Retrieval

$$img^* = \arg\min_{img_j \in K} d(p_k^{text}, p_j^{vis})$$

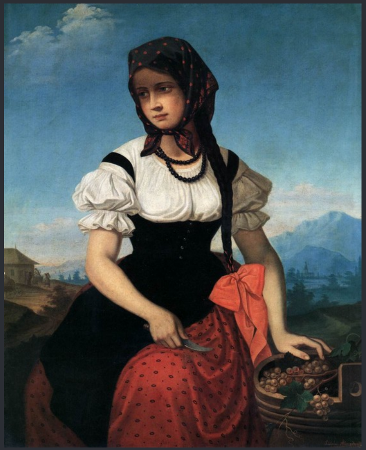
The painting depicts a still-life with roses, tulips and other flowers resting on a ledge. It demonstrates the elegance, refinement, and technical brilliance cultivated during the painter's formative years in Italy.



Text2Art Challenge

Image-to-Text Retrieval

$$com^*, att^* = \underset{com_j, att_j \in K}{\arg\min}\ d(p_j^{text}, p_k^{vis})$$



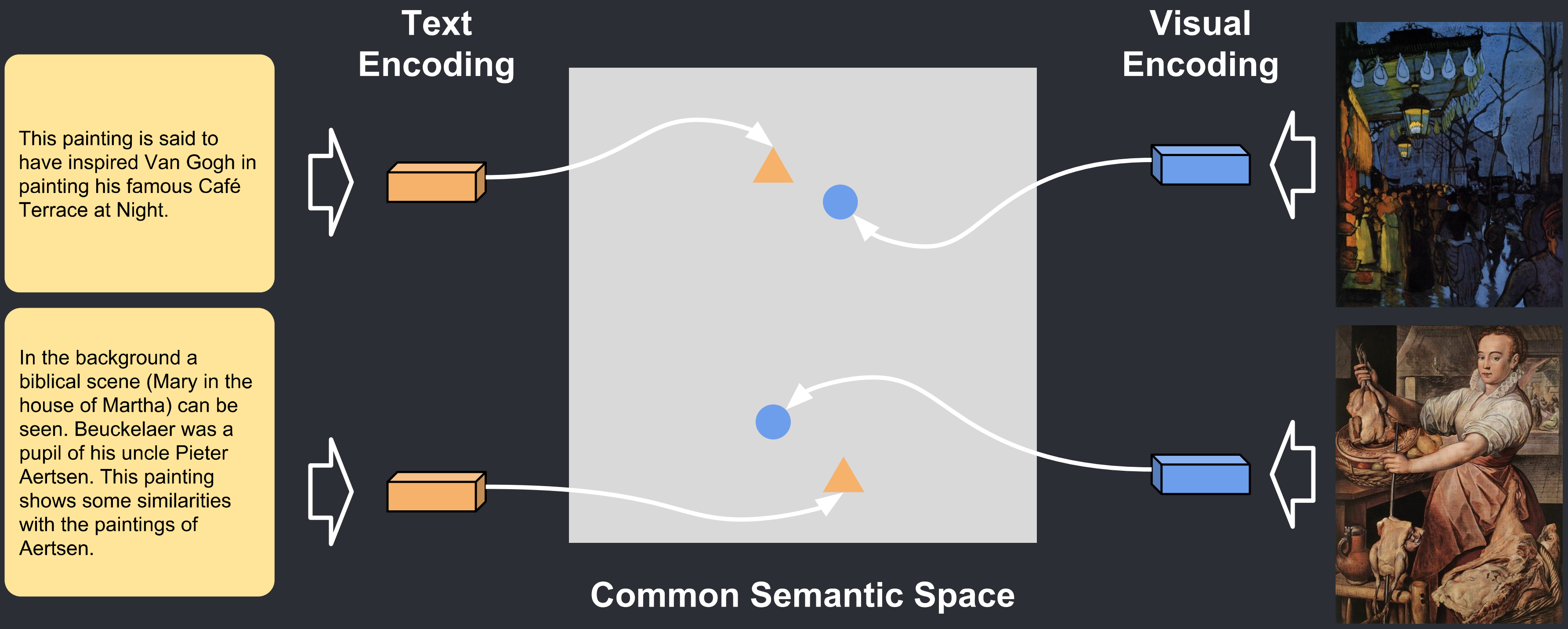
The painting depicts a still-life with roses, tulips and other flowers resting on a ledge. It demonstrates the elegance, refinement, and technical brilliance cultivated during the painter's formative years in Italy.

In Croatia, Bosnia and Herzegovina, and in northern Serbia, depending on the kind of harvest, people celebrate harvest season by dressing themselves with fruits of the harvest.

This landscape depicts ships moored off a rocky coastline with fishermen unloading their catch.

Models

We study 3 fundamental parts: visual encoding, text encoding and multi-modal transformation



Text Encoding

Visual Encoding

This painting is said to have inspired Van Gogh in painting his famous Café Terrace at Night.

In the background a biblical scene (Mary in the house of Martha) can be seen. Beuckelaer was a pupil of his uncle Pieter Aertsen. This painting shows some similarities with the paintings of Aertsen.

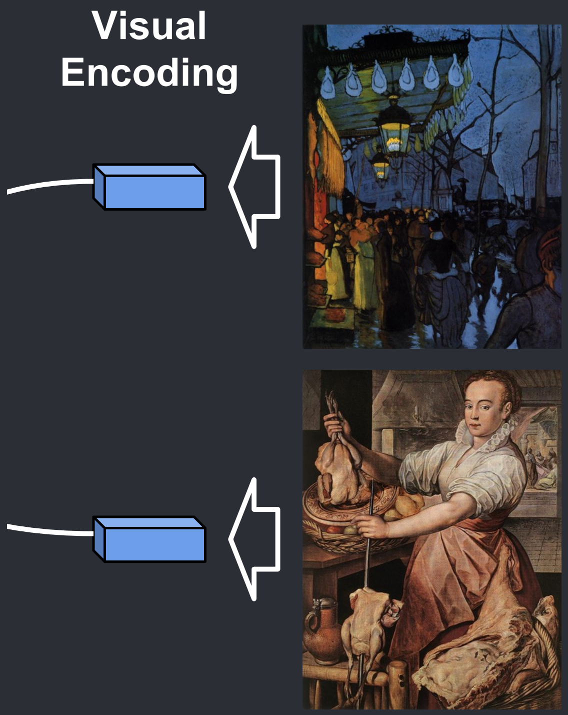
Common Semantic Space

Models

Visual Encoding

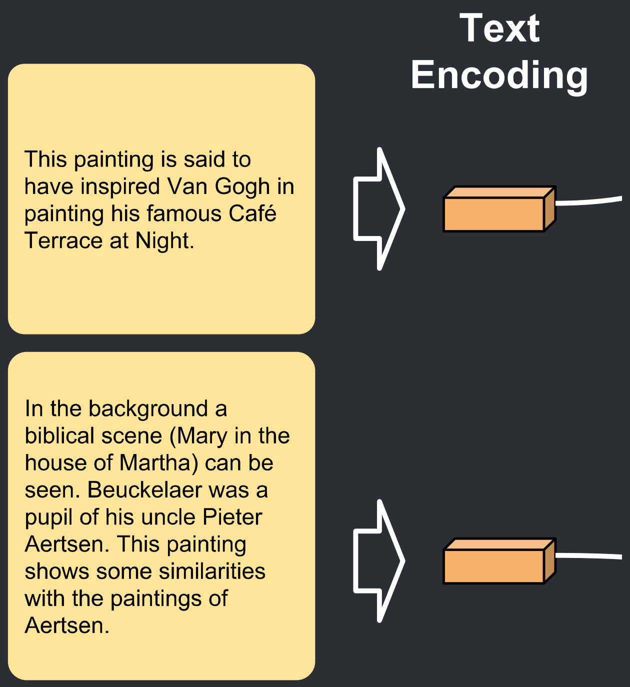
We consider the following visual encoders:

- VGG16 (Simonyan and Zisserman, 2014)
- ResNets (He et al. 2016)
- RMAC (Tolias et al. 2016)



Models

Textual Encoding



Text Encoding

This painting is said to have inspired Van Gogh in painting his famous Café Terrace at Night.

In the background a biblical scene (Mary in the house of Martha) can be seen. Beuckelaer was a pupil of his uncle Pieter Aertsen. This painting shows some similarities with the paintings of Aertsen.

We encode titles and comments independently and concatenate their vectors.

We consider the following text encoders:

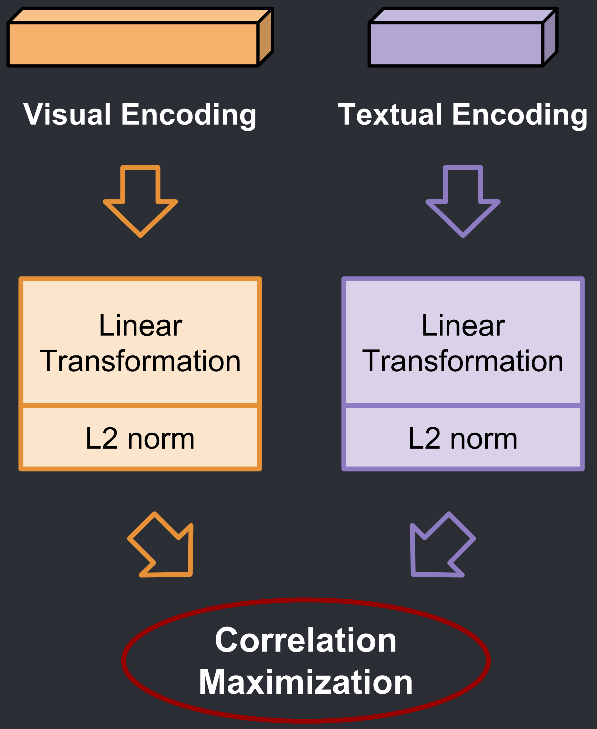
- BOW (bag-of-words)
- MLP (multilayer preceptron)
- RNN (recurrent neural networks)

Models

Multi-Modal Transformation

We map visual and text encodings into the common semantic space using the following methods:

CCA, CML and AMD

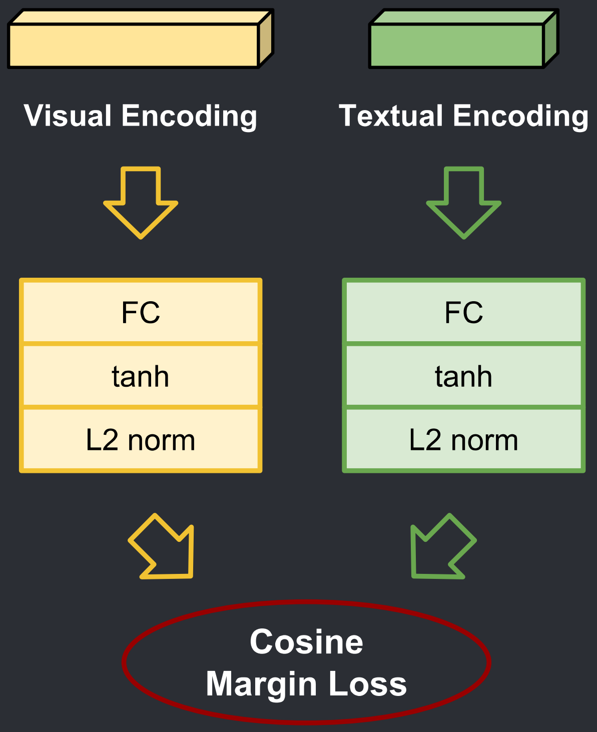


Models

Multi-Modal Transformation

We map visual and text encodings into a common semantic space using the following methods:

CCA, CML and AMD



Visual Encoding

Textual Encoding

FC

tanh

L2 norm

FC

tanh

L2 norm

Cosine Margin Loss

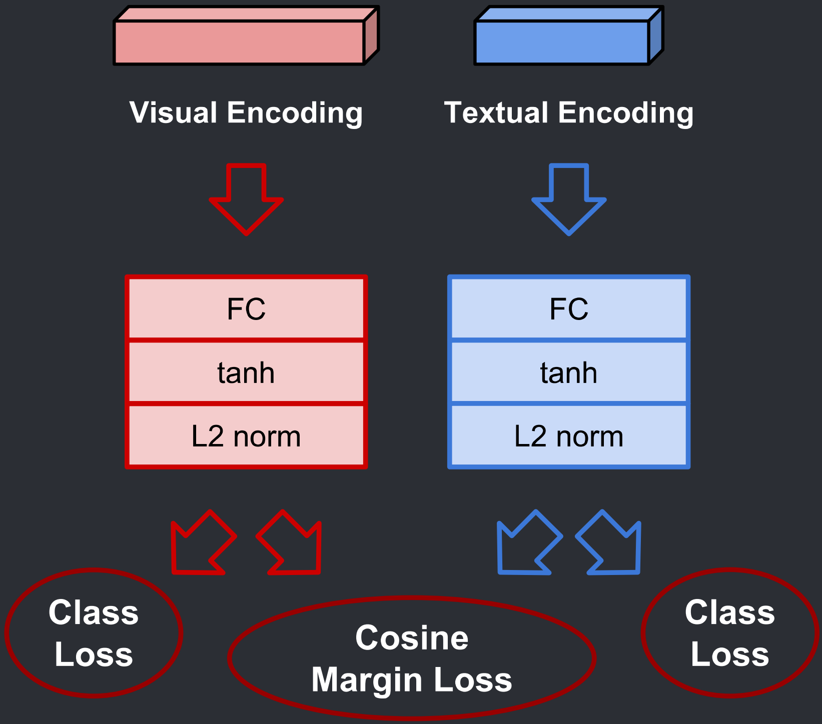
$$L_{CML}(p_k^{vis}, p_j^{text}) = \begin{cases} 1 - \cos(p_k^{vis}, p_j^{text}), & \text{if } k = j \\ \max(0, \cos(p_k^{vis}, p_j^{text}) - m), & \text{if } k \neq j \end{cases}$$

Models

Multi-Modal Transformation

We map visual and text encodings into a common semantic space using the following methods:

CCA, CML and AMD

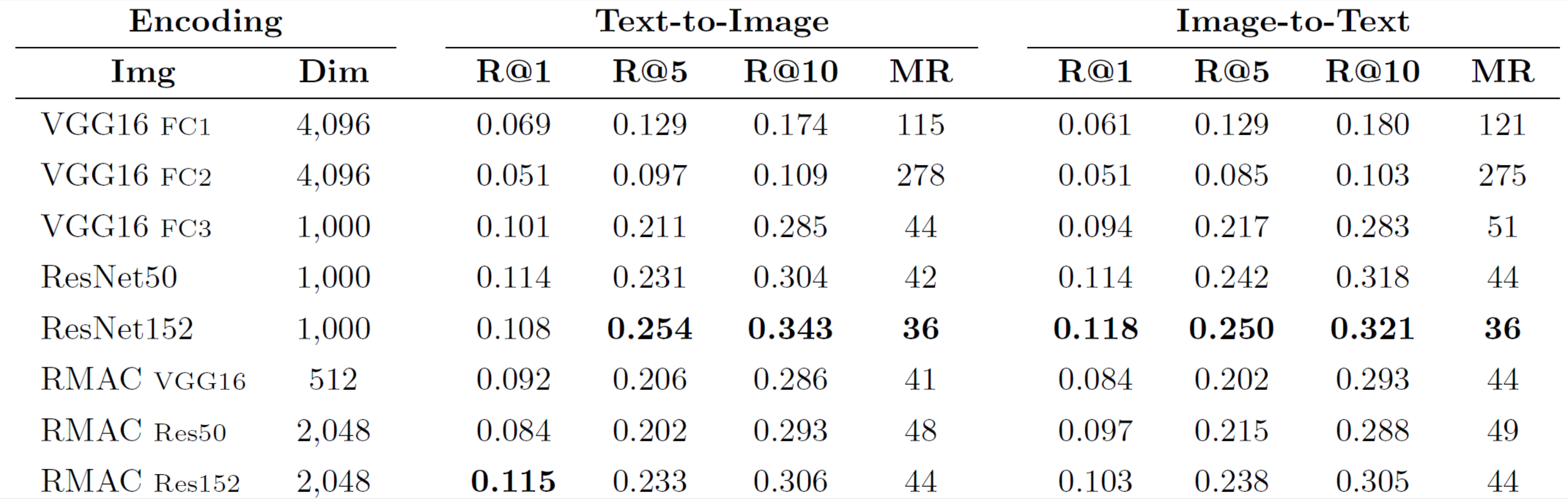


$$L_{AMD}(p_k^{text}, p_j^{vis}, l_{p_k^{text}}, l_{p_j^{vis}}) = (1-2\alpha)L_{CML}(p_k^{text}, p_j^{vis})$$

$$+\alpha L_{META}(p_k^{text}, l_{p_k^{text}}) + \alpha L_{META}(p_j^{vis}, l_{p_j^{vis}})$$

Evaluation

Visual Encoding

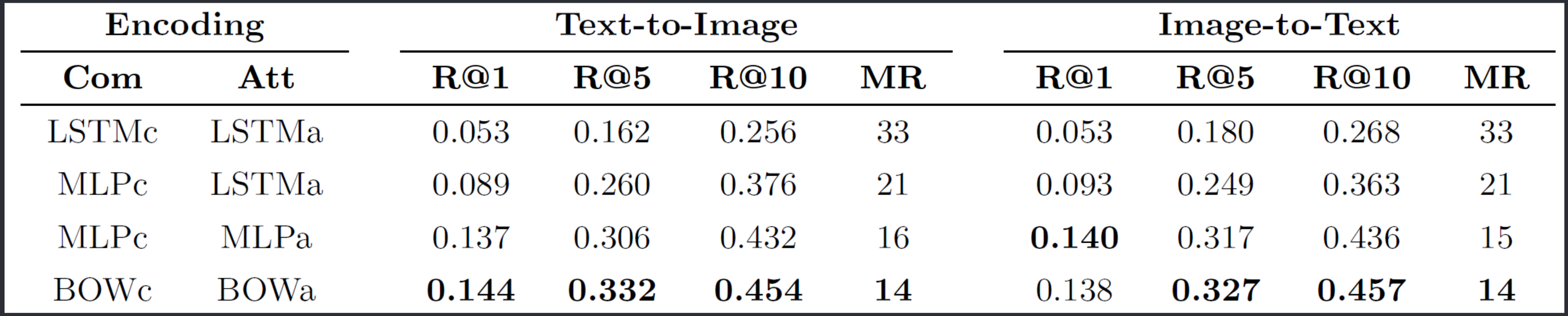


Encoding		Text-to-Image				Image-to-Text			
Img	**Dim**	**R@1**	**R@5**	**R@10**	**MR**	**R@1**	**R@5**	**R@10**	**MR**
VGG16 FC1	4,096	0.069	0.129	0.174	115	0.061	0.129	0.180	121
VGG16 FC2	4,096	0.051	0.097	0.109	278	0.051	0.085	0.103	275
VGG16 FC3	1,000	0.101	0.211	0.285	44	0.094	0.217	0.283	51
ResNet50	1,000	0.114	0.231	0.304	42	0.114	0.242	0.318	44
ResNet152	1,000	0.108	**0.254**	**0.343**	**36**	**0.118**	**0.250**	**0.321**	**36**
RMAC VGG16	512	0.092	0.206	0.286	41	0.084	0.202	0.293	44
RMAC Res50	2,048	0.084	0.202	0.293	48	0.097	0.215	0.288	49
RMAC Res152	2,048	**0.115**	0.233	0.306	44	0.103	0.238	0.305	44

ResNet152 is the best visual encoder

Evaluation

Textual Encoding

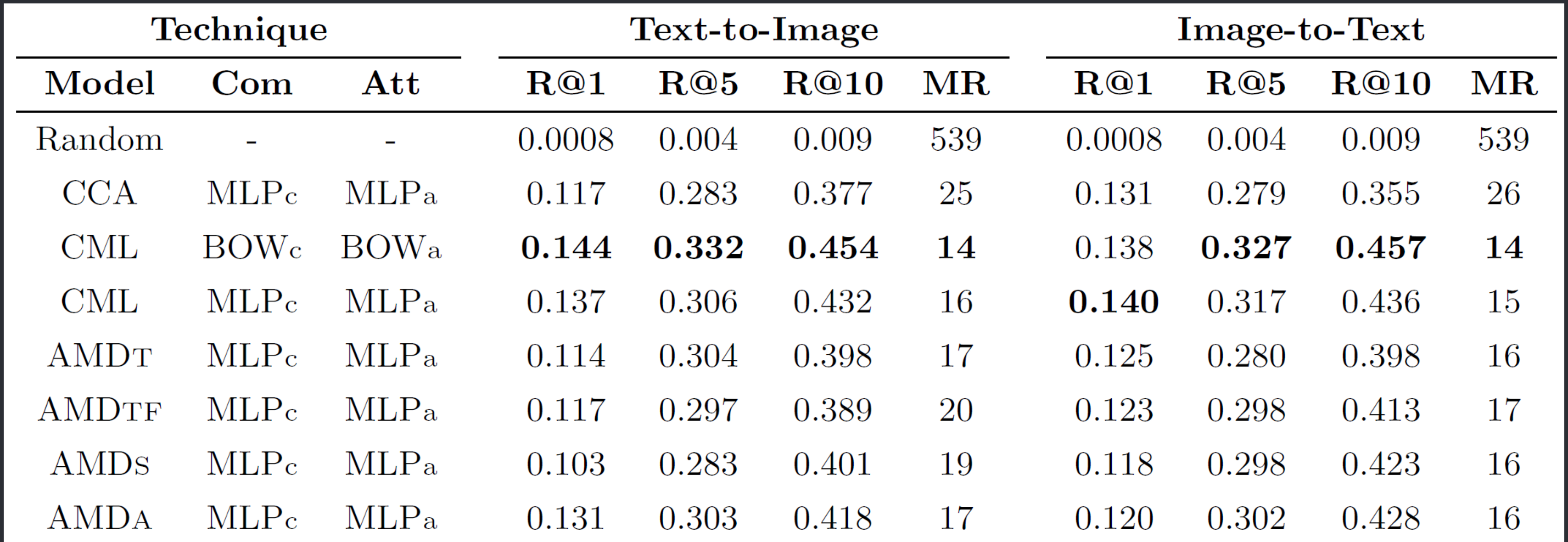


Encoding		Text-to-Image				Image-to-Text			
Com	**Att**	**R@1**	**R@5**	**R@10**	**MR**	**R@1**	**R@5**	**R@10**	**MR**
LSTMc	LSTMa	0.053	0.162	0.256	33	0.053	0.180	0.268	33
MLPc	LSTMa	0.089	0.260	0.376	21	0.093	0.249	0.363	21
MLPc	MLPa	0.137	0.306	0.432	16	**0.140**	0.317	0.436	15
BOWc	BOWa	**0.144**	**0.332**	**0.454**	**14**	0.138	**0.327**	**0.457**	**14**

Simple BOW performs better than recurrent models, as observed in other multi-modal retrieval work (Wang et al. 2018)

Evaluation

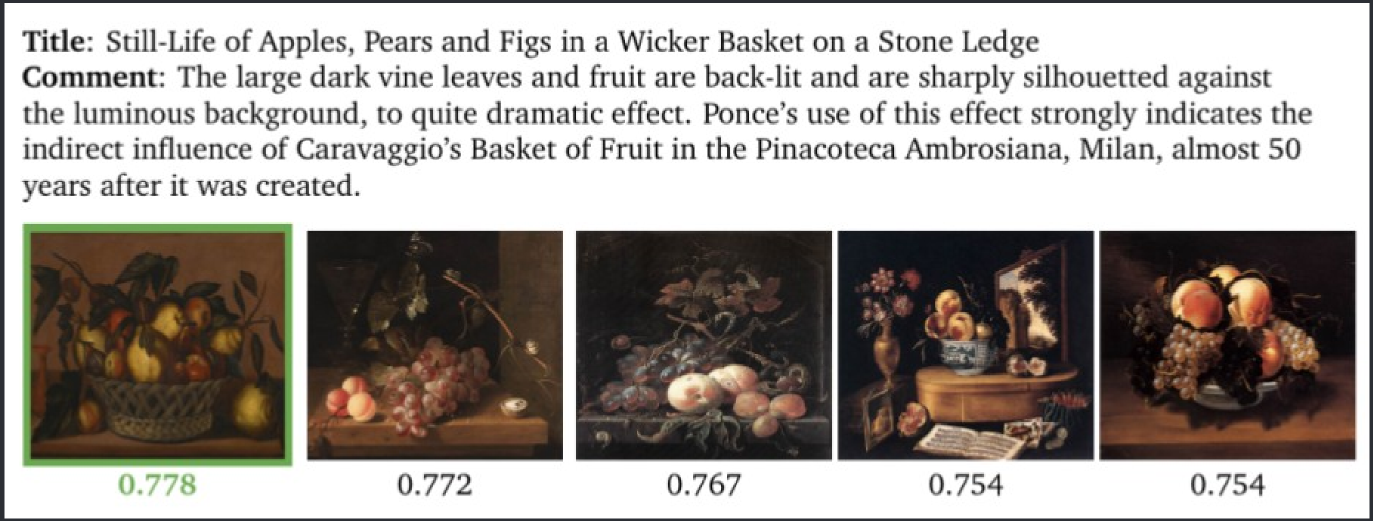
Multi-Modal Transformation



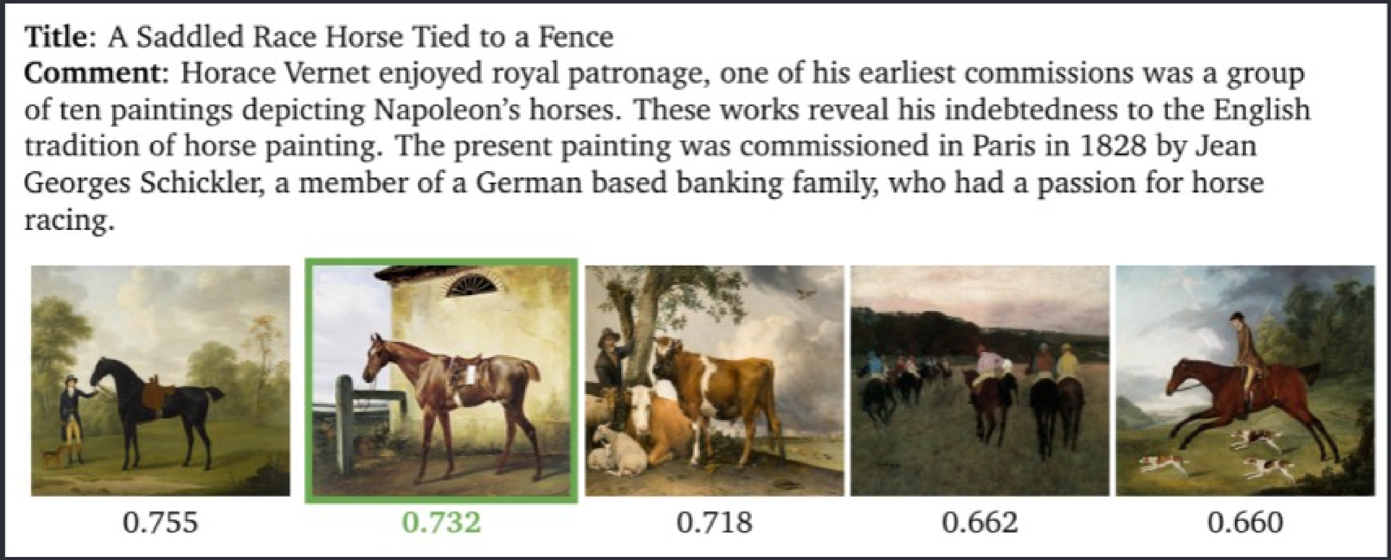
Technique			Text-to-Image				Image-to-Text			
Model	**Com**	**Att**	**R@1**	**R@5**	**R@10**	**MR**	**R@1**	**R@5**	**R@10**	**MR**
Random	-	-	0.0008	0.004	0.009	539	0.0008	0.004	0.009	539
CCA	MLPc	MLPa	0.117	0.283	0.377	25	0.131	0.279	0.355	26
CML	BOWc	BOWa	**0.144**	**0.332**	**0.454**	**14**	0.138	**0.327**	**0.457**	**14**
CML	MLPc	MLPa	0.137	0.306	0.432	16	**0.140**	0.317	0.436	15
AMDT	MLPc	MLPa	0.114	0.304	0.398	17	0.125	0.280	0.398	16
AMDTF	MLPc	MLPa	0.117	0.297	0.389	20	0.123	0.298	0.413	17
AMDS	MLPc	MLPa	0.103	0.283	0.401	19	0.118	0.298	0.423	16
AMDA	MLPc	MLPa	0.131	0.303	0.418	17	0.120	0.302	0.428	16

CML is the best model

Qualitative Results



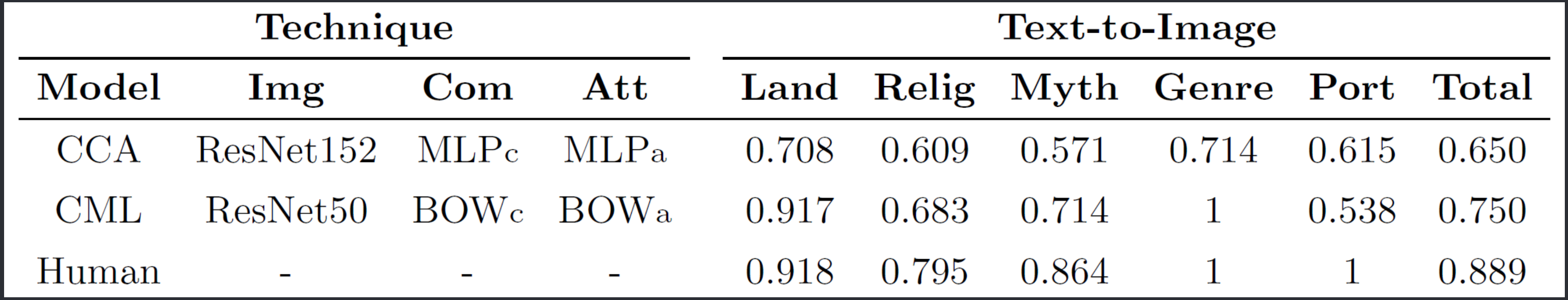
Title: Still-Life of Apples, Pears and Figs in a Wicker Basket on a Stone Ledge
Comment: The large dark vine leaves and fruit are back-lit and are sharply silhouetted against the luminous background, to quite dramatic effect. Ponce's use of this effect strongly indicates the indirect influence of Caravaggio's Basket of Fruit in the Pinacoteca Ambrosiana, Milan, almost 50 years after it was created.



Title: A Saddled Race Horse Tied to a Fence
Comment: Horace Vernet enjoyed royal patronage, one of his earliest commissions was a group of ten paintings depicting Napoleon's horses. These works reveal his indebtedness to the English tradition of horse painting. The present painting was commissioned in Paris in 1828 by Jean Georges Schickler, a member of a German based banking family, who had a passion for horse racing.

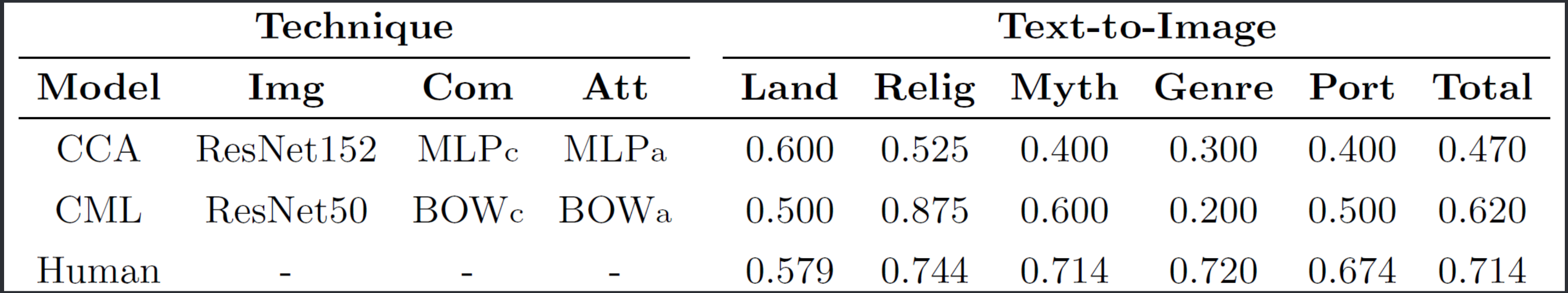
Human Evaluation

Easy



Technique				Text-to-Image					
Model	**Img**	**Com**	**Att**	**Land**	**Relig**	**Myth**	**Genre**	**Port**	**Total**
CCA	ResNet152	MLP$_c$	MLP$_a$	0.708	0.609	0.571	0.714	0.615	0.650
CML	ResNet50	BOW$_c$	BOW$_a$	0.917	0.683	0.714	1	0.538	0.750
Human	-	-	-	0.918	0.795	0.864	1	1	0.889

Difficult



Technique				Text-to-Image					
Model	**Img**	**Com**	**Att**	**Land**	**Relig**	**Myth**	**Genre**	**Port**	**Total**
CCA	ResNet152	MLP$_c$	MLP$_a$	0.600	0.525	0.400	0.300	0.400	0.470
CML	ResNet50	BOW$_c$	BOW$_a$	0.500	0.875	0.600	0.200	0.500	0.620
Human	-	-	-	0.579	0.744	0.714	0.720	0.674	0.714

Summary

- SemArt dataset for semantic art understanding

Summary

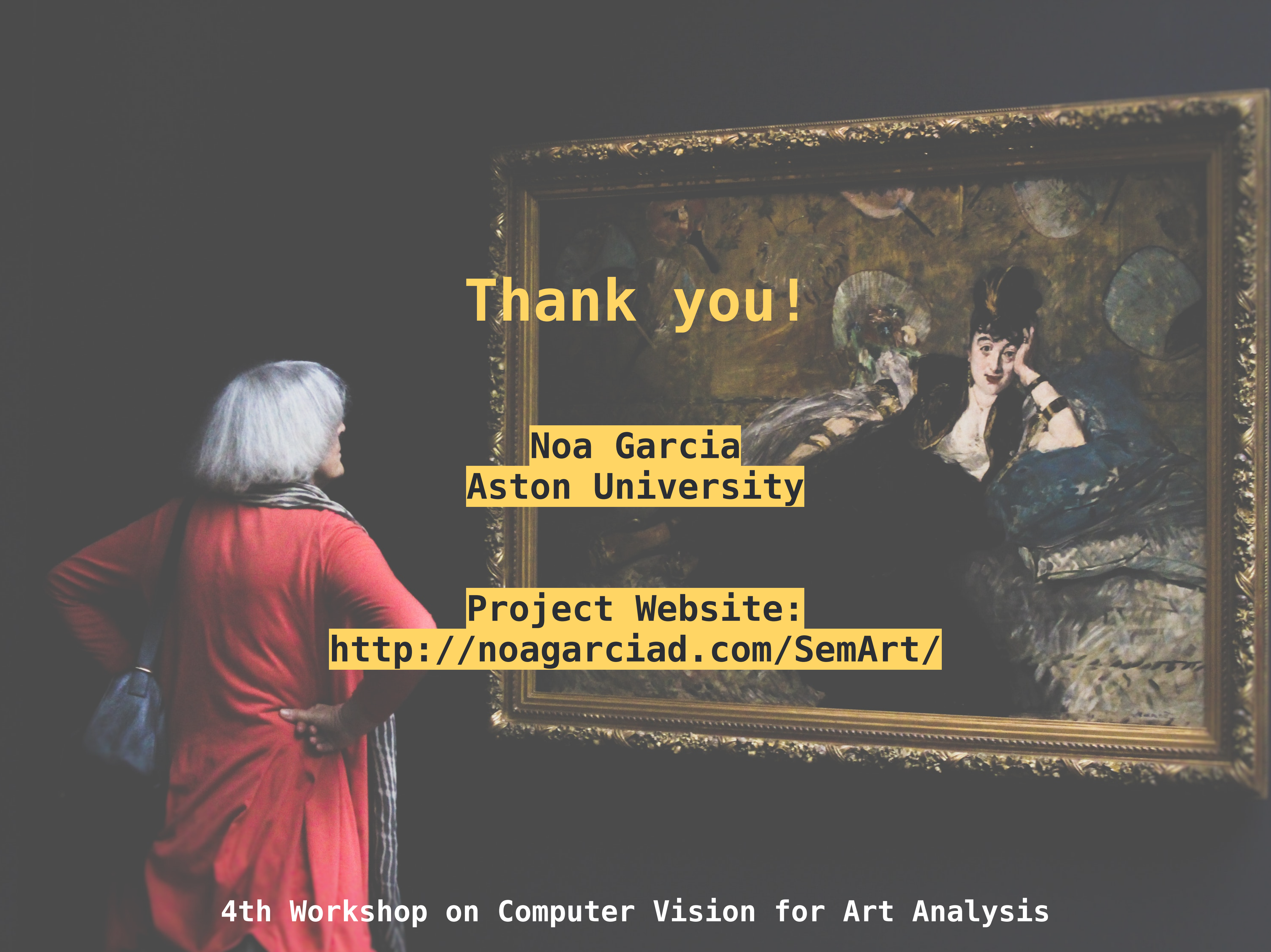
- SemArt dataset for semantic art understanding
- Text2Art challenge as a retrieval task

Summary

- SemArt dataset for semantic art understanding
- Text2Art challenge as a retrieval task
- Best model based on ResNet, BOW and CML

Summary

- SemArt dataset for semantic art understanding
- Text2Art challenge as a retrieval task
- Best model based on ResNet, BOW and CML
- Not that far from human performance



Thank you!

Noa Garcia
Aston University

Project Website:
http://noagarciad.com/SemArt/