Painting-to-3D model alignment via discriminative visual elements

Mathieu Aubry[1], Bryan Russell[2], Josef Sivic[1]

[1] INRIA, WILLOW project-team, École Normale Supérieure
[2] Intel Labs

To appear in Transactions on Graphics (ToG)

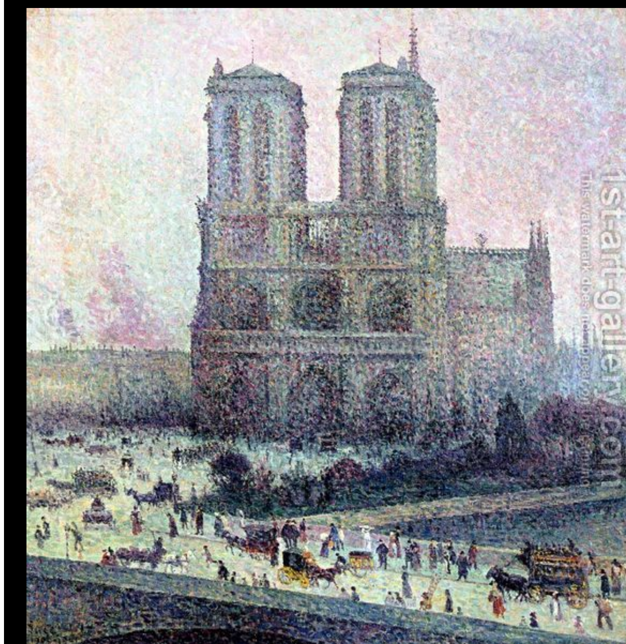
Goal

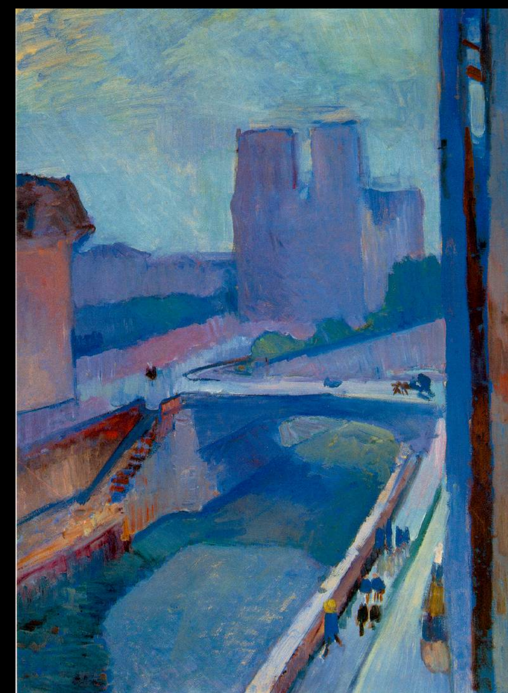
Inputs: paintings, drawings, historical photographs, reference 3D model

Output: recovered artist/camera viewpoints

Why do this?

There are many non-photographic depictions of our world





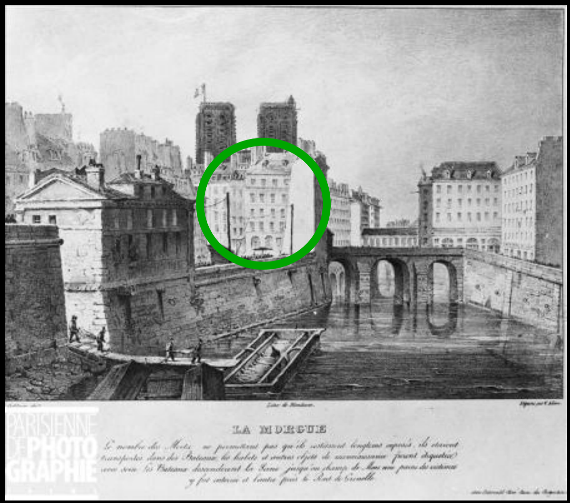
Ultimate goal: to reason about these depictions

Applications

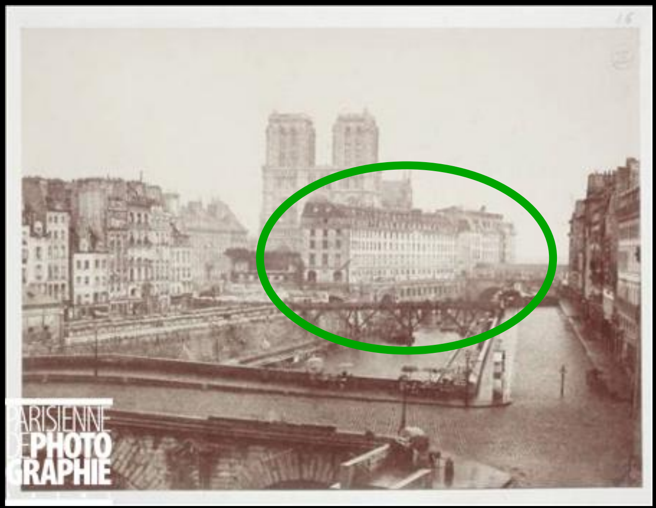
New ways to access archives for

archaeology, history or architecture

Example: evolution of a particular place over time



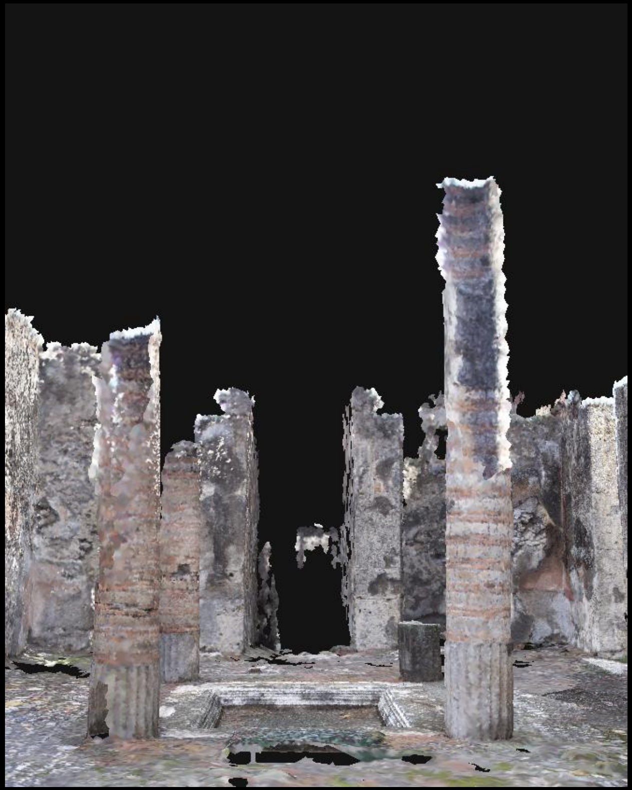
1830

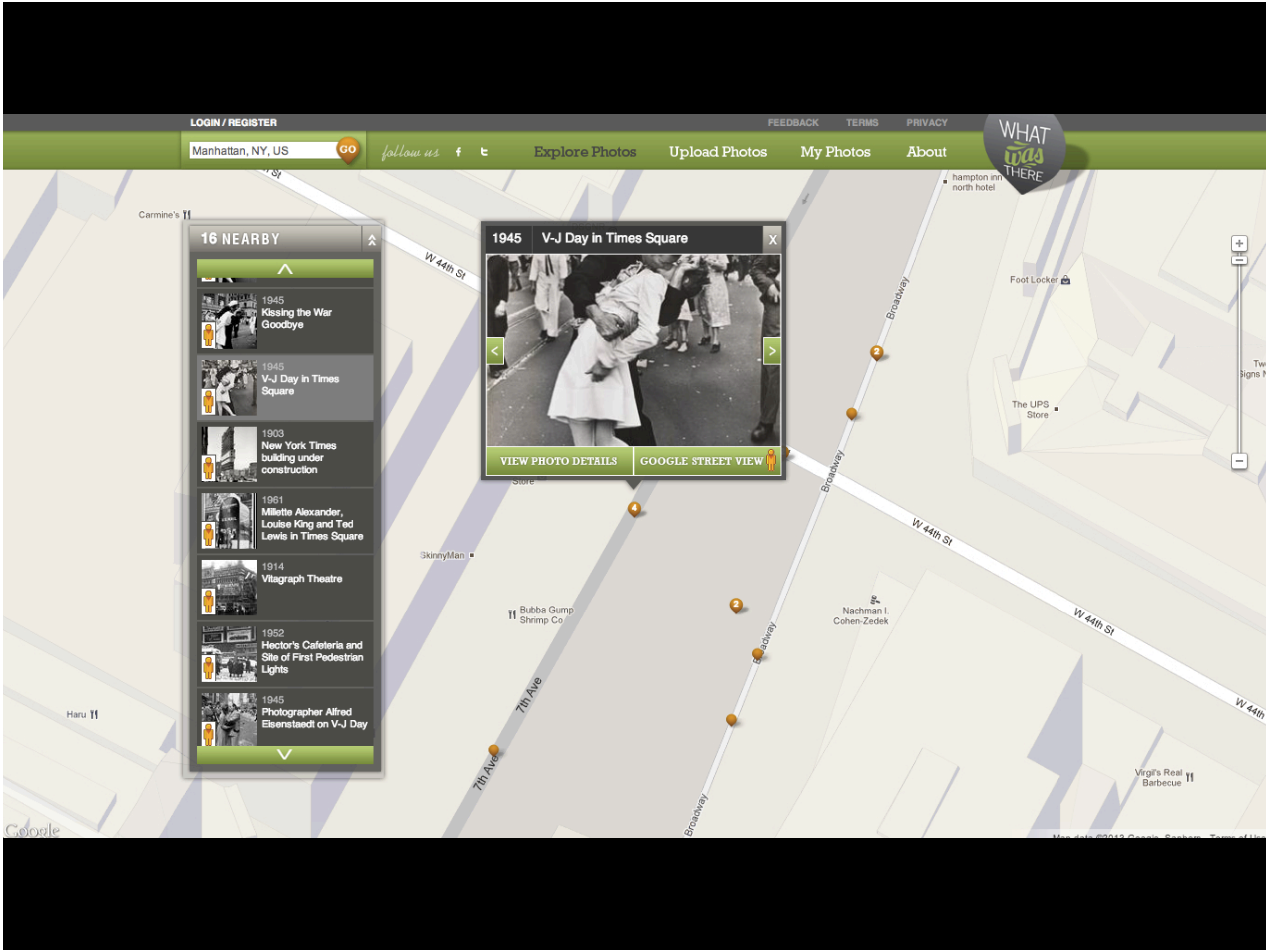


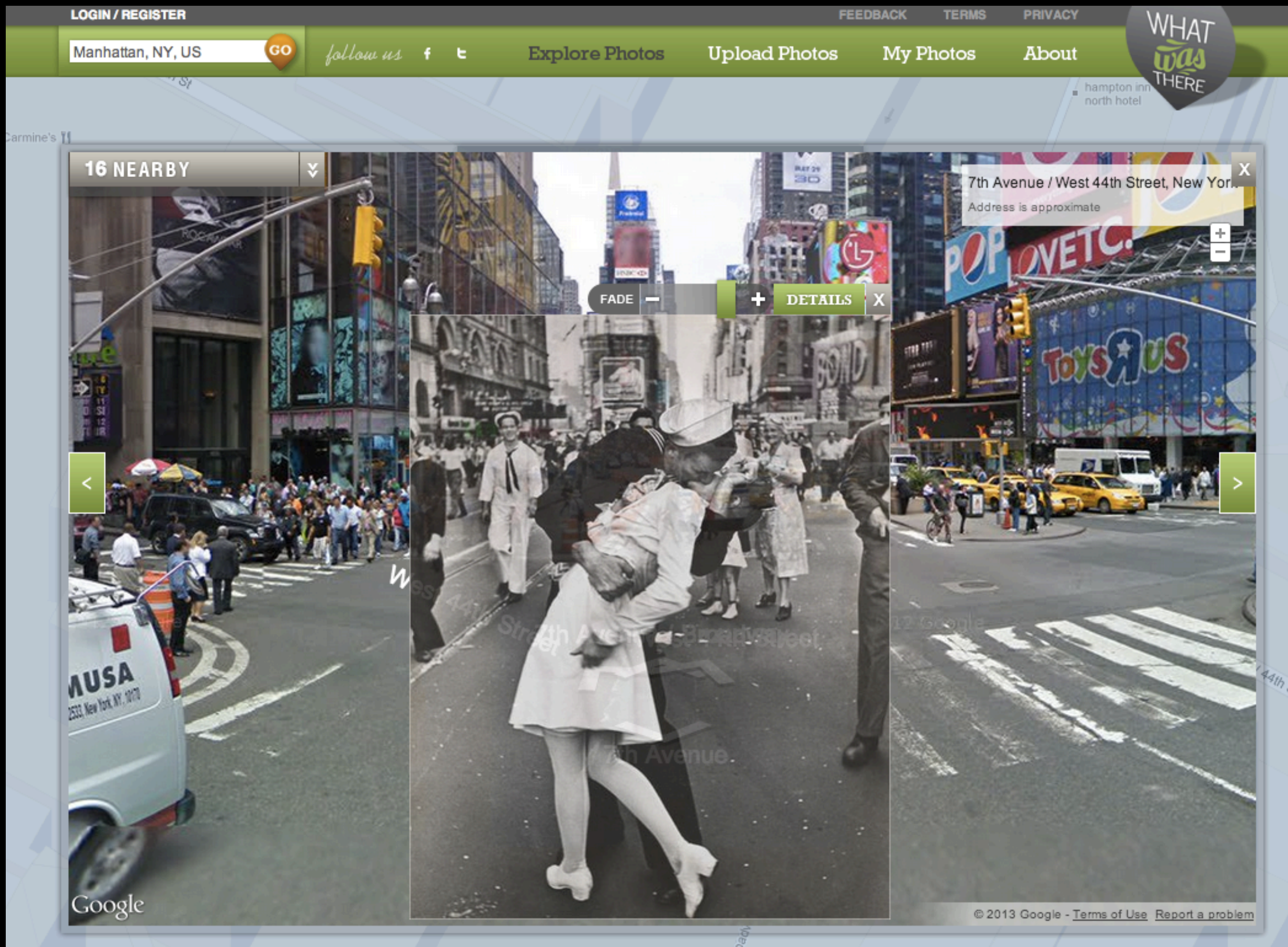
1852

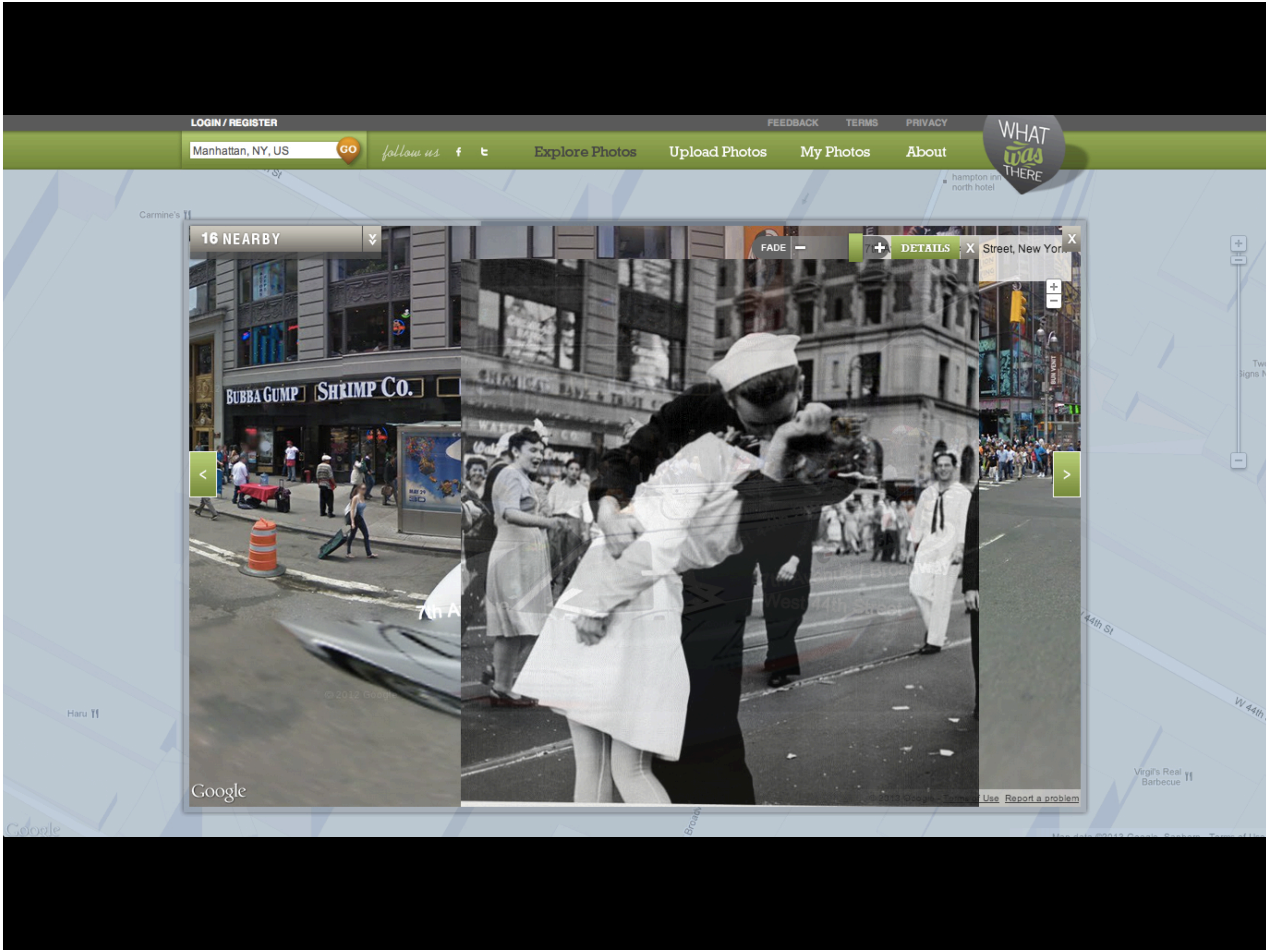
1900

Application : archaeology




LOGIN / REGISTER
FEEDBACK
TERMS
PRIVACY
WHAT was THERE
Manhattan, NY, US
GO
follow us
Explore Photos
Upload Photos
My Photos
About
16 NEARBY
1945
Kissing the War Goodbye
1945
V-J Day in Times Square
1903
New York Times building under construction
1961
Millette Alexander, Louise King and Ted Lewis in Times Square
1914
Vitagraph Theatre
1952
Hector's Cafeteria and Site of First Pedestrian Lights
1945
Photographer Alfred Eisenstaedt on V-J Day
1945
V-J Day in Times Square
X
VIEW PHOTO DETAILS
GOOGLE STREET VIEW
Carmine's
hampton inn north hotel
Foot Locker
The UPS Store
W 44th St
Broadway
7th Ave
SkinnyMan
Bubba Gump Shrimp Co
Nachman I. Cohen-Zedek
Haru
Virgil's Real Barbecue
Google


LOGIN / REGISTER
FEEDBACK
TERMS
PRIVACY
Manhattan, NY, US
GO
follow us
Explore Photos
Upload Photos
My Photos
About
WHAT was THERE
hampton inn
north hotel
16 NEARBY
7th Avenue / West 44th Street, New Yor
Address is approximate
FADE
DETAILS
X
Google
© 2013 Google - Terms of Use Report a problem

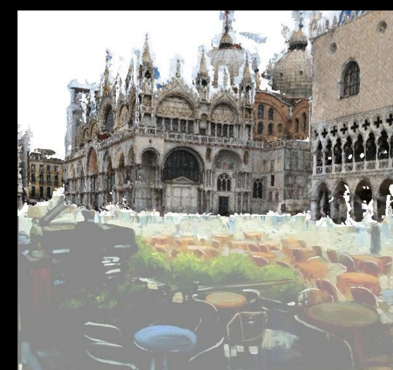


LOGIN / REGISTER
FEEDBACK
TERMS
PRIVACY
Manhattan, NY, US
GO
follow us
Explore Photos
Upload Photos
My Photos
About
WHAT was THERE
16 NEARBY
FADE
DETAILS
Street, New Yor
BUBBA GUMP SHRIMP Co.
Google
Report a problem

Problem statement

Inputs

Output



3D model

Painting

Camera parameters $\boxed{\Theta}$

Camera center, rotation, principal point, focal length

Let's try to run Bundler…

Step 1: Compute putative correspondences using SIFT key point matching

Difficulty in finding correspondences

Color, geometry, illumination, shading, shadows and texture may be rendered by the artist in a realistic, but "non physical" manner

- 121 putative matches total across 563 photographs using SIFT matching
- 0 correct putative matches

Difficulty in finding correspondences

Local feature matching using SIFT:

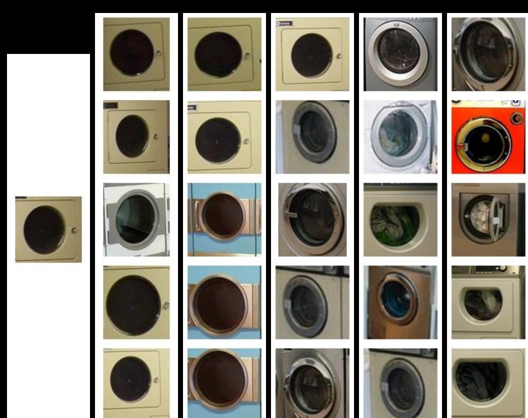
Figure from [A. Shrivastava, T. Malisiewicz, A. Gupta, A. Efros
Data-driven Visual Similarity for Cross-domain Image Matching
SIGGRAPH Asia 2011]

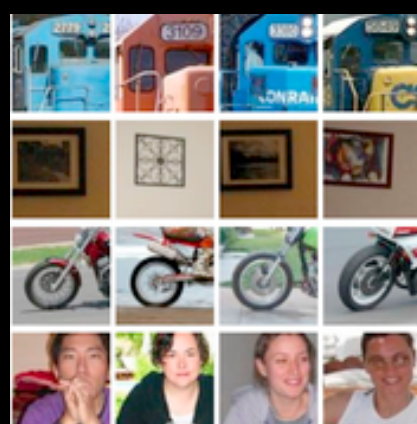
See also:
[Hauagge & Snavely CVPR 2012]
[Chum & Matas CVPR 2006]
[Russell, Sivic, Ponce, Dessalles 2011]

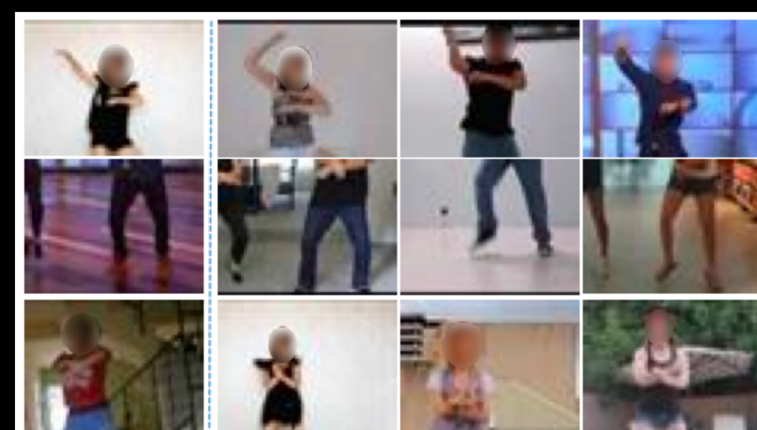
Related work: “mid-level” visual elements

Learn a **vocabulary of discriminative visual elements** that characterize a city.

[Doersch, Singh, Gupta, Sivic, Efros, What makes Paris look like Paris?, SIGGRAPH 2012]

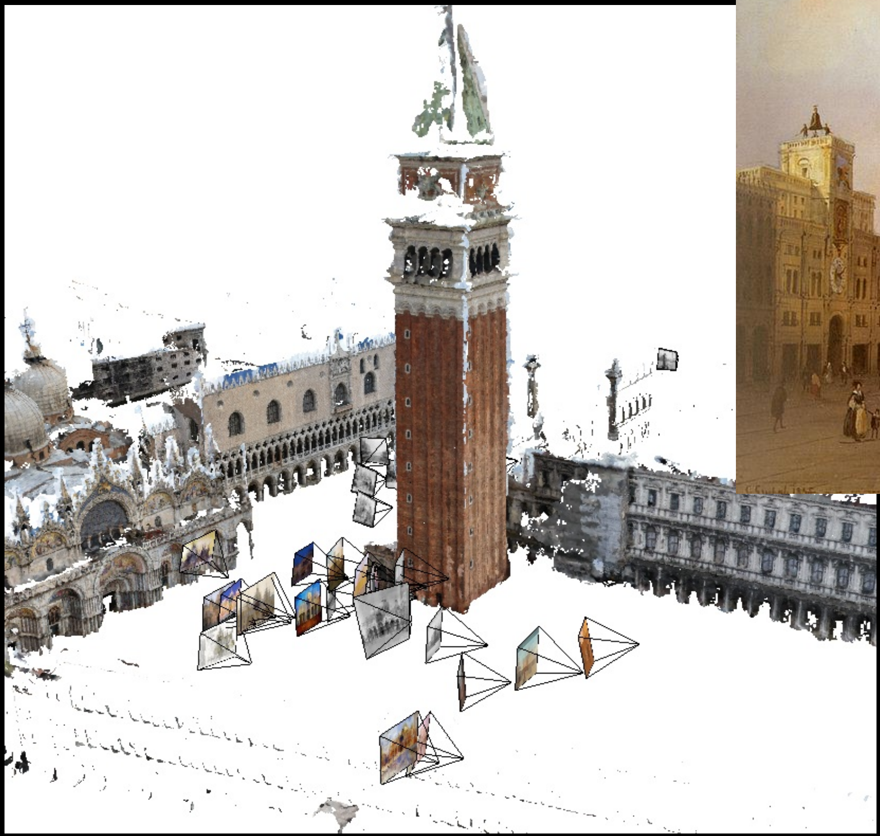






See also [Singh et al. ECCV 2012], [Juneja et al. CVPR 2013], [Jain et al. CVPR 2013], …

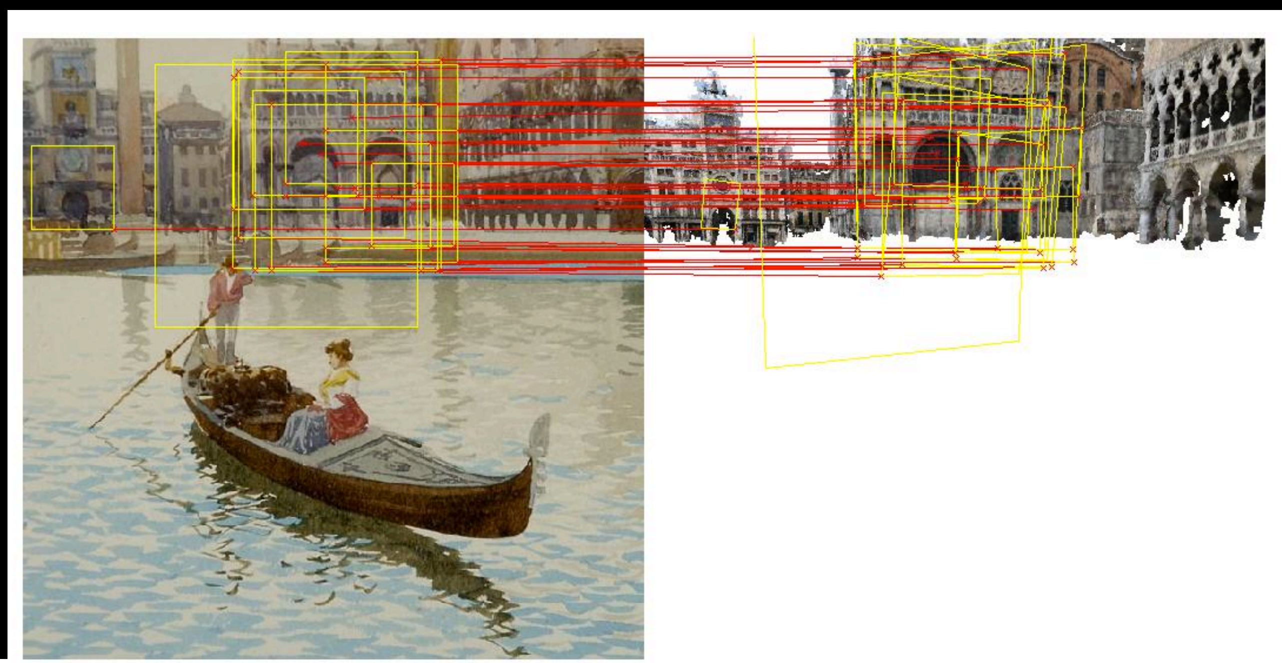
How to match a painting to a 3D model?



High level ideas

- Summarize a 3D model with a set of discriminative elements – "view-dependent distinct 3D fragments"

- Recover the viewpoint of a painting by matching visual elements.



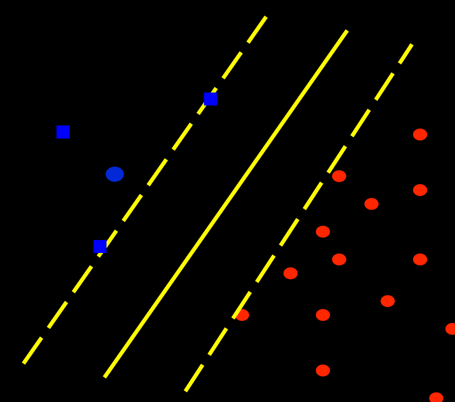
Challenges

- How can we select the set of meaningful visual elements out of all possible ones in the 3D model?

  Select the discriminative and reliable ones.

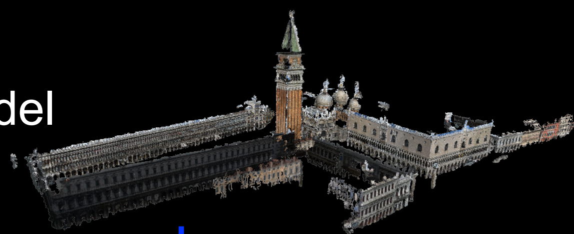
- How to compare a visual element in the 3D model and in the painting?

  Treat as an object detection task.



Algorithm outline

3D model

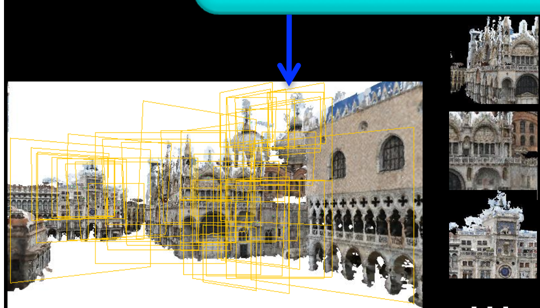


depiction

Rendering representative views

Finding discriminative visual elements

Filtering elements unstable across viewpoint

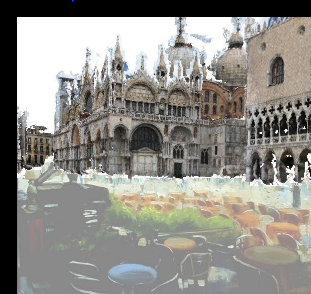


. . .

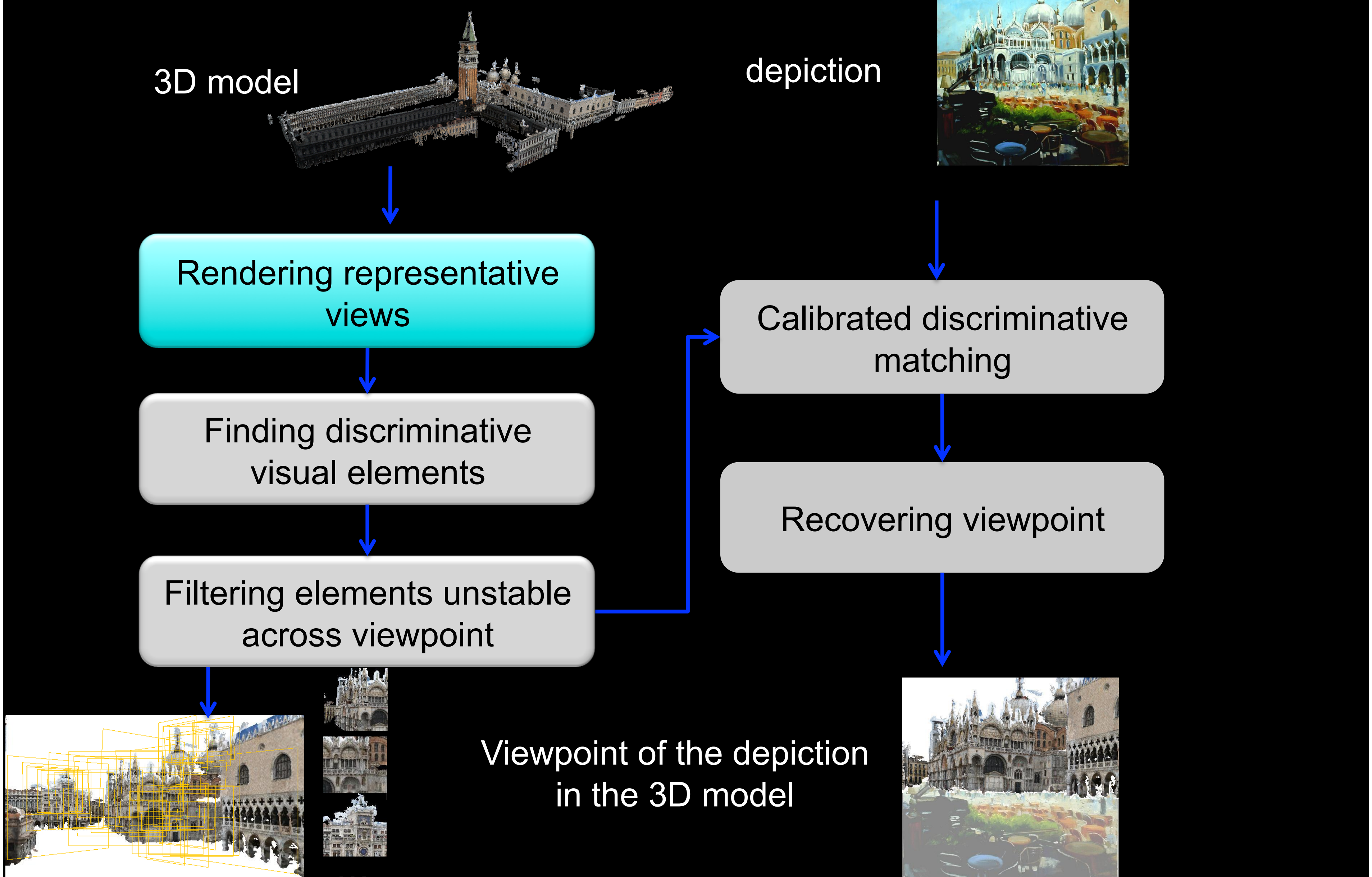
Calibrated discriminative matching

Recovering viewpoint

Viewpoint of the depiction in the 3D model

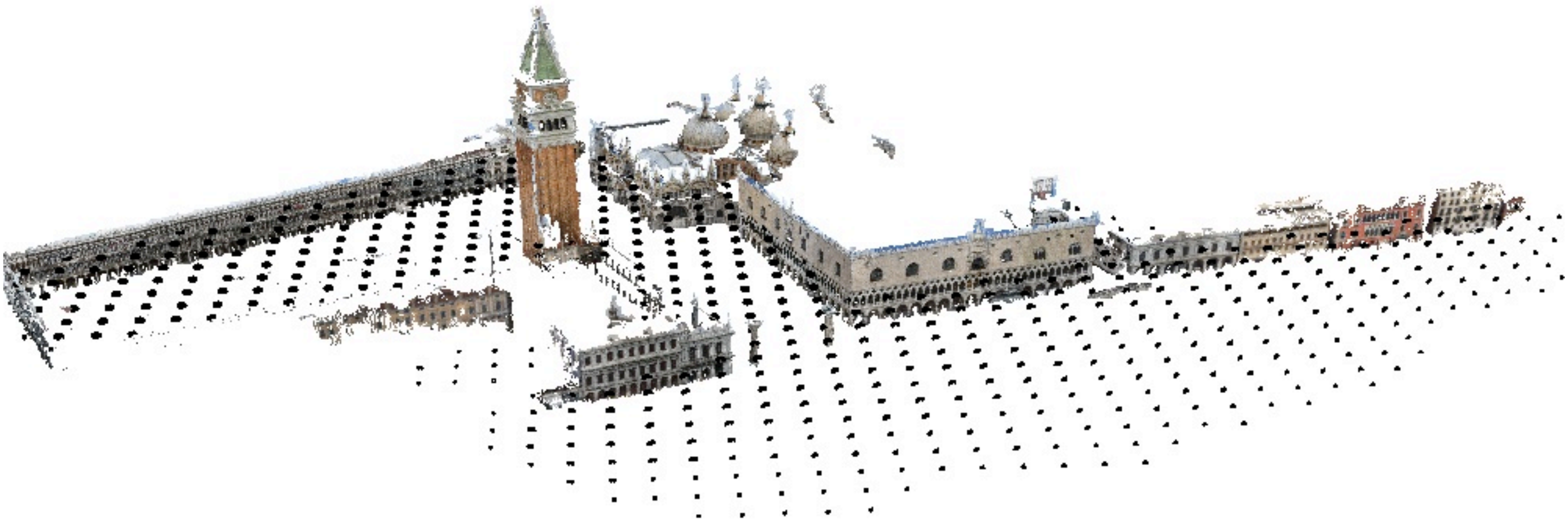


Algorithm outline



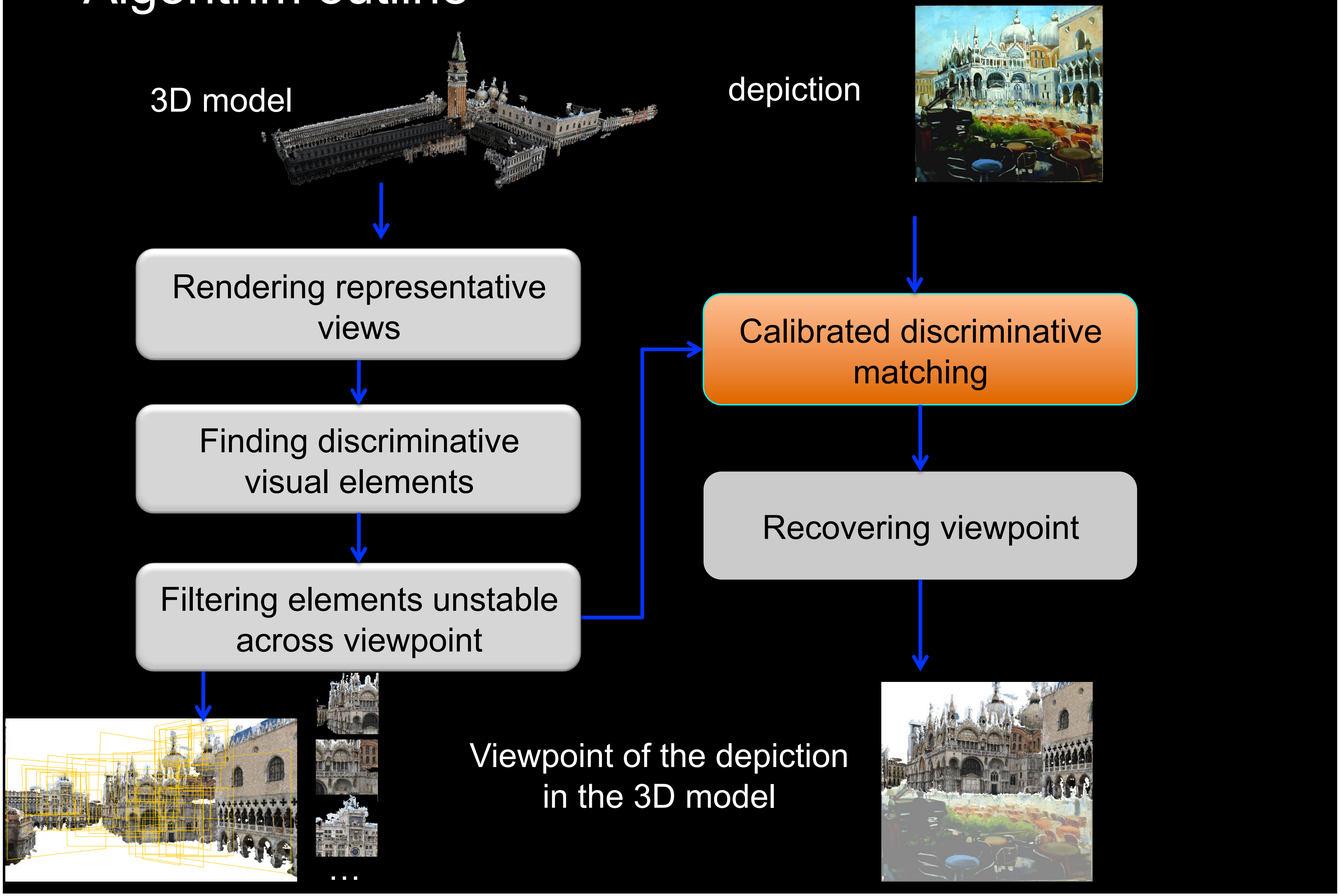
Rendering representative views

Synthesize ~10,000 viewpoints for an architectural site

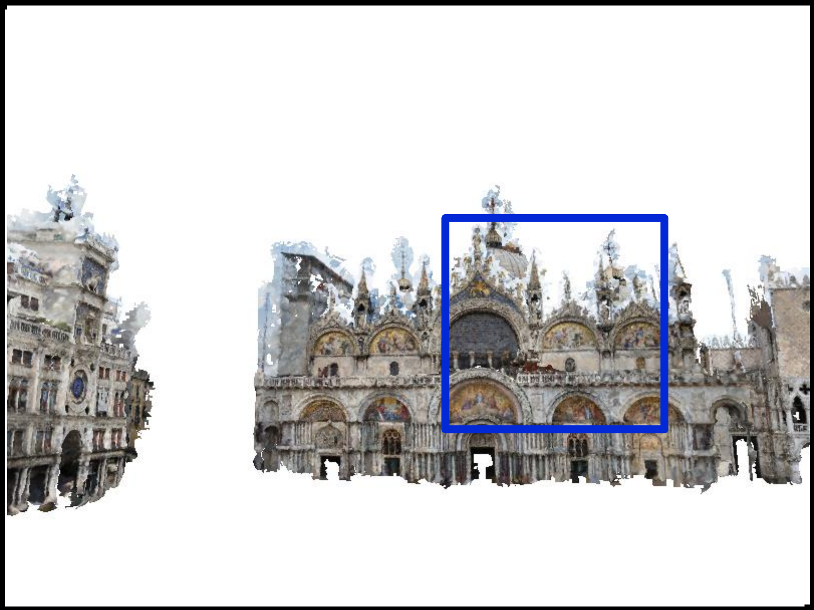


See also: [Irschara et al. CVPR 2009], [Baatz et al. ECCV 2012]

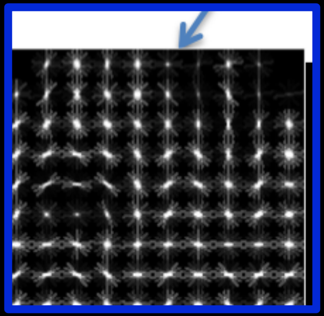
Algorithm outline



Matching as discriminative classification

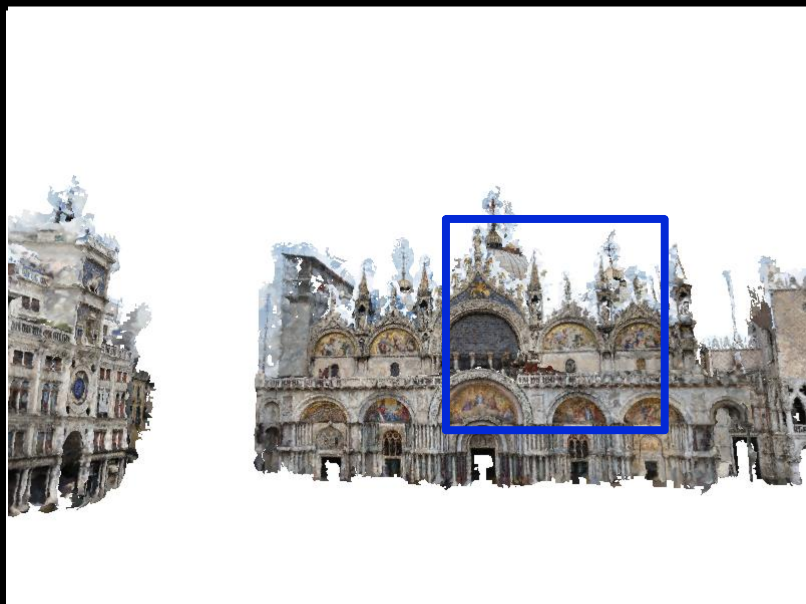


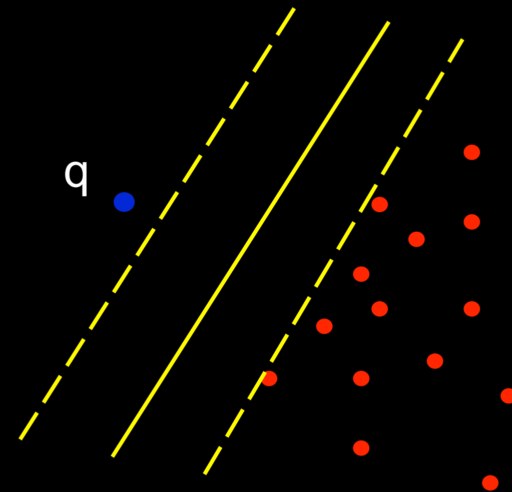
Query region q:



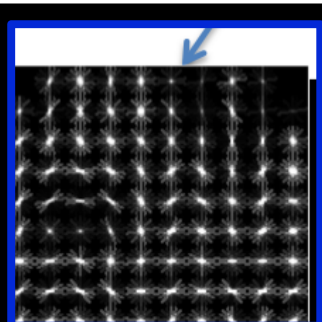
1. Represent query region q using HOG descriptor

Matching as discriminative classification



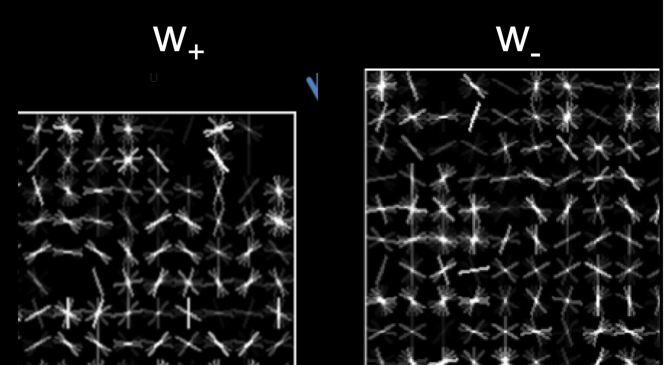


Query region q:



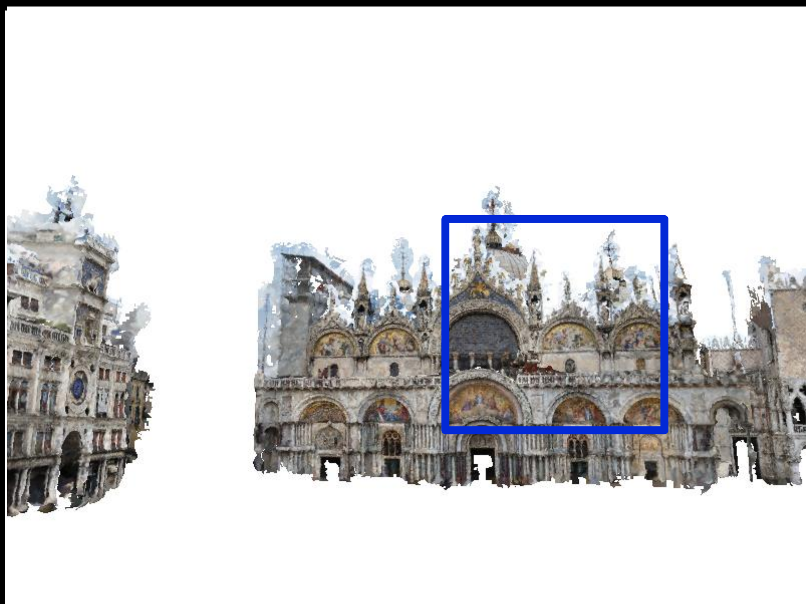
See also exemplar SVM by [Malisiewicz et al., ICCV'11], [Shrivastava et al.'11]

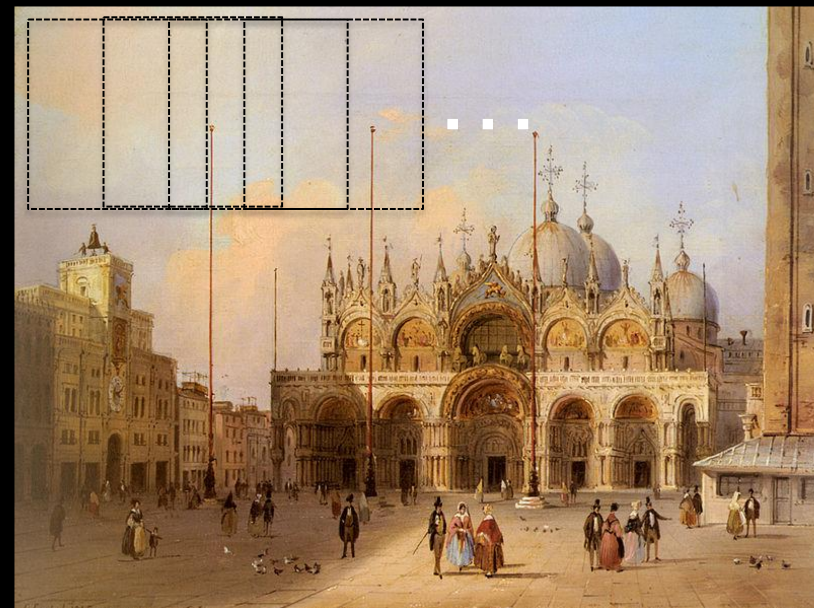
Here used for weighted matching



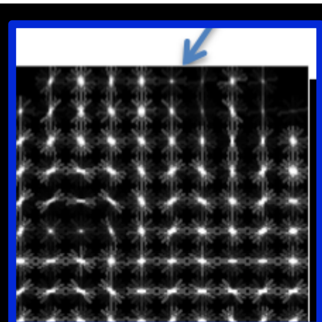
1. Represent query region q using HOG descriptor
2. Train a linear classifier $f(x) = w^T x + b$ using q as a positive example and large number of negatives

Matching as discriminative classification



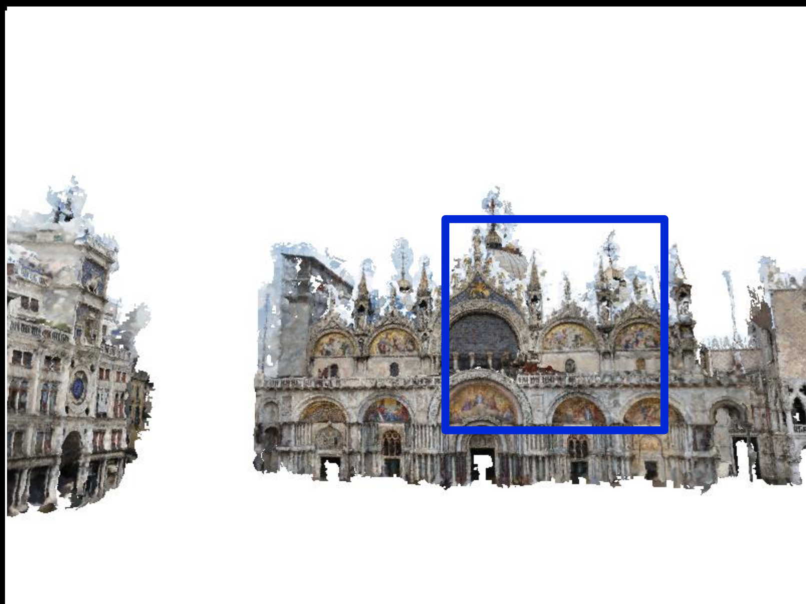


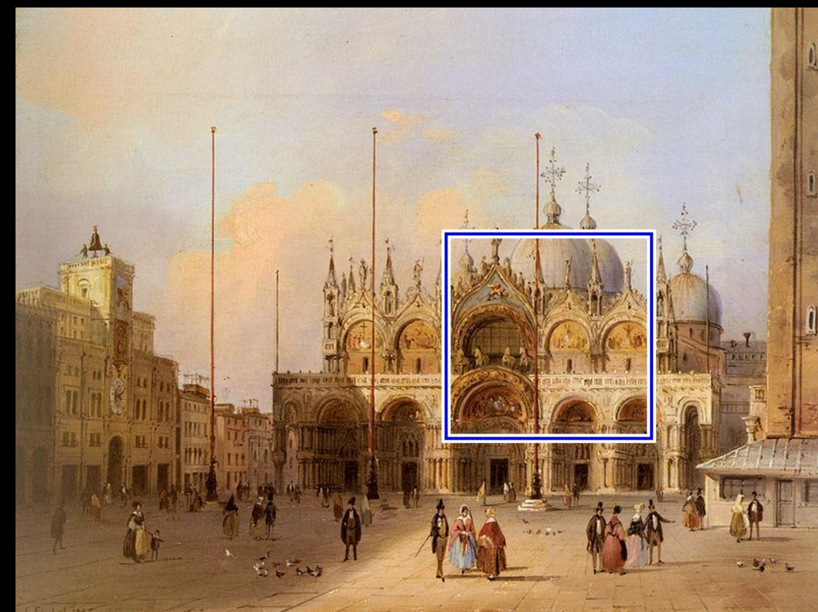
Query region q:



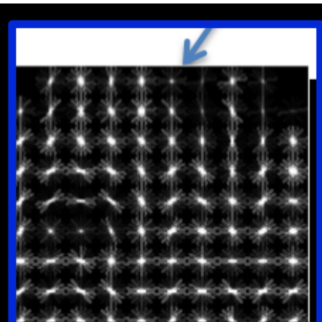
1. Represent image region using HOG descriptor x
2. Train a linear classifier $f(x) = w^T x+b$
3. Find the best match in the painting maximizing the classification score f(x)

Matching as discriminative classification

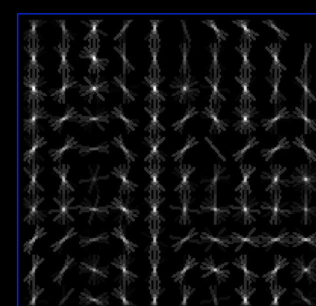




Query region q:

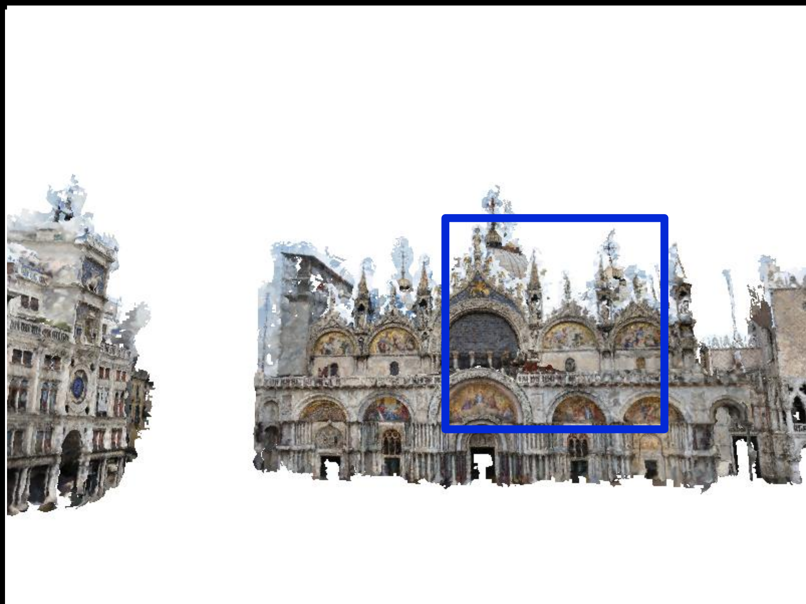


Best match:

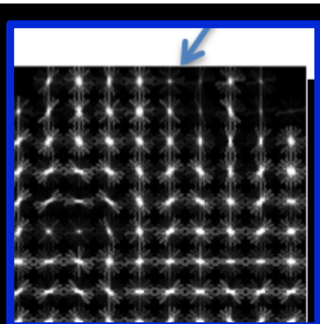


1. Represent image region using HOG descriptor x
2. Train a linear classifier $f(x) = w^T x + b$
3. Find the best match in the painting maximizing the classification score f(x)

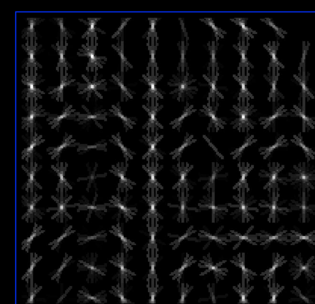
Matching as discriminative classification



Query region q:



Best match:



Discriminative visual element: trained classifier $f(x) = w^T x + b$

How to choose discriminative visual elements representing architectural site?

See also [Doersch et al. SIGGRAPH 2012] [Singh et al. ECCV 2012], [Juneja et al. CVPR 2013]

Algorithm outline

3D model

depiction

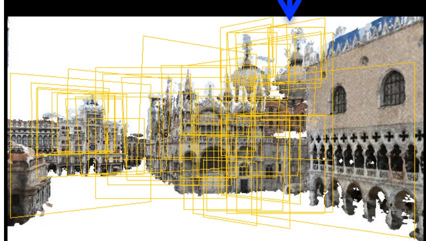
Rendering representative views

Finding discriminative visual elements

Filtering elements unstable across viewpoint

Calibrated discriminative matching

Recovering viewpoint

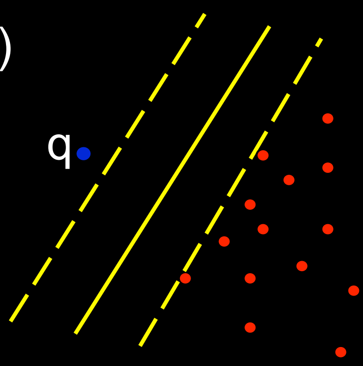


. . .

Viewpoint of the depiction in the 3D model

Finding discriminative visual elements

- Train classifier for each candidate region q:
  - {q,+1}, {x_i,-1} for i = 1..N (set of "generic" negatives)



$$E(w,b) = L\left(1, w^T q + b\right) + \frac{1}{N}\sum_{i=1}^{N} L\left(-1, w^T x_i + b\right)$$

  - Example: hinge loss (e-SVM)

$$L(y, s(x)) = (y - s(x))_+$$

  - Example: square loss

$$L(y, s(x)) = (y - s(x))^2$$

Finding discriminative visual elements

For square loss E can be minimized in closed form
[Bach&Harchaoui 2008; Gharbi et al. 2012; Hariharan et al. 2012]

$$w_{LS} = \frac{2}{2 + \|\Phi(q)\|^2} \Sigma^{-1}(q - \mu),$$

$$b_{LS} = -\frac{1}{2}(q + \mu)^T w_{LS},$$

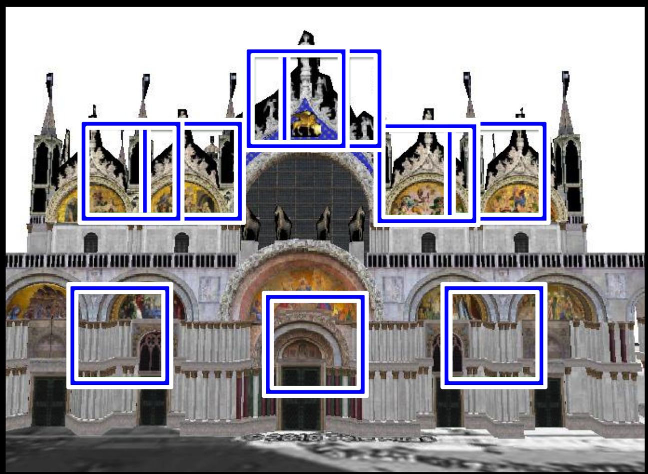
$$E^*_{LS} = \frac{4}{2 + \|\Phi(q)\|^2},$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ denotes the mean of the negative examples, $\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^\top$ their covariance and

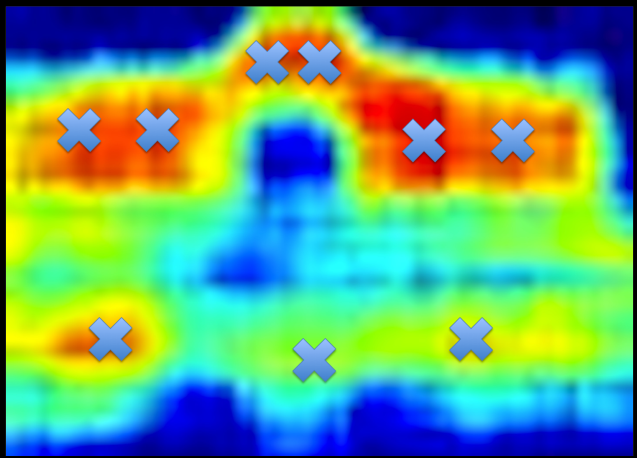
$$\Phi(q) = \Sigma^{-\frac{1}{2}}(q - \mu).$$

Finding discriminative visual elements

- Train classifiers for all candidate regions in synthesized views
  - Can be done in closed form [Gharbi et al. 2012; Hariharan et al. 2012]
- Score each classifier by its training cost E.
- Keep only the top N most discriminative visual elements.



Original image

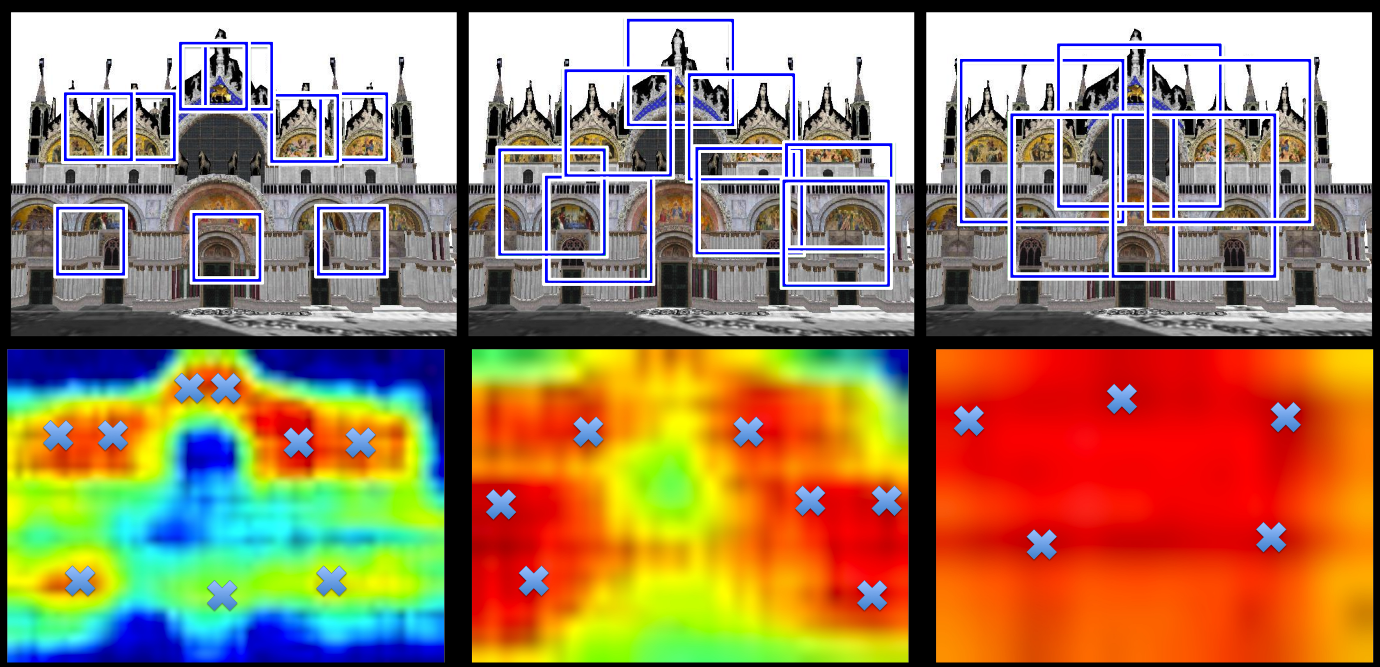


Discriminative score: 1 / Energy

Note: Can be thought of as a generalization of local feature detection.

Finding discriminative visual elements

- Train classifiers for all candidate regions in synthesized views
  - Can be done in closed form [Gharbi et al. 2012; Hariharan et al. 2012]
- Score each classifier by its training cost E.
- Keep only the top N most discriminative visual elements.

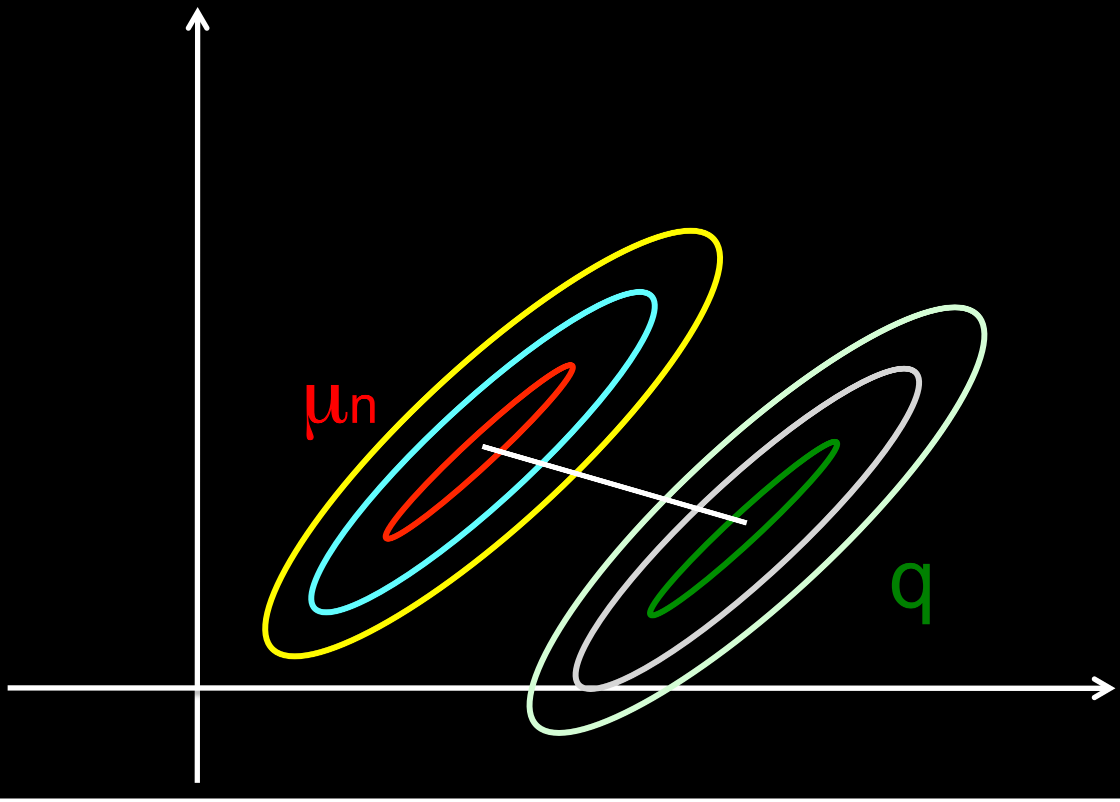


Note: Can be thought of as a generalization of local feature detection.

Related: Linear Discriminant Analysis

Let's

- consider a simple probabilistic model
- assume that the positive and negative class have Gaussians distribution
- assume that they same variance.



Related: Linear Discriminant Analysis

A log likelihood ratio test with this probabilistic mode leads to a classifier

$$s_{LDA}(x) = w_{LDA}^T x + b_{LDA}$$

With

$$w_{LDA} = \Sigma^{-1}(q - \mu_n)$$

$$b_{LDA} = \frac{1}{2}\left(\mu^T \Sigma^{-1} \mu - q^T \Sigma^{-1} q\right)$$

Note: $w_{LDA} = \left(1 + \frac{1}{2}\|\Phi(q)\|^2\right) w_{LS}$ $\rightarrow$ $s_{LDA} = \alpha s_{LS} + \beta$

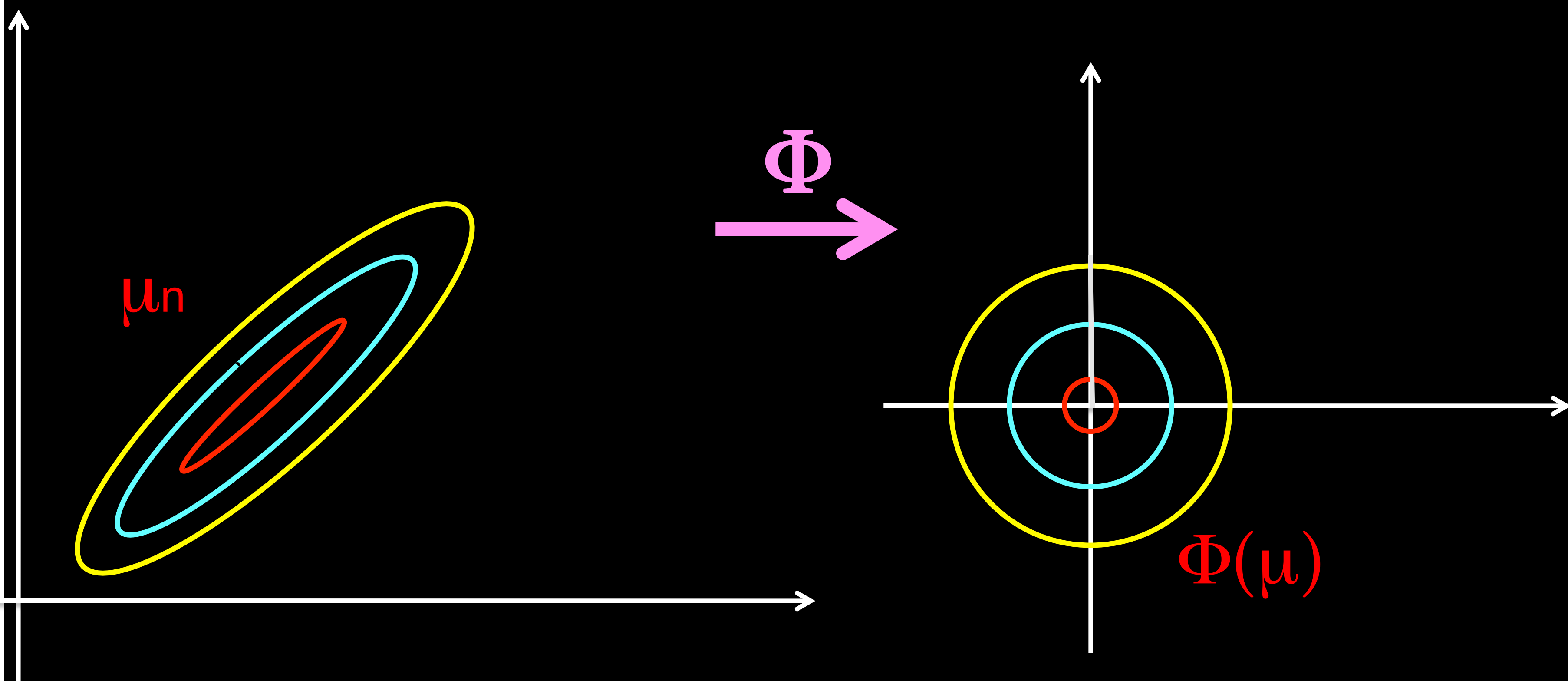
$$\alpha = 1 + \frac{1}{2}\|\Phi(q)\|^2$$

$$\beta = b_{LDA} - \alpha b_{LS}$$

Duda, Hart, Stork, 2001
Hariharan, Malik, Ramanan 2012
Gharbi, T. Malisiewicz, S. Paris, F. Durand, 2012

“Whitening interpretation”

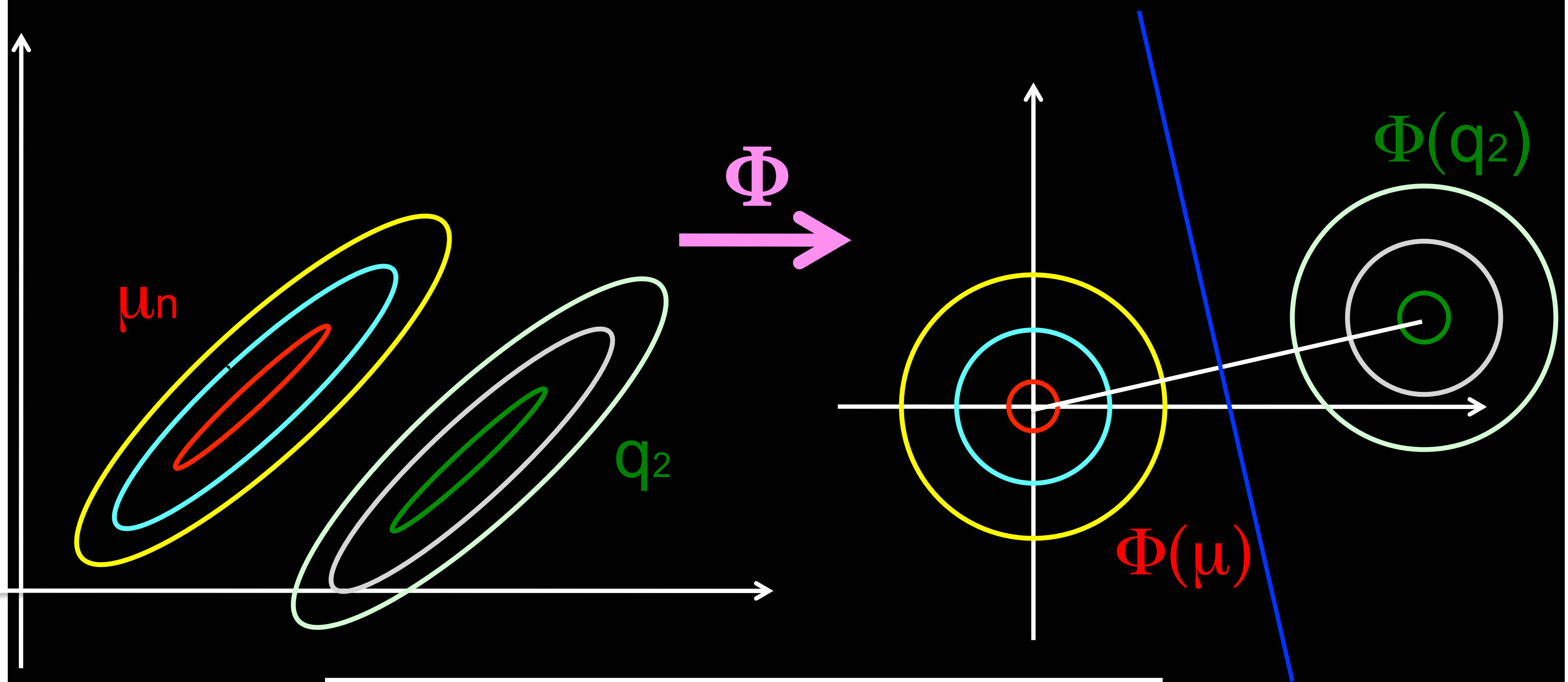
Our detection and matching can be interpreted in the ‘whitened space’:



$$\Phi(x) = \Sigma^{-\frac{1}{2}}(x - \mu)$$

“Whitening interpretation”

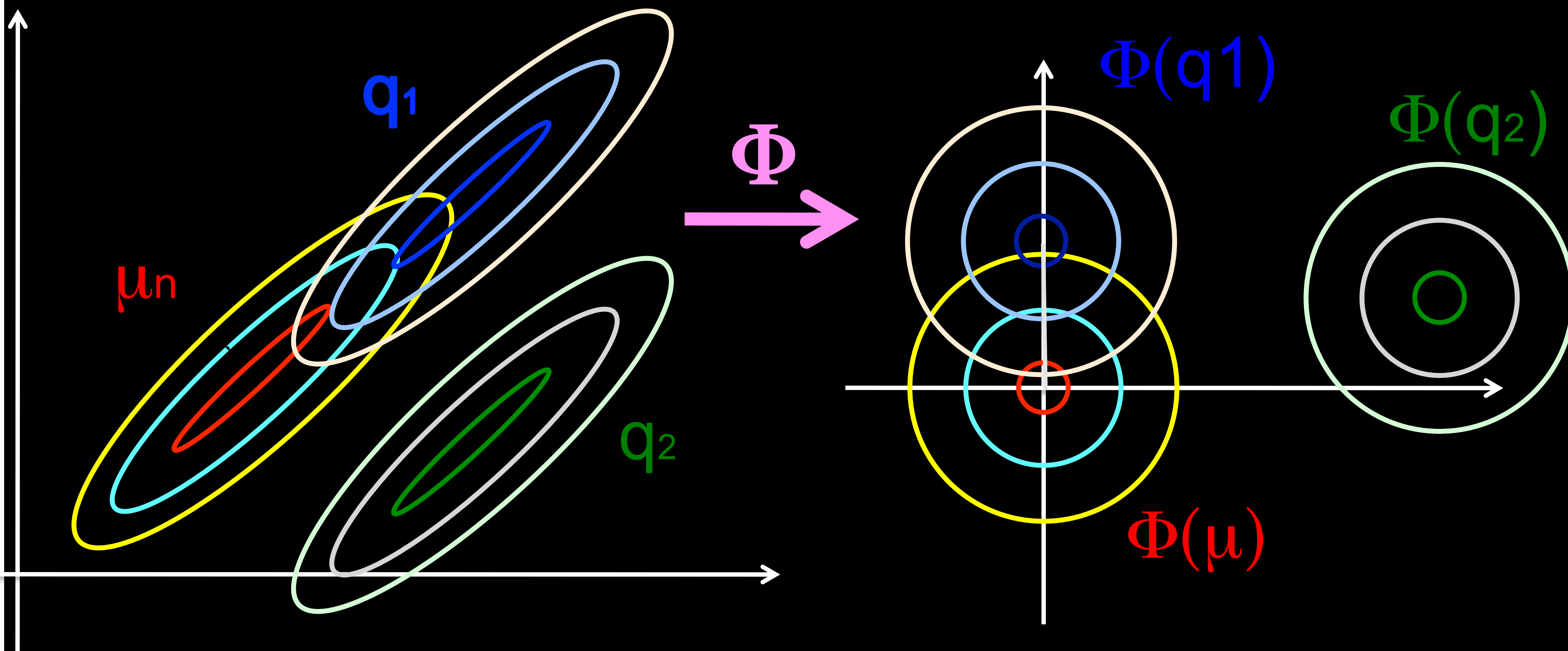
Matching :



$$s_{LDA}(x) \;=\; \Phi(q)^T\Phi(x) - \frac{1}{2}\|\Phi(q)\|^2$$

“Whitening interpretation”

Detection:



Big || Φ(q) || = discriminative

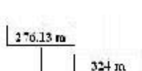
Calibrated discriminative matching

The LDA score improves over the LS score, but overrates low-contrast matches. Thus we add a constant such that the score of a zero HOG is 0.

$$\begin{aligned} s_{calib}(x) &= s_{LDA}(x) - s_{LDA}(0) \\ &= (q - \mu)^T \Sigma^{-1} x. \end{aligned}$$

Results:

Matching method	mAP (“desceval”)
Local symmetry [Hauagge and Snavely 2012]	0.58
Least squares regression (Sec. 4.2.2)	0.52
LDA (Sec. 4.2.3)	0.60
Ours (Sec. 4.2.5)	**0.77**



Algorithm outline

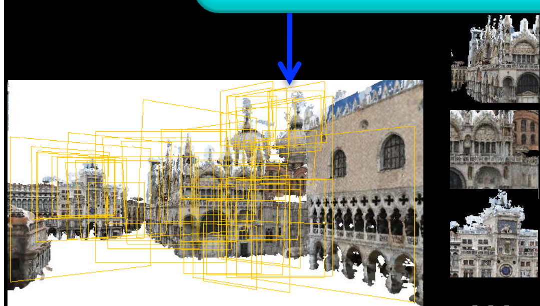
3D model

depiction

Rendering representative views

Finding discriminative visual elements

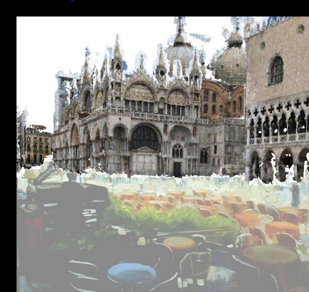
Filtering elements unstable across viewpoint



Calibrated discriminative matching

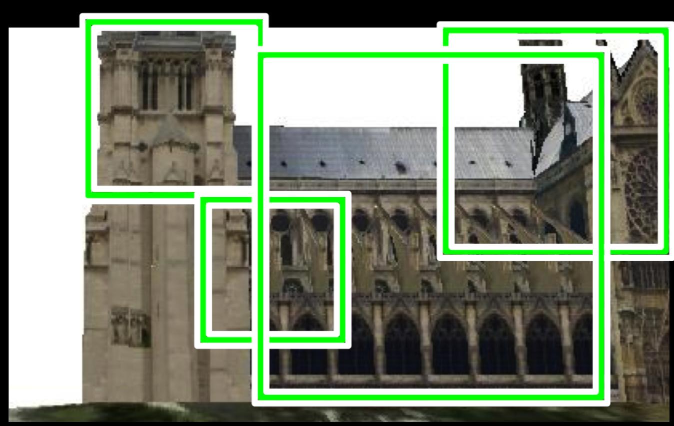
Recovering viewpoint

Viewpoint of the depiction in the 3D model

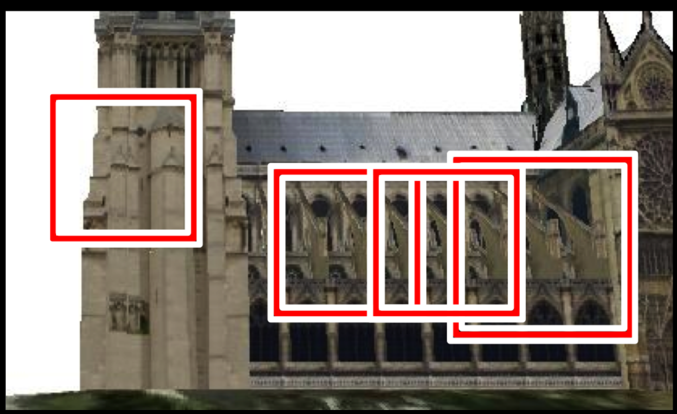


Filtering elements unstable across viewpoint

- Filter out elements unstable across viewpoint.
- 3D model provides ground truth matches in near-by views
- Require elements to be reliably detectable in near-by views



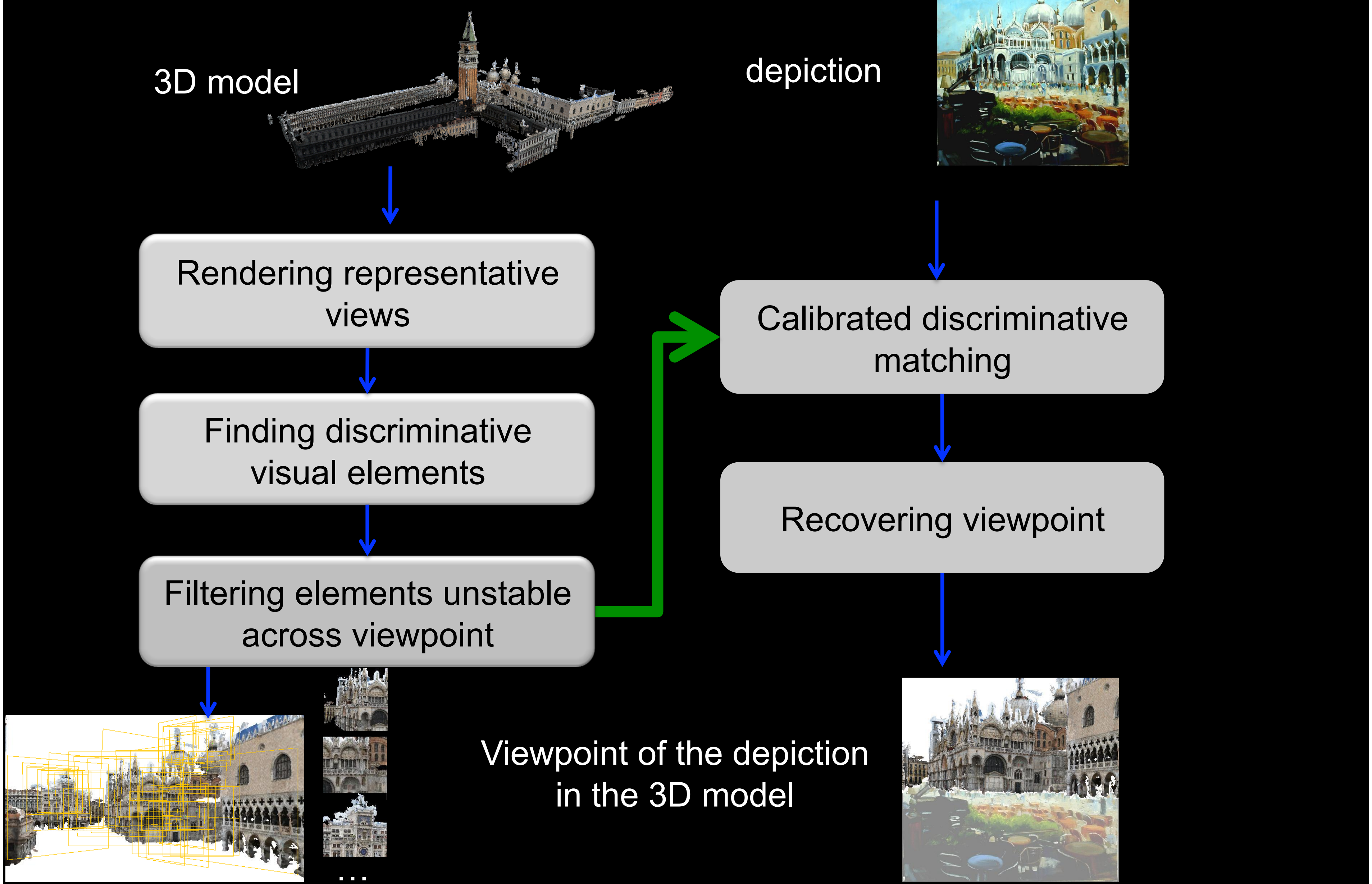
Top stable elements



Top unstable elements

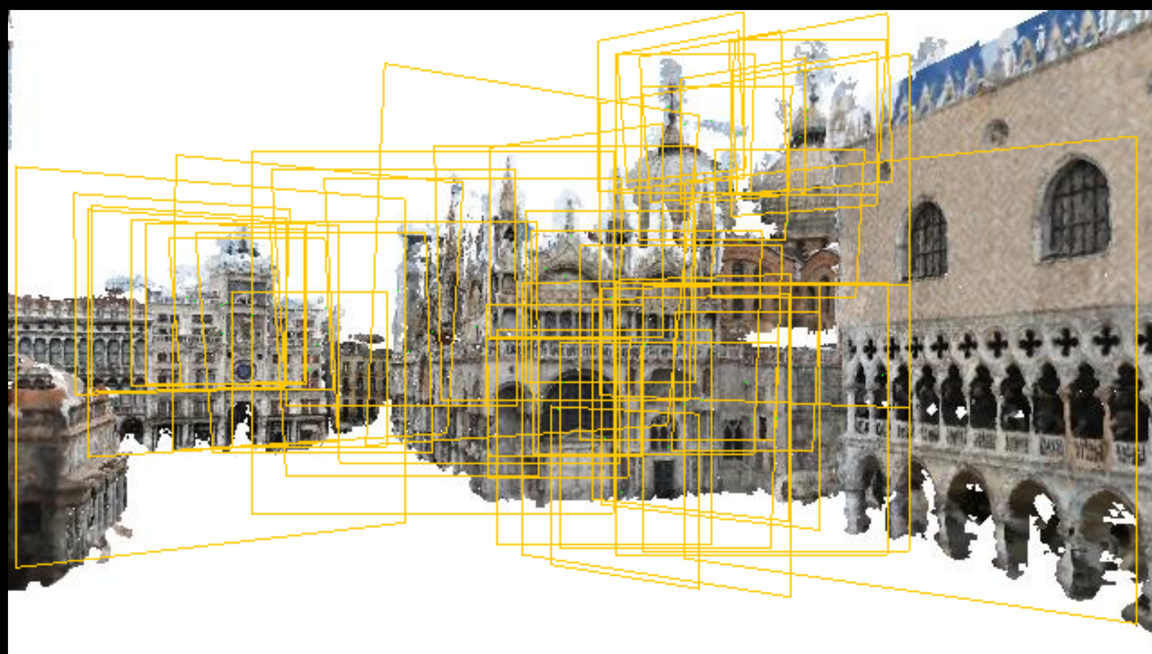
See also [Doersch et al. SIGGRAPH 2012] [Singh et al. ECCV 2012], [Juneja et al. CVPR 2013]

Algorithm outline



Summary : discriminative visual element

- Back-project learnt discriminative elements onto the 3D model



See also [Doersch et al. SIGGRAPH 2012] [Singh et al. ECCV 2012], [Juneja et al. CVPR 2013]

Algorithm outline

3D model

depiction

Rendering representative views

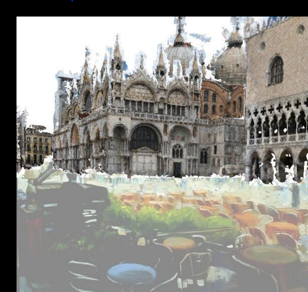
Finding discriminative visual elements

Filtering elements unstable across viewpoint

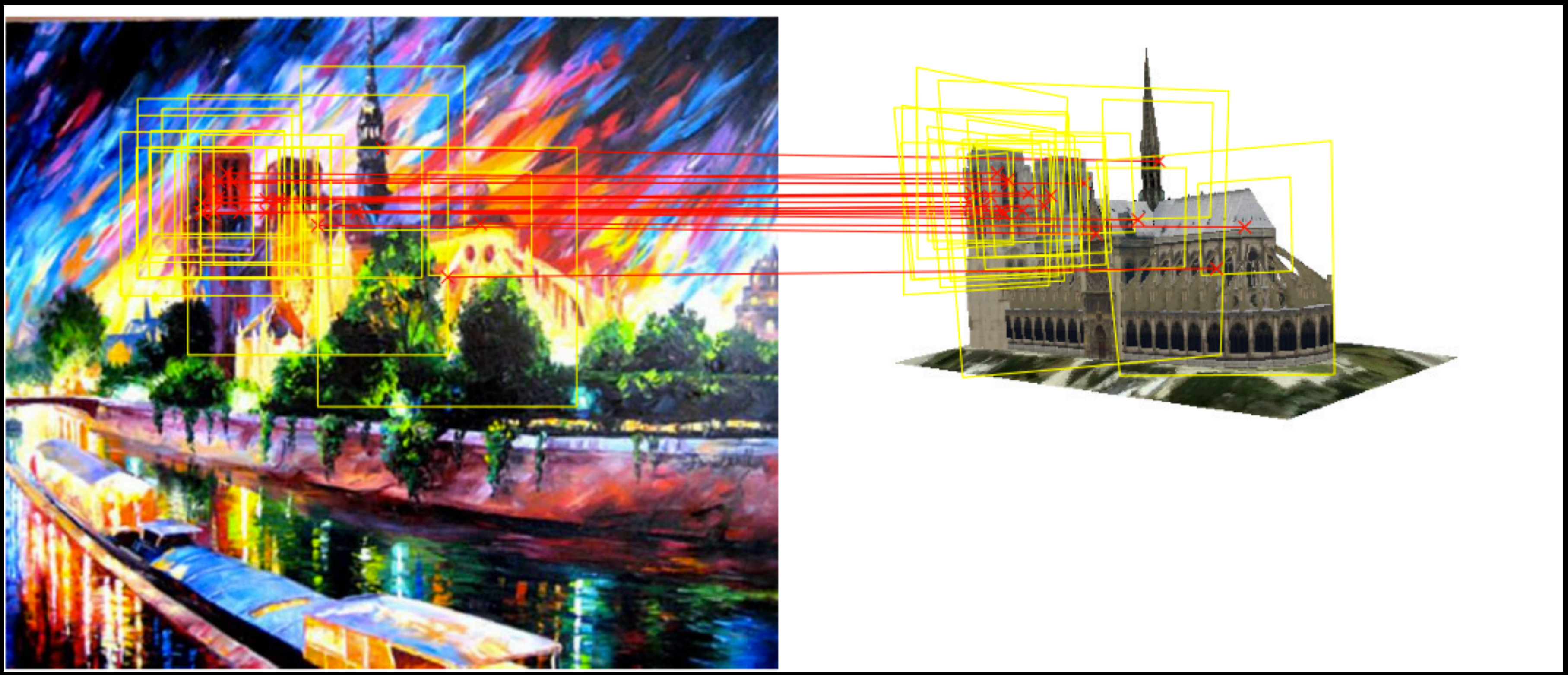
Calibrated discriminative matching

Recovering viewpoint

Viewpoint of the depiction in the 3D model



Recovering viewpoint: RANSAC



Algorithm summary

3D model

depiction

Rendering representative views

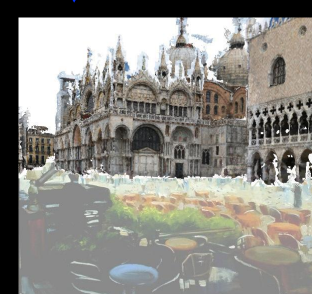
Finding discriminative visual elements

Filtering elements unstable across viewpoint

Calibrated discriminative matching

Recovering viewpoint

Viewpoint of the depiction in the 3D model



Experiments

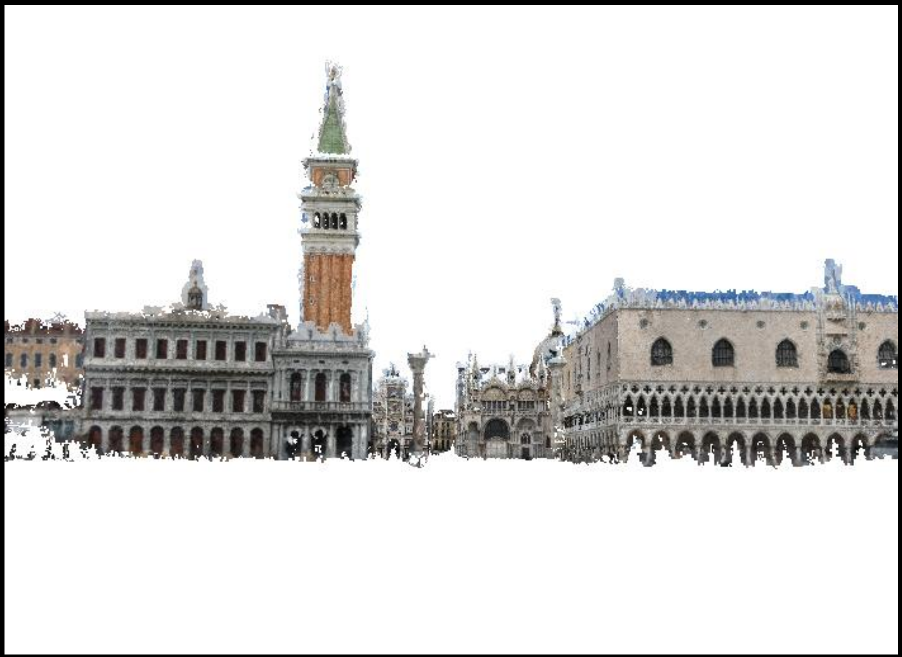
3D architectural sites

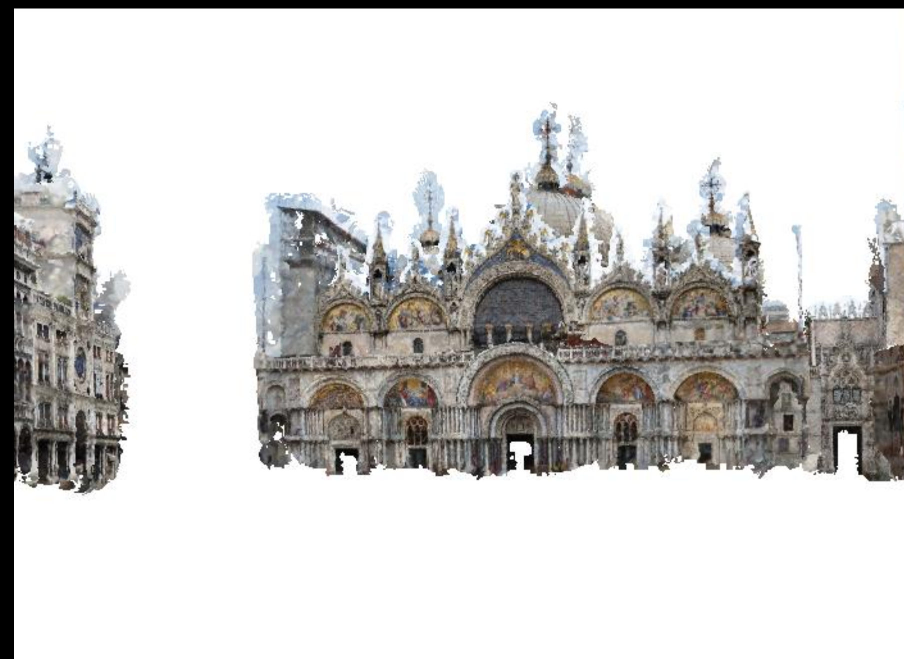
- Venice (PMVS reconstruction from “Rome in a day” photographs)
- Venice (3D CAD model)
- Trevi Fountain (3D CAD model)
- Notre Dame of Paris (3D CAD model)

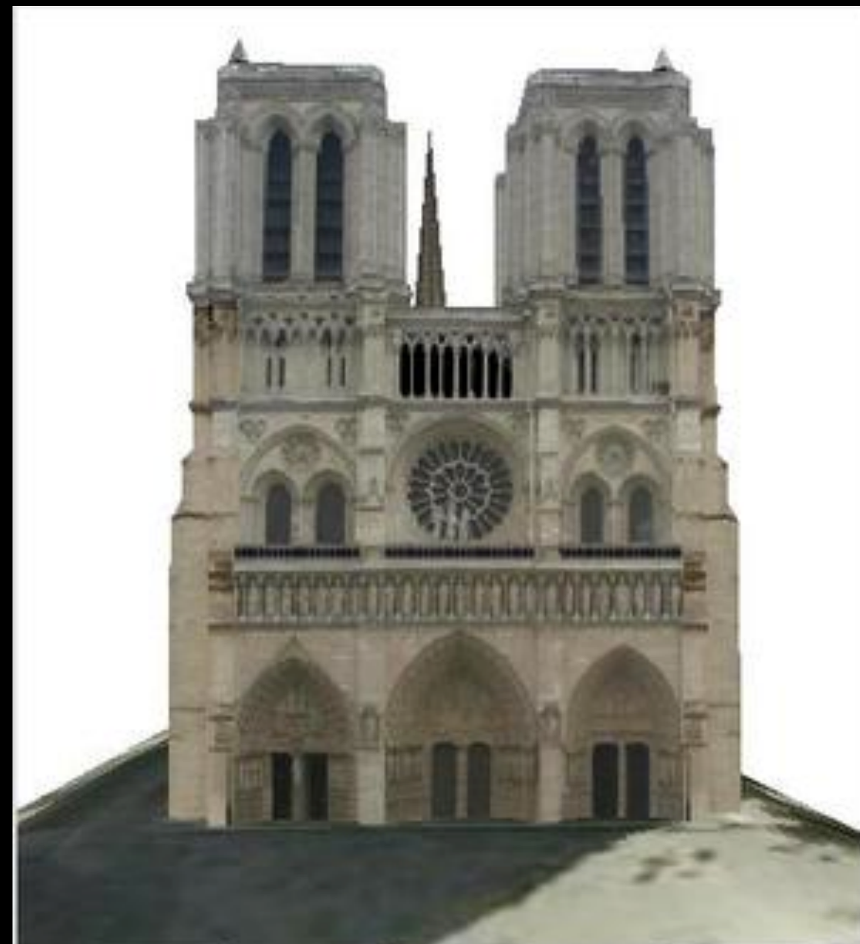
337 “Test queries”

- 85 historical photographs
- 147 paintings
- 60 drawings
- 45 engravings

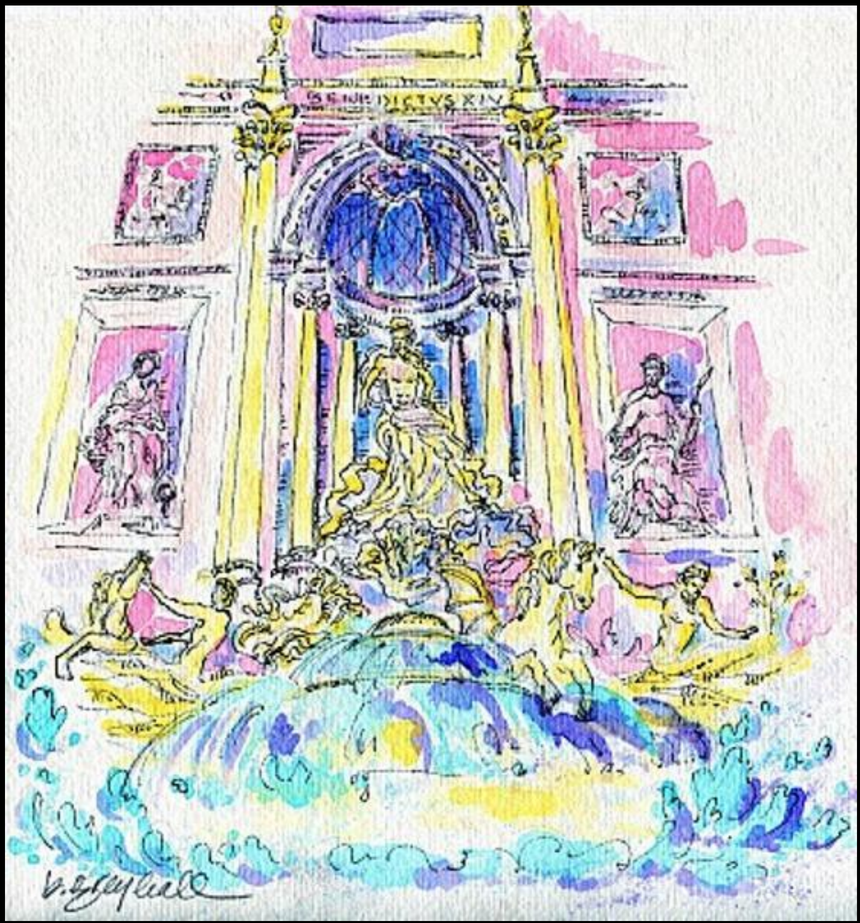
Results: historical photographs

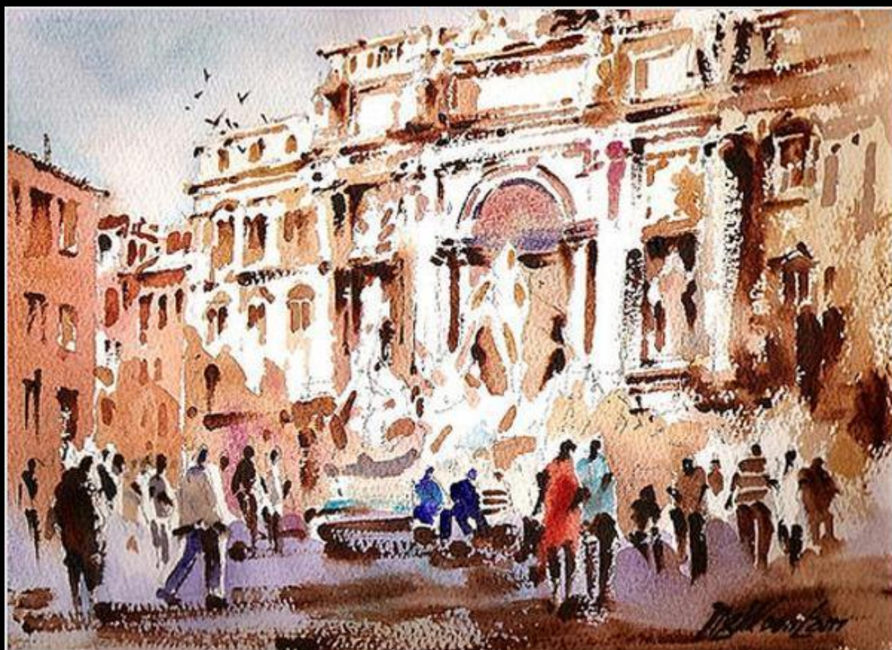


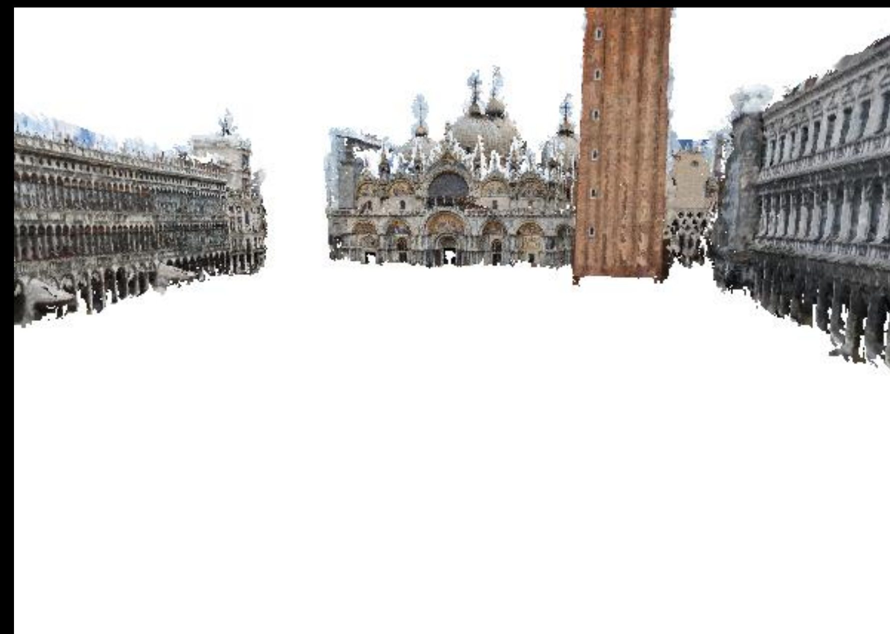


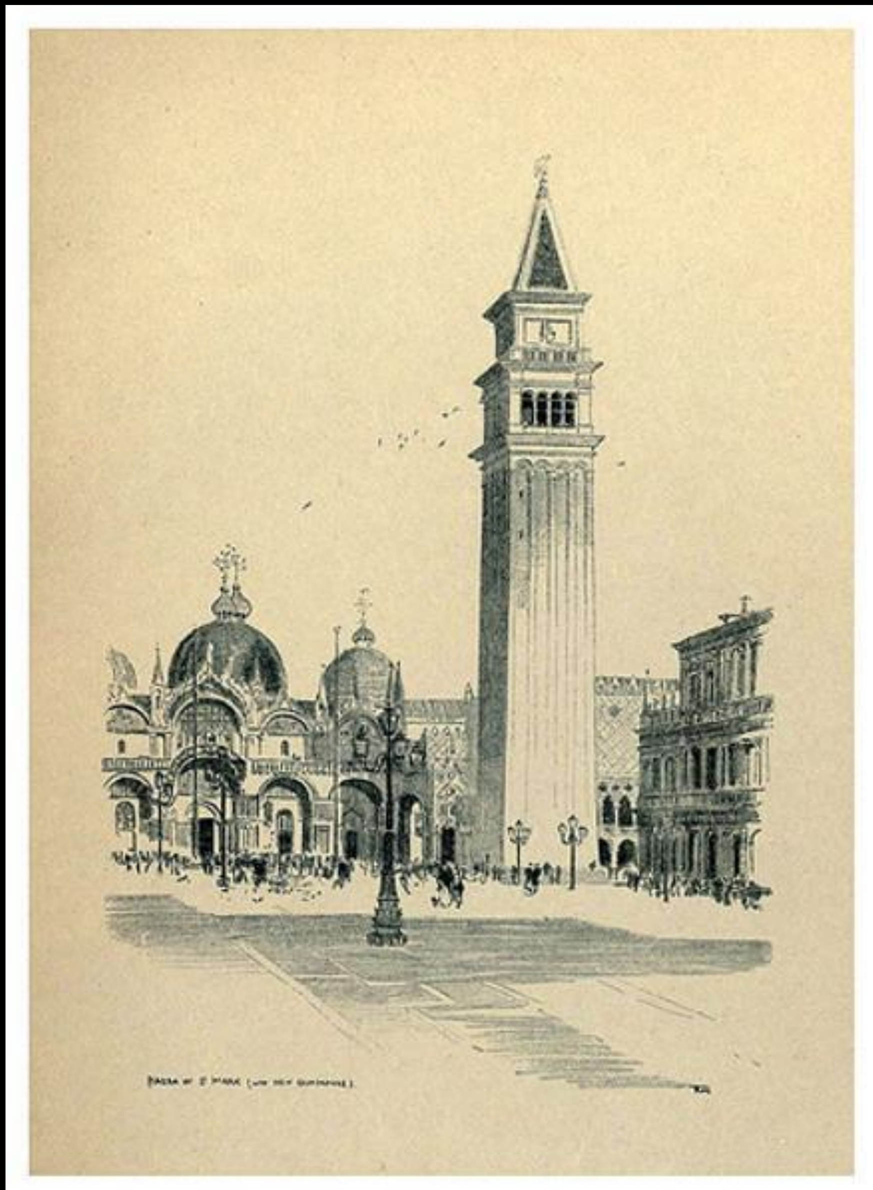
Results: paintings and drawings



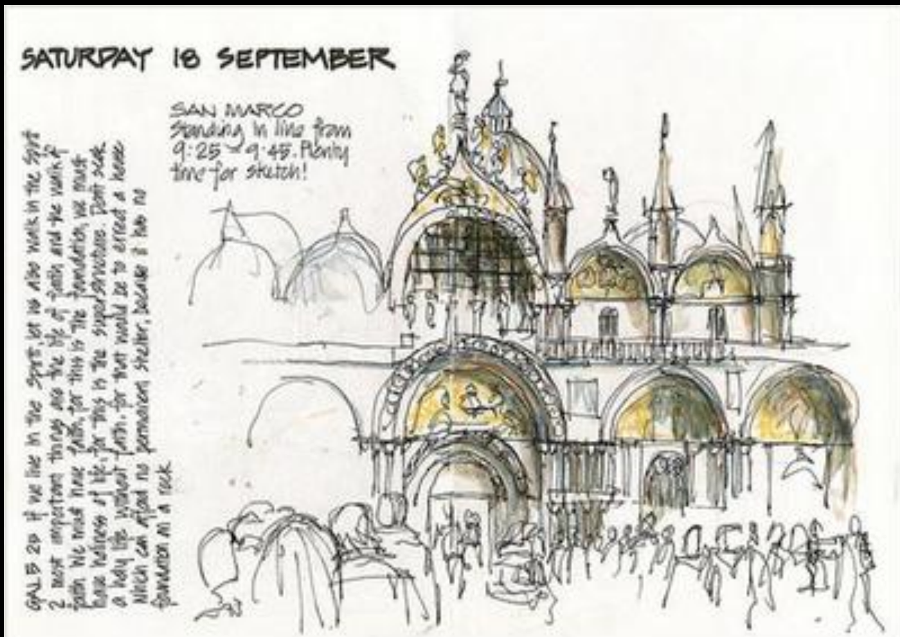



5/10
Piazza San Marco, Venice

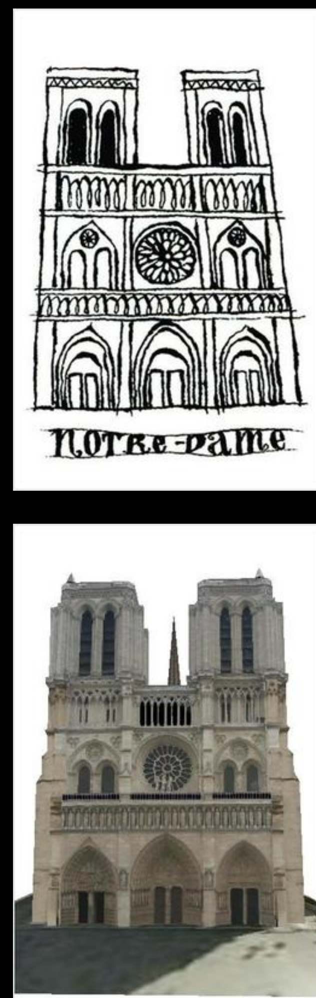




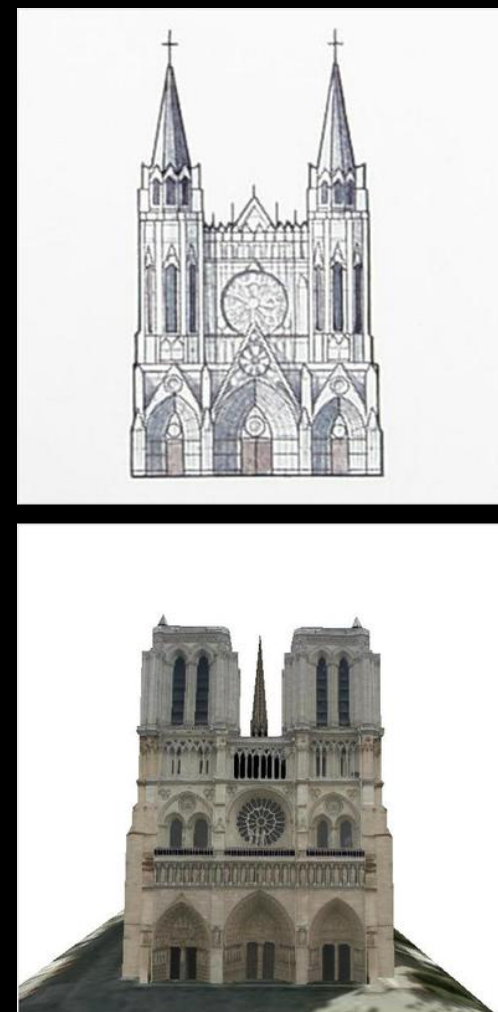

1st-art-gallery.com


SATURDAY 18 SEPTEMBER
SAN MARCO
Standing in line from
9:25 – 9:45. Plenty
time for sketch!

Scene distortion



Drawing errors



Different scene

Failures

Extreme change in depiction styles (smeared watercolor)

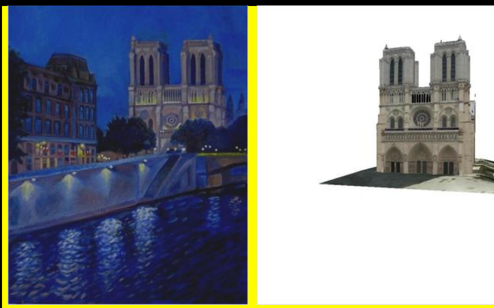
Part of the architectural site not covered by 3D model

Extreme geometric distortion

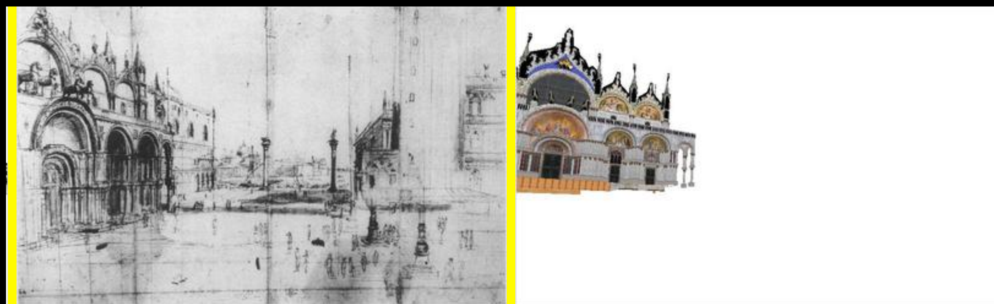
Viewing

Quantitative evaluation

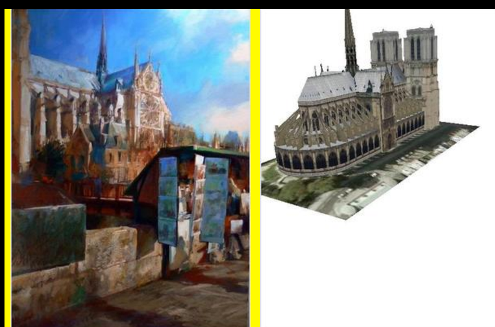
Quantitative evaluation - user study



(a) Good match



(b) Coarse match

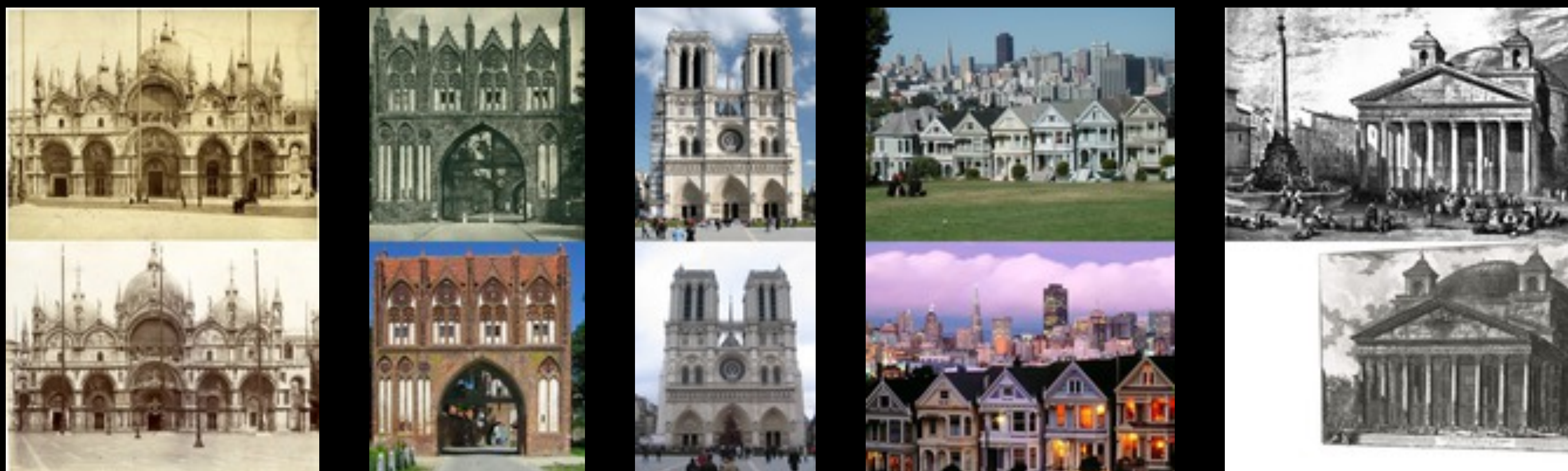


(c) No match

	Good match	Coarse match	No match
SIFT on rendered views	40%	26%	33%
Viewpoint retrieval [Russell et al. 2011]	1%	39%	60%
Exemplar SVM [Shrivastava et al. 2011]	34%	18%	48%
mid-level painting visual elements	33%	29%	38%
3D discrim. visual elements (ours)	**51%**	**21%**	**28%**

NB: the performance of SIFT baseline drops if we don't consider photographs, when our algorithm results remain the same.

Comparison on benchmark dataset of [Hauggage and Snavely, 2012]



Matching method	mAP ("desceval")
Local symmetry [Hauagge and Snavely 2012]	0.58
Least squares regression (Sec. 4.2.2)	0.52
LDA (Sec. 4.2.3)	0.60
Ours (Sec. 4.2.5)	**0.77**

Fly-through video

Conclusions and open questions

- Automatic painting/image-to-3D model alignment is possible for a range of depiction styles
- We represent a 3D model by a compact set of visually distinct mid-level scene elements extracted from rendered views
- How to efficiently index paintings and historical photographs for visual search?
- How to model and cope with drawing error?