Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time

Paper Presentation By Bhavin Modi

Slides By- Yong Jae Lee, Alexei A. Efros, and Martial Hebert
Carnegie Mellon University / UC Berkeley
ICCV 2013

Long before the age of “data mining” …

when?
(historical dating)

where?
(botany, geography)

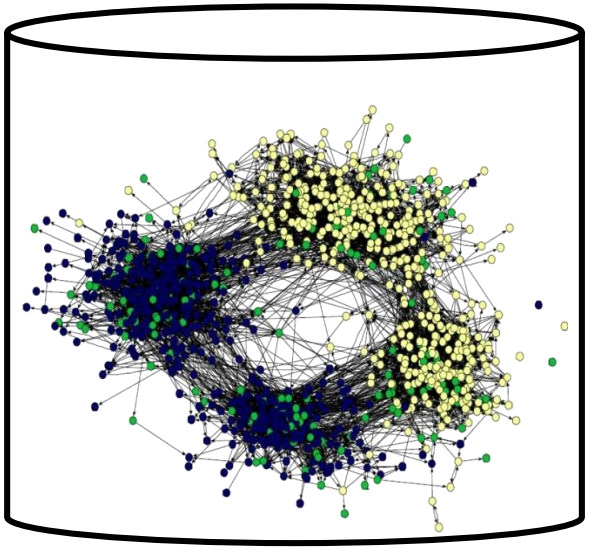
when?

1972

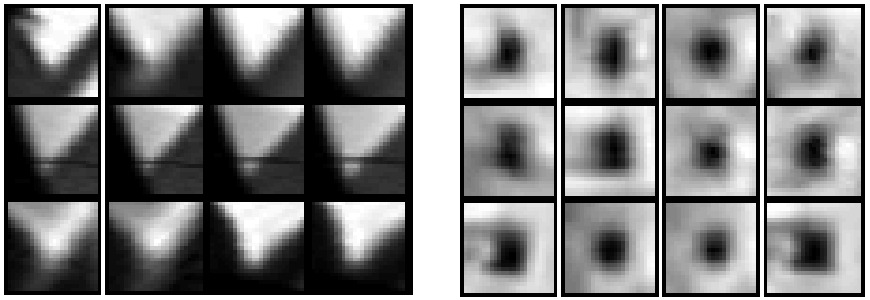
THE DISH
BIASED & BALANCED

"The View From Your Window" challenge

Visual data mining in Computer Vision



Visual world



Low-level "visual words"

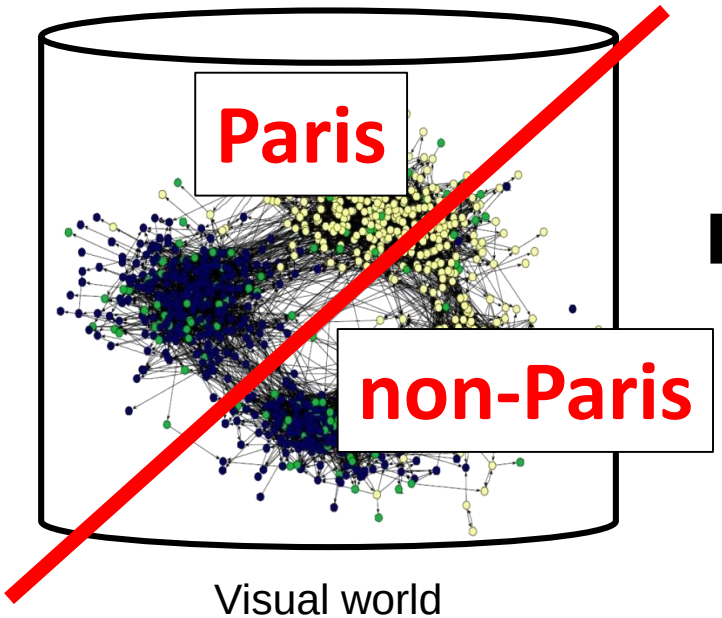
[Sivic & Zisserman 2003, Laptev & Lindeberg 2003, Czurka et al. 2004, …]

Object category discovery

[Sivic et al. 2005, Grauman & Darrell 2006, Russell et al. 2006, Lee & Grauman 2010, Payet & Todorovic, 2010, Faktor & Irani 2012, Kang et al. 2012, …]

- Most approaches mine *globally consistent* patterns

Visual data mining in Computer Vision



Visual world

Mid-level visual elements

[Doersch et al. 2012, Endres et al. 2013, Juneja et al. 2013, Fouhey et al. 2013, Doersch et al. 2013]

- Recent methods discover *specific* visual patterns

Problem

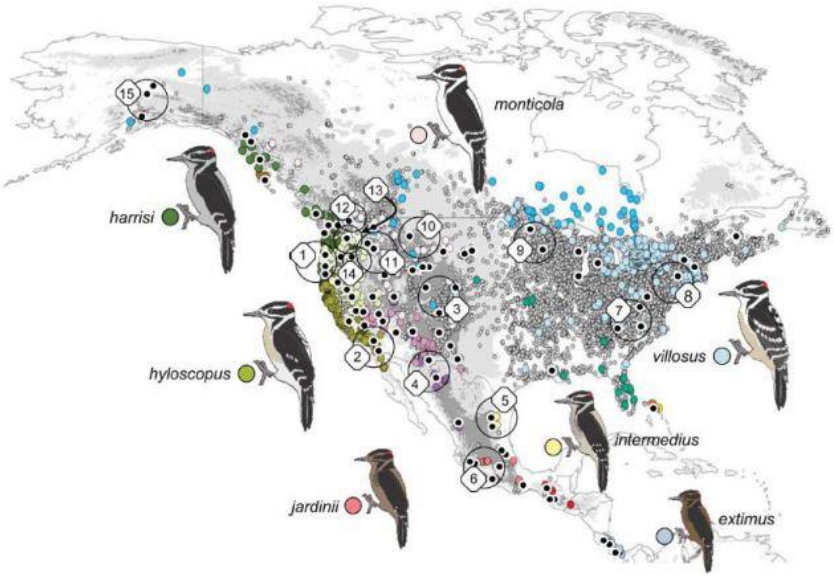
- Much in our visual world undergoes a *gradual change*

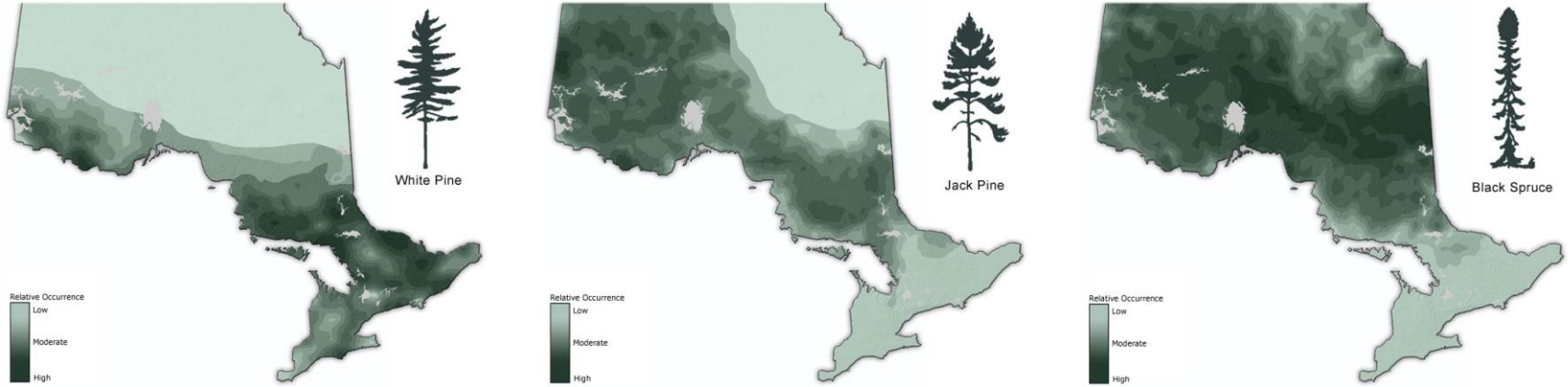
Temporal:

1887-1900 | 1900-1941 | 1941-1969 | 1958-1969 | 1969-1987

- Much in our visual world undergoes a *gradual change*

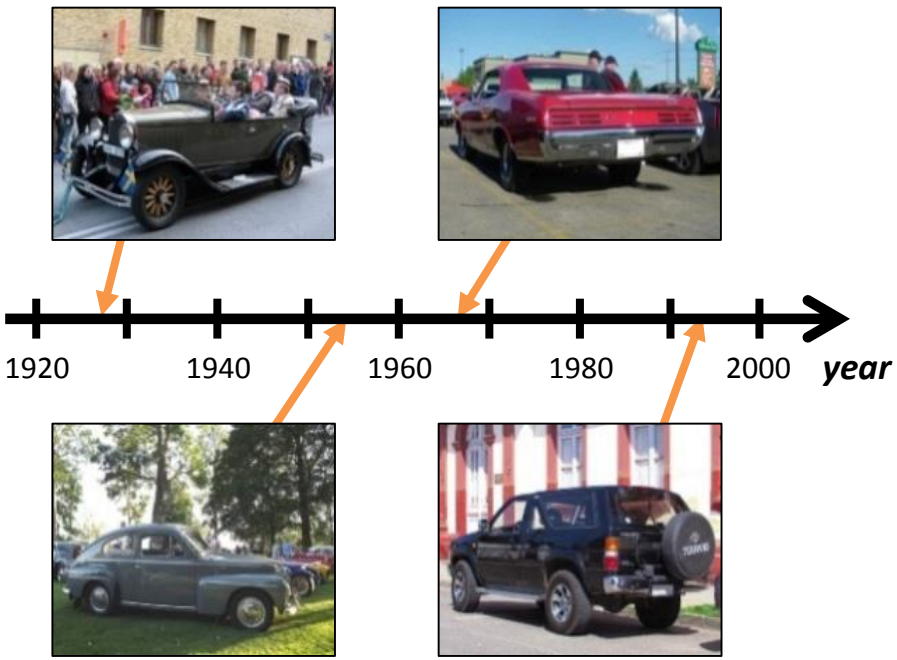
Spatial:





Our Goal

- Mine mid-level visual elements in temporally- and spatially-varying data and model their "visual style"



when?

Historical dating of cars

[Kim et al. 2010, Fu et al. 2010, Palermo et al. 2012]

where?

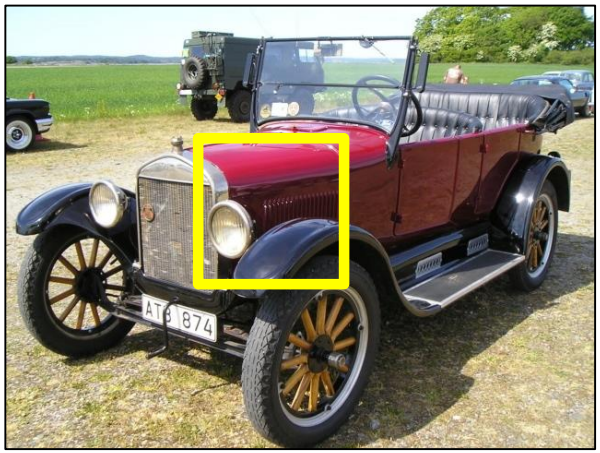
Geolocalization of StreetView images

[Cristani et al. 2008, Hays & Efros 2008, Knopp et al. 2010, Chen & Grauman. 2011, Schindler et al. 2012]

Key Idea

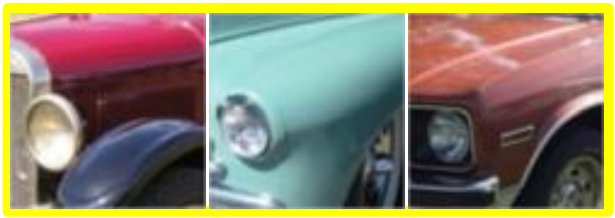
1) Establish connections

1926



1947

1975



1926 1947 1975

"closed-world"

2) Model style-specific differences

Approach

Unsupervised Discovery of Mid-Level Discriminative Patches

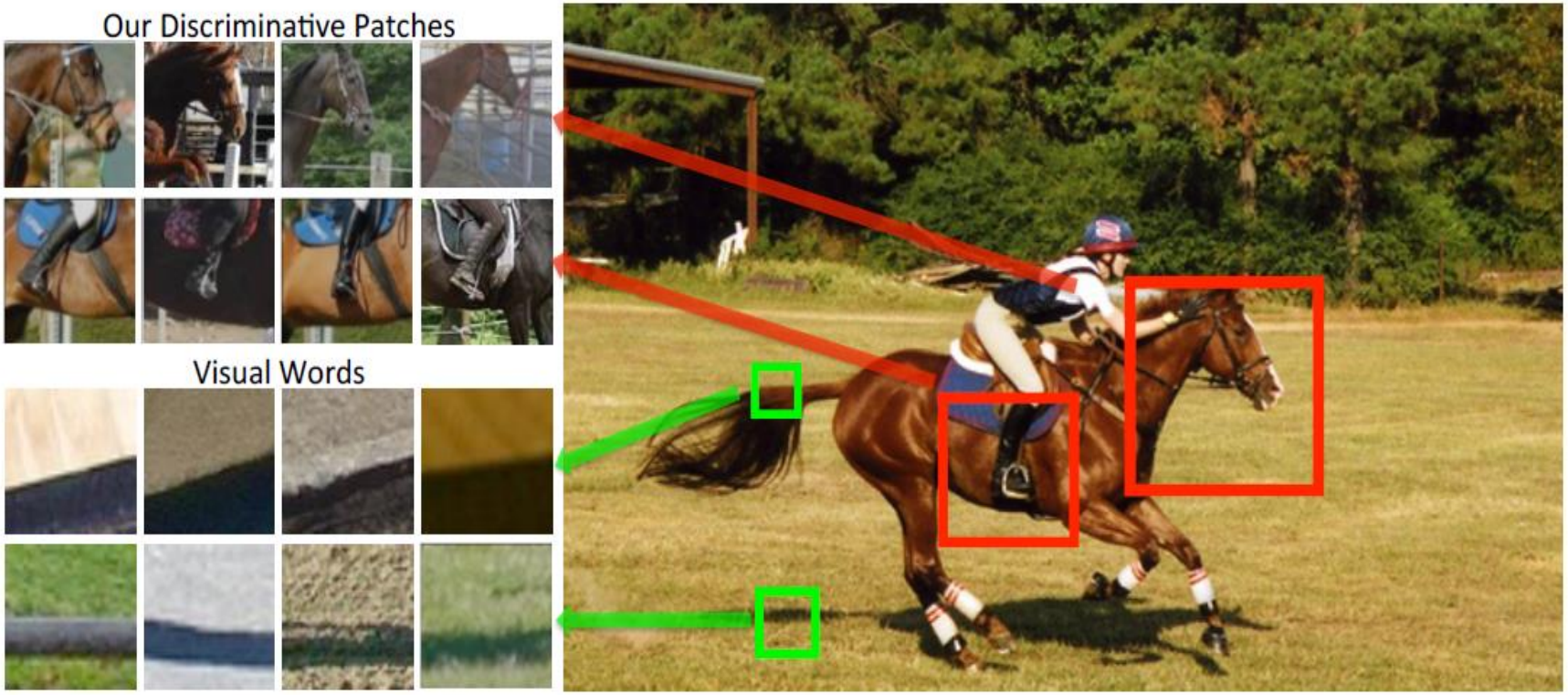
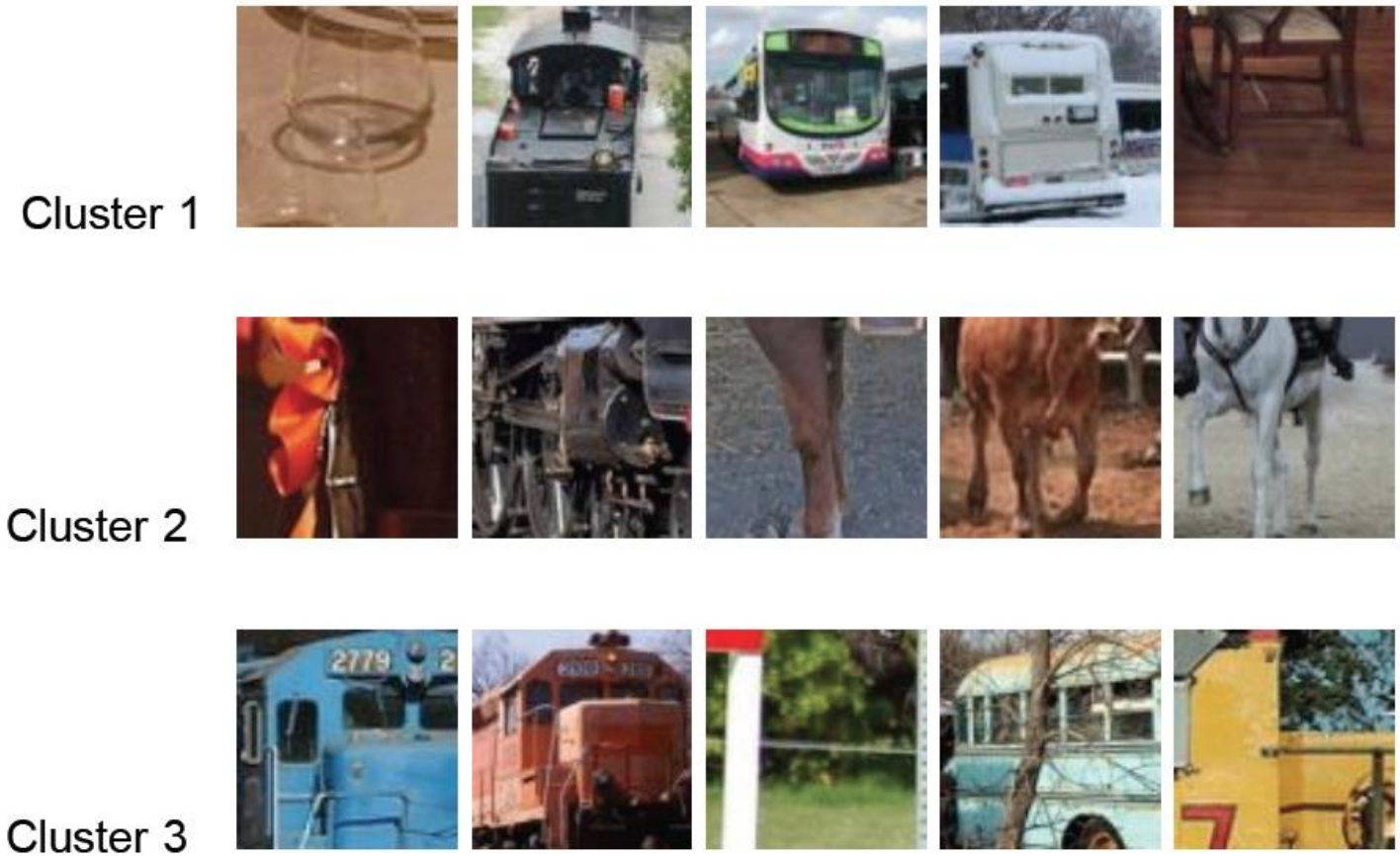


Fig. 1. The top two detected Visual Words (bottom) vs. Mid-level Discriminative Patches (top), trained without any supervision and on the same large unlabeled dataset.

Can we get nice parts without supervision?

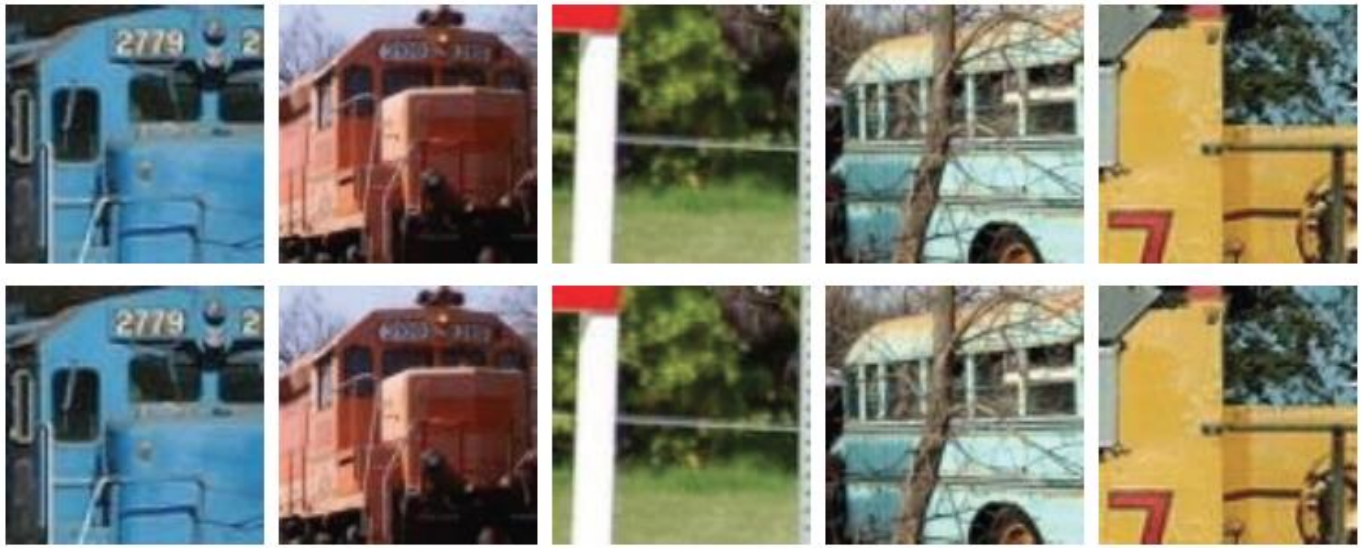
- Idea 0: K-means clustering in HOG space



Still not good enough

Initial

Final



- The SVM memorizes bad examples and still scores them highly
- However, the space of bad examples is much more diverse
- So we can avoid overfitting if we train on a training subset but look for patches on a validation subset

Why K-means on HOG fails?

- Chicken & Egg Problem
  - If we know that a set of patches are visually similar we can easily learn a distance metric for them
  - If we know the distance metric, we can easily find other members

Idea 1: Discriminative Clustering

- Start with K-Means
- Train a discriminative classifier for the distance function, using all other classes as negative examples
- Re-assign patches to clusters whose classifier gives highest score
- Repeat

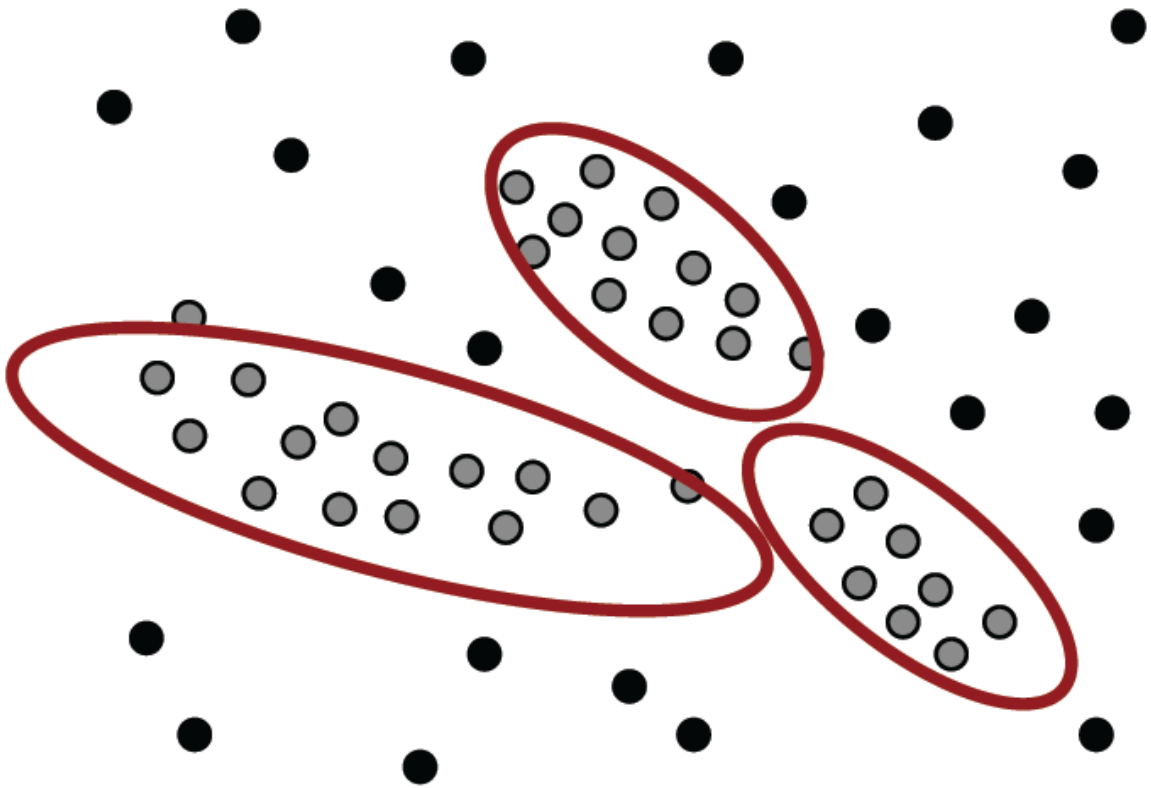
Idea 2: Discriminative Clustering+

- Start with K-Means or kNN
- Train a discriminative classifier for the distance function, using Detection
- Detect the patches and assign to top k clusters
- Repeat

Can we get good parts without supervision?

- What makes a good part?
  - Must occur frequently in one class (representative)
  - Must not occur frequently in all classes (discriminative)

Discriminative Clustering+



Discriminative Clustering+

Initial

Final

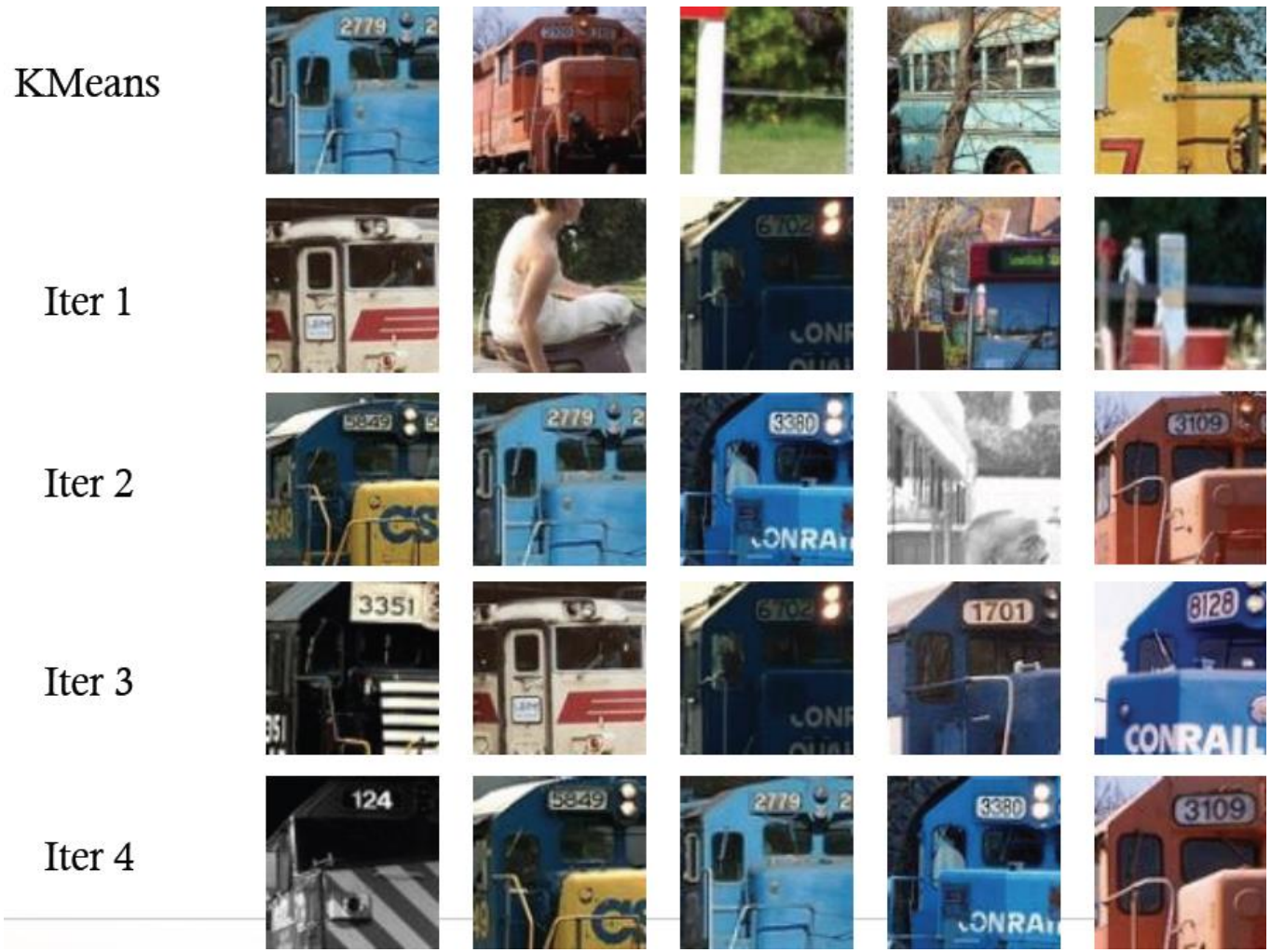
Initial

Final

Idea 3: Discriminative Clustering++

- Split the discovery dataset into two equal parts (training and validation)
- Train on the training subset
- Run the trained classifier on the validation set to collect examples
- Exchange training and validation sets
- Repeat

Discriminative Clustering++



Algorithm 1 Discover Top n Discriminative Patches

Require: Discovery set $\mathcal{D}$, Natural World set $\mathcal{N}$

$\mathcal{D} \Rightarrow \{D_1, D_2\}$; $\mathcal{N} \Rightarrow \{N_1, N_2\}$ ▷ Divide $\mathcal{D}, \mathcal{N}$ into equal sized disjoint sets

$S \Leftarrow rand\_sample(D_1)$ ▷ Sample random patches from D_1

$K \Leftarrow kmeans(S)$ ▷ Cluster patches using KMeans

while not $converged()$ **do**

  for all i such that $size(K[i]) \geq 3$ **do** ▷ Prune out small ones

    $C_{new}[i] \Leftarrow svm\_train(K[i], N_1)$ ▷ Train classifier for each cluster

    $K_{new}[i] \Leftarrow detect\_top(C[i], D_2, m)$ ▷ Find top m new members in other set

  end for

  $K \Leftarrow K_{new}$; $C \Leftarrow C_{new}$

  $swap(D_1, D_2)$; $swap(N_1, N_2)$ ▷ Swap the two sets

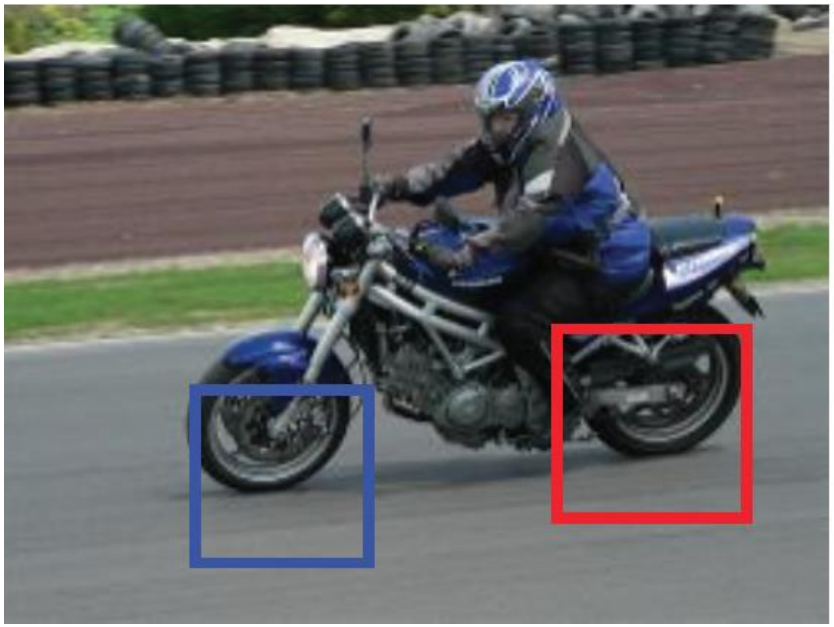
end while

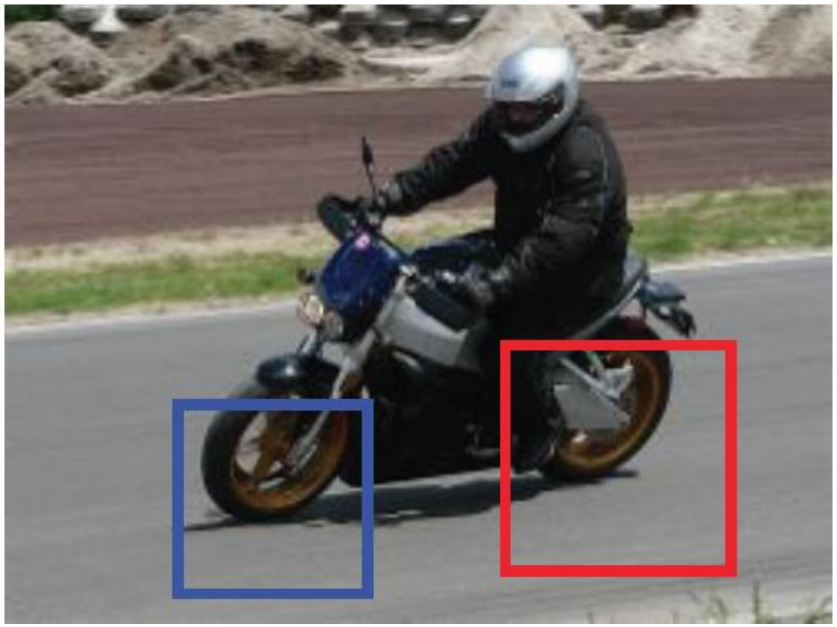
$A[i] \Leftarrow purity(K[i]) + \lambda \times discriminativeness(K[i])\ \forall\ i$ ▷ Compute scores

return $select\_top(C, A, n)$ ▷ Sort according to scores and select top n patches

Doublets: Discover second-order relationships

- Start with high-scoring patches
- Find spatial correlations to other (weaker patches)
- Rank the potential doublets on validation set





Doublets

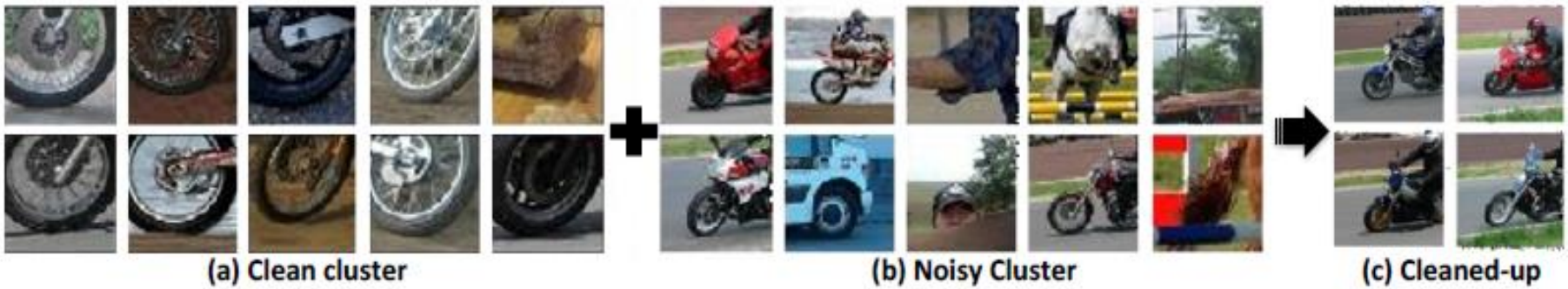


Fig. 5. Cluster clean-up using "doublets". A visually non-homogeneous cluster (b) that has learned more than one concept, when coupled into a doublet with a high quality cluster (a), gets cleaned up (c).

Fig. 6. Examples of discovered discriminative "doublets" that were highly ranked.

AP on MIT Indoor-67 scene recognition dataset

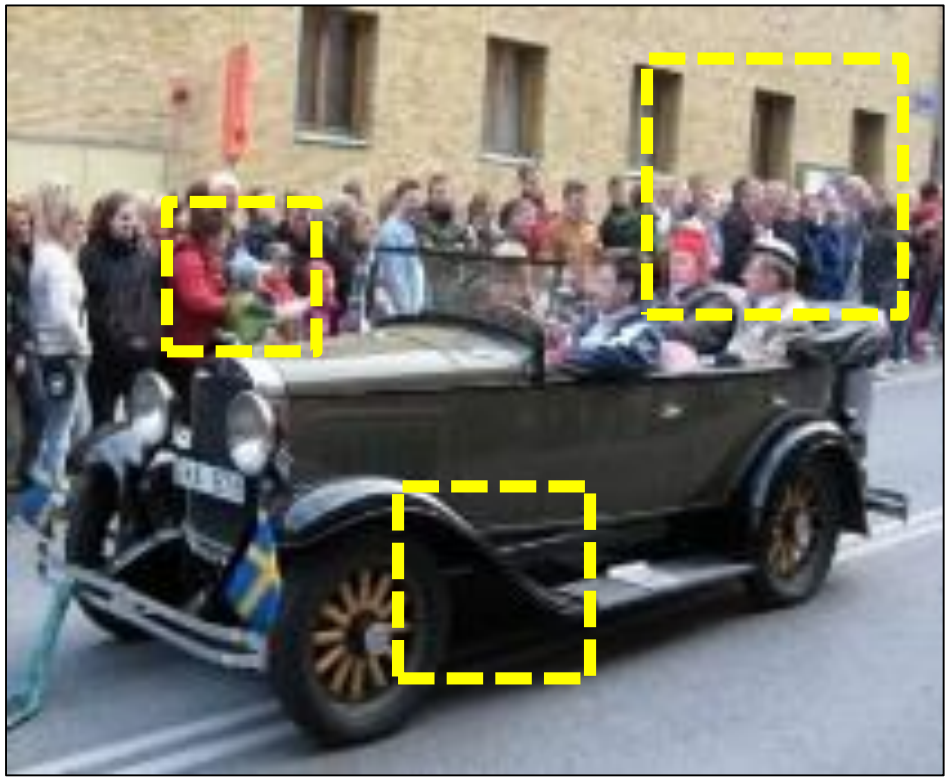
GIST	29.7	Ours+GIST	44.0
Spatial Pyramid HOG ($SPHOG$)†	29.8	Ours+SP	46.4
Spatial Pyramid SIFT (SP)†	34.4	Ours+GIST + SP	47.5
ROI-GIST [5]	26.5	Ours+DPM	42.4
Scene DPM [4]	30.4	Ours+GIST+DPM	46.9
MM-Scene [37]	28.0	Ours+SP+DPM	46.4
Object Bank [3]	37.6	GIST+SP+DPM [4]	43.1*
Ours	**38.1**	Ours+GIST+SP+DPM	**49.4**

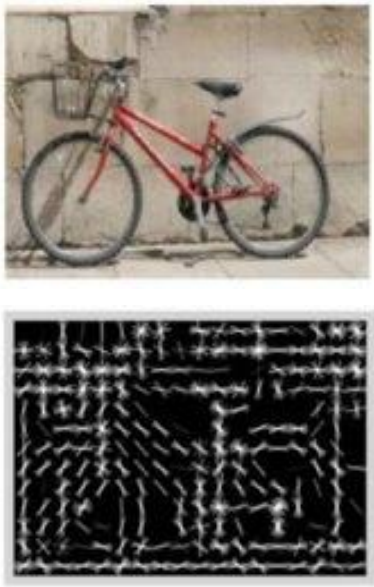
Table 1. Quantitative Evaluation: Average Classification on MIT Indoor-67 dataset. *Current state-of-the-art. †Best performance from various vocabulary sizes.

Coming Back

Mining style-sensitive elements

- Sample patches and compute nearest neighbors



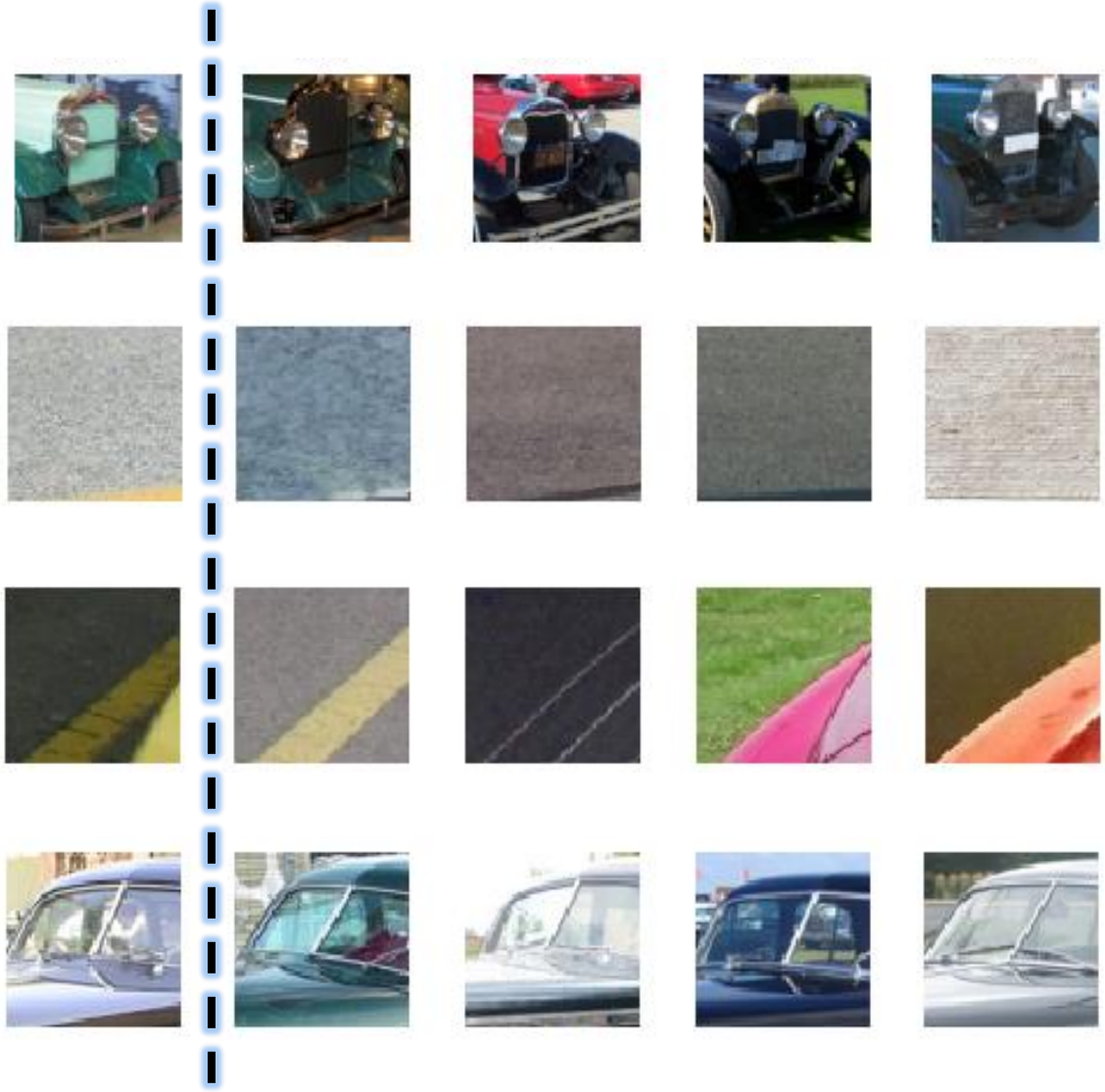


[Dalal & Triggs 2005, HOG]

Mining style-sensitive elements

Patch

Nearest neighbors



Mining style-sensitive elements

Patch

Nearest neighbors

style-sensitive

Mining style-sensitive elements

Patch

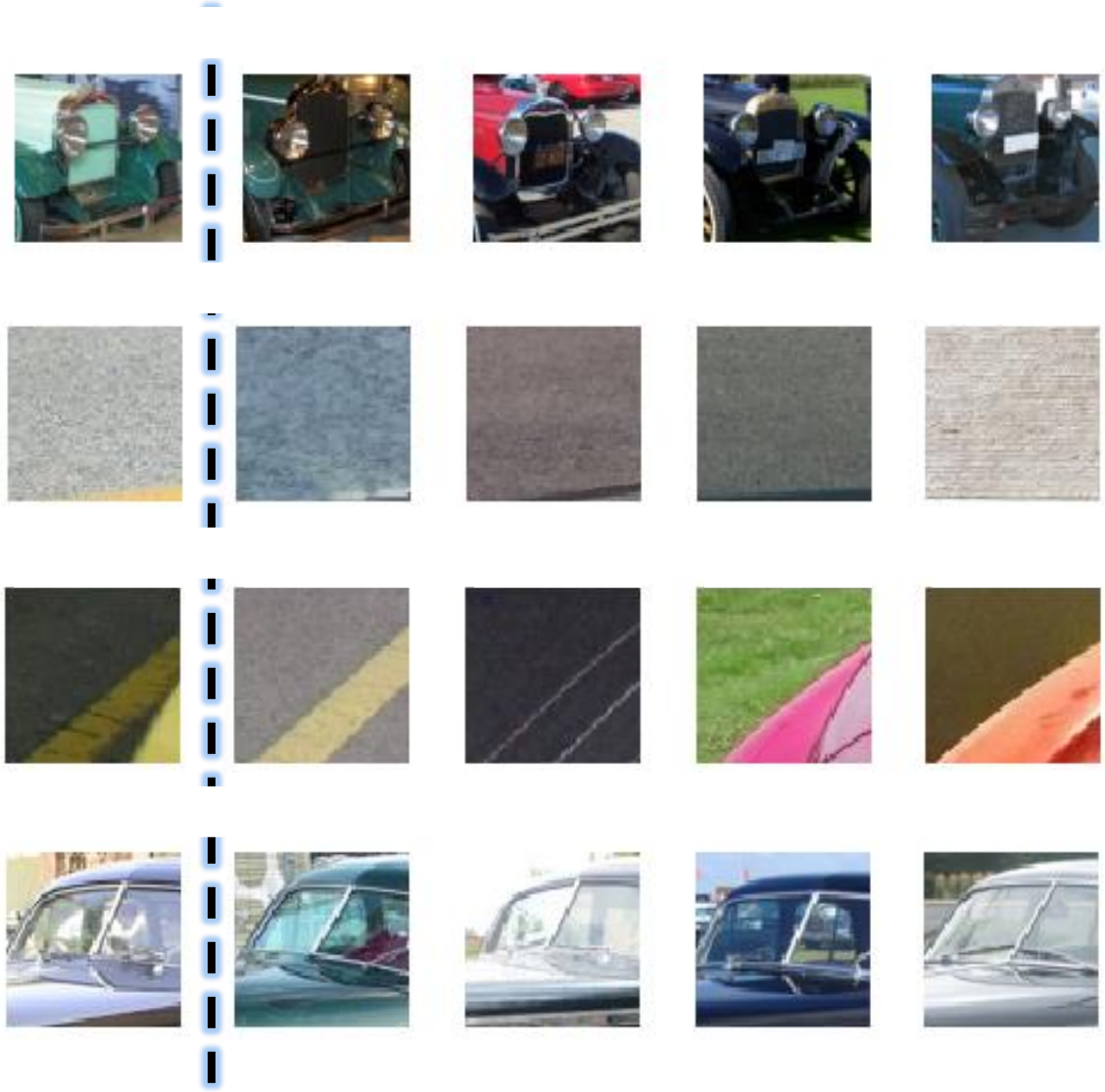
Nearest neighbors

style-*in*sensitive

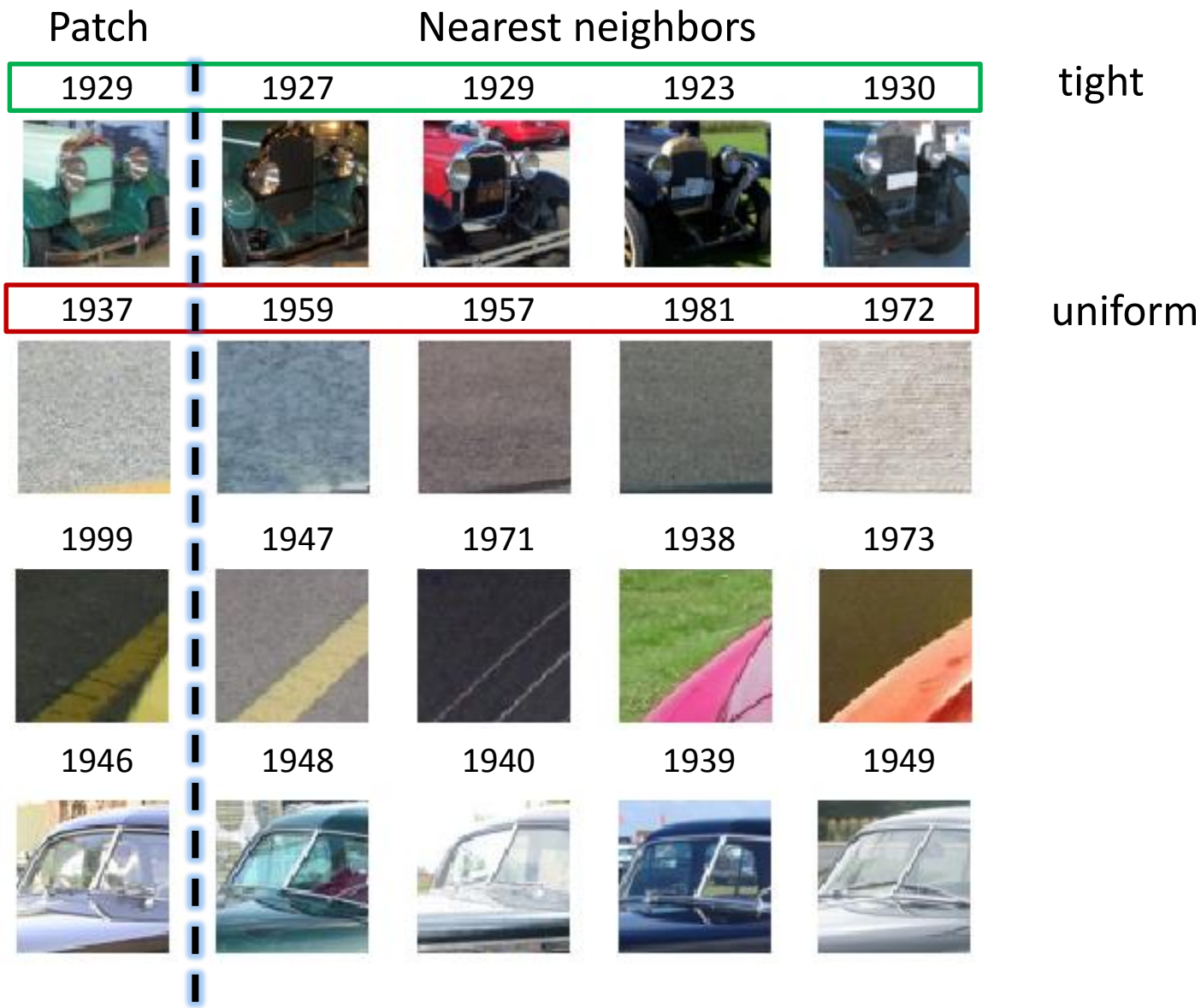
Mining style-sensitive elements

Patch

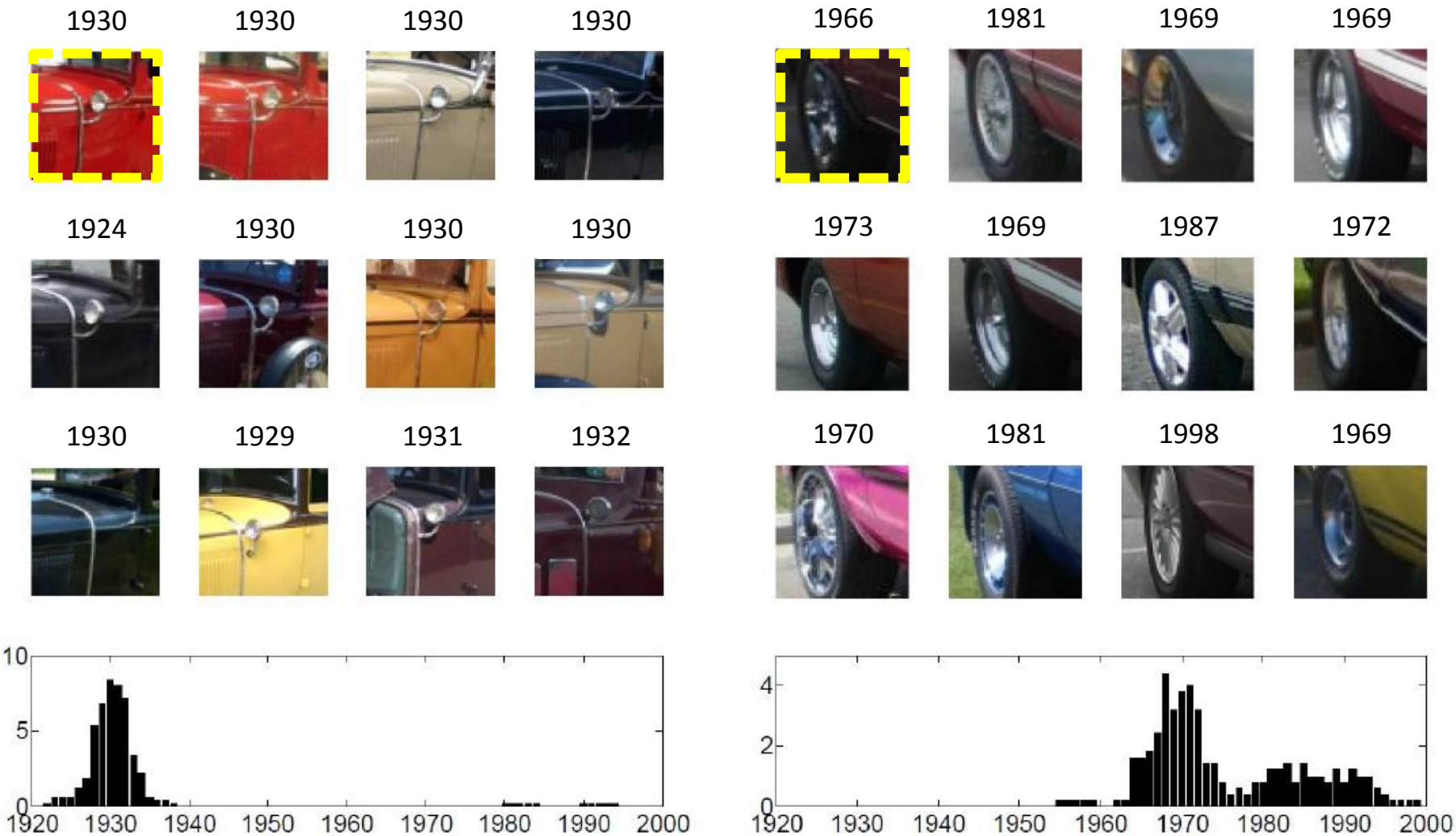
Nearest neighbors



Mining style-sensitive elements

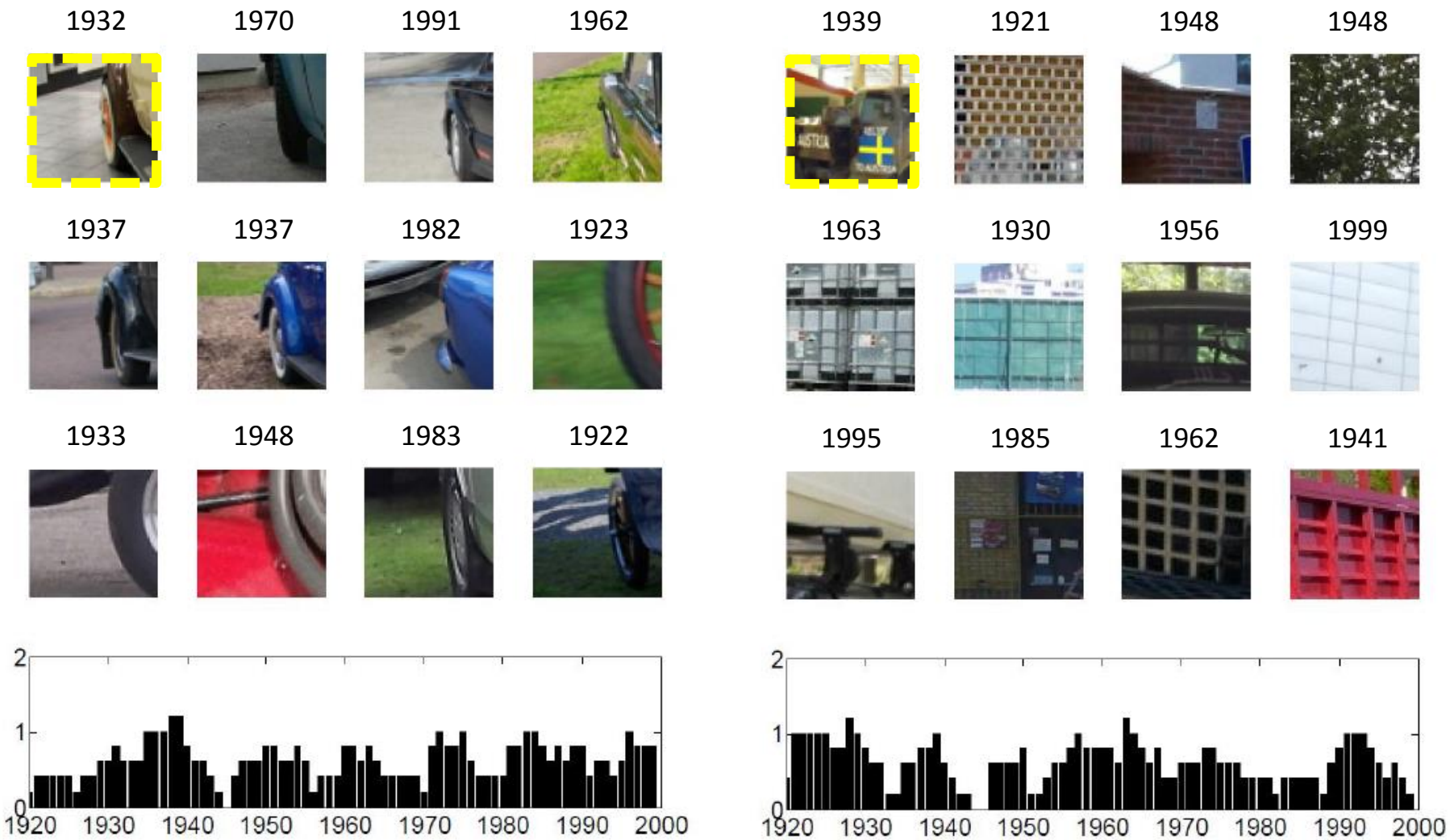


Mining style-sensitive elements



(a) Peaky (low-entropy) clusters

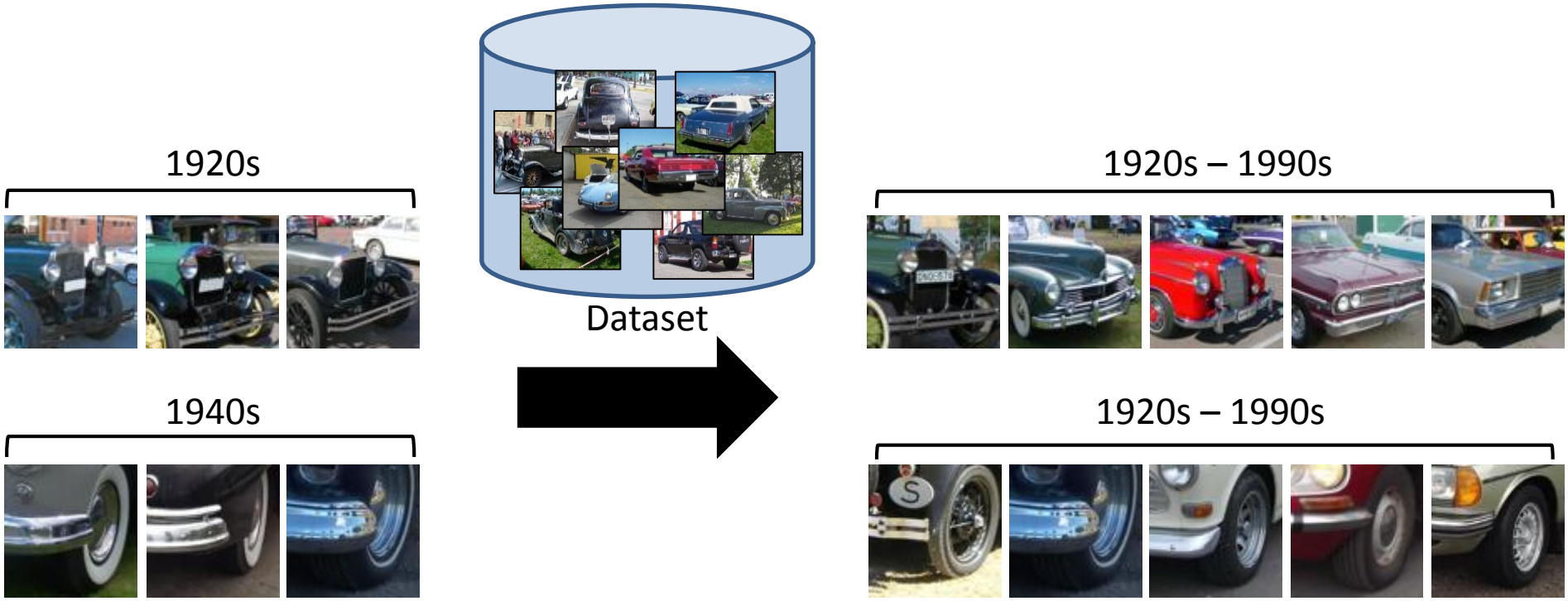
Mining style-sensitive elements



(b) Uniform (high-entropy) clusters

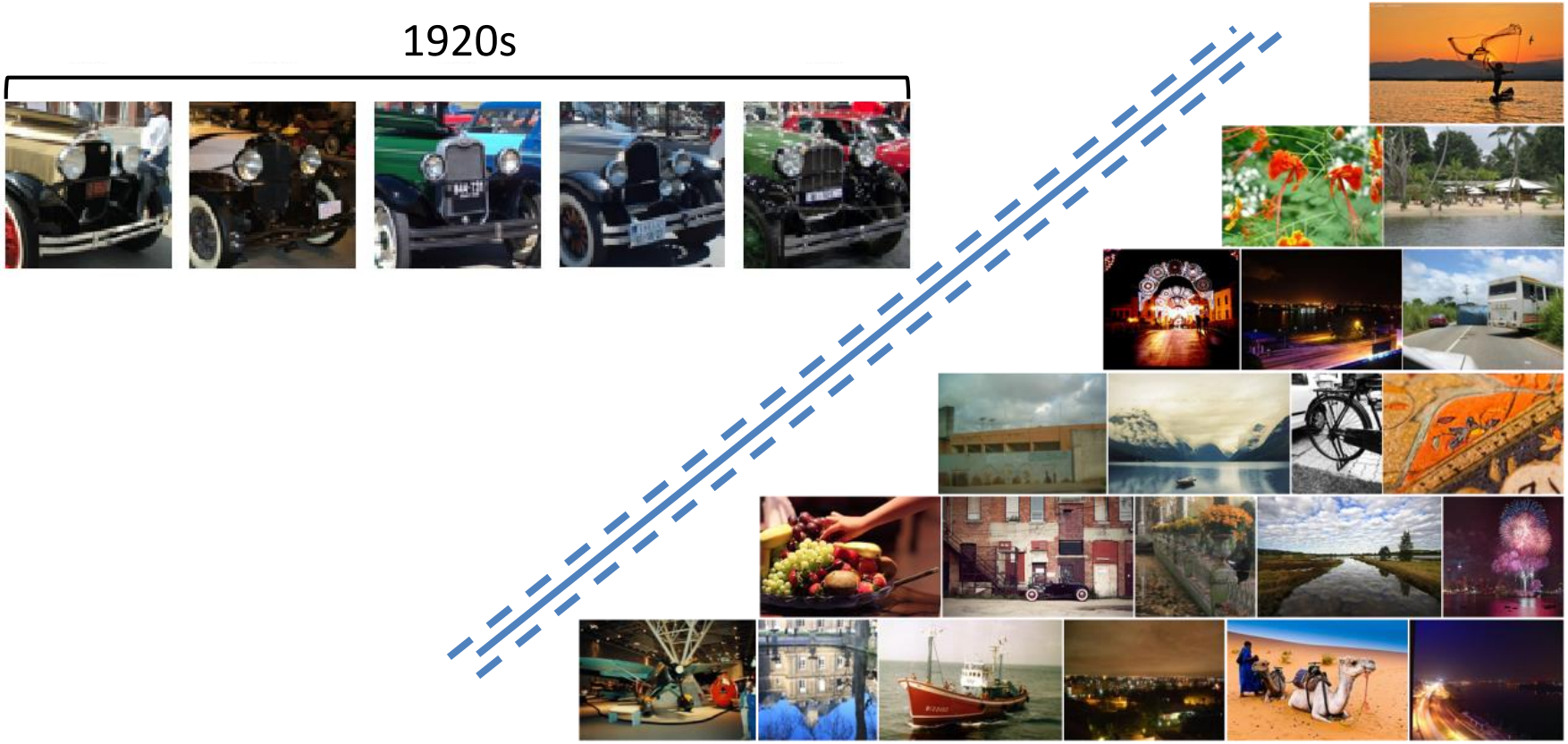
Making visual connections

- Take top-ranked clusters to build correspondences



Making visual connections

- Train a detector (HoG + linear SVM) [Singh et al. 2012]



Natural world “background” dataset

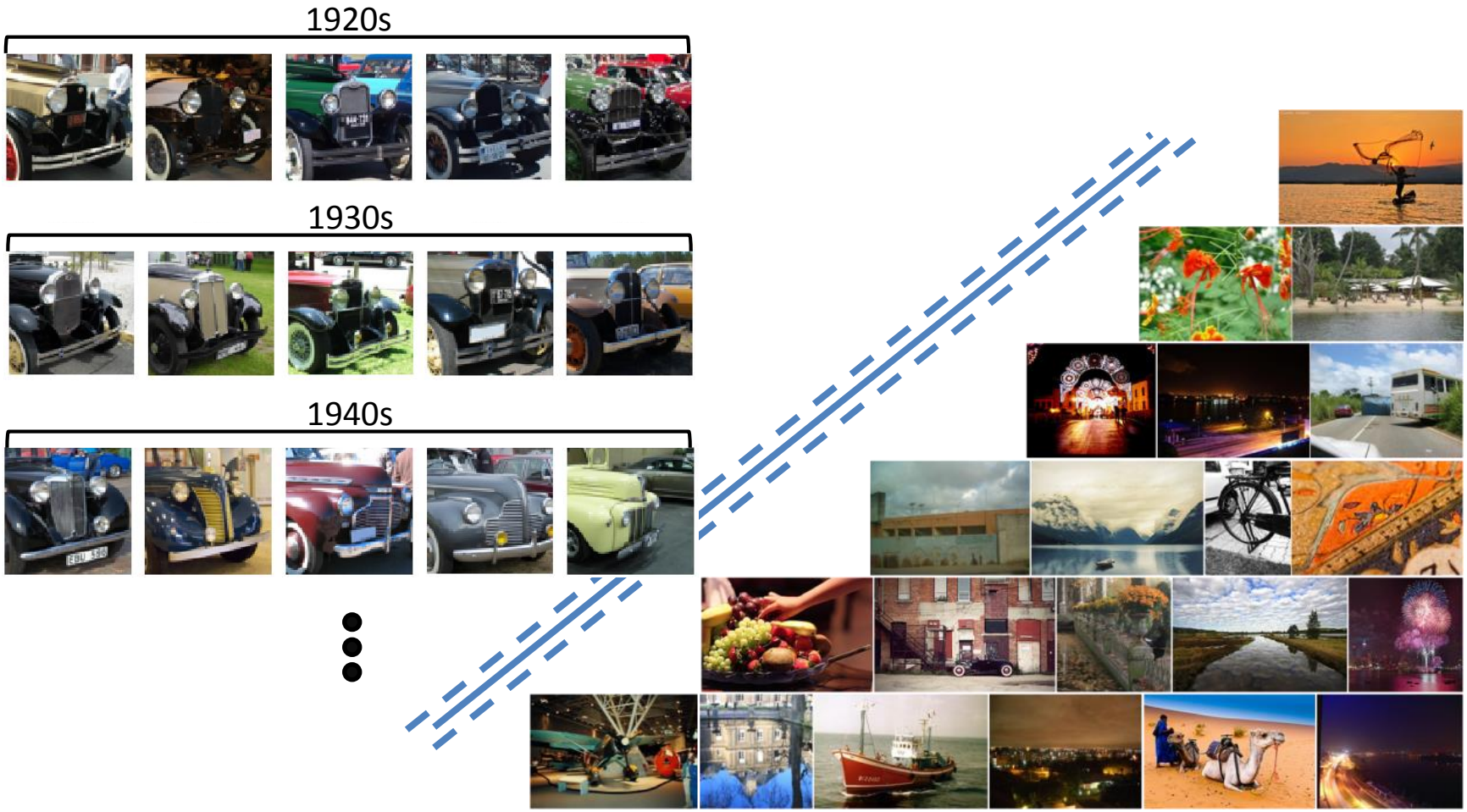
Making visual connections

Top detection per decade

[Singh et al. 2012]

Making visual connections

- We expect style to change gradually…



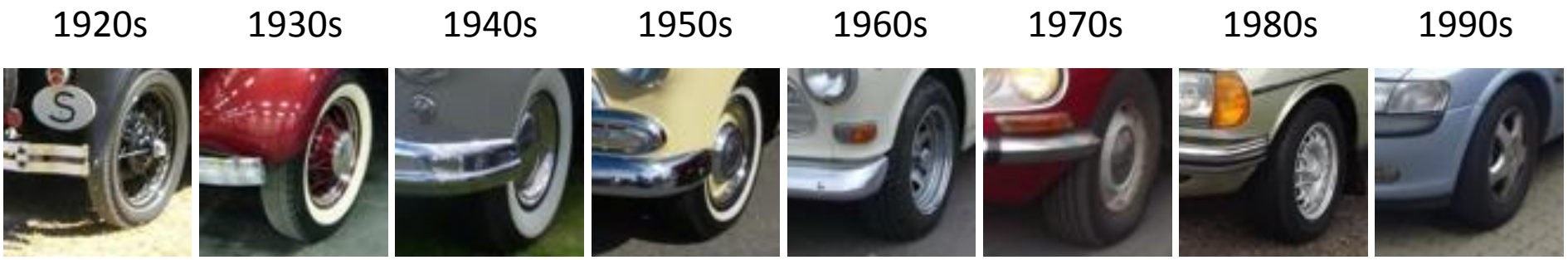
Natural world “background” dataset

Making visual connections

1920s 1930s 1940s 1950s 1960s 1970s 1980s 1990s

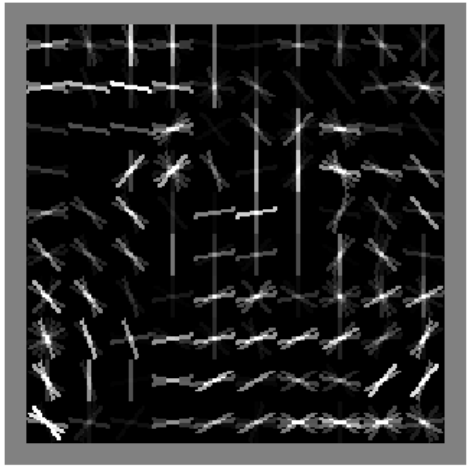
Top detection per decade

Making visual connections

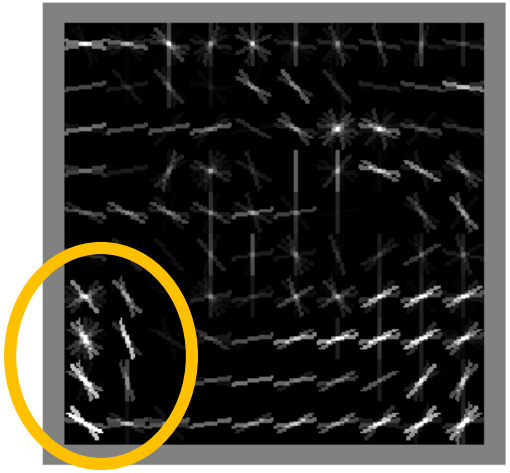


Top detection per decade

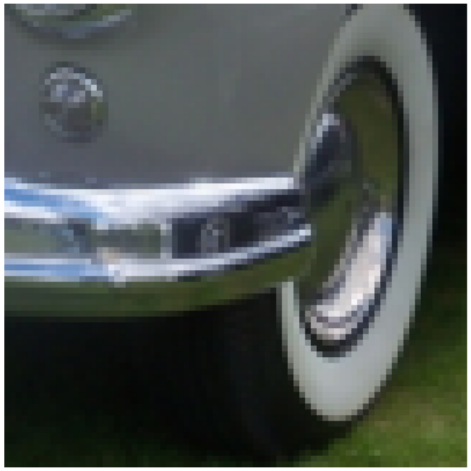
Making visual connections

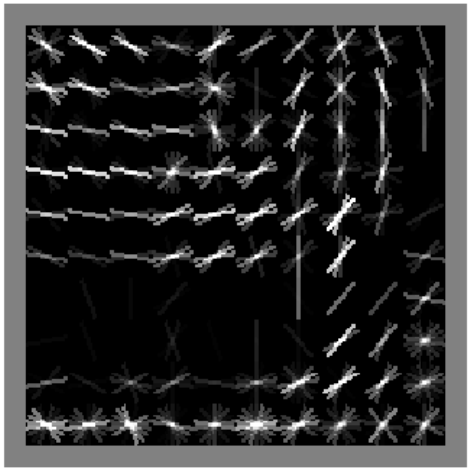


Initial model (1920s)

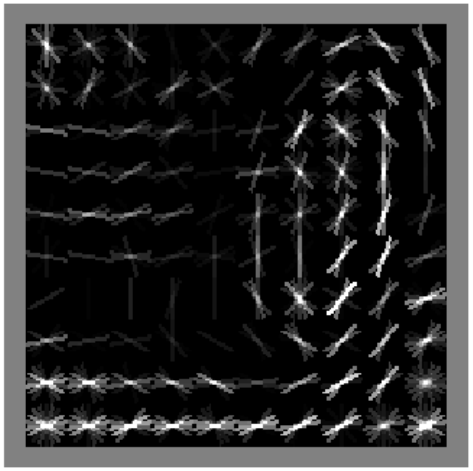


Final model



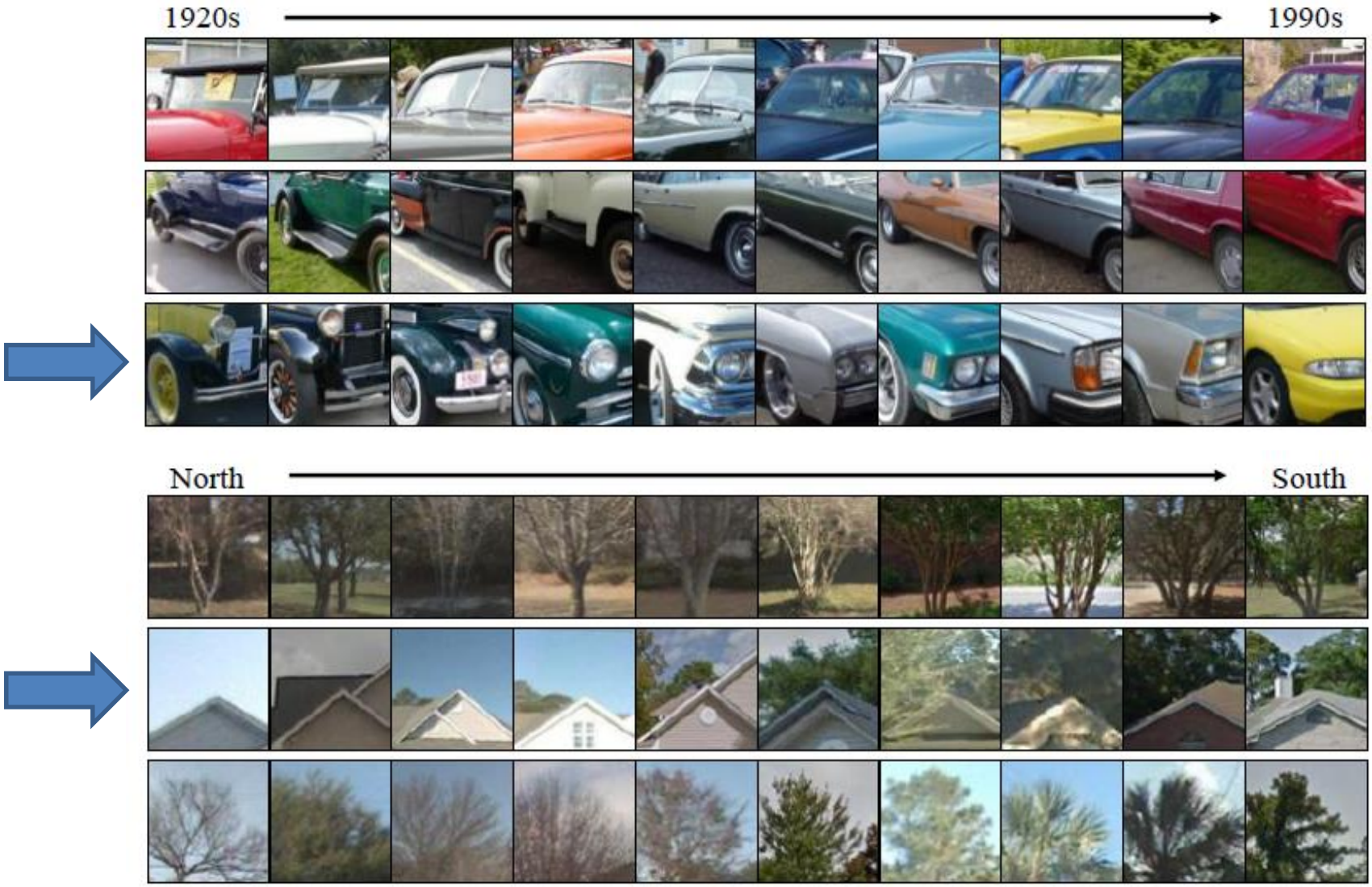


Initial model (1940s)



Final model

Results: Example connections



Training style-aware regression models

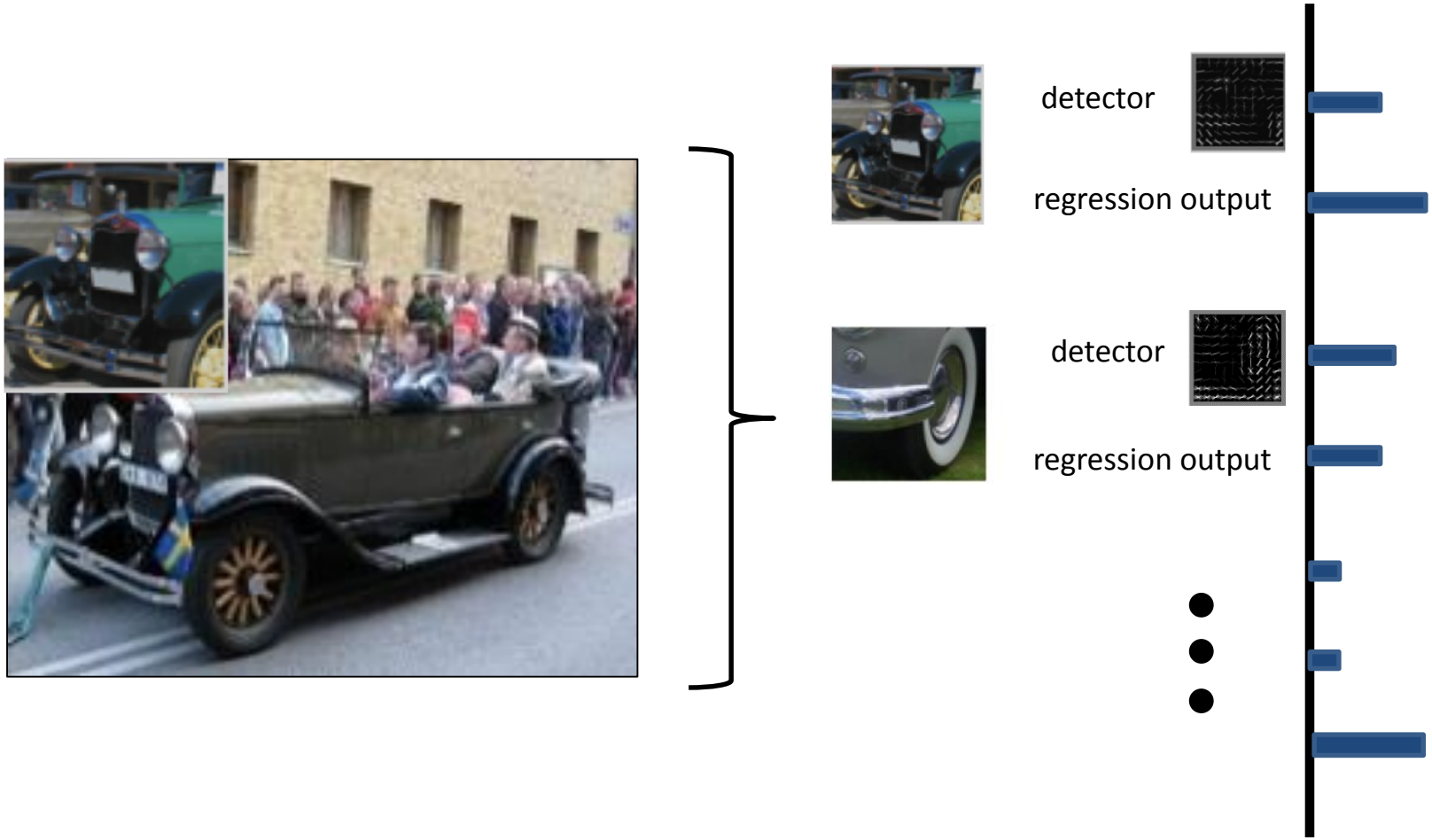
Regression model 1

Regression model 2

⋮

- Support vector regressors with Gaussian kernels
- Input: HOG, output: date/geo-location

Training style-aware regression models



- Train image-level regression model using outputs of visual element detectors and regressors as features

Figure 8. We visualize the styles that a single style-aware regressor has learned by averaging the predictions for each decade.

Results

Results: Date/Geo-location prediction

Crawled from www.cardatabase.net

Crawled from Google Street View

- 13,473 images
- Tagged with year
- 1920 – 1999

- 4,455 images
- Tagged with GPS coordinate
- N. Carolina to Georgia

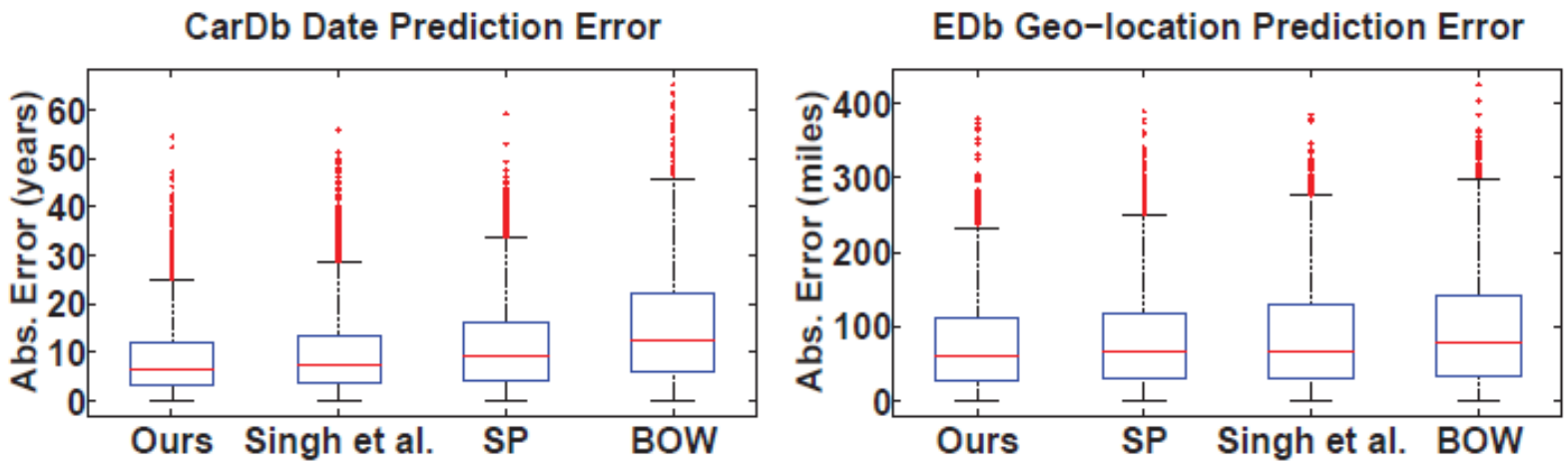


Figure 6. Box plots showing date and location prediction error on the CarDb and EDb datasets, respectively. Lower values are better. Our approach models the subtle stylistic differences for each discovered element in the data, which leads to lower error rates.

Results: Date/Geo-location prediction

Crawled from www.cardatabase.net

Crawled from Google Street View

	Ours	Singh et al. [23, 4]	SP [13]	BOW
CarDb (years)	**8.56**	9.72	11.81	15.39
EDb (miles)	**77.66**	87.47	83.92	97.78
IMCDb (years)	**13.53**	15.32	17.06	18.65

Table 1. Mean absolute error on CarDb, EDb, and IMCDb for all methods. The result on IMCDb evaluates cross-dataset generalization performance. Lower values are better.

Results: Learned styles

Average of top predictions per decade

Extra: Fine-grained recognition

Mean classification accuracy on Caltech-UCSD Birds 2011 dataset

Ours	Zhang et al. CVPR 2012	Berg, Belhumeur CVPR 2013
41.01	28.18	56.89

Zhang et al. ICCV 2013	Chai et al. ICCV 2013	Gavves et al. ICCV 2013
50.98	59.40	62.70

weak-supervision

strong-supervision

Conclusions

- Models *visual style*: appearance correlated with time/space
- First establish visual connections to create a *closed-world,* then focus on *style-specific differences*

Thank you!

References

1. http://techtalks.tv/talks/style-aware-mid-level-representation-for-discovering-visual-connections-in-space-and-time/59414/
2. http://web.cs.ucdavis.edu/~yjlee/projects/style_iccv2013.pdf
3. http://graphics.cs.cmu.edu/projects/discriminativePatches/discriminativePatches.pdf
4. http://www.cs.berkeley.edu/~rbg/ICCV2013/KeypointParts.pptx