# When Was That Made?

Sirion Vittayakorn Alexander C. Berg Tamara L. Berg
University of North Carolina at Chapel Hill
sirionv, aberg, tlberg@cs.unc.edu

## Abstract

*In this paper, we explore deep learning methods for estimating when the objects were made. Temporal estimation of objects is a challenging task which requires expertise in the object domain. With temporal information of objects, historian, genealogists, sociologist, archaeologist or conservationists can study the past through the objects. Toward this goal, we utilize features from existing deep networks and fine-tune new networks for temporal estimation task. The results demonstrate that the deep learning approach outperforms both a color-based baseline and visual data mining approach which is the previous state of the art method for the temporal estimation. To gain the insights into the deep network performance, we provide the analyses of neuron activations and their entropy including neuron temporal sensitivity, neuron activity and the correlation between discriminative parts from the deep network and the data mining approach. Finally, we demonstrate the potential of the temporal estimation pipeline for an interesting application such as fashion trend analysis.*

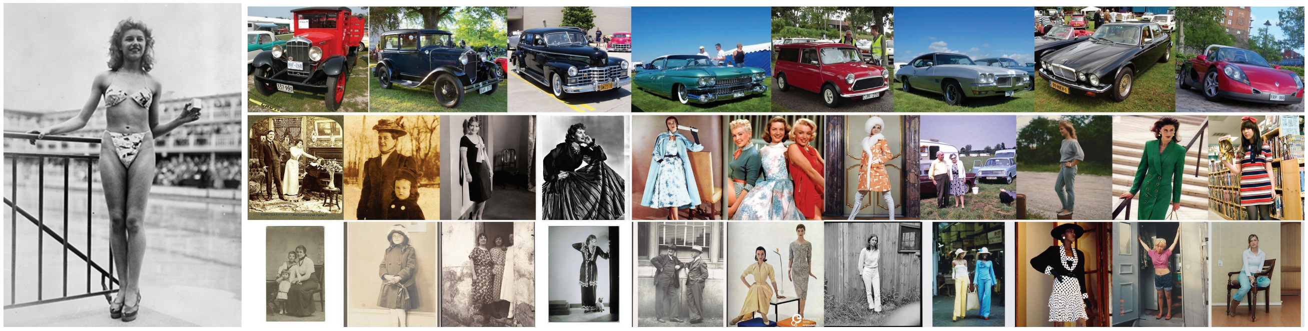
## 1. Introduction

Historical photographs are the valuable resource for learning about the past. With the temporal information about the photograph, ones can extract various type of knowledge from a photograph based on their own interest. For example, from Figure 1a, the designer might see a first bikini in the fashion industry from Louis Réard in 1946. The economist might see the consequence of the economic fallout, lack of fabric resources for textile industries, after World War II from the same photograph.

Dating the photographs is a challenging task, without the expertise, the task is very difficult. In January 2008, The Library of Congress was asking the public for annotating their collections through the collaboration project with Flickr called The Commons [2]. The Commons is a designated area of Flickr where the cultural heritage institutions can share their historical photographs and Flickr users are invited to describing the photographs through tags or comments. In October 2008, although The Commons received the overwhelming positive response from the community with 10.4 million views and 98.5% of photographs have at least one community-provided tag, only 5.8% of the tags are time period. To determine the period of the photographs, many literature propose various manual approaches based on visual clues from objects that appear in the photograph such as clothing [31, 34, 36], hair styles [9, 37] or photographic artifact [25]. However, the manual approaches are impractical for large data and some of them are domain specific. Due to these limitations, an automatic approach for temporal estimation is required.

Inspired by the visual clues of the vintage color photographs, [26] suggest a feature which captures temporal information based on the evaluation of color imaging processes over time. Focusing on visual clues of vintage cars, [22] propose a data-driven approach to automatically discover the mid-level visual patterns that correlate with time or space. In this work, we take an alternative, purely discriminative, approach to the temporal estimation task using deep-learning based methods and achieve the greater accuracy than previous methods. Our first approach, training the discriminative models on top of existing Convolutional Neural Network (CNN) features, yields the average mean absolute error (MAE) of $7.55 \pm 0.51$ years compared to the previous state of the art of 8.56 years in [22] on car images. By adapting and fine-tuning the existing networks to directly estimate the time, we improve the performance to 3.97 years. Moreover, our experiments demonstrate that the deep learning approach is domain independent as it can easily adapt to the new domain, two novel clothing datasets, and achieves the better performance than the baselines.

To our knowledge this is the first time deep networks have been used to estimate the time period of objects. Due to the complicated structure of the deep neuron networks, their learned representations are not easily interpretable. To gain insights into what the deep networks have learned, the analyses of neuron activity and the comparison between temporal discriminative parts from deep network and the mid-level discovery approach [22] are presented. Finally, we demonstrate the potential of our temporal estimation pipeline for an appealing application such as the analysis



(a)

(b)

Figure 1: (a) Micheline Bernardini modeling Réard's bikini at the Piscine Molitor in July 5, 1946. (b) Examples images from CarDb (first row), Flickr clothing dataset (second row) and Museum dataset (third row), sorted by time.

of the influence of vintage fashion on runway collections.

In summary, our contributions are: 1) Deep learning approaches to estimate when an object was made, 2) Two novel datasets of 67,771 dated photographs of clothing items made between 1900 and 2009; one collected from Flickr, and the other collected from museums, 3) Analyses of what the temporal estimation networks have learned, and comparison to the mid-level patterns learned by the previous state of the art approach [22], and 4) the analysis of the influence of vintage styles on fashion show collections.

The rest of the paper is organized as follows. First, we review related works (Sec 2). Then, we describe three dated object datasets, one existing, and two novel datasets (Sec 3). Next, we describe our deep learning based approaches to estimate when objects were made (Sec 4) and evaluate these models on the temporal estimation task (Sec 5). After that, we present several neuron activity analyses of what our networks have learned and how our approach compares to the other temporal prediction approach (Sec 6). Finally, we apply our pipeline to explore the influence of vintage fashion on the fashion show images (Sec 7).

## 2. Related work

This section will review the previous work relevant to deep learning for visual recognition, visual analyses of deep networks, and visual data mining.

**Deep Learning:** Convolutional neural networks (CNNs) have been one of the driving forces toward improving performance of many tasks. In the past recent years, CNNs have almost taken over the recognition community after demonstrations with their remarkable object classification [21, 32, 18] and detection [35] performance on benchmark dataset [6]. The feature representations learned by these networks on ImageNet data have been shown to generalize well to other image classification tasks [8] as well as related tasks such as object detection [14, 30], pose estimation and action detection [16], or fine-grained category detection [40]. Moreover, in a somewhat related task to ours, S. Karayev *et al.* [20] show that using a pre-trained network [8] as a generic feature extractor, produces a better classifier for photo and painting style than hand-crafted features. We are not aware of any prior work on modeling historical visual style using deep-learning based methods.

**Analysis of CNNs:** Unlike hand-crafted features such as SIFT [24] or HOG [5], the representation learned by CNNs is not obviously interpretable. For many tasks the CNN is used as a black-box algorithm where it is not always clear why the CNN is outperforming previous approaches. Several recent works have attempted to peer into this box, to better understand the representations learned by CNNs. P. Fischer *et al.* [13] compare the learned representation with SIFT in a descriptor matching task. M.D. Zeiler *et al.* [39] propose several heuristic visualization techniques for units in the network. J. Long *et al.* [23] study the effectiveness of CNNs activation features for tasks requiring correspondence. Recently, B Zhou *et al.* [41] present a technique to visualize learned representations of each unit in the network. Here they focus on a large dataset of scene images and show that object detection is embedded in the network as a result of learning. We use variants of this approach to evaluate what our temporal networks have learned.

**Visual Data Mining:** Visual data mining approaches have been applied to various tasks: 1) discover the object categories from unsupervised data [17, 27, 11, 33], 2) identify the discriminative parts of actions [28], cities [7], or objects [22], and 3) discover the local attributes [3, 10, 29] in fine-grained recognition task. While the previous works focus on mining for the visual elements which are common among samples from the same category but discriminative from others, [22] goes beyond simply detecting recurring visual elements by modeling the transformation of these visual elements over space or time. Unlike the others, we do not only utilize the deep learning approach for temporal estimation task, but also analyze the CNNs representation in comparison to the existing data mining approach [22].

## 3. Datasets

In this work, We use three different datsets: 1) the Car Database (CarDb) [22], 2) a large novel collection of clothing photographs with associated dates from an online photo-sharing website, and 3) a novel collection of clothing photographs from museum collections. The example images of each dataset are shown in Figure 1b.

**Car Database:** CarDb [22] contains 13,474 photos of cars made from 1920 to 1999 resulting 8 temporal classes, collected from cardatabase.net.

**Flickr Clothing Dataset:** We initially collect more than 100,000 clothing related images together with their corresponding meta-data from a wide variety of 50 groups focused on vintage fashions (e.g., *"Fashions Past - Best and Worst"*, *"As She Was"*, etc.) from flickr.com. However, some of these images contain drawings of fashion items or depict other fashion related images without clear examples of clothing items. To remove these images we apply a face detection algorithm [42] to automatically filter out images without a depicted person. The remaining images are manually inspected to remove additional non-photographic content such as artwork, painting, and advertisements. From the meta-data such as title, description, and tags we automatically extract a potential decade label such as 90s, 1965, 1954-1957, 1920s, etc. Then, we quantized the labels into an 11-bin histogram with dates ranging from 1900-2009. The most frequent label will be assigned as temporal label of an image. The final dataset contains 58,350 clothing photographs with corresponding meta-data, including photo id, user id, title, description, tags, longitude, latitude, number of views and groups, etc.

**Museum Dataset:** Since the textual information of Flickr clothing dataset is provided by the user, the dates associated with objects can sometimes be noisy. Therefore, we collect an additional dataset of clothing related photographs from 2 different museums: the Metropolitan Museum of Art, and Europeana Fashion (a museum network co-funded by 22 partners from 12 European countries which represent leading European institutions and some of the largest collections in the fashion domain). The Museum dataset has 9,421 images between 1900 to 2009, showing clothing worn on people. Since museum collections have been curated by experts, their date labels are reliable. We use this dataset as an alternative test set to evaluate clothing date models trained on the larger Flickr clothing dataset. As this Museum dataset has a somewhat different domain than the Flickr clothing dataset, it is also used to evaluate the model generalization, i.e. training on one dataset and testing on another.

## 4. Approaches

We pursue two approaches for temporal estimation. First, we evaluate classifiers trained on features from a pre-trained network. Then, we explore adapting the network to directly predict time period, fine-tuning for this task.

**Pre-trained models:** We start with two CNN models pre-trained on 1.2 million labeled images from ImageNet. The first network, AlexNet, was originally described by [21] and the latter, VGG, was described by [32], both networks are implemented as part of the Caffe framework [19]. For each network, we extract the learned representation from the second fully-connected layer and use this 4096 dimensional vector as our visual representation. We experiment with two different classification methods: a linear Support Vector Machines (SVM) [12] with fixed $C_{svm} = 0.1$, and Support Vector Regressors (SVR) [4] with fixed $\epsilon = 0.1$ and set $C_{svr} = 100$.

**Fine-tuned models:** Network fine-tuning has proven successful for adapting networks to new tasks beyond their original purpose such as object detection [15], pose estimation and action detection [16], or fine-grained category detection [40]. In this work, we are interested in fine-tuning the original object classification model for the temporal estimation task.

From each pre-trained model, described by [21, 32], we fine-tune 3 different models. The first model, the fine-tuned Car model, is fine-tuned on 10,130 training images and tested on 3,343 images from the CarDb (same train/test split as [22]). The second model, the fine-tuned Clothing model, uses $3/4$ of the images from the Flickr clothing dataset as training set. Moreover, since our datasets depict vintage photographs from 1900-2009, historic color cues, e.g black and white color or sepia tones in early photos and color photographs in later photos, appear in both datasets. To evaluate performance without any color cues, we also fine-tune a third network using only black/white images from the Flickr clothing dataset. Finally, both models are evaluated during testing on the portion of the Flickr clothing dataset and on the Museum dataset. To fine-tune each model, we change the output of the last fully connected layer from 1000 classes to 11 temporal classes (one per decade) for the Flickr clothing dataset and 8 temporal classes for CarDb. We trained our models using stochastic gradient descent with a batch size of 50 examples, momentum of 0.9, weight decay of 0.0005 and decrease the learning rate of the models to 0.00001. Finally, we trained the networks for 50,000 cycles.

## 5. Temporal Estimation Experiments

We evaluate performance of our models compared to Lee *et al.* [22] and F. Palermo *et al.* [26].

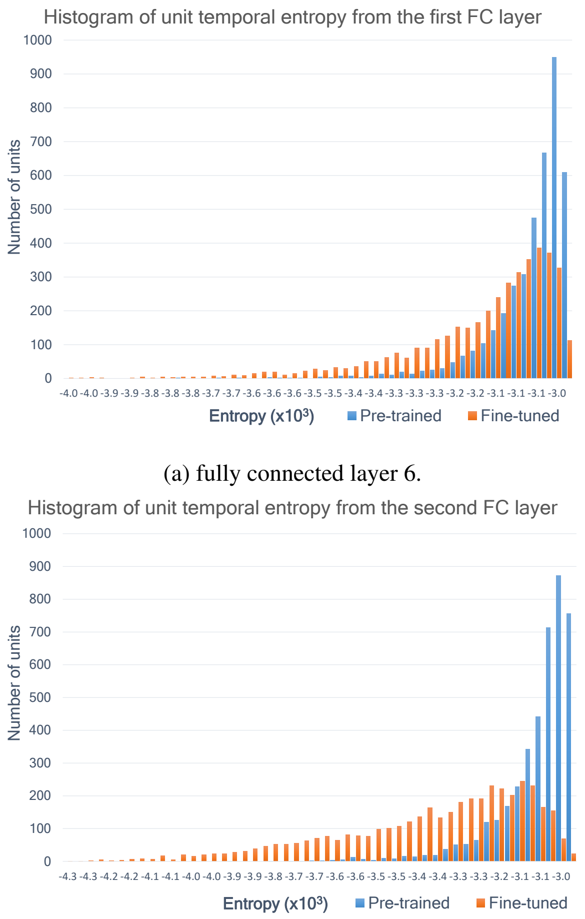
**Pre-trained Model Performance:** For the pre-trained

|  | **CarDb** | **Clothing** | **Museum** |
|---|---|---|---|
| Lee *et al.* [22] | 8.56 | 17.74 | 19.56 |
| Palermo *et al.* [26] | - | 17.21 | 21.21 |
| alexNet + SVM | 7.77 | 12.99 | 17.33 |
| alexNet + SVR | 8.10 | 16.73 | 22.00 |
| VGG-16 + SVM | 6.90 | 12.60 | 16.35 |
| VGG-16 + SVR | 7.43 | 15.87 | 18.76 |
| alexNet (BW/FT) | 6.78 | 17.16 | 17.96 |
| alexNet (FT) | 6.17 | 12.88 | 16.43 |
| VGG-16 (BW/FT) | 4.27 | 13.66 | 16.40 |
| VGG-16 (FT) | **3.97** | **11.54** | **14.23** |

Table 1: The mean absolute error (years) training and testing on CarDb and training on Flickr clothing dataset and testing on held out clothing and Museum dataset.

models, we extract features from last fully-connected layer, then train SVM and SVR models to estimate when an object was made. We evaluate performance on both the car and clothing datasets. Table 1 shows mean absolute error (MAE) in years across features and classifiers, including comparisons to previous works [22] on all datasets. Though we are relying on pre-trained models in this approach, we already outperform the previous state of the art on both datasets. Moreover, for clothing, the results show that even though the domain shift is quite evident (the results on the Museum dataset are worse than results on the held out portion of the Flickr clothing dataset), the deep learning feature is beneficial for the temporal estimation task. Evaluations achieve error reductions of $0.95 \pm 2.47$ years and $3.19 \pm 2.06$ years on the Museum and Flickr clothing collections respectively compared to the baseline [22].

**Fine-tuned Model Performance:** The fine-tuned models consistently outperform the pre-trained models on both object categories. For cars, the MAE decreases around $2.48 \pm 0.86$ years. Similar trends apply for clothing, the fine-tuned model decrease MAE around $2.86 \pm 2.42$ years on the Museum dataset and $1.67 \pm 1.94$ years on the nosier Flickr clothing dataset. Moreover, the results confirm that the fine-tuned network learns visual elements beyond color by outperforming two color-based baselines. The first baseline, by F. Palermo *et al.* [26] which proposed temporally discriminative features related to the evolution of color imaging processes over time, achieves worse MAE on both datasets compare to our fine-tuned networks. Finally, we also compare our approach with a second baseline, where we fine-tune the network using only black/white images from Flickr clothing dataset. The results show that the B/W model achieves about $2.53 \pm 1.2$ years higher MAE in this task. These results emphasize that even though color is an important clue, our fine-tuned network is able to learn temporally sensitive features of an object beyond color.



(a) fully connected layer 6.

(b) fully connected layer 7.

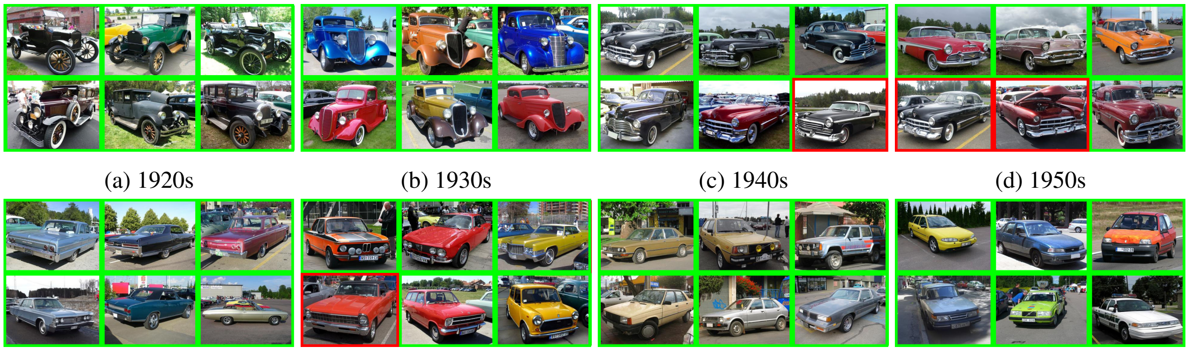
Figure 2: Histogram of unit temporal entropy from 2 different fully connected layers from alexnet(blue) and fine-tuned network(orange).

## 6. Deep Network Analyses

Based on these quantitative results, we find that deep learning methods are promising for temporal estimation task. However, unlike the patch discovery methods [22], the learned representations from deep networks are not immediately interpretable. Thus, in this section, we provide some analyses of the CNN networks to gain additional understanding about what the fine-tuned networks have learned.

### 6.1. Temporally-sensitive units

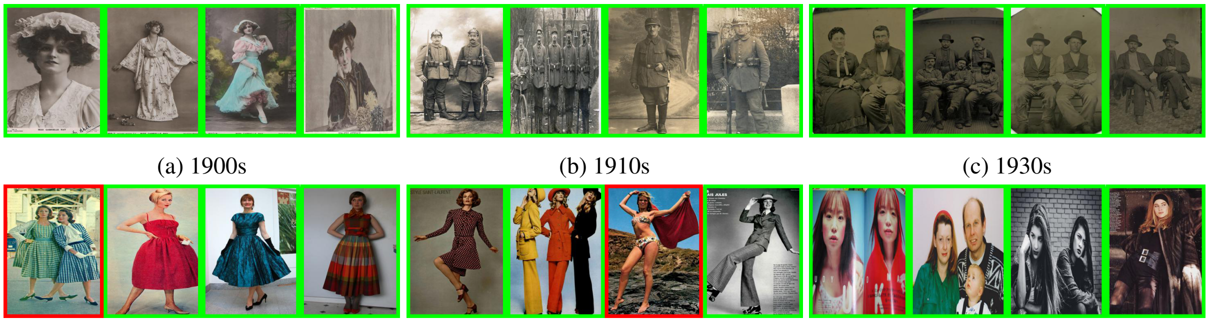
Since the fine-tuned network outperforms the pre-trained network, we hypothesize that there may be interesting differences in the network before and after fine-tuning. One potential difference that we investigate is the temporal sensitivity of units. To explore the temporal sensitivity of units in both networks, for each unit, images are ranked by their maximum activation. Then we bin the top $N =$



(a) 1920s (b) 1930s (c) 1940s (d) 1950s

(e) 1960s (f) 1970s (g) 1980s (h) 1990s

Figure 3: Top 6 images with maximum activation from the low entropy units in each decade. The green rectangle indicates an image is from the same decade as the decade with *the lowest entropy* for a given unit, while the red indicates otherwise.



(a) 1900s (b) 1910s (c) 1930s

(d) 1950s (e) 1970s (f) 1990s

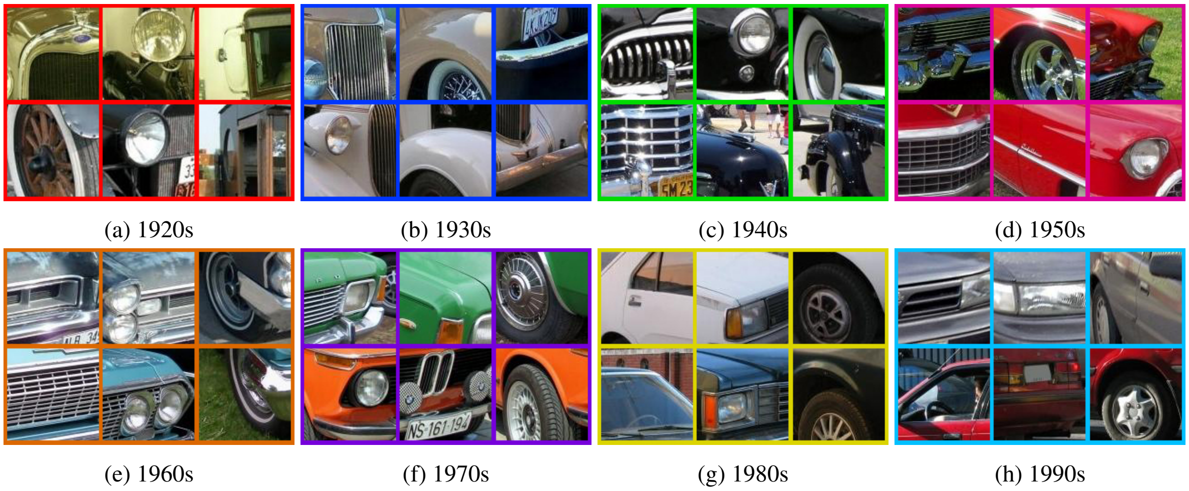
Figure 4: Top 4 images with maximum activation from the low entropy units in each decade. The green rectangle indicates an image is from the same decade as the decade with *the lowest entropy* for a given unit, while the red indicates otherwise.

500 maximum activation images into a temporal histogram, by decade, and compute the entropy of each histogram: $E(u) = -\sum_{i=1}^{n} H(i) \cdot log_2 H(i)$ where $H(i)$ denotes the histogram count for bin $i$ and $n$ denotes the number of quantized label bins. Lower entropy values indicate higher temporal sensitivity. Finally, we compute the entropy histogram of all units from 2 Fully-Connected layers as shown in Figure 2a (first FC layer) and Figure 2b (second FC layer). Both histograms show that the fine-tuned network has more low entropy units than the pre-trained network which indicate that units have been fine-tuned to capture a temporally discriminative feature for a specific time period. These results are visible in both layers, but are more pronounced in second layer, which makes intuitive sense since this layer is closest to the end temporal estimation. Qualitative examples of top images ranked by their maximum activation from the low entropy units are shown in Figure 3 for the CarDb and Figure 4 for clothing dataset.

### 6.2. Unit activation analysis

Beyond more highly-tuned temporal sensitivity in units, we would like to understand whether the network has learned to detect temporally discriminative object appearances or not. Thus, we investigate the unit activation patterns to better understand whether the temporally sensitive regions correspond to semantic elements of objects. To do so, we follow the data-driven approach proposed by [41] to estimate the learned receptive fields (RFs) of units. To estimate a unit's RF, images are ranked by their maximum activations for that unit, and the top $K$ images are selected to identify image regions that led to these high activations. To recover the high activation regions within an image, each image is replicated many times with a small occluder of size 11x11 placed at one of about 5,500 locations in a dense grid (stride 3 pixels) in the images. Each occluded image is evaluated by the same network and the change in activation versus the original is calculated. Those differences are combined into a discrepancy map over the image. The intuition behind this approach is that if there is a large discrepancy between activations before and after occlusion, then the occluded region is important for activating that unit.

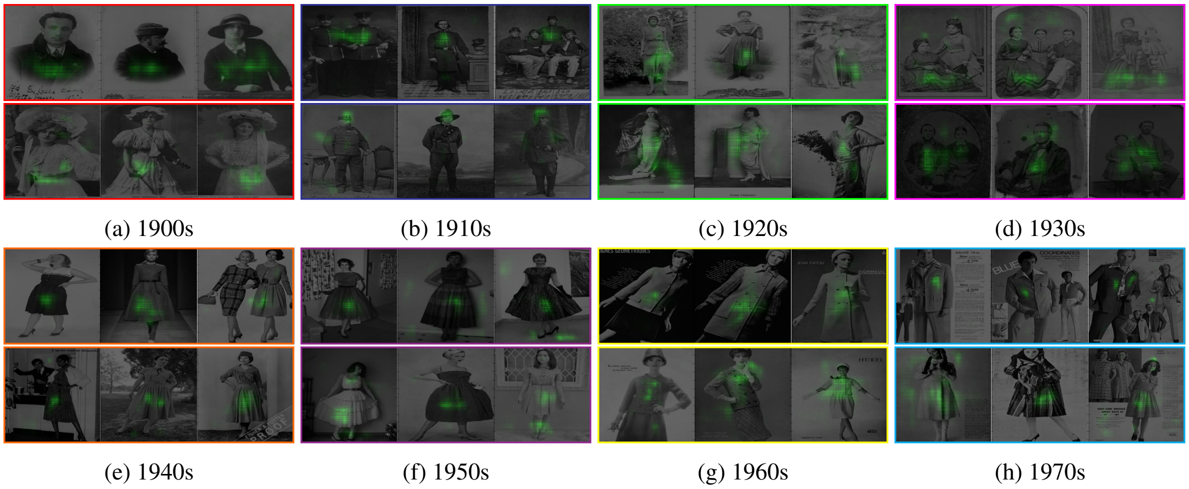
In order to investigate the most informative regions in an image, we focus on image regions which highly contribute



(a) 1920s (b) 1930s (c) 1940s (d) 1950s

(e) 1960s (f) 1970s (g) 1980s (h) 1990s

Figure 5: In each rectangle, 3 image regions, from 1920s to 1990s, with the maximum activations from 3 different units from fc7 of the fine-tuned network on CarDb are shown.



(a) 1900s (b) 1910s (c) 1920s (d) 1930s

(e) 1940s (f) 1950s (g) 1960s (h) 1970s

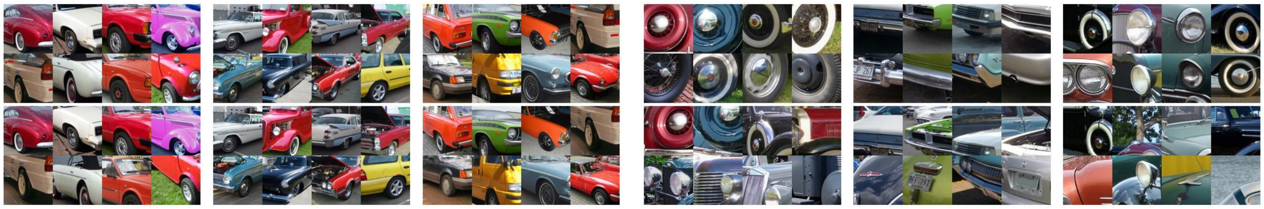
Figure 6: Each rectangle shows 3 images with the maximum activation regions, in green highlight, of units which have the lowest entropy in each decade of the fine-tuned network on Flickr Clothing dataset.

to the prediction decision. Therefore, we first select all true positive images from the fine-tuned network. For a given image, we rank the units based on their contribution to the prediction decision and assign the image to the top $N$ units. For each unit, we compute the discrepancy map of assigned images, following [41]. Figure 5- 6 show image regions that caused the maximum activation for the given unit from the last FC layer from CarDb and Flickr clothing dataset. Results indicate that units in the fine-tuned network respond to temporally informative parts such as front bumpers, headlights, or wheels for cars and cinched-in waists (40s-50s), mod dresses (60s) and leisure suits (70s) for clothing.

### 6.3. Discriminative part correlation

So far we have observed temporal sensitivity adaptation and indications that nodes in the network have tuned their activations to particular object parts. These results lead us to new interesting questions. Are the parts discovered by the network discriminative in time? Do these visual elements correspond to the style-sensitive elements discovered by [22]?

To identify correspondences between the visual elements learned by our fine-tuned network and the style-sensitive elements proposed by [22], we search for units which have a similar behavior as their style-sensitive detectors. We look for two behavioral factors: (1) high responses on similar sets of images, and (2) similar localization patterns. To im-



(a) high correlation

(b) poor correlation

Figure 7: For each block, the top two rows show style-sensitive patches from [22], while the bottom two rows show regions with the maximum activation from the same images. While (a) shows unit activations which are highly correlated to style-sensitive patches, (b) shows unit activations which are poorly correlated to style-sensitive patches

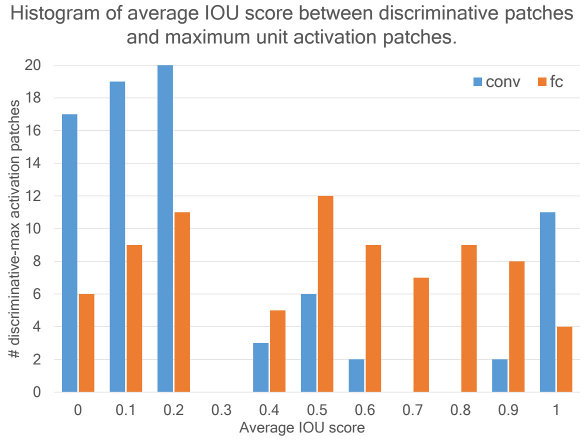


Figure 8: The average IoU score between style-sensitive patches and maximum activation patches.

plement this, we generate two image rankings: the first one is based on the maximum activation for a given unit $u$ over images, and the other is based on the maximum detector confidence for a given detector $d$ over images. Then, we defined the correlation $C$ between unit $u$ and generic detector $d$ as $C(u, d) = \frac{|A_n \cap D_n|}{n}$ where $A_n$ are the set of top $n$ images from the activation based ranking and $D_n$ are the set of top $n$ images from the detection based ranking ($n = 30\%$ in all experiments). For a given detector, we rank all units in each layer by the correlation score.

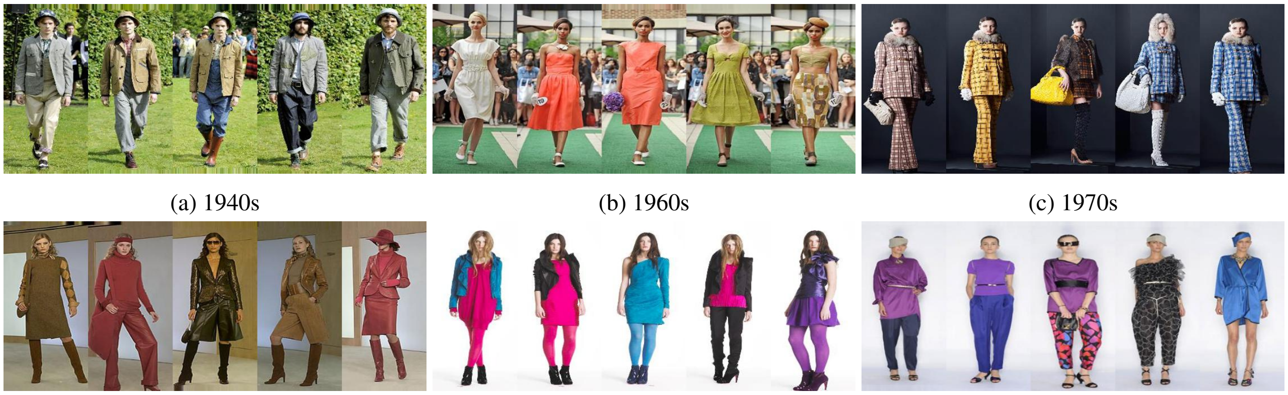
We evaluate this correlation on units in the last convolutional layer (conv) and a second fully connected layer (fc). The average correlation scores between style-sensitive detectors and their top 5 correlated units from conv and fc are 0.521 and 0.543 respectively, indicating that about half of the units in both layers overlap with style-sensitive detectors. Finally we compute the average Intersection-over-Union score (IoU) between our maximum activation patches and the style-sensitive patches. More specifically, for each correlated unit, we rank images by the maximum activation. Then, we sample 80 × 80 pixel patches with maximum activation and compute the IoU score between this patch and the style-sensitive patch from the correlated detector. In this experiment, we sample a maximum activation patch from the top 20 images/unit with 5 correlated units/style-sensitive detector. The average IoU scores of 20 × 5 style-sensitive and maximum activation patches are shown in Figure 8.

These results, again, emphasize that units in the network are fine-tuned to temporally sensitive parts of an object. Additionally, we find that only 7.5% of the patches have average IoU score $< 0.1$ while 61.25% of the patches have the average IoU score $\geq 0.5$, confirming that the style-sensitive parts from [22] are automatically discovered by our network. Qualitative examples of high/poor correlation patches are shown in Figure 7. While Figure 7a shows the maximum activation patches from units which are highly correlated with style-sensitive parts proposed by [22], Figure 7b shows visual elements which have low correlation to [22]. We posit that these additional patches are also temporally sensitive, and contribute to our improved performance.

## 7. The influence of vintage fashion

Denim miniskirts/jackets, ripped distressed jeans, tracksuits, trench coats, the list can go on and on. People look back to the past and say “That’s ridiculous! What were they wearing?”. Yet somehow the fashion world recycles these trends, makes a few tweaks, and voila! they keep coming back. Designers have to look for inspiration from somewhere, and what better place to be inspired by than the past?

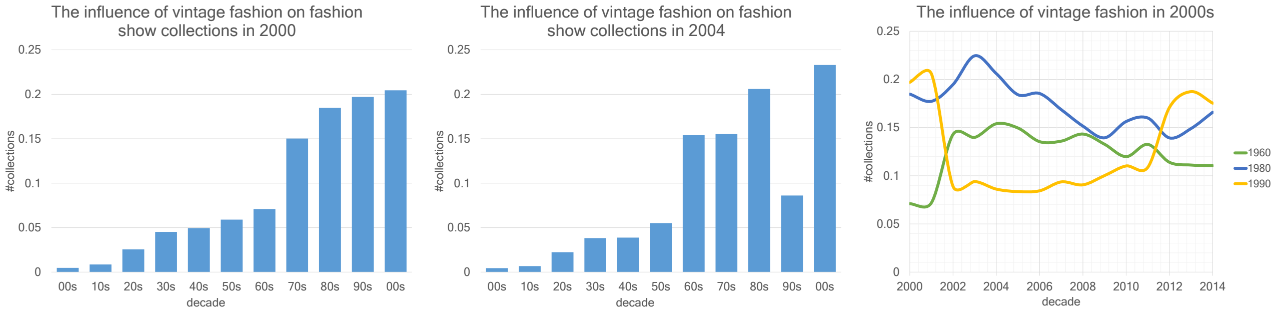
Therefore, we demonstrate the use of our learned models for predicting the influence of fashion from past decades on the fashion collections. In particular, we evaluate our models on a Runway dataset [38] containing 300k images from fashion shows over the years of 2000 to 2015. To estimate the influence of the vintage fashion on fashion collections, we first apply our fine-tuned model to estimate when the outfits were made. Then, we define the inspiration date of the collection as the decade with the highest probability among images from that collection. To evaluate our approach, we collect human judgments on the same task using



(a) 1940s (b) 1960s (c) 1970s

(d) 1970s (e) 1980s (f) 1980s

Figure 9: Predicting vintage influence in fashion collections. (a)-(f) indicate the decade of predicted influence.



(a) 2000 (b) 2004 (c) 2000-14

Figure 10: The influence of vintage fashion (1900s-2000s) in fashion show collections from (a) 2000 and (b) 2004. Figure (c) shows the influence of 1960s, 1980s and 1990s fashion through out 2000s - 2010s.

Amazon Mechanical Turk. For each assignment, 5 annotators are shown 5 fashion show images per collection with the associated collection description (written by the fashion experts), then the annotators are asked to identify the decade that inspired these images. We randomly pick up to 200 collections per predicted decade and remove collections with low human agreement (less than 3 of 5 agree). This leaves us with 300 collections for evaluation. On these collections, our model achieves the MAE of 8.6 years compared to non-expert judgments. Moreover, we also eveluated our approach based on fashion experts' descriptions. To do so, we filtered the runway collections where there exist both the specific words (e.g., inspire, influence, motivate or encourage) and the temporal words (e.g., 60s, seventies or edwardian) in the description. Then, 3 annotators are asked to identify the decade that inspired the collection from the text. The low human agreement (less than 2 of 3 agree) collections are removed, and 300 collections are randomly picked for the evaluation. Based on the fashion expert descriptions, our approach achieves the MAE of 9.70 years. The examples of our temporal prediction are shown in Figure 9.

Finally, we also observe temporal influence trends on fashion show collections over time. To do so, we explore the classification confidence of collections from the same year as shown in Figure 10a - 10b. If we look at the classification confidence of a particular vintage decade across years, we spot some interesting trends. For example, from Figure 10c, we can see that 90s fashion had a strong influence during both early 2000s and early 2010s while during the mid-2000s a revival of 60s and 80s fashion occurred [1].

## 8. Conclusions

In this work, we first explore CNN approaches to automatically estimate when objects were made, evaluated on an existing dataset of cars and two new datasets of vintage clothing photographs. Then, we provide several analyses of what the networks have learned, including exploring the temporal sensitivity of nodes, as well as examining node activations and comparison to discriminative parts learned by the data mining approach [22]. Finally, we propose an application for temporal estimation task of clothing in the real world scenario.

## References

[1] 2000s in fashion. https://en.wikipedia.org/wiki/2000s_in_fashion. Accessed: 2016-05-13. 8

[2] Library of congress photos on flickr. https://www.loc.gov/rr/print/flickr_pilot.html. Accessed: 2016-09-24. 1

[3] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 2

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. 3

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893 vol. 1, June 2005. 2

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2

[7] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM ToG*, 31(4), 2012. 2

[8] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014. 2

[9] M. Doyle. *An Illustrated History of Hairstyles 1830-1930*. A Schiffer book. Schiffer Pub., 2003. 1

[10] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, pages 3474–3481. IEEE, 2012. 2

[11] A. Faktor and M. Irani. clustering by compositionunsupervised discovery of image categories. *tPAMI*, 36(6):1092–1106, 2014. 2

[12] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008. 3

[13] P. Fischer, A. Dosovitskiy, and T. Brox. Descriptor matching with convolutional neural networks: a comparison to sift. *arXiv preprint arXiv:1405.5769*, 2014. 2

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587. IEEE, 2014. 2

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 3

[16] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. R-cnns for pose estimation and action detection. *arXiv preprint arXiv:1406.5212*, 2014. 2, 3

[17] K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *CVPR*, 2006. 2

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 2

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 3

[20] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *BMVC*, 2014. 2

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012. 2, 3

[22] Y. J. Lee, A. A. Efros, and M. Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *ICCV*, 2013. 1, 2, 3, 4, 6, 7, 8

[23] J. L. Long, N. Zhang, and T. Darrell. Do convnets learn correspondence? In *NIPS*, 2014. 2

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2

[25] P. Messier. Notes on dating photographic paper. 2005. 1

[26] F. Palermo, J. Hays, and A. A. Efros. Dating historical color images. In *ECCV*, pages 499–512, 2012. 1, 3, 4

[27] N. Payet and S. Todorovic. From a set of shapes to object discovery. In *ECCV*. 2010. 2

[28] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012. 2

[29] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*. 2012. 2

[30] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*, 2013. 2

[31] J. Severa. *Dressed for the Photographer: Ordinary Americans and Fashion, 1840-1900*. Kent State University Press, 1995. 1

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 2, 3

[33] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005. 2

[34] N. Storey. *Military Photographs & how to Date Them: Neil Storey*. Countryside Books, 2009. 1

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[36] M. Taylor. *Uncovering Your Ancestry Through Family Photographs*. PBS Ancestor. Betterway Books, 2000. 1

[37] M. Taylor. *Hairstyles: 1840-1900*. Picture Perfect Press, 2014. 1

[38] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg. Runway to realway: Visual analysis of fashion. In *WACV*, 2015. 7

[39] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*. 2014. 2

[40] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based R-CNNs for fine-grained category detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 2, 3

[41] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 2, 5, 6

[42] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *ICCVW*, 2013. 3