

# From Culture to Clothing: Discovering the World Events Behind A Century of Fashion Images (Supplementary File)

Wei-Lin Hsiao  
UT-Austin

kimhsiao@cs.utexas.edu

Kristen Grauman  
UT-Austin

grauman@cs.utexas.edu

This supplementary file contains:

- Training details for all methods in trend forecasting.
- Style trends for legs, neck, sleeves regions.
- Top Granger-causal relations from Vintage and GeoStyle.
- More examples of discovered cultural influences on clothing styles.
- Year distribution for Vintage dataset and New York Times dataset.
- Qualitative examples of timestamping through retrieval.
- Example failure cases in influence-based forecasting.
- Interface and protocol for user study on quality of visual clusters.

## I. Training details for methods in trend forecasting

In the trend forecasting experiment in Sec. 4.2 in the main paper, we compare our proposed influence-based method with 5 other baselines. Here, we describe training details for each of them below. Let  $\{x_{i,t}\}, t = 1, \dots, T_{train}$ , be the time series for style  $i$  in training set. All methods predict future trends using the training time series.

**Last-baseline.** This baseline uses the immediate previous value as the predicted value for the future trend:

$$\hat{x}_{i,t+1} = x_{i,T_{train}}, t \geq T_{train} \quad (1)$$

**Linear-baseline.** This baseline fits the training time series with a linear function (slope  $m$ , y-intercept  $b$ ):

$$m = \frac{x_{i,T_{train}} - x_{i,1}}{T_{train} - 1}, b = x_{i,1}, \quad (2)$$

and predicts future values as:

$$\hat{x}_{i,t+1} = mt + b, t \geq T_{train} \quad (3)$$

**Mean-baseline.** This baseline aggregates the mean value from training time series, and predicts future values using the mean from training:

$$\hat{x}_{i,t+1} = \frac{\sum_{j=1}^{T_{train}} x_{i,j}}{T_{train} - 1}, t \geq T_{train} \quad (4)$$

**Exponential smoothing (EXP).** This baseline exponentially decreases weights for past observations, so latest time-points have higher weights than earlier time-points:

$$\hat{x}_{i,t+1} = \alpha x_{i,t} + (1 - \alpha) \hat{x}_{i,t}. \quad (5)$$

where  $\alpha \in [0, 1]$ . For our experiment,  $\hat{x}_{i,t}$  is set to  $x_{i,T_{train}}$ . We report the best numbers for all settings by using  $\alpha = 0.30$  for the neck, torso, legs regions,  $\alpha = 0.2$  for the sleeves region on the Vintage data, and  $\alpha = 0.7$  for the GeoStyle [3] data.

**Autoregression (AR).** Like the previous EXP baseline, this method also weights the past observations to predict future values. Instead of exponentially decreasing the weights through time, it learns weights by fitting the training series:

$$\operatorname{argmin}_{\alpha_{i,m}, \forall m} \|\hat{x}_{i,t+1} - x_{i,t+1}\|^2, \quad (6)$$

where

$$\hat{x}_{i,t+1} = \sum_{m=0}^{q_1-1} \alpha_{i,m} x_{i,t-m}, t = 1, \dots, T_{train} - 1, \quad (7)$$

|          |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Vintage  | 96.6 | 36.2 | 31.7 | 24.0 | 20.6 | 17.3 | 16.8 | 14.3 | 13.9 | 13.4 | 12.3 | 11.9 | 11.6 | 11.2 | 10.2 | 12.7 | 9.9  | 9.8  | 9.4  | 9.0  |
| GeoStyle | 16.6 | 15.8 | 15.6 | 14.1 | 14.1 | 13.9 | 13.9 | 13.6 | 13.6 | 13.5 | 13.3 | 13.3 | 13.3 | 13.3 | 13.0 | 12.8 | 12.7 | 12.7 | 12.7 | 12.6 |

Table 1: Top 20 F-values for Granger-causal relations on Vintage (F-critical-value= 3.98) and GeoStyle (F-critical-value= 2.48). Since the F-values are greater than the F-critical value, they are statistically significant.

| Neck         |           | Torso      |             | Arms              |            | Legs      |             |
|--------------|-----------|------------|-------------|-------------------|------------|-----------|-------------|
| Style        | Topic     | Style      | Topic       | Style             | Topic      | Style     | Topic       |
| cheetah      | wilson    | everyday   | today       | southwest-pattern | wilson     | skirt     | clinton     |
| polka-dot    | greenberg | cable-knit | earlier     | tassel            | greenberg  | lace      | southampton |
| dotted       | jess      | velvet     | advertise   | fringed           | jess       | galaxy    | streak      |
| perforated   | pageant   | slit       | settlement  | geo               | pageant    | sheath    | campus      |
| animal-print | embrace   | pintuck    | continue    | batwing           | embrace    | baroque   | tumultuous  |
| <hr/>        |           |            |             |                   |            |           |             |
| halen        | crew      | everyday   | data        | southwest-pattern | blizzard   | checkered | kennedy     |
| surfer       | nixon     | cable-knit | change      | tassel            | copenhagen | grid      | harold      |
| van          | junior    | velvet     | temperature | fringed           | danish     | plaid     | bullet      |
| love         | burger    | slit       | wave        | geo               | primarily  | gingham   | wound       |
| palm tree    | member    | pintuck    | phenomenon  | batwing           | weird      | swiss     | tragedy     |

Table 2: Top 2 Granger-causal relations in each body region in **Vintage**: each relation is a Granger-causality from a cultural factor (topic) to a clothing style. We show the top detected attributes/categories in each style, and the top words in each topic.

and  $q_1$  is the window size that AR weights past time points in.  $q_1$  is set to 2 and 4 on the Vintage and GeoStyle [3] datasets. A window size of 2 corresponds to 8 (10) years in the vintage data, and a window size of 4 corresponds to 1 month in the GeoStyle data. It is safe to assume that information contained in past observations earlier than this range may already be covered in this window. (Past observations older than 8 or 10 years may be irrelevant for predicting yearly trends; likewise, past observations older than a month may be irrelevant for weekly trends.) Each style  $i$  learns the regression weights  $\alpha_{i,m}$ s separately. At test time, AR makes its prediction by:

$$\hat{x}_{i,t+1} = \sum_{m=0}^{q_1-1} \alpha_{i,m} \hat{x}_{i,t-m}, t \geq T_{train} \quad (8)$$

**Cultural influence (ours).** Finally, our proposed cultural-influence-based model builds on AR by including the external time series  $y_{l,t}$  from mined textual topics  $l \in C_i$  (described in Sec.3.6. in the main paper), where  $C_i$  is the set of influential topics for style  $i$ . This set of topics is obtained by performing Granger-causality tests [2] on all style-topic pairs in the dataset.

Like AR, this model also learns weights that best fit the training series:

$$\underset{\alpha_{i,m}, \beta_{i,m,l}, \forall m, \forall l}{\operatorname{argmin}} \quad \|\hat{x}_{i,t+1} - x_{i,t+1}\|^2, t = 1, \dots, T_{train} - 1 \quad (9)$$

where

$$\hat{x}_{i,t+1} = \sum_{m=1}^{q_1} \alpha_{i,m,l} x_{i,t-m} + \sum_{m=0}^{q_2-1} \beta_{i,m,l} y_{l,t-m}, \quad (10)$$

and  $q_2$  is the window size for the external time series  $\{y_{l,t}\}$ .  $q_2$  is set to 2 and 26 on the Vintage and GeoStyle [3]

datasets, respectively. A window size of 26 corresponds to 6 months in the GeoStyle data, and this window size for external time series allows transferring seasonal characteristics with arbitrary lags on external series  $\{y_{l,t}\}$  to target series  $\{x_{i,t}\}$ . Each style  $i$  that is paired with one of its Granger-causal topics  $l$  learns the regression weights  $\alpha_{i,m,l}$ s and  $\beta_{i,m,l}$ s separately. At test time, where  $t \geq T_{train}$ , this model makes prediction by ensembling predictions from all models  $l \in C_i$ :

$$\hat{x}_{i,t+1} = \frac{1}{|C_i|} \sum_{l \in C_i} \left( \sum_{m=1}^{q_1} \alpha_{i,m,l} x_{i,t-m} + \sum_{m=0}^{q_2-1} \beta_{i,m,l} y_{l,t-m} \right). \quad (11)$$

## II. Style trends for legs, neck, and sleeves regions

In Sec.3.3. in the main paper, we describe our approach for discovering clothing styles in a century of fashion images, and obtaining the style trends by computing their popularity trajectories. The timeline of the top styles in the torso region is shown in Fig.4 in the main paper. Here, we show the timelines for other body regions: legs region in Fig. 3, neck region in Fig. 4, and sleeves region in Fig. 5. Each color represents a style, while the area a style occupies shows the frequency of that style at a time delta. A general trend that seems to hold for all regions is that styles in later times expose more skin regions, like illusion necklines, off-shoulder cuts, strappy design, short skirts/pants, *etc.* This is likely due to more liberal and open-minded views on clothing.

| Style | Chicago-cyan                                   | Sofia-cyan                                     | Guangzhou-cyan   | Madrid-dress                                 | Toronto-cyan                              | Tokyo-dress  | NYC-cyan                                    | Chicago-cyan  | Berlin-cyan                                      | Chicago-cyan                              |
|-------|--|--|--|--|---|--|---|---|--|---|
| Topic | race<br>season<br>finish<br>indiana<br>kennedy | reid<br>seth<br>geography<br>savage<br>webster | wan<br>unmistakable<br>miranda<br>rebellion<br>dermatology | driver<br>mile<br>car<br>formula<br>distance | budget<br>year<br>cut<br>spend<br>billion | stake<br>franklin<br>roosevelt<br>crown<br>triplet | office<br>police<br>shoot<br>fatal<br>bronx | donald<br>rumsfeld<br>ponder<br>blumenthal<br>claudia | store<br>retail<br>department<br>confirm<br>shop | budget<br>year<br>cut<br>spend<br>billion |

Table 3: Top 10 Granger-causal relations in **GeoStyle**: each relation is a Granger-causality from a cultural factor (topic) to a clothing style. Each style corresponds to a detected attribute in a city. For each topic, we show the top words in it.

### III. Full lists of top Granger-causal relations

To verify the statistical significance of our Granger-causal relations, we report the range of top F-values on both Vintage and GeoStyle data in Sec.4.2 in the main paper. In more details, F-tests are conducted by the following procedure [1]: the null hypothesis is rejected if the F(-value) calculated from the data is greater than the critical value of the F-distribution for some desired false-rejection probability. In our setting, we use false-rejection probability 0.05, which gives our model an F-critical-value 3.98 on Vintage and 2.48 on GeoStyle. The full list of top 20 F-values on both datasets is in Tab. 1. Since the F-values are greater than the F-critical value, they are statistically significant. Aside from the F-values, in Tab. 2 and Tab. 3 we also show the style-topic pairs of the top Granger-causal relations.

### IV. More examples of detected influences

Fig. 6 shows another six influences detected by our model, where Fig.(a-c) is on the Vintage data, and Fig.(d-f) is on the GeoStyle data. Fig. 6a is a topic about finance and annual reports. It constantly grew throughout the years, and influenced formal styles in skirts. Fig. 6b is a topic about conferences, and had peaks when influential summits took place. It affects stylish-business clothing, like buttoned, woven sleeves, or herringbone-patterned collars with lapels or pin tucks. Fig. 6c is a love/romantic-centric style. Interestingly, it may have influenced bridal styles like sparkling and glitter designs with full, elegant skirts.

For the GeoStyle data, Fig. 6d is a topic about fashion (e.g., Chanel, Givenchi) and design (e.g., Karl Lagerfeld). It may have influenced styles of high-end dresses in Paris, France. Fig. 6e is a topic about music, and potentially influenced the popularity of wearing scarves in Milan, Italy. Fig. 6f is a topic about sports games (e.g. soccer team Chelsea). A number of sports teams in Europe have blue in their team colors (e.g., Barcelona, Real Madrid, Chelsea, etc.), and this topic influenced wearing blue clothing in Madrid, Spain.

### V. Year distribution on newly collected datasets

As described in Sec.3.1. in the main paper, we collect new datasets for this study: image data from Flickr and

news articles from New York Times (NYT). The year distribution of the number of instances for Vintage data is in Fig. 1, and for NYT is in Fig. 2. For both datasets, later decades have more instances than earlier ones. The earlier decades (prior to 1980s) are used as training, while the denser later decades (after 1980s) are heldout for testing. Each testing sample for the image dataset still has hundreds of clothing instances per time point.

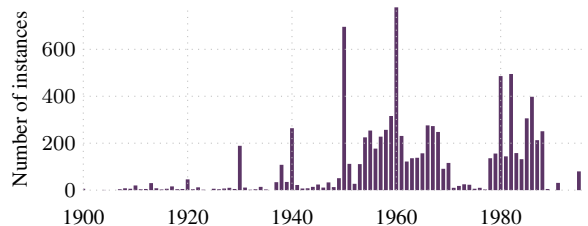


Figure 1: The distribution of clothing instances per year.

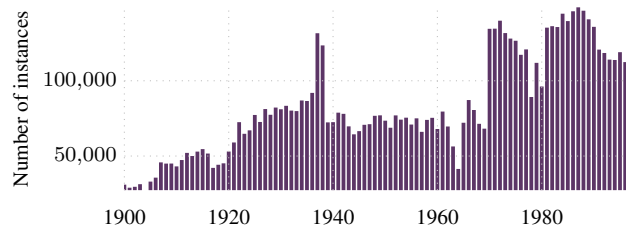


Figure 2: The distribution of news articles per year.

### VI. Qualitative examples of timestamping through retrieval

Quantitative results for timestamping a photo through retrieval is shown in Tab. 2 in the main paper. Here, we show qualitative examples comparing retrieval results using visual features only or also including inferred cultural features (approach described in Sec. 3.7 of the main paper) in Fig. 7. These are examples where visual features alone are not enough for accurately predicting query photos' date labels, and including cultural features helps. While all the retrieved photos look stylistically similar (color, pattern, fit, etc.) to the query, retrieved photos that include inferred cultural features are more temporally sensitive, with clothing styles unique to the query photo's time period (date label).

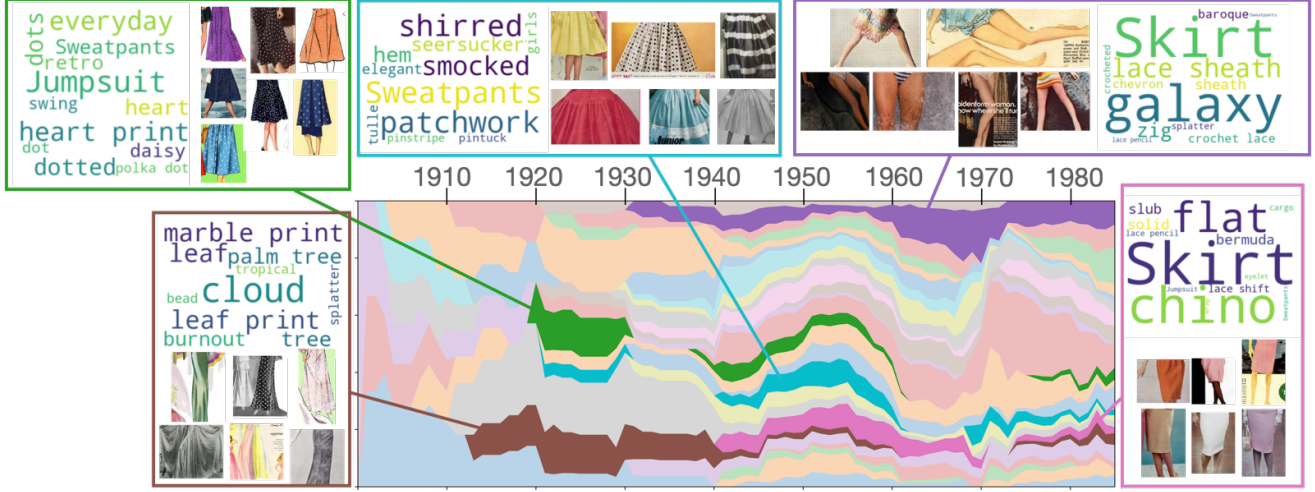


Figure 3: Timeline of the top styles in the **legs** region: Styles in later times generally accentuate more natural curves in the legs region, either exposing more skin areas (purple box) or with tighter cuts (pink box).

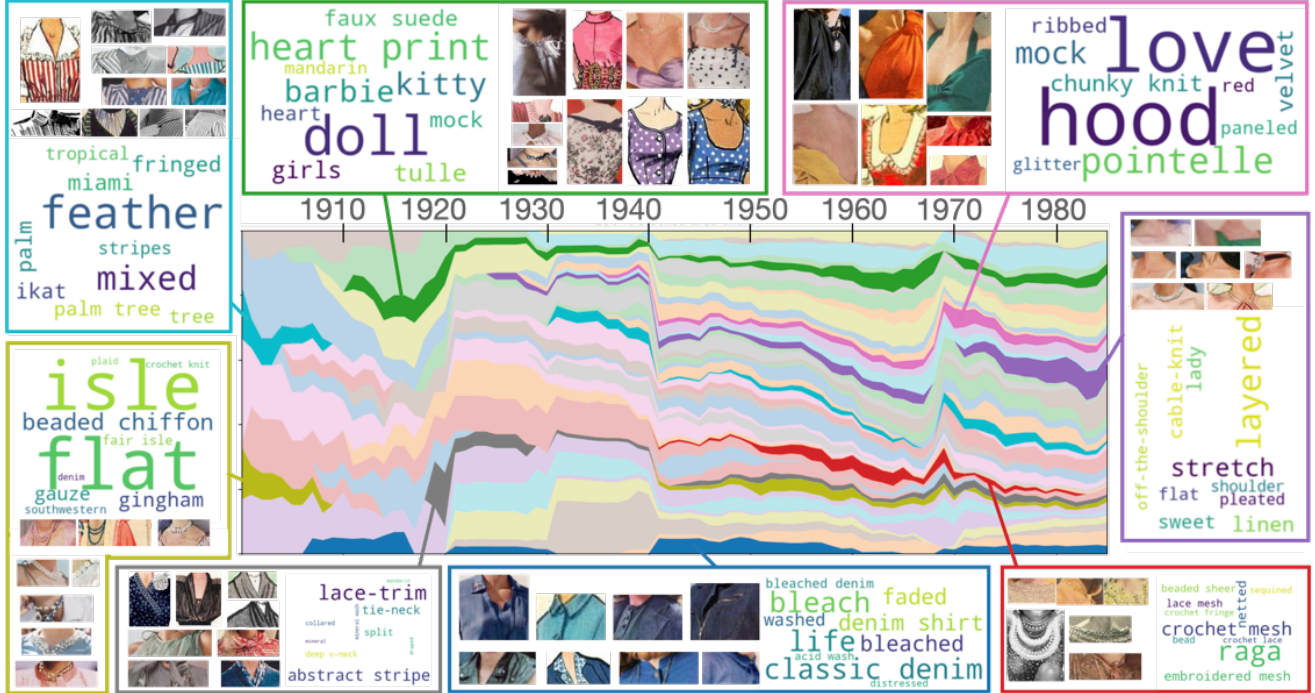


Figure 4: Timeline of the top styles in the **neck** region: Styles in earlier times generally have higher necklines (sky-blue box and olive-green box), while those in later years have deeper cuts (purple box and pink box).

## VII. Examples of failure cases in influence-based forecasting

Quantitative results for applying our detected influences on trend-forecasting are in Tab.1. in the main paper, and qualitative examples of how cultural influences help predict more accurate trends are in Fig.6. in the main paper. In summary, including cultural influences in autoregression

(AR) to predict future trends improves 57% of the styles on the Vintage data when comparing to AR, and improves 80% of the styles on the GeoStyle data. Fig. 8 shows qualitative examples when including cultural influences could not help AR (Fig. 8(a-b)), or perform even worse (Fig. 8(c-d)). Trend forecasting is an extremely challenging task, so cultural influences detected on the training series may no longer hold true on the test series, sometimes containing no



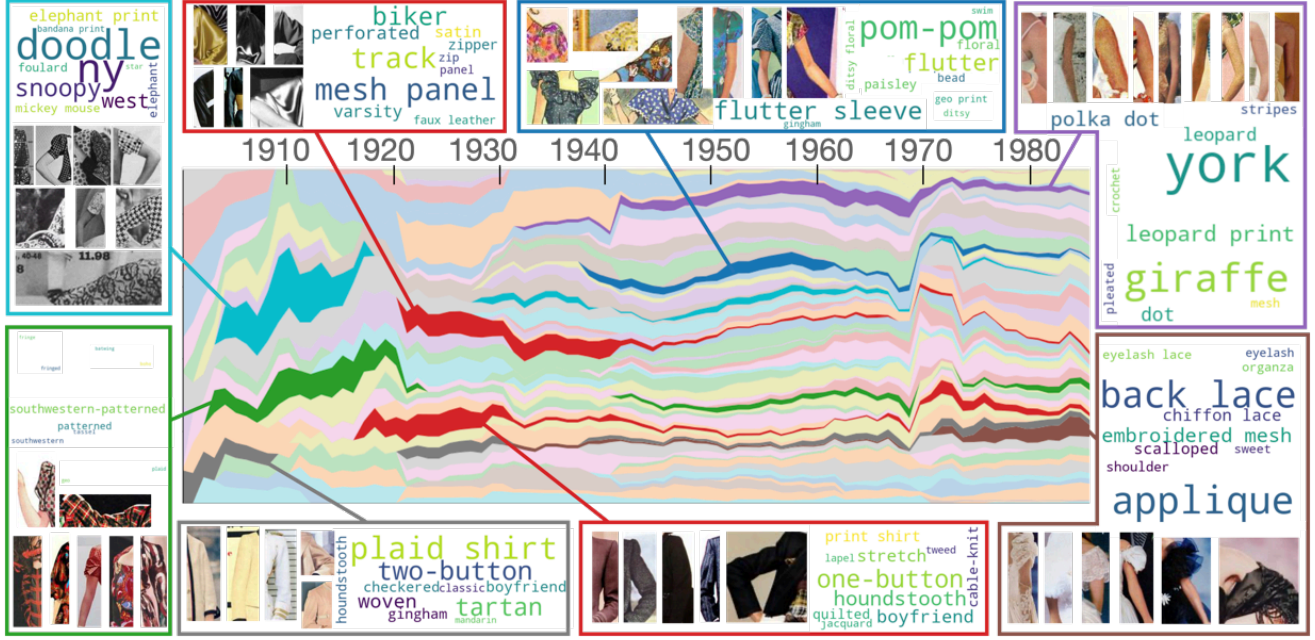


Figure 5: Timeline of the top styles in the **left arm** region: Styles in earlier times have more busy textures but conventional cuts (green box and sky-blue box), while those in later times have fancier cuts but solid patterns (purple box).

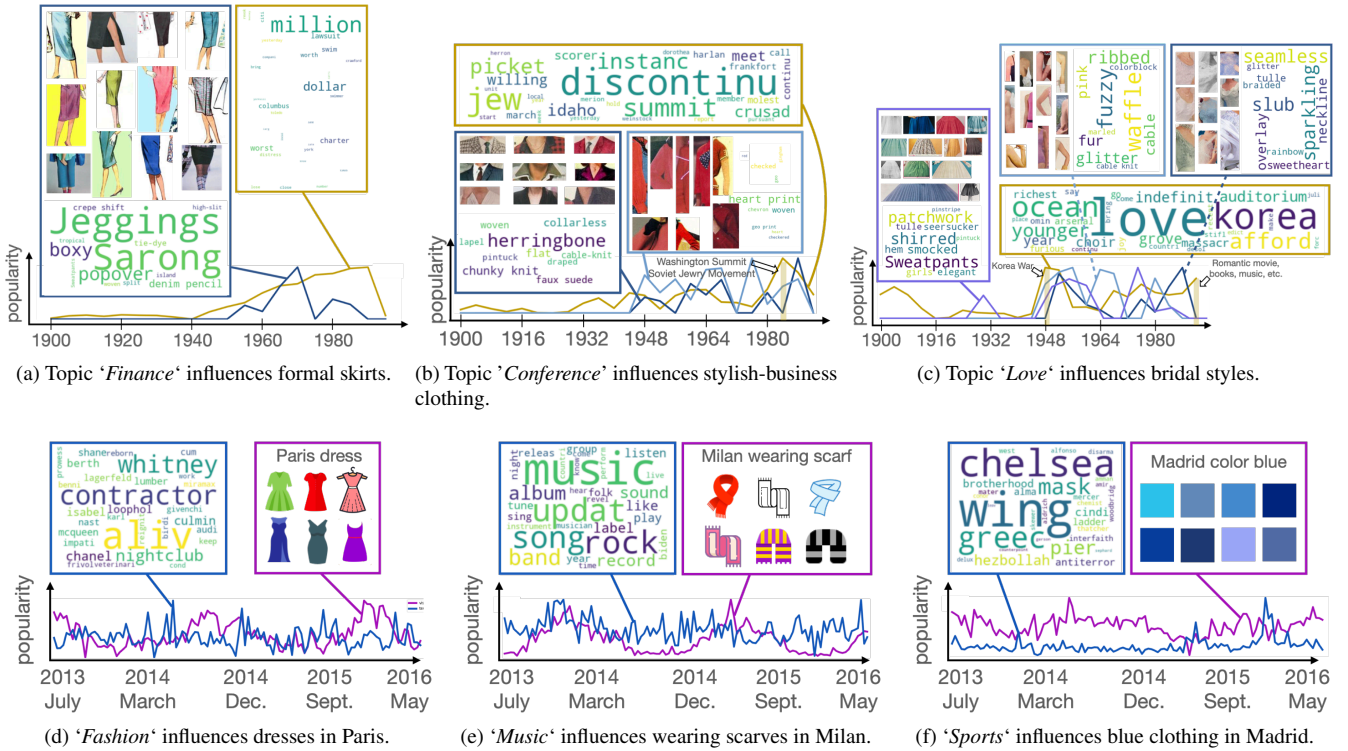


Figure 6: **More examples of discovered influences.** Fig. (a-c) are on the Vintage dataset, while Fig. (d-f) are on the GeoStyle dataset.

useful information, or even outliers that result in unstable predictions. In those cases, more naive baselines like mean (Fig. 8(b)) or linear (Fig. 8(a)) predictions perform better

than both vanilla AR and AR including cultural influences.

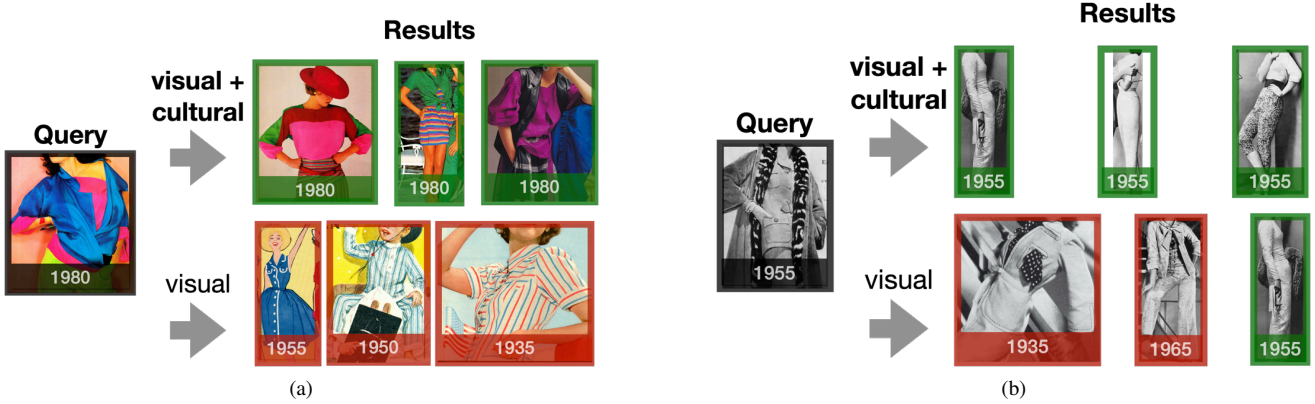


Figure 7: **Examples of timestamping through retrieval:** Query instance is on the left, and retrieved results are on the right. The top row in each example is retrieved using visual features augmented with our inferred cultural features, while the bottom row uses visual features only. True date labels are shown on the bottom of each photo. Temporally consistent photos retrieved are bounded by green boxes, while temporally inconsistent ones are bounded by red boxes. While all retrieved results are stylistically similar (color, pattern, fit, *etc.*) to the query, those retrieved by visual with cultural features capture more temporally-sensitive styles, *i.e.*, styles unique at that time.

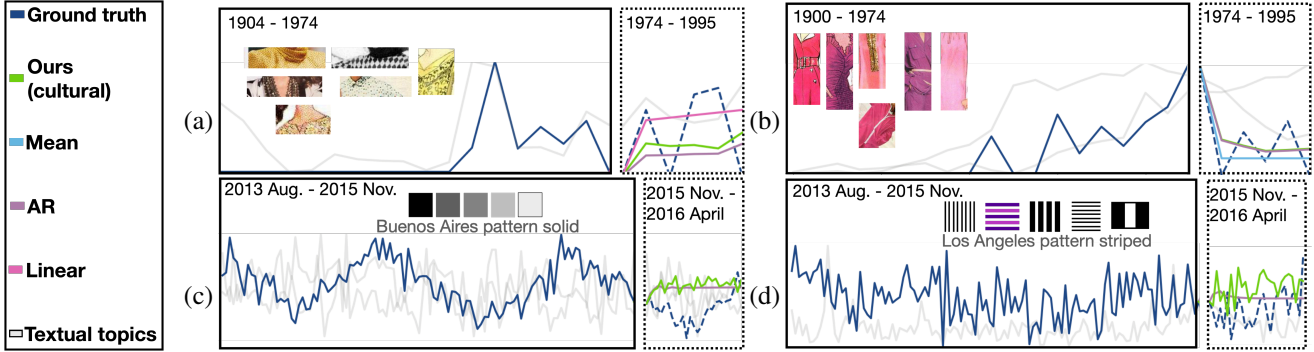


Figure 8: **Examples of failure cases** when considering cultural influences in trend forecasting. Top row on the Vintage data, and bottom row on the GeoStyle data. Future trends predicted by including cultural influences in AR are inaccurate, because the cultural time series did not offer useful information during test time. In some of these cases, more naive baselines like mean or linear predictions perform better than both vanilla AR or AR including cultural influences.

## VIII. User study for visual cluster’s quality

To verify whether the automatically discovered visual clusters correspond to meaningful clothing styles, we conduct a user study with the following protocol: for each visual cluster, we compare its most centroid 20 images with 20 randomly sampled images (as shown in Fig. 10), and ask human subjects to select the option that is more coherent in terms of *clothing styles*. Our instruction clarifies which factors are relevant or irrelevant to clothing styles (as shown in Fig. 9).

The user study is conducted on the Amazon Mechanical Turk platform, and each pair of comparison is answered by 5 to 7 Turkers, in total 156 unique Turkers.

75% of the time, human judges find the clusters to ex-

hibit coherent clothing styles that they can describe (as reported in the main paper), and example descriptions they gave for what they see in the selected images are: ‘*Most of the pictures include sleeveless tops for women. Most of the models in the picture seem to wear a necklace.*’, ‘*Group A is more coherent because many patches in it have formal suits.*’, ‘*A group is more coherent because of the pink and pastel color prevalence.*’, ‘*I chose B because some patches show jackets or dresses with long sleeves.*’, ‘*Image patches are very coherent in terms of showing sleeves of garments, mostly short sleeves. Most patches of clothing appear to be cut from a fine fabric such as linen or silk and expensive-looking.*’

The human judges not only find that most of our automatically discovered clusters are coherent, but explanations

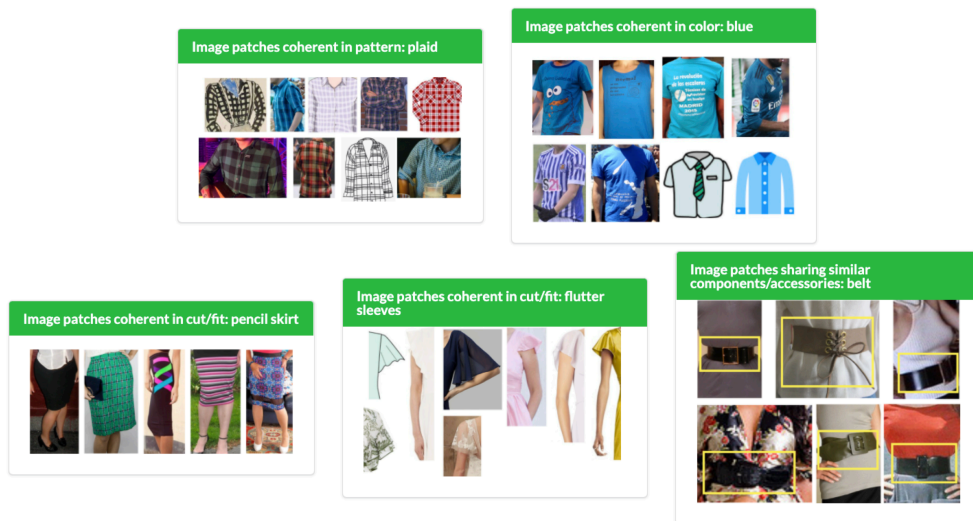
for their decisions also show that they were not based on differences in photography techniques. The fact that human judges can see and describe the coherence of the styles discovered by our method is evidence that we do find meaningful styles despite the very wide span of time (100 years) in the Vintage photos.

## References

- [1] F-test Interpretation. <https://en.wikipedia.org/wiki/F-test>, 2021. [Online; accessed 4-March-2021]. 3
- [2] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 1969. 2
- [3] Utkarsh Mall, Kevin Matzen, Bharath Hariharan, Noah Snavely, and Kavita Bala. GeoStyle: Discovering fashion trends and events. In *ICCV*, 2019. 1, 2

## Instructions

Task: We will show 2 groups of image patches. Tell us whether group A or group B contains image patches that are more coherent, in terms of clothing styles. Examples of images with **coherent styles**:



Please try to ignore irrelevant factors:

- Ignore photography style (colored vs gray-scale).
- Ignore image artifacts.
- Ignore body pose of the human wearer.
- Ignore whether an image is a sketch or a real photograph.

Examples of images that are **NOT** coherent in terms of clothing styles:

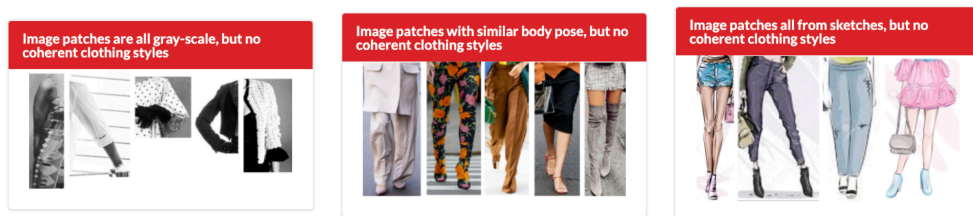


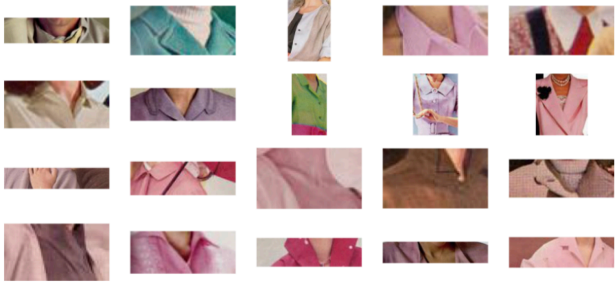
Figure 9: Instruction page for user study on whether a visual cluster consists of coherent clothing styles.



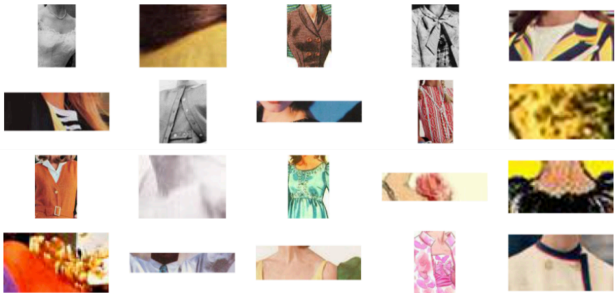
## Task

Please select the group that contains more **coherent** patches, in terms of **clothing styles**.

☐ A



☐ B



Your confidence in the above answer:

☐ very confident  
(clearly more  
coherent)

☐ somewhat  
confident  
(slightly more  
coherent)

☐ not confident  
(similarly  
coherent/incoherent)

Explanation for your choice:

Figure 10: Task page for user study on whether a visual cluster consists of coherent clothing styles: one of the options is our algorithm's discovered cluster, the other is a random grouping of images.