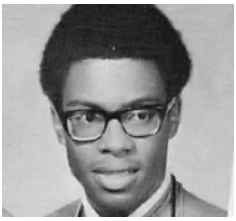
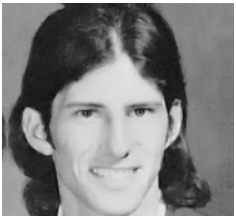
The Yearbook Dataset



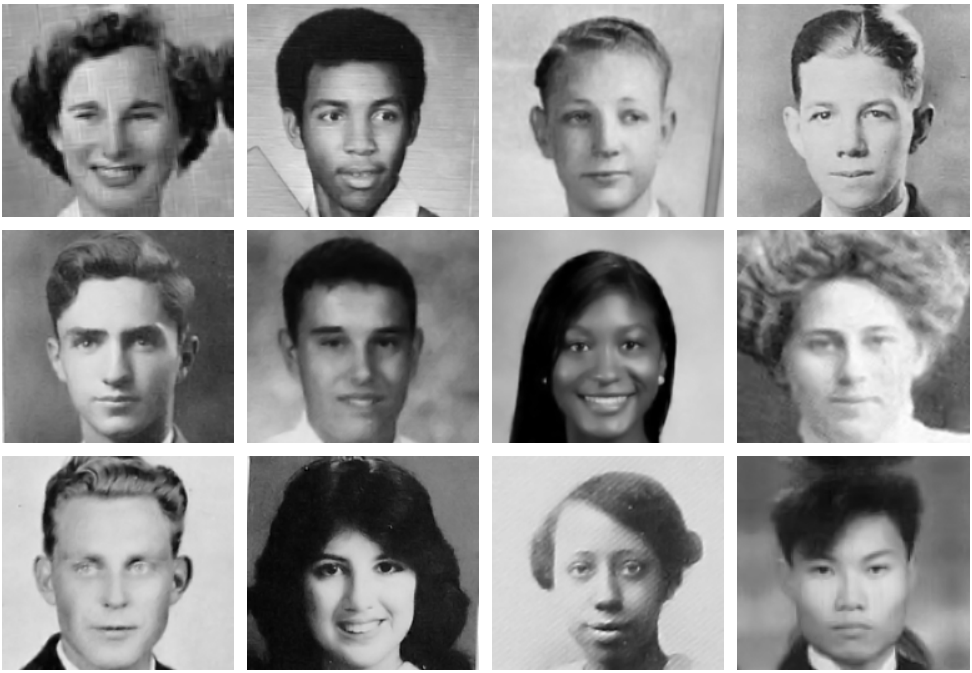


a publicly available dataset of 37,921 American high school yearbook portraits.

Who is in the data?

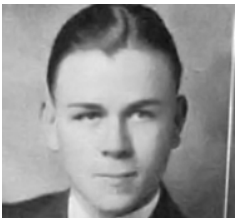
The dataset contains yearbook portraits of 37,921 American high school senior yearbook portraits spanning 1905 to 2013 across 115 schools in 26 states from across the United States. The yearbook scans were conducted by local libraries and uploaded to the internet.

46.4% of the photos come from the 100 largest cities according to US census 53.4% women and 46.6% men.



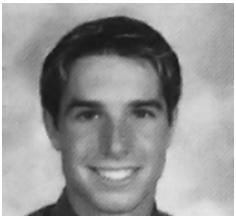
Who collected the data?

The dataset was gathered, cleaned, and published by a team of PhD candidates let by Shiry Ginosar at the Berkeley AI Research lab. A monetary grant from the NSF and a hardware grant from Nvidia supported the project.



How did they collect it?

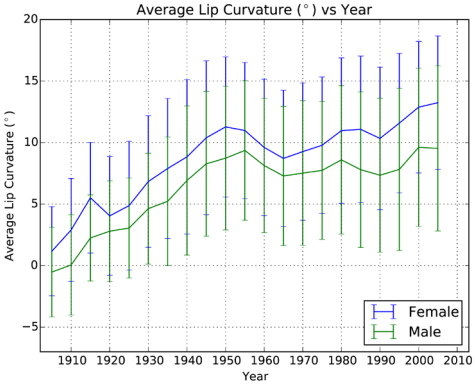
They downloaded publicly available yearbooks from various online resources such as the Internet Archive and numerous local library websites. They originally collected 168,055 photos ranging from 1905–2013 across 128 schools in 27 states. After removing non-frontal facing photos, they were left with 37,921 photographs from 115 high schools in 26 states. The digitization was originally done by local libraries with yearbooks in their collections.



Why did they collect it?

The data was collected for the progression of machine learning. The resulting output was a convolutional classification model to date the age of a photo and a method to visualize which pixels are important to the dating model.

The dataset itself is also intended to contribute to future historical analysis.

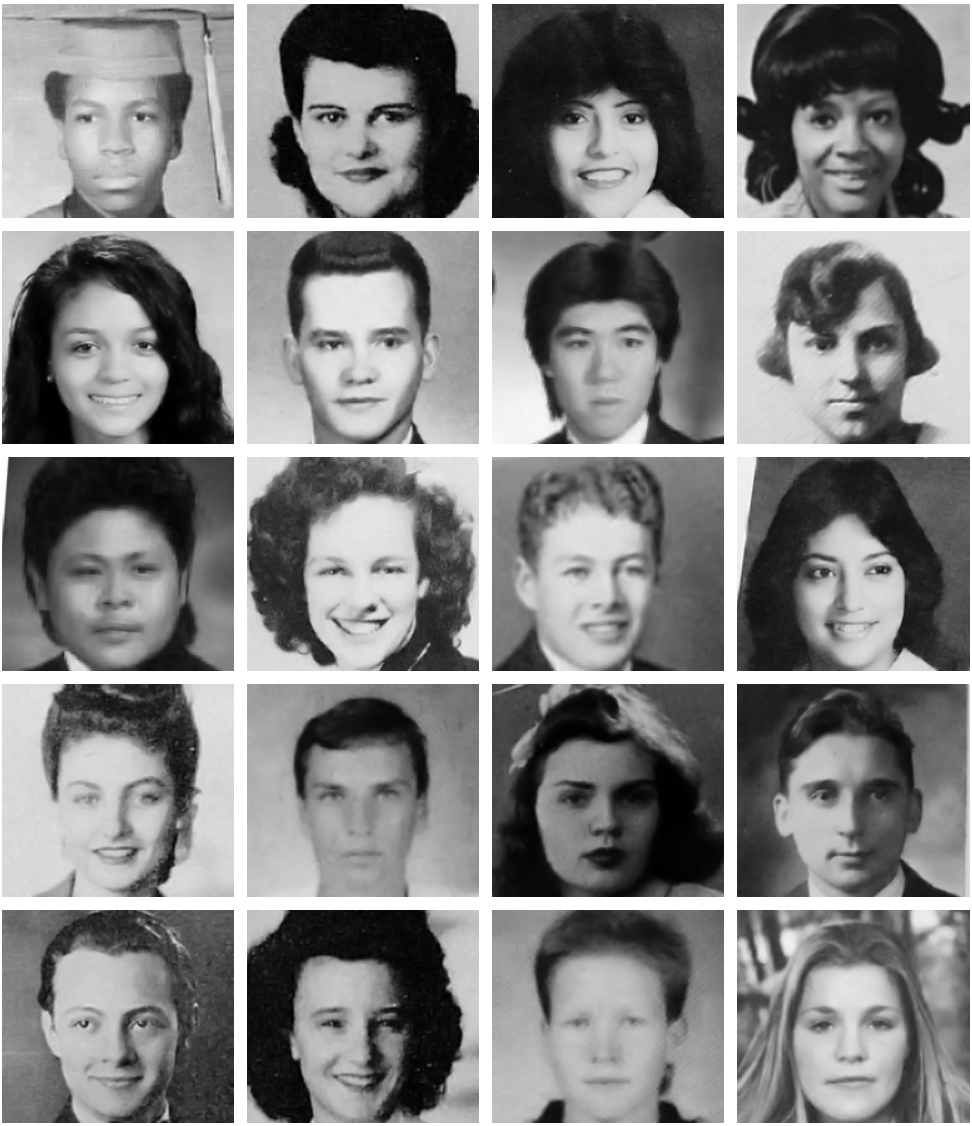


"Smiles increasing over time, but women always smile more than men: Male and female Average lip curvature by year with one STD error bars."

What is missing?

Images of people who did not go to high school or who did not continue to their senior year of high school. Less than 10 percent of American 18-year-olds graduated from high school in the 1900s, but this rose to more than 50 percent at the end of the 1960s. African Americans are under represented in the dataset due to discrimination and segregation in school systems.

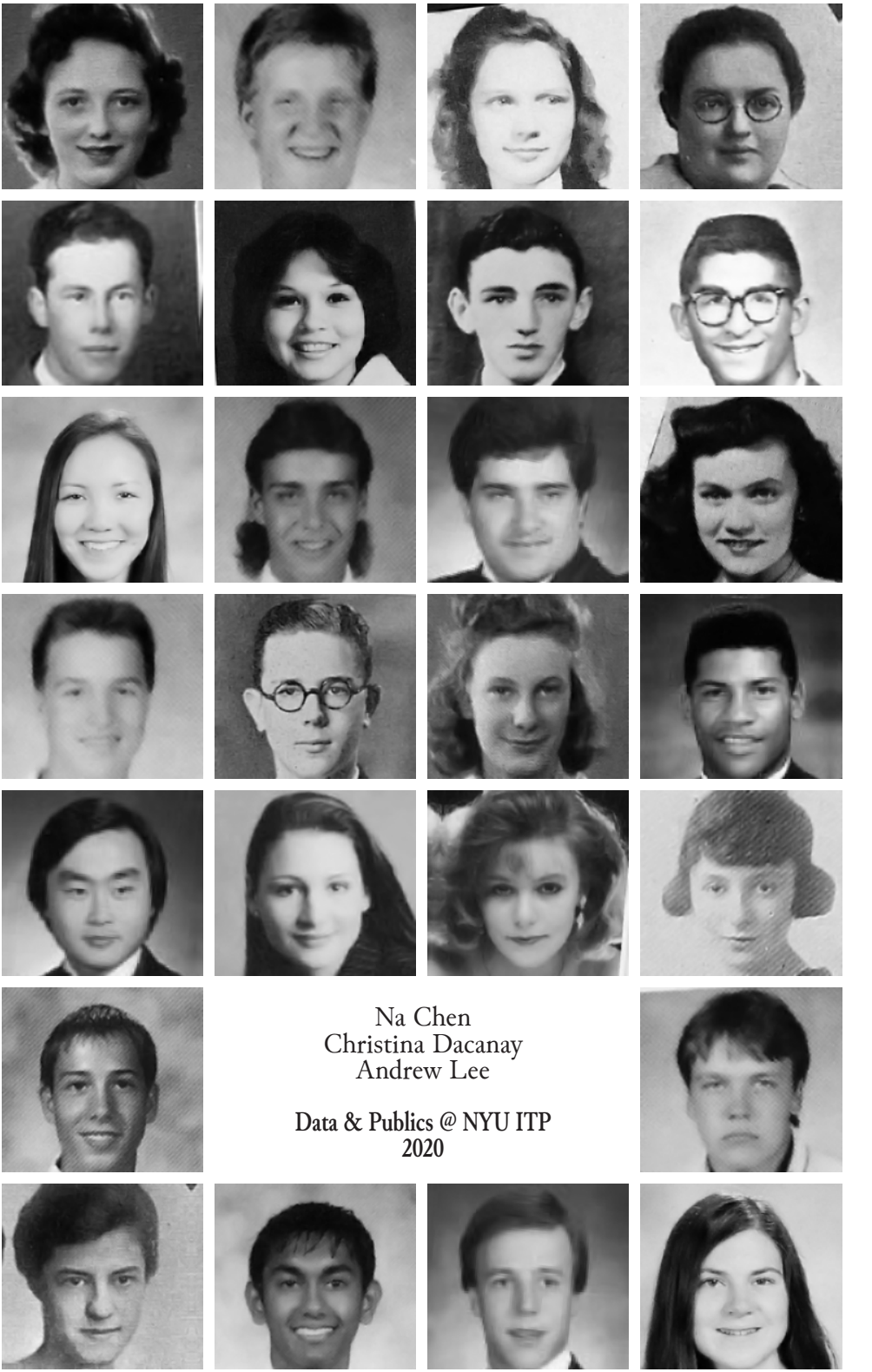
This dataset also excludes yearbooks that were not available online, and portraits of senior highschoolers from the rest 23 states.



Who has access?

A version of the dataset has been made publicly available to download from from Dropbox. A research paper was published in 2015[1]. More information about the project can be found on UC Berkeley's website[2].

[1]A Century of Portraits: *A Visual Historical Record of American High School Yearbooks* by Shiry Ginosar, Kate Rakelly, Sarah M. Sachs, Brian Yin, Crystal Lee, Philipp Krahenbuhl, and Alexei A. Efros
[2]http://people.eecs.berkeley.edu/~shiry/projects/yearbooks/yearbooks.html



Na Chen
Christina Dacanay
Andrew Lee

Data & Publics @ NYU ITP
2020