Just Look: Knowing Peers with Image Representation*

Tomasz Kaczmarek† Kuntara Pukthuanthong‡

May 10, 2023

Abstract

We present a novel approach to assess firm similarity by analyzing two million images. We leverage machine learning techniques to identify graphical objects that best represent companies' operations, forming Image Industry Classifications (IIC). IIC mirrors investor-defined peer groups and, akin to the brain's visual processing superiority, outperforms SIC, GICS, NAICS, and text-based similarity in delivering the greatest diversification benefits and industry momentum profits. This success is attributed to high investor agreement within IIC, leading to substantial aggregated demand and supply effects on stock prices. IIC excels in industries characterized by expectation-driven value, growth, and intangibility.

Keywords: Images, Industry classifications, Diversification, Industry momentum

JEL Codes: G00, G11, G12

*Code and data are available upon request. We will post and update the data of industry classifications on our websites.

†Tomasz is from the Department of Investment and Financial Markets, Institute of Finance, Poznan University of Economics and Business, Poland. Email: tomasz.kaczmarek@phd.ue.poznan.pl. Polish National Agency partly funded Tomasz's research for Academic Exchange within Bekker Programme, grant number BPN/BEK/2021/1/00404/U/DRAFT/00001, and National Science Centre, Poland, grant number 2021/41/N/HS4/02344.

‡Kuntara is from Trulaske College of Business at the University of Missouri, Columbia, MO, USA. Email: pukthuanthongk@missouri.edu

1 Introduction

Capital market research frequently requires enterprises to be separated into more homogeneous groups, with industry classifications being the most popular means of attaining this objective. Academic researchers have employed industry groups to restrict the scope of their analysis, identify control firms, acquire performance benchmarks, and offer descriptive statistics on sample firms. Meanwhile, investors look for peers to diversify portfolios, create investment strategies, or prepare multiple valuations. A classification method for industries divides businesses into finer subgroups in all these applications. However, as the similarity is not universal, commonly available classifications are based on different criteria of similarity and create various groups of similar companies. They identify peers through comparable production processes, supply-based methods, or evaluation of financial statements. Alternative approaches also compare firms with their business descriptions in 10-K reports or their common analysts' coverage.[1]

However, none of the existing classifications search for firms' similarities by an image that is the most efficient and intuitive to the human mind. In an age with an overload of information, using the natural abilities of our mind increases our efficiency in processing information. First, human brains can process images faster than other objects. Potter, Wyble, Hagmann, and McCourt (2014) find that the human brain can process images that the eye sees for as little as 13 milliseconds. A total of 90% of the information transmitted to the brain is visual, and images are processed at 60,000 times the speed of text; this suggests that pictures capture more information than words, text, and numbers. This phenomenon is

[1]Standard Industry Classification (SIC) codes have been used since 1939. However, they are being replaced with North American Industry Classification System (NAICS) numbers. Simultaneously, the Global Industry Classifications Standard (GICS) system—established jointly by Standard & Poor's (S&P) and Morgan Stanley Capital International (MSCI)—is gaining widespread acceptance. This is especially true among financial practitioners and shown by Bhojraj, Lee, and Oler (2003) to outperform industries grouped by SIC and NAICS. As a result of these modifications, prominent data providers—such as S&P Compustat—now carry all three sets of industry codes. In addition, financial researchers have explored their answer to the industry classification problem (for instance, Fama and French's FF algorithm), text-based industry group from 10-K reports (Hoberg & Phillips, 2016; Lewellen, 2012), analyst-based industries (Kaustia & Rantala, 2021; Ramnath, 2002).

known as the picture superiority effect, which states that pictures and images are more likely to be remembered than words. Images are also superior to text because reading is relatively inefficient for people; our brain sees words as images we must first recognize. Second, a single image can convey a narrative requiring more words. Images should convey any idea more effectively than text or voice alone. Images catch a higher level of attention and engage viewers more than text and numbers.

Photos also contain more information and transmit it better than text and numbers. First and foremost, a screenshot of photos conveys the intricacies of how the products function. Second, they depict various emotions of the people and scene in a single snapshot which cannot be conveyed as effectively and efficiently through text or numbers. Obaid and Pukthuanthong (2022) find that news photos are more persuasive and have a higher substantial predictive power than text. Third, visuals give refined information that may be missed or to taxing for the human brain to process when seeing lengthy text or dull figures. Fourth, photos can freeze a moment in time. They can show a fleeting expression or gesture and capture the essence of a moment that may not be fully captured by other forms of data.

Within a snapshot, a photo captures multi-dimensions of information and is an efficient and effective form of communication. These qualities make photographs useful for clustering industries.

In this study, we propose a new measure of firm's similarity based on graphical objects that best characterize the nature of their business and define a new image-based industry classification (IIC).[2] Typically, the objects represent companies' products but can also visualize the final products in the supply chain, raw materials, or any other representation related to the business. We demonstrate IIC's strengths from the R^2 of financial ratios regressions and the benefits that its use provides for building investment portfolios. The IIC is less bureaucratic and more agile than existing industry classifications. Images are more captivating and require less processing than just text or numbers, as seen by the predominance

[2]In this study, we use photos, images, photographs, and pictures interchangeably. Our IIC classifies all of them.

of images in most children's books. IIC is also realistic, not restricting firms from belonging to one industry. IIC's strength is for firms with distinct products; thus, it is adaptable to innovations in the form of new product offerings or significant improvements.

IIC performs well because it functions like the human brain, which groups similar-appearing objects (Shen, Horikawa, Majima, & Kamitani, 2019). Both Branthwaite (2002) and Dewan (2015) demonstrate that visual communication significantly influences human decisions. As a result of the brain's outstanding power to comprehend images, when an investor learns new information about a firm and examines its competitors, they associate the image of the given goods and their rival's offer. Numerous investment trading tactics influence the short-term fluctuations of stock prices. We demonstrate that the categorization of stocks based on picture similarity mimics the natural brain response. Industry momentum tactics best illustrate the reactivity of humans to photographs. The IIC should be more critical in this era where image-based communication on television, the internet, and mobile devices is pervasive.

We extend Hoberg and Phillips's (2016) (henceforth, HP) pioneer work to identify competitors who offer comparable products. Instead of language, where certain words may have multiple meanings, we employ images whose interpretation is more straightforward (Branthwaite, 2002; Dewan, 2015). IIC is predicated on the notion that photographs capture non-linearity and complex relations and show their product offerings to consumers beyond what text and numbers can. Our methods average R^2 is at least comparable to other classification techniques—demonstrating its reasonably high quality. Nevertheless, the cross-correlation of stock returns across industries is modest—as shown by the highest Sharpe and Calmar ratios generated by the diversification strategies exploiting our IIC.

Ideally, practical industry classifications should meet the following criteria (Kaustia & Rantala, 2021). First, they should be characterized evenly by observable and unobservable traits. Second, businesses within a group would have genuine economic linkages instead of merely statistical similarities. Third, they should adjust quickly to both business recomposition and economic structure changes. Fourth, they should be straightforward to comprehend and implement. Fifth, the cost of implementation should be reasonable. Finally, firms should be allowed to fall into multiple industries. IIC meets these requirements. Images capture fundamentals (product resemblance) and unobservables (color, taste, and status—which are intangibles and sentiment). Product similarity necessitates economic language, not simply financial ratios similarity. As stated previously, the primary advantage of IIC is that it captures changes in product similarity because human brains process images faster than words and fundamentals; hence, IIC is more dynamic than the alternative measures. Investors can apply IIC to diversify their portfolios' risk or maximize the Sharpe ratio, and to gain from the risk-adjusted and unadjusted industry momentum. We provide the data on our website to lower implementation costs for other academics and practitioners, and will continue to update the list over time.

In terms of implementation, we collect photographs from Google Images. We have chosen Google Images as our significant data source because of its plethora of images, especially firm-level photos.Another common source researchers might use is news. Analogous to text, there is a limitation of photos for individual firms. Only superstar corporations are often cited in the press, whereas most are not. The majority are logos rather than meaningful images. While we can collect images from Form 10-K and shareholder filings, the number is limited. Generally, just one or two images accompany a full report. Worst yet, many businesses provide graphs and tables with no visuals.

Moreover, these images are skewed and only favorably portray firms. They are often unconnected to businesses and can resemble ads; hence, grouping sectors using these images from annual reports is incorrect. Our solution based on machine learning is an ideal complement to Google's vast image archive. The more a machine can access images, the greater its learning capacity.

We employ machine-learning algorithms to build image-based industries with more than two million photos. We use a multi-view clustering approach that, to the best of our knowledge, has not been applied to finance; state-of-the-art computer vision algorithms like Deep Convolutional Neural Network (CNN) based EfficientNet (Tan & Le, 2019); and several other machine learning solutions such as transfer learning, k-means clustering, or object recognition—among others.

First, most photos downloaded from Google are unrelated to enterprises' products offering. From more than two million downloaded photos, we remove nearly 80% of pictures before starting the industries' formation. To clean photos, we detect logotypes, faces, and photos that are contextually unrelated to our task. Second, building industries based on image similarity comes with a significant challenge. How do we identify photos that correctly represent the products the companies offer? There are many photos with random objects in a dozen photos depicting each company. For example, images of Coca-cola can present the company's famous trucks, suggesting it should be classified as transportation, or highly popular iPads may indicate that the electronics manufacturers industry should include most of the companies listed on the NYSE. Therefore, out-of-the-box and unsupervised algorithms would create industries concentrated over random objects and unintentionally classify many companies. At the same time, semi-supervised or supervised learning could only be used to assign companies to predefined industries based on other than image criteria. Our methodology solves these problems as we both create industries based on companies' product offerings represented with images and distill a correct product representation for each company. Finally, to ensure our study is free from any look-ahead bias, we verify the economic significance of industry image classification (IIC) and the performance of proposed applications in the out-of-the-sample setting. For instance, we predict IIC for 2018 to 2019 with photos from 2016 to 2017 and 2020 to 2021 with pictures from 2018 to 2019.

We validate our measure using R^2 of the firm's financial ratio regression on the same ratio of the firm's corresponding industry. Our benchmarks include SIC, NAICS, and GICS -based classification with granularity closest to IIC. From these popular classifications, we focus on the GICS-based because Bhojraj et al. (2003) show that it is the top performer among the industrial classifications compared to SICs and NAICs. Moreover we compare results with Hoberg and Phillips' (2016) transitive text-based classification which, like the IIC, is created based on quantitative methods and uses the companies' product offerings as the main measure of similarity. IIC exceeds both HP and GICs for market-to-book, price-to-book, forecasted stock returns, market leverage, PE, and Tobin Q ratios among the 10 market indicators. For all 16 accounting ratios, IIC outperforms HP in 9 ratios, ties with them on dividend payout, debt-to-equity, SG&A to employees, EPS growth, RNOA and ROE, and only underperforms for asset turnover. IIC excels GICs in 11 ratios, ties with them on assets, debt-to-assets, and RNOA, and only underperforms for SG&A to employees and asset turnover.

Notably, IIC performs well for the ratios that capture expectation and growth—as it is a winner for forecasted market returns and forecasted EPS. Amazingly, IIC has R^2 of SG&A growth and sales growth even 10 percent points higher than that of HPs and GICs; therefore, IIC also captures growth. IIC is ranked first for both types of ratios.

In addition to the conventional market and accounting indicators, we present ratios that capture intangibles; although they are challenging to value, they are the highest value-added asset for firms. Pukthuanthong (2006) shows that innovation and human capital are the primary value drivers of technology companies; however, they are not easily measured from widely accessible sources. Human capital, frequently expressed by salary and education, is only available for top management and the board of directors in any financial report. She exploits the wealth of information in the financial report using R&D growth to capture innovation growth, R&D per unit of sales to gauge the timing of businesses' innovation investments,[3] and SG&A per employee to capture the firm's human capital. Companies with more human capital should have a higher value. IIC surpasses HPs and GICS in all intangibles ratios except SG&A to employees.

[3] The earlier a firm invests in R&D in its existence, the greater its R&D-to-sales ratio. Her sample consists of biotech companies, where she discovers that the sooner they invest in R&D for their pipeline, the higher their market value.

Our R^2 results corroborate the premise that our image-based IIC captures both observables and unobservables, in contrast to the current industrial classification systems that appear to focus primarily on observables. Small stocks do not influence our results, as our performance remains robust after eliminating them in a separate test.

In response to industry agility, we evaluate industry dynamics as the frequency with which a particular business is reclassified among industries. Our measurement is the most dynamic compared to the other 18 industry classes. A total of 22% of IIC-classified businesses switch industries yearly. The regularity with which businesses release new items and enhancements corresponds to our industry's high level of dynamic activity. IIC is more than twice as dynamic as HP and NAICS and more than four times as dynamic as the other classifications. To illustrate this, we show two examples. SIC (NAICS) reclassified Chevron from oil and gas extraction (petroleum and coal products manufacturing) to automotive dealers and gasoline service stations (gasoline stations) in 2021. Our image classification does the same reclassification but more than two years earlier, based on photos from 2018 to 2019. ACCO Brands Corporation was reclassified by SIC from fabricated metal products, with machinery and transportation equipment exceptions, to wholesale trade-non-durable goods in 2021; meanwhile, photo representation demonstrated a similar industry change from 2018 to 2019. The dexterity of industry classification is crucial for real-time investment.

We next present the two applications that utilize industry groups: diversification and momentum. For diversification purposes, our industry 25-unique has the highest annual returns, Sharpe ratio, and Calmar ratio for equal- and value-weighted portfolios, maximum Sharpe ratio optimized portfolios, and CVaR ratio optimized portfolios. The closest similar rivals are the six-digit GICs, which have the lowest yearly standard deviation, maximum drawdown, and one-year loss. The outcome validates that the industries identified by IIC are distinctive and potentially provide diversification advantages.

Turning to investment strategies based on industry classifications, we adhere to the techniques proposed by Moskowitz, Grinblatt's (1999) industry momentum and Barroso, Santa-

Clara's (2015) volatility-adjusted momentum, and short-term reversal. We demonstrate that our 25 unique classifications delivers the highest Sharpe ratios for equally weighted industries and second the best after GICS for value weighted. In addition, strategies built on IIC demonstrate the highest robustness to changes in strategy parameters involving modifications to the holding period, how companies are weighted in industries, or the use of momentum or short-term reversal strategies. Finally, we also generate pseudo-random portfolios and show that the industry momentum effect associated with IIC is directly due to industry momentum rather than the momentum of individual companies.

Finally, we show the mechanism behind our image categorization. We claim that a promising industry categorization should observe high investors' agreement within an industry categorization. The more the agreement of investors within an industry, the more significant the influence of aggregated demand and supply on stock prices, and the more advantageous it is to use that industry categorization in investing applications. We demonstrate that the IIC provides reasonable agreement within an industry. This characterization describes the benefit of adopting IIC in diversification and momentum strategies.

This project aims to develop a new categorization metric for industries. We do not assert that we are the best in this field. Our IIC thrives in the industries that manufacture products or services, such as oil drilling, waste management, etc. Our IIC does not perform well with a business that images, such as financial services, consulting firms, abstract technologies, and multi-product conglomerates, cannot represent.

Our contributions are fourfold. First, we contribute to the industry classification literature. We explain in detail in the next section the distinction across methods. Second, we join a group of recent papers that utilize machine learning and images for investment applications. Obaid and Pukthuanthong, 2022 extract sentiment from news images and show it surpasses text in predicting stock returns. Jiang, Kelly, and Xiu, 2020 employ graphs to predict returns from trend strategies. They find that their strategies outperform trading technical trend profit. Third, we participate in neuroscience. We show that the human brain can cluster photos better than text and numbers. We also contribute the behavioral economics in that photos might cause overreaction, which has been a critical explanation of the well-known momentum trading strategies (see Daniel, Hirshleifer, and Subrahmanyam, 1998).

2 Comparative analysis of industry classification metrics

Financial economics research has long addressed the categorization of industries. Kahle and Walkling (1996) compare the informativeness of Standard Industrial Classification (SIC) codes obtained from the Center for Research in Security Prices (CRSP) and Compustat databases, and Fama and French (1997) create new industry classifications based on grouping existing four-digit SIC codes. Krishnan and Press (2003) compare SIC codes to NAICS codes, and Bhojraj et al. (2003) also compare various fixed industry classifications. Although these studies are informative and suggest that existing static classifications can be used better, they do not explore whether the underlying core methodology can be improved.

Although it is simple to utilize current industry categories such as SIC or NAICS for research purposes, these metrics have at least two limitations. First, neither substantially reclassifies enterprises over time as the product market develops. IIC, on the other hand, is more dynamic and purely based on the images of products. The different classifications take some time to clean, digest and analyze. Second, photos represent firms' products strictly and thus are more agile when there is an advancement in product and service offerings. Hundreds of new technology and web-based companies were classified as "business services" by the SIC around the end of the 1990s.

Both text- and analyst-based groups, described below, have an advantage over SIC, NAICS, and GICS as they are company-specific and subject to yearly revision. On the other hand, SIC, NAICS, and GICS are transitive—making them appropriate for many applications .[4]

[4] Transitivity implies that for any firm, A and B, in the same industry, a firm C in A's industry is also

The text-based groups pioneered Lewellen (2012) and Rauh and Sufi (2012), it was formalized and extended by Hoberg and Phillips (2016) (HP). We focus on comparing our IIC with the text-based classifications by HP, as they are the only group that provides the data. HP determines the industries based on the company description section of 10K filings. They have established at least 1,000 characters for company descriptions, corresponding to around 100 words. Thus, the minimal number of words demonstrating commonalities between companies must include at least 100 items. Most of these terms are not applicable from the standpoint of industry categorization, lowering the number of phrases that convey essential information.

On the other hand, the typical length of a company description is more than 100 words. Therefore HP employs around 100 words to categorize stocks within the industry, and each variate is a dummy variable that provides only two possible states.[5]

We identify industries based on a collection of photographs. In contrast to HP, the average number of images per firm is far more than 100. Approximately 100 variables are utilized to categorize a stock according to its industry. Unlike HP, however, each of our variables is a three-dimensions matrix with dimensions 224-224-3; this provides incomparably more information for identifying a company's product offerings. This is an advantage of computer vision capability.[6]

in B's. HP provides two types of classifications where one is transitive, and the other is not. Classifications that meet the transitivity condition are universal and widely applicable. For example, when diversifying a portfolio across industries, the classification must be transitive, and industries must be unrelated.

[5]To be precise, when HP compares company A to B, each word is used as a measure of similarity – "1" means the same word as in Company B, and "0" means it is not the same as in B.

[6]For HP, the dimension for one word is one because only one variable can take two states: 0 or 1. In the case of the one photo, we have three dimensions. We work on the image size 224-224-3, where the two first dimensions represent the position of a pixel (row and column or height and weight); the third demonstrates the intensity of three colors-where for RGB, it is red, green, and blue.

Regarding possible states, one image 224-224-3 represents 224x224x3 combinations; this is the starting point where we configure the CNN. Then we have all the CNN features that reduce it to make it trainable. We do not think 224x224x3 can be regarded as one. One pixel does not mean too much, but indeed, an enormous number of combinations of objects can be represented in photos with dimensions 224-224-3—which we cannot quantify. It also depends on colors (Do two images of the same objects with different colors represent the same objects or different ones?) or object size in the photo. Therefore, we would hesitate to quantify it precisely but argue that the dimensions HP and is used as input are incomparable. For further explanation about pixels, see https://levelup.gitconnected.com/pixels-arrays-and-images-ef3f03638fe7

Turning to the analyst-based group by Ramnath (2002) and Kaustia and Rantala (2021), we focus on the latter. The industry-based group is not a focus of Ramnath (2002). Kaustia and Rantala (2021) apply Kaustia and Rantala (2015)'s methodology to identify firms with common analysts to construct peer groups. Their classification is non-transitive in that there can be non-mutual peer connections—where firm A is firm B's peer, but firm B is not firm A's. Their strategy provides the company's peers, not industries, and cannot utilize these applications as we do. Therefore, we do not include their peer groups in this study. Notably, their high granularity of industry groups likely accounts for their excellent R^2 scores.

3 Methodologies and Data

The objective of our study is to demonstrate that images capture the relatedness of firms based on their product offerings to customers using unsupervised image clustering. We use unsupervised machine learning to automatically seek similarities in product space—represented as related image objects—and create clusters sharing similar photos. Clusters divide the total universe of photos into distinct groups. We make a transitive industry classification by linking photos with stocks that they represent. Finally, to reach our study objective, we compare the economic relatedness of firms in each Image Industry to the competitive industry classification techniques based on SIC, NAICS, and GICS. To build our Image Industry classification, we take three central notions. First, there is a photo set

$$A_i = \{m \in Nand\, m \geq m\_min_i\}, \tag{1}$$

where N denotes any natural number consisting of M_i photos connected with the firm i representing the firm's i main business activity, where m_min_i states the minimum number of photos in the set A_i. In other words, we assume that each company has a set of photos that correctly identifies the main products it offers customers. Second, there are $Nind$ firms i whose photos represent similar objects. We can identify these objects and create a set of photos.

$$\mathrm{B}_{ind} = \sum\nolimits_{i=n\_min}^{N_{ind}} A_i, \tag{2}$$

where n_{min} denotes the minimum number of firms in industry ind. Note that sets of photos A_i are disjointed for unique classifications where each stock is allocated to a single industry. Equation 2 shows that each industry is represented with photos from the set B_{ind}. Third, we can define all photos $x_{k,i,m}$ used to create Image Industries as set

$$\mathrm{Q} = \sum\nolimits_{k=1}^{K} \sum\nolimits_{i=n\_\min}^{N} \sum\nolimits_{m=m\_\min_i}^{M} x_{k,i,m}, \tag{3}$$

and cluster them into K sets of photos, where each cluster represents photos of one industry.

The m_min_i depends on three parameters: 1) the number of clusters K; 2) the total number of photos per firm i; and 3) the confidence level of omitting the false discovery. We calculate m_min_i empirically with 10,000 trials. In each trial, we randomly classify all photos describing each firm i into K classes and observe the density of how many times elements fall into a single class. After all trials, we construct the cumulated probability and estimate m_min_i with a 95% confidence level. In other words, for each firm and clustering granularity, we define the minimum number of photos m_min_i that need to be clustered to one class to omit with a 95% confidence level the false discovery of classifying a stock to a false class. The definition of m_min_i is vital because our photo cleaning procedure, described in Section A.1, removes various images for various companies. As a result, some stocks can be represented with 100+ photos, while others by only several dozen. Therefore, defining how many images from stock A_i clustered into the industry B_{ind} is enough to say that stock A_i belongs to industry B_{ind}. If we set m_min_i equally for all companies, we could encounter two potential problems. First, setting m_min_i too small would classify stocks with many photos as false industries. They would be classified correctly to industries with high photo representation but might, by chance, be classified to industries with relatively low photo representation. Second, with too high m_min_i, we would not assign the companies the small number of photos represents to any industry. A firm might be represented with only several photos; however, we should classify it as a member when all of them are related to a single industry. Finally, our simulation results in m_min_i in the range of three to 10 photos per stock.

In the next part of this section, we start by discussing the data we use. Then, we discuss the technology used to construct Image Industries. Furthermore, we demonstrate how we create a time-varying classification and the out-of-sample setting. Finally, we explain the methodology to evaluate the economic value of Image Industries.

3.1 Data

We define the stock universe as all firms from NYSE, AMEX, and NASDAQ that we have obtained from CRSP. We take shares with codes 10 and 11 and exclude financial and service firms (SIC codes range from 6,000 to 8,900). We exclude financial and service firms because their photos do not represent products offered to customers. We collect all photos representing firms directly from the Google search engine via Python API. The set of images retrieved from Google for each query depends on the ranking Google defines. The algorithm ranks photos based on the reliability of their upload sources. For example, images displayed first by Google are assessed to originate from the most credible sources. Given that Google indexes thousands of images for each listed company, the first 100 photos each year come only from highly evaluated sources; the majority originate from newspapers, company websites, or Wikipedia.[7] To achieve time-varying classification, we download photos for each year separately. We use the following search phrase: '{Company Common Name} products after: {year}-01-01 before: {year}-12-31.' Refinitiv retrieves data for firm names with the field 'Company Common Name.' Google provides the history of indexed photos from 2008.

[7] The order of images displayed by Google depends on two ranks. The first is a universal PageRank that estimates the ranks of the most reliable content sources and displays them according to the ranks (Brin & Page, 1998). The second is VisualRank which is dedicated to image (Jing & Baluja, 2008). It uses a combination of information from PageRank extended with input from the image to create a ranking of images per query.

We retrieve photos for each year from 2009 to 2019.[8] In each query, we require 100 photos. Finally, we collect close to 2 million photos and group our data sample into four periods: 2009 to 2013, 2014 to 2015, 2016 to 2017, and 2018 to 2019. We create groups to achieve a more comprehensive photo representation for each period. The sample covers, on average, 2,250 stocks per year.

Photos downloaded from Google need a thorough cleaning. Pictures do not convey meaningful information about a firm's business activity (e.g., faces, logotypes, or landscapes). We perform a cleaning procedure that eliminates most photos but significantly improves photo quality. We implement two primary cleaning techniques: mechanical and contextual. In mechanical cleaning, we eliminate photos representing logotypes, people's faces, and graphs. In contrast, we aim to remove photos unrelated to any products in contextual cleaning. The timestamp on each image is a crucial component that allows us to cluster businesses and develop a dynamic measure of our categorization. Our photographs contain time stamps; however, the time stamps are based on the upload time—not the time the photos were taken. As the Data section explains, this is not a significant issue in our context. We demonstrate a detailed procedure of photo cleaning in Appendix A.1.

In addition, we build two collections of company-level financial ratios. First, we download WRDS Industry Financial Ratio set that includes over 70 financial ratios from seven categories: Capitalization, Efficiency, Financial Soundness/Solvency, Liquidity, Profitability, Valuation, and Others.[9]. WRDS financial ratio set is a collection of academic researchers' most commonly used financial ratios. We use it to input the multi-view clustering procedure described in Section 3.2.1. Second, we create a collection of 19 stock-level ratios that we use to evaluate the economic performance of our industries. We use the same ratios as Kaustia and Rantala (2021) and group them into ratios using market information and ratios based

[8]We do not need photos from 2020 to 21 because we examine IIC in the OOS setting. Photos retrieved from 2009 to 2013 are used to train the model, and photos from 2014 to 2019 to create industries. As our testing sample finishes with the end of 2021, we use photos from two years, from 2018 to 2019, to predict industries for 2020 to 21. Photos from 2020 to 2021 will be used to define industries for 2022 to 2023.

[9]The dataset, as well as its detailed description, can be downloaded directly from Wharton Research Data Services (WRDS) https://wrds-www.wharton.upenn.edu/pages/grid-items/financial-ratios-firm-level/

on accounting data only. Table ?? in the Internet Appendix demonstrates a detailed calculation methodology for each ratio. We download the SIC and NAICS codes from CRSP, GICS from Compustat, and industries based on text classification from Hoberg and Phillips (2016) to compare our Image Industries with other classification techniques. Classifications from CRSP are time-varying, making them more comparable to our technique.[10]

Our total data sample covers the years 2009 to 2021. We divide the 13 years of data into the five years training sample (2009 to 2013) and eight years of the out-of-sample testing sample (2014 to 2021). To achieve a more extensive representation of photos per period, we group years into two-year periods. Thus, we have four periods (2014 to 2015, 2016 to 2017, 2018 to 2019, and 2020 to 2021) in the testing sample. We use the period of 2014 to 2015 to create predictions for 2016 to 2017 and carry out updates for the following years with the rolling window procedure. Therefore, finally, we have data for the out-of-sample testing from 2016 to 2021.

We build image-based industries in two significant steps. First, we start with data from the training sample to define industries. The established industries are constant for the entire out-of-sample testing period.[11] Our procedure of industry definition is demonstrated in Section 3.2. Second, as companies dynamically change their product offerings depending on market conditions, we update the composition of each industry on a biennial basis—which we describe in Section 3.3.

3.2 Image industry - definition

Our procedure for defining image industries aims to find similar objects in images that depict products offered by companies. Standard machine learning algorithms in computer vision are designed to identify similar items. However, our task has two objectives. We both

[10] Compustat delivers static classifications. CRSP gives dynamic but only for SIC and NAICS. Therefore we have dynamic from CRSP (SIC, NAICS) and static from Compustat (GICS). See Table 6

[11] Given that access to the photos' history is limited and the testing period lasts six years, changes in the design of the industries would not significantly affect the study results. Nevertheless, our methodology allows us to update the composition of industries as the period of photo availability lengthens.

want to find familiar objects and identify which represent the products companies offer. For example, since laptops, desks, or suits are frequent objects in photos describing companies, a typical unsupervised machine learning algorithm would identify these objects and creates clusters; our task is different. We do not want an industry consisting of all companies using laptops. Computers can be the basis for building an industry, but only among the companies that produce them. That is why we need a dedicated methodology that allows parallel identification of similar objects, a typical machine learning job that isolates those that characterize the company's products.

To achieve this, we use two customized machine learning-based solutions. First, we define image industries with unsupervised multi-view clustering. We characterize each company's business activity with two views: an illustration of offered products and a comprehensive set of financial ratios. Combining these two sources of information provides a more explicit message, something akin to using sight and hearing simultaneously. Thus, the connection of financial indicators supports searching for similarities between graphic objects.

Furthermore, we must highlight that our image industries are designed with an unsupervised algorithm. Unlike semi-supervised or supervised algorithms, we do not impose any pre-configuration of how industries should look. Instead, we let photos supported with financial ratios build clusters that define industries.

Second, to identify photos of offered products among all the images related to the company, we build a mechanism of guided clustering. Below, we demonstrate how we define image industries on the training sample with unsupervised multi-view and guided clustering.

3.2.1 Multi-view clustering

Clustering involves identifying and assigning similar instances to clusters or groups of similar cases (Géron, 2019). Equal to the classification task, each instance is assigned to a group. However, clustering is an unsupervised machine learning technique where classification is a supervised task. Clustering is a popular technique used in various applications, e.g., customer segmentation, dimensionality reduction, or search engines. The most common clustering technique is K-means, whose main drawback is that it cannot separate clusters that are not linearly separable in input space.[12] The popular algorithm that tackles this problem is spectral clustering; this is accomplished by constructing a graph from the data points with edges between them representing similarities and solving a relaxation of the normalized min-cut problem on this graph (Ng, Jordan, & Weiss, 2001; Shi & Malik, 1997).[13]

Although spectral clustering is a well-suited algorithm for image clustering, the similar objects identified in photos might not be related to firms' business activity. This is because image clustering may recognize photos of different everyday objects—starting from company products to, e.g., standard desks used by a company employee. To resolve this dilemma, we use a multi-view clustering approach that creates clusters based on information from several sources. In our clustering task, we want to identify clusters sharing comparable photos of product offerings to customers. Therefore, we add a second view to the clustering algorithm based on each company's financial ratios. In particular, we use Co-regularized Multi-view Spectral Clustering—introduced by Kumar, Rai, and Daume (2011)—where a modified objective function implicitly combines graphs from multiple data views to achieve better clustering.[14] The multi-view approach combines information about everyday objects on photos with similar financial ratios. As a result, the link between images sharing objects related to product offerings is enhanced while the connection for economically unrelated things is weakened.

Figure 1 visualizes the architecture that we apply to define IIC. We need many layers because our architecture comprises images, not only numbers. The precision of object recognition based on multiple layers is incomparably higher than with small networks. Any transfer learning method will always be based on multilayer, pre-trained models.

The first view relates to the image and the second to a financial situation where we

[12]Stuart Lloyd proposed k-means at Bell Labs in 1957; however, it was published outside of the company in 1982. In 1965, Edward W. Forgy published a similar algorithm.

[13]The detailed description of spectral algorithms is available in von Luxburg (2007).

[14]To implement multi-view clustering, we use the mvlearn Python package of Perry et al. (2021).

use a comprehensive set of more than 70 WRDS Industry Financial Ratios to represent a diversified profile of a company's financials. In particular, we create a 70-dimensional vector for each company with financial ratios—where each dimension represents one ratio. All financial data is from the end of 2012, the end of our training sample. As for the image, to acquire a numerical representation of a single photo, we use the 4096d vector extracted by the second last layer of VGG19 Convolutional Neural Network (Simonyan & Zisserman, 2014) and reduce its dimension to 20 with Principle Component Analysis. To use a numerical representation of three photos, we concatenate them after PCA.[15] So, finally, the representation of an image view is a 60-dimensional vector. But finding a selection of several photos to use for multi-view clustering is not straightforward. After all cleaning procedures, we have 204k images, which gives around 100 photos per firm. To create Image Industries, we need each firm to have three similar photos that properly characterize its product offerings.[16] To identify three representative images, we introduce guided clustering.

3.2.2 Guided clustering

Typical clustering is a one-step procedure where all photos from the universe are divided into a predefined number of clusters. Although we have thoroughly cleaned the photo database, we do not need all photos to create Image Industries. Some photos may create well-defined industries representing company products' offerings, while others may not. To resolve this issue, we propose a guided clustering technique to link information from photos with an economic measure that can demonstrate if stocks characterized by photos from a cluster are economically related. We link photos with firms and evaluate each cluster's inter-industry correlations and its ability to explain cross-sectional variations in firm-level stock returns,

[15] To be specific, we join three vectors together. Each of the three vectors is 20-dimensional. After concatenation, we have one 60-dimensional vector.

[16] We define the required number of photos as three because—from one side—the more similar photos we have, the universal representation we get. However, on the other side, the input vector with a representation from photos must have the same length for each firm-level observation. Therefore, we must take the shortest representation from all possible observations to achieve an equal level. As the minimum number of photos per stock is three, the maximum equal number of photos we can collect is also three.

market-based financial ratios, and accountancy measures. In particular, we first estimate the correlation matrix for every cluster to evaluate the inter-industry correlation of all stocks in the cluster. Then, we adopt a methodology from Bhojraj et al. (2003) and evaluate clusters by regressing a set of financial ratios used by Kaustia and Rantala (2021) on clusters' industry means. We estimate the average adjusted R^2 cross-sectionally across firms within each cluster. R^2 is estimated from a cross-sectional regression below for each firm i:

$$\mathrm{v}ble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t, \tag{4}$$

where the dependent variable *vble* represents 10 ratios using market information: 1) MARKET to BOOK; 2) PRICE to BOOK; 3) MONTHLY RET; 4) FORECASTED MONTHLY RET; 5) MARKET LEVG; 6) EV to SALES; 7) PE; 8) BETA; 9) MARKET CAP; and 10) TOBIN'S Q for each firm i at month t, and 16 ratios using accountancy information: 1) TOTAL ASSETS; 2) NET SALES; 3) DIV PAYOUT; 4) PROFIT MARGIN; 5) DEBT to EQUITY; 6) SALES GROWTH; 7) R&D EXPENSE to SALES; 8) R&D GROWTH; 9) SG&A to # EMPLOYEES; 10) SG&A GROWTH; 11) FORECASTED EPS; 12) EPS GROWTH; 13) DEBT to ASSETS; 14) RNOA; 15) ROE; and 16) ASSET to SALES, for each firm i at quarter t, the independent variable $vble_{ind,t}$ is the average of this variable for all firms in cluster *ind* excluding firm i at month t.[17] Both inter-industry correlations and the average adjusted R^2 enable a selection of clusters with stronger economic relations between its stock members. Clustering many photos in one step has its drawbacks, as it is challenging to pre-define the required cluster number.

To omit this problem, we create a loop where in each repetition, we: 1) cluster all photos into an arbitrarily chosen 100 clusters; 2) link photos with firms they represent using m_min_i from Equation 3 and with the minimum number of firms n_{min} from Equation 2

[17] Full list of ratios along with the calculation details is presented in Table A1 in the Appendix. To eliminate the possible effect of outliers, we winsorize at 1% for all ratios except MONTHLY RET and the binary variable DIV PAYMENT. Also, we perform only one-side winsorization on SCALED R&D EXPENSE for the highest values because many firms report zero.

as 25;[18] 3) calculate inter-industry correlation and the average adjusted R^2 for stocks from each cluster; 4) create a cluster rank estimated as an average of normalized inter-industry correlations and the average adjusted R^2 for all clusters; 5) choose a cluster with the highest rank; 6) remove photos representing the best cluster; and 7) repeat the procedure. Figure 2 demonstrates that the inter-industry correlation and the average adjusted R2 of the cluster with the highest rank decrease with a rising repetition number. The procedure lasts 130 trials when there are not enough photos to create any cluster representing 25 firms.

Furthermore, repetitions beyond 100 deliver clusters with very low economic values. Finally, as for many repetitions, the economic measures are low, we decide to eliminate all clusters where inter-industry correlation is below 0.275 (marked as a horizontal line on Figure 2, which is at the average level of Fama French industries at the same period). This procedure defines 65 clusters that represent similar photos and economically related firms.

The guided clustering technique allows us to define missing input data to the Multi-view Spectral Clustering algorithm. So next, we can perform multi-view clustering with a 70-dimensional financial representation and 60-dimensional image representation for each stock. The algorithm allows the definition of a desired number of clusters. This research presents the results for two different granularities of 25- and 50-image industries.[19] We do not want to define more industries, as each should have enough photos to train the CNN model. The uneven distribution of companies and photos in each industry, as presented in Figure 3, suggests a higher granularity than 50. We would generate industries with too few images. Thus, the CNN model described in the next section would not have enough data to be trained.

[18]Image Industry definition is a preliminary step to train CNN network, as described in Section 3.3. We need a balanced number of photos per cluster for CNN training that should count to around 100. For the guided clustering procedure, we set the minimum number of stocks in one industry (n_{min}) to 25. From Equation 3, we have the minimum number of photos per stock m_min_i of three. We tested empirically that each cluster counts at least 100 photos with these parameters.

[19]The 70-dimensional vector is a set of financial ratios from WRDS as described in Section 3.2.1, where a 60-dimensional vector is a numerical representation of 3 photos representing each company.

3.3 Image industry - time-varying setup

The collection of photos describing a firm change each year. New products and the withdrawal of existing ones are reflected in the image set representing a company. This dynamic creates a time-varying setup where firms' significant economic activity may change. The fast development of computer vision enables us to create reliable models for a classification task to track product offering changes and dynamically assign companies to Image Industries that are the most closely related in photo representation. Below, we demonstrate a technique to classify firms into Image Industries and track the time dynamics.

Gradually, computers outperform the human brain in more tasks. In 1996, IBM's Deep Blue supercomputer beat the chess world champion Garry Kasparov. Later, in 2016, AlphaGo won against the legendary Go player Mr. Lee Sedol (Silver et al., 2016). Progress in computer vision has been recently achieved, where during the ImageNet Challenge 2015, the first CNN-based neural network achieved higher accuracy in object classification tasks than a human.[20] Superhuman performance stimulates a growing number of applications. For example, Yu, Wang, Majumdar, and Rajagopal (2018) use CNN photo models to identify solar panels installations, Scoles (2019) for finding the locations of slave camps, Jean et al. (2016) for poverty levels in underdeveloped countries, Jiang et al. (2020) for stock price trends, and Obaid and Pukthuanthong (2022) for financial market sentiment.

Our task requires a model that identifies similar product offerings on a large set of photos to classify firms into industries. We build a photo classification model based on EfficientNet (Tan & Le, 2019). EfficientNet achieves much better accuracy and efficiency than the previous winning ConvNets of the ImageNet contest. The central assumption of the algorithm is that balancing network depth, width, and resolution can lead to better performance.

We use TensorFlow, a popular machine-learning Python-based open-source library, to im-

[20] PReLU-Net, developed by He, Zhang, Ren, and Sun (2015), surpassed the best human-level accuracy by achieving 4.94% top-five error rate—in comparison to 5.1%—for classifying ImageNet data.

plement the CNN-based classification model. We start with the pre-trained EfficientNet-B0, provided in Tensorflow, and use transfer learning to fine-tune the model to our application. Training a CNN network from scratch is not efficient (Géron, 2019). EfficientNet is trained with more than one million images from the ImageNet database. The network classifies photos into 1,000 object categories—such as different types of machines, materials, equipment, etc. Next, we replace the top hidden and output layer of EfficientNet with a new hidden and output layer with the number of outputs equal to the number of Image Industries. Finally, we train the model with the training set defined in Section 3.2.

As we have a relatively large number of images to use in transfer learning, we use a learning rate of 0.0001.[21] To improve performance, we use several regularization techniques. First, we use dropout (Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012)—where in every training step, a part of the neurons is entirely ignored but may be active during the next step. Second, we implement performance scheduling by reducing a plateau and early stopping procedures. After the validation error reaches a minimum, the algorithm starts with a learning rate reduction and finishes the training.

We build the testing sample with photos separated into two-year periods from 2014 to 2021. This period covers three classification updates based on industry predictions prepared with photos from 2014 to 2015, 2016 to 2017, and 2018 to 2019. For each prediction period, we have around 70,000 photos. In our task, we do not need to classify each picture into one of the Image Industries. Instead, it is better to classify only photos with high classification probability. Therefore, in analogy to the procedure we use during contextual photo cleaning, we let CNN distinguish clear photos from unclear with a 95% probability and classify to Image Industries only photos representing objects within a class. Finally, we get classifications for around 16k photos per period and classify firms by linking images to companies with m_min_i—defined in Equation 3.

[21] The learning rate defines the step size at each iteration while moving towards a minimum of a loss function (Géron, 2019).

3.4 Non-unique and unique image industries

CNN classifies photos to the desired number of industries. However, the distribution of images representing one company in different industries can vary greatly. For instance, a firm manufacturing hydraulic components may have photographs published online showing cars being used to transport products to customers. However, photos like that appear incidentally and are classified randomly among image industries. Therefore, for a company to be considered to be operating in a particular industry, we require a certain number of images associated with it to be allocated to that industry. Equation 3 defines m_min_i as the required number of photos. However, for companies with a diversified product range, the m_min_i criterion may be met for more than one industry. Thus, the classification created along with the methodology from Section 3.3 is not unique—meaning that one company can be classified into many industries. In contrast, typical industry classifications such as SIC, NAICS, and GICS are unique. In addition, the industry classification applications we demonstrate in the following chapters require a unique classification—assigning each company to only one industry. Therefore, to create a scheme analogous to the standard unique approach and test portfolio applications, we create two unique classifications: transformations from the non-unique ones of 25 and 50 industries.

The picture will not assist in deciding which industry may better reflect the dominant product group in the company. We need a different modality to determine where the best fit lies. Therefore, to create a unique classification for each company that is classified into more than one industry, we examine its homogeneity with industry peers and select an industry where it is the highest. For this purpose, we again use the financial indicators proposed by Kaustia and Rantala (2021); for each, we estimate the regression described in Equation 4. For a company belonging to two industries, we get two average R^2s, and the highest score indicates the industry that we choose to build a unique classification. From an economic perspective, the image of the selected industry correctly defines the company's product offer; the financial indicators assign the company to the industry in which the peers have the most similar financial ratios.

3.5 Out-of-sample setting

We verify the economic significance of IIC and the performance of proposed applications in the out-of-the-sample setting. To build the out-of-sample (OOS) approach, all photos we use need a timestamp. We download it with a google search algorithm that we use. It downloads photos from defined uploading periods. For example, suppose we create a query to download photos of a company Noble Energy Inc from 2016 as 'Noble Energy Inc products after 2016-01-01 before 2016-12-31.' In that case, the downloaded images could not have been uploaded to the internet any later than in 2016. This protects us from any ahead-look bias. On the other hand, they might have been taken earlier than in 2016 and then uploaded to the internet in 201; this is not a problem because it is free from look-ahead bias.

In the first step, we use photos with timestamps from 2009 to 2013 to create clusters, define IIC, and train the CNN model that we described in Section 3.2. Next, along with the procedure from Section 3.3, we take photos from 2014 to 2015 and use the trained CNN model to predict IIC. Finally, we test the performance of industries predicted with these photos on financial data from 2016 to 2017. Thus, we do not use photos uploaded to the internet from 2016 to 2017 to test IIC performance for this period. This OOS procedure is repeated three times. We also predict IIC for the years 2018 to 2019 with photos from 2016 to 2017 and for years 2020 to 2021 with pictures from 2018 to 2019. This setting builds an OOS approach for the results we demonstrate in the following sections.

4 Results

This section describes the relatedness of firms clustered with photos that illustrate their product offerings. We start by reporting the characteristics of Image Industries formed with unsupervised clustering machine-learning techniques. Then, we examine IIC's usefulness and limitations in explaining cross-sectional variations in firm-level stock returns, market-based valuation multiplies, and financial ratios by comparing them to other industry classification methods; this includes SIC, NAICS, GICS, Fama French industries, as well as Hoberg and Phillips's (2016).

4.1 Firm relatedness visualized by image

We build the Image Industries with a large set of photos that, after intensive cleaning, reaches 219k for the testing sample. It gives around 4,000 (8,000) photos per industry in Image Industries 50 (25). Figure 3 shows the final average number of photos per industry in each prediction period. The typical industry is represented by several hundred images per period, where the number of images per class varies significantly. The detailed human-made evaluation of all photos per industry is certainly beyond the scope of this research. Nevertheless, we wish to indicate several observable characteristics even after looking at only dozens of images.

Figure 5 illustrates the random representations of three sectors from Image Industries 50 characterized by high inter-industry correlation. The convergence of objects describing randomly selected companies is apparent. For example, most photos representing Industry 32 show offshore oil wells. However, objects recognized within this Industry include sailing tankers and refineries—as well. We can see the same tendency in Industry 6 (engines, valves, or tubes) or Industry 3 (trucks, vans, or motorhomes). It demonstrates the first characteristic of Image Industries is selective similarity aggregation, where one class collects a set of similar objects but not identical ones. If neural networks could aggregate only identical objects, clusters would strongly shred and eliminate the products that do not have identical shapes.

The photo representation of industries with lower inter-industry correlations demonstrates a higher object diversity. Each of the three Industries signified in Figure 6 links varied objects, sometimes even not so intuitive to human eyes. This suggests that Image Industries may capture some sophisticated object relations. This feature may work as an advantage for Image Industries' by identifying hidden common photo representations and as a disadvantage when finding similarities that are unrelated to company product offerings.

Next, we shift our attention from the micro level to the bigger picture. Table 1 demonstrates the characteristics of unique Image Industries. First, our Image Industries classify 54.7% to 62.6% of our stock sample—translating into a monthly average of 923 to 1,083 firms. Photo representation of unclassified firms is of non-distinguishing quality to indicate what class it should be assigned to.[22] Second, with rising granularity, the number of classified stocks decreases. This suggests that the effective allocation of companies to sectors becomes less evident with an increasing sectors number.

Table 2 shows in Panel A the inter-industry correlation of our image classification with 25 classes, NAICS Industries, and 4-digit GICS is around 26%; meanwhile SIC based FF 30 and Industry_MG achieve a correlation of about 28%. HP, with 25 classes, gets the lowest correlation of 24%. The classifications with higher granularity perform higher (Panel B). This time the best scores get 6-digit GICS, 3-digit NAICS, and 2-digit SICs. Image industries 50 ranks close to FF 48 and again better than HP with 50 classes. We require each industry to have at least five stocks for correlation estimation.[23]

4.2 Economic homogeneity of firms clustered with image

To verify economic relations between companies that are classified into a single industry, we imply the methodology proposed by Bhojraj et al. (2003). We create equal-weighted industry portfolios and verify the ability of these portfolios in explaining contemporaneous firm-level indicators ($vble_t$). We estimate regressions along with Equation 4 and use the same set of ratios as Kaustia and Rantala (2021) extended by FORECASTED MONTHLY RET, TOBIN'S Q, R&D GROWTH, SG&A to # EMPLOYEES, SG&A GROWTH, FORECASTED EPS, and EPS GROWTH. We pre-process along with the winsorization method

[22]Image Industry 25 classifies on an average similar number of stocks as the typical analyst approach introduced by Kaustia and Rantala (2021) who categorize an average of 1,075 stocks for years 1983 to 2013.

[23]A total of 11 out of our Image Industries' 50 (both non-unique and unique) comprise fewer than five stocks; thus, there are 39 industries left in both classifications.

that was described in Section 3.2. Table A1 in the Appendix shows details of ratios calculation. We calculate $vble_{ind,t}$ monthly frequency for 10 indicators based on market information. We estimate the quarterly industry averages for 16 that use only the accountancy information, consistent with the reporting frequency. We calculate regression for industries that have at least five members. The average adjusted R^2 value shows the explanatory power of the industry average on the used ratio of the firm.

We compare four classifications based on Image Industries to SIC, NAICS, GICS, and transitive Hoberg and Phillips (2016) classifications. Bhojraj et al. (2003) show that the granularity of industry classification has a significant impact on economic homogeneity tests. The higher the number of industries in the classification the higher the R^2. Therefore, we compare the IIC of 25 and 50 industries to schemes that have the most similar numbers of industries. The 2-digits SIC has over 50 classes which is too high to compare with IIC with 25 classes. Instead, we use a classification proposed by Moskowitz and Grinblatt (1999) that, through the aggregation of SIC codes, creates 20 industries. Furthermore, to define comparable classifications to IIC 25 (50) we use SIC-based Fama-French classifications with 30 (48) classes, 20 Industry NAICS sectors (3-digit NAICS), 4-digit (6-digit) GICS and Hoberg and Phillips (2016) classification with 25 (50) classes. Table 2 demonstrates the average yearly number of industries with at least five stocks per classification.

Tables 3 and 4 demonstrate a pair-wise results comparison that stems from the same set of firms; i.e., those classified with Image Industries Unique—as well as with all other classification techniques. Table 3 compares Image Industries with variables based on market information. The average rank for all four classifications based on image varies from the first to the last. The homogeneity of image classifications is particularly strong when evaluating six variables: FORECASTED MONTHLY RET, MARKET LEVG, MARKET to BOOK, PE, PRICE to BOOK, and TOBIN'S Q. It is also high in the case of MARKET CAP.

On the contrary, it is weaker than its competitors when we assess it with three remaining variables: MONTHLY RET, EV to SALES, or BETA. In most cases, the nonunique Image

Industries 50 classification achieves the highest R^2 out of image-based classifications. Its relative strength is particularly evident for three indicators where the average rank is the worst.

In terms of competitors, the best R^2 is achieved for classifications with higher granularities demonstrated in Panel B. Bhojraj et al. (2003) argue that a classification system with fewer industry partitions has an inherent disadvantage when the total number of firms is constant. When we compare two schemes—where one has 10 classes consisting of six members per firm, and the other has three classes with 20 members each—we can expect the first scheme to achieve a higher R^2 than the second, even with the random allocation of firms for each scheme. The image-based industries have a low number of classes. Our R^2 is higher than the most granular schemes.

Now, let us compare two particular groups of classifications. First, Hoberg and Phillips (2016) prepare their text-based classifications for different classes. They show that their 10K-based 300 Fixed Effects scheme outperforms both SIC and NAICS-based classifications.[24] We report their 10K-based system performance for two granularities closest to image classifications.[25] In particular, Icode 25 Industries and Icode 50 Industries represent a similar number of classes to Image Industries 25 and 50. Our classifications based on 25 (50) classes perform no worse than the text-based counterparts in 80% (80%) of observations. Second, Bhojraj et al. (2003) compares the most used industry classification schemes and demonstrates that GICS-based classification outperforms SIC and NAISC systems. We compare the Image Industry 25 (50) to four(six)-digit GICS with a similar number of industries of 22 (58). The R^2 for classification based on the image with 25 classes is higher for six out of 10 ratios (in the case of BETA, it is equal) when for IIC with 50 classes is higher for five (equal for BETA and MARKET to BOOK), implying that the best image industry scheme

[24] HP show results of only 300 because they state it is the typical granularity for the entire SIC and NAICS classification (with all numbers).

[25] We download data for Hoberg and Phillips (2016) classifications from https://hobergphillips.tuck.dartmouth.edu/ For the time we download data, the classification is available through the end of 2019. Our sample ended in 2021; therefore, we use the latest available data to create classes for the 2020 to 2021.

performs at least equal to the best commonly used industry classification.

Next, we extend our comparison with ratios based on accountancy information only. Table 4 reports the results. Again, the average rank varies from the first to the last, where image-based classification gets the best homogeneity for NET SALES, SALES GROWTH, TOTAL ASSETS, R&D EXPENSE to SALES, R&D GROWTH, SG&A GROWTH, FORECASTED EPS, EPS GROWTH, ROE, and the worst for ASSET to SALES. The image and text-based comparison show the relative overperformance of similar classifications in 78% of observations. The Image Industries performs better than the GICS in 84% of cases.

Finally, we change the approach to build the stock universe and match our classification to competitors with all available firms. Tables A2 in the Appendix reports the results for market-based indicators. The overall rankings are similar to those reported for the reduced stock universe. There is one significant positive rank changes for image classifications in BETA and three negatives for PRICE to BOOK, FORECASTED MONTHLY RET and PE. Accountancy-based ratios are available in Table A3 in the Appendix—ratios for DEBT to EQUITY, DEBT to ASSETS, RNOA, and ROE are relatively more substantial. We also test how the elimination of microcaps impacts results. Tables A4 and A5 in the Appendix demonstrate the exclusion of micro-caps doesn't change the relative position of image industries in rankings.

Overall, the homogeneity of firms clustered with images reaches the top-performing GICS classification. Furthermore, the high variation in R^2 rankings suggests that image-based industries differ significantly from competitors. The following example illustrates the advantage of IIC as a classification based on product similarities. Drill Quip INC is classified with IIC 50 Unique to industry 32, characterized by an exceptionally high inter-industry correlation of 0.503. Such a high correlation level means that companies classified in industry 32 are characterized by intense homogeneity. Drill Quip INC is a worldwide oil and gas drilling and production equipment manufacturer. The IIC classified this company into industry 32 through photos of offshore drilling platforms. Close-up images also depict products of, e.g., Chevron, Atwood Oceanics, or Murphy Oil Corp—so these firms are peers along with IIC. Instead, two-digit SIC classifies Dril Quip INC to Industrial and Commercial, three-digit NAICS to machinery manufacturing, and six-digit GICS to energy equipment and services. Because of this—in the case of classifications not based on product similarities—Dril Quip INC is not a peer of Chevron, Atwood Oceanics, or Murphy Oil Corp .

In contrast, HP classifications that are also product related agree with IIC about the Dril Quip INC peers. This example shows that product-oriented classifications categorize companies that need drilling platforms to extract oil to the same industry as drilling platform producers. Meanwhile, other classifications break such companies down by different industries—where companies' performances are linked to a demand for different, unrelated products. In the next Section, we verify how these features translate to the sample applications of the proposed image-based classification.

5 Applications

We propose two methods for using image-based industry classifications to build investment portfolios. The first method presents the relative benefits of using the image to achieve portfolio diversification; meanwhile, the second one shows the possibilities of using the IIC to create an industry momentum strategy.

The are several requirements for an industry classification to be applied to construct a portfolio. First, the classification can't be excessively granular and should reflect the primary industries in the market. For example, diversifying a portfolio based on an overly granular scheme is undesirable because some industries representing subindustries will be highly correlated. Scattering companies among such sectors will not ensure portfolio diversification. Second, the industries used to build investment strategies should be well diversified. The dominance of single companies in industries eliminates the industry effect by making industries identical to single companies. Third, a classification should be transitive and unique to split the total stock universe into disjointed sets of firms.

Our Image Industry 25 Unique meets this requirement. Table 1 demonstrates the number of stocks per industry for IIC with 25 (Panel A) and 50 classes (Panel B). IIC 50 is too granular because the average number of stocks per industry is below 20, and 11 industries have less than five stocks. Therefore we do not use it to diversify portfolios or create industry momentum strategies. We base our applications on IIC with 25 classes that, on average, have 43 stocks per industry, with only two industries with less than ten stocks.[26] Analogous to the comparative analysis in the previous section, we compare the results of applications based on IIC 25 to five classifications with a similar number of industries.

5.1 Diversification benefits

Investors use industry classifications to diversify their investment portfolios. This is because two stocks from different industries are expected to be less correlated than those from the same industry. Therefore, portfolio construction often limits the maximum share of stocks allocated to each class. A good industry classification should match related firms into classes. We compare the portfolio diversification benefits for different industry classification schemes. To achieve that, we create sample portfolios—where for each portfolio, we randomly select one stock in each month from every industry and set the portfolio weights: 1) equal-weight; 2) value-weight; 3) mean-variance optimized to maximize Sharpe Ratio; and 4) optimized to minimize the conditional value-at-risk (CVaR). To ensure that a stock we randomly select represents the actual industry, we perform 500 trials per industry and weighing method. The performance is averaged across the stocks we pick from those 500 trials.

Table 5 reports the results for stocks classified with Image Industry 25 Unique. The random portfolios built with Image Industries 25 Unique perform the best in the Sharpe ratio out of all classification methods. The relative overperformance of image-based schemes

[26]Moskowitz and Grinblatt (1999) make a similar decision and create a dedicated classification based on SICs with 20 classes to apply to industry momentum. Typical 2-digit SIC classification has more than 50 classes and is too granular for the industry momentum strategy.

is definitive, and only the four-digit GICS often achieves comparable results in terms of risk reduction. All in all, classifications based on an image deliver substantial diversification benefits for equally weighted, value-weighted, maximum Sharpe, and minimum risk portfolios. Table 6 shows that image-based classifications are the most dynamic compared to competing classifications. The high volatility of classifications corresponds well with the dynamics that are observable in the market, where individual companies often modify their product offerings to remain competitive. The high dynamics of assigning companies to industries support the diversification benefits of using the image to build portfolios.

5.2 Industry momentum

Moskowitz and Grinblatt (1999) demonstrate that persistence in industry return generates significant profits that may account for much of the profitability of individual stock momentum strategies. They construct a well-balanced industry classification based on two-digit SIC codes that consist of 20 classes and build a strategy that longs in industries with the highest industry momentum and short in the lowest. Their system achieves significantly positive results that demonstrate evidence of industry momentum. We follow their direction and compare industry momentum strategies performance based on different industry classification techniques. In particular, we verify the performance in image-based industries to the momentum-based strategy.

We construct a momentum strategy in a similar vein as Moskowitz and Grinblatt (1999) by sorting industry portfolios based on their past six-month value-weighted returns and investing equally in the top three industries while taking short positions equally in the bottom three industries and holding these positions for six months. Furthermore, we extend this setting by lengthening the holding period to nine and twelve months and estimating the same strategies but with equally-weighted returns. Finally, we also form a short-term reversal strategy that invests long for six, nine, or twelve months in three industries with the lowest one-month value or equally-weighted return. [27]

Table 7 reports the results comparing Sharpe ratios of strategies built with industry classifications used in the previous section.Of the six alternative classifications and 12 settings tested, Image Industry 25 Unique ranks first five times, second six times, and third once. The strategies built with four-digit GICS dominate value-weighted portfolios, and IIC is the best for equally weighted settings. In addition, IIC demonstrates the most robust results. Its Sharpe ratio is solid for the momentum, reversal, value, and equally weighted portfolios. Interestingly, strategies based on four-digit GICS are only robust for value-weighted portfolios. The nine and twelve months equally-weighted industry momentum strategy built with GICS performs poorly.[28]

Next, we create "random" industry portfolios and compare their Sharpe ratios with Image Industry 25 Unique to confirm performance robustness from the images-based industry momentum strategies. To construct a "random" industry, we follow the methodology of Moskowitz and Grinblatt (1999) and replace every actual stock in the image-based scheme with another stock with almost the same six-month return. We find similar stocks by ranking six-month returns and picking a replacement stock that differs by "n" ranks. Comparing the results between proper and random strategies shows if the firms' industry membership drives a strategy performance or if it is random and comes only from the firm-level momentum. Table 8 demonstrates the simulation results, where Sharpe ratios of Image Industry 25 Unique are higher than those of any random portfolio.

[27]In addition to the long short-term reversal strategy, we build a long-short version. In our sample period, we find that the short leg of the strategy performs quite unpredictably. The results of this strategy are available upon request. The design of the short-term reversal strategy based solely on long positions is in line with market practice (see, e.g., VESPER U.S. LARGE CAP SHORT-TERM REVERSAL STRATEGY ETF).

[28]Table A6 reports the results of industry momentum strategy extended with volatility targeting mechanism as proposed by Barroso and Santa-Clara (2015). Moreover, for this application, IIC shows its high application benefits.

6 Underlying mechanism

Depending on how market players identify their peers, the industry's categorization considerably impacts the trading of stocks. We suggest a suitable industry classification should have a high degree of agreement (disagreement) regarding stocks categorized in the same industry (different industries). The more the agreement within an industry, the more significant the impact of aggregated demand/supply of investors on stock prices and the greater the advantages of any market-based industry categorization application. This phenomenon is the basis for any application for industry categorization based on market pricing (Diermeier, Ibbotson, & Siegel, 1984; Ibbotson, Diermeier, & Siegel, 1984; Merton, 1973).

Sections 5.1 and 5.2 show that the image industry classification generates the highest portfolio diversification benefits and high profitability from the industry momentum strategy compared to all other classification schemes. This section presents the theories underlying the concept and the empirical evidence corroborating the views.

To illustrate, when a business publishes material information, the market responds to that company's stock price and the stocks of other firms seen by those investors as its competitors. However, investors employ diverse methods to determine each company's comparables, and their investment choices are not uniform. Consider two different investors. The first one thinks that firm 1 has peers in companies 2 and 3. The second one has a different opinion and believes that firm 1's competitors are companies 2 and 4. Now, firm 1 discloses positive information that might result in a favorable position for all companies in company 1's industry (or company 1's peers). Therefore, the first investor may be enticed to acquire shares of firm 1 and companies 2 and 3 that are comparable. The positive news may lure the second investor; however, their investing selections will differ. The second investor may purchase stock in companies 1, 2, and 4. The demand for shares of companies 1, 2, 3, and 4 will vary because the first and second investors would evaluate their peers differently. Specifically, the demand for shares of businesses 1 and 2 will arise from the combined purchase choices of the first and second investors. Still, the demand for shares of firms 3 and 4 will be lower due to the first and second investor's purchasing decisions—respectively.

We can extend this inference and say that different views about peers by two investors create two separate industry classifications. One classification has a first cluster consisting of stocks 1, 2, and 3 and the other with stocks 1, 2, and 4. However, our investors do not limit their peers' perceptions only to those four stocks; they have a view on peers for many firms. To generalize, investor's A opinions on peers define classification X while the views of investor B define classification Y. Additionally, the perception of those investors is not unique. N investors share the same opinion on industry classification as investor A (we can define it as group A_n), and m investors share a view of investor B (group B_m).

Now, let us look into firms classified into any of the two industries of classification X. Investors A_n have a homogeneous demand-supply for shares within the same sector but a heterogeneous one for claims classified into two distinct sectors. This results in a higher agreement (disagreement) about stocks ranked in the same (different) industry. Furthermore, the higher the aggregated demand/supply from investors A_n, the higher the market agreement (disagreement) when stocks are classified into the same (different) industry. This phenomenon is a foundation of the rationale for any industry classification application based on market prices (Diermeier et al., 1984; Ibbotson et al., 1984; Merton, 1973).

Comparing the quality of classifications X and Y should be based on their abilities to capture aggregate demand/supply from investors A_n and B_m. Suppose the perception of A_n and B_m is more congruent. In that case, the classification will significantly impact stock prices and generate good returns from the market-based price application—as shown in Sections 5.1 and 5.2. Referring to the relationships we described earlier, the aggregated demand/supply related to classification based on the image is high. The following Section attempts to provide the reasons for this.

6.1 Cognitive psychology and the concept of similarity

According to the cognitive psychology theory on similarity, investors A_n and B_m disagree about firm connections because we cannot consider similarity a universal concept.[29] Therefore, it requires a frame of reference. Goodman (1972) argues that the similarity comparison process systematically fixes respects. This concept states that any two things share an arbitrary number of predicates and differ from each other in a random number of ways. Therefore, searching for similarity requires non-arbitrary assumptions to constrain the predicates.

Investors A_n and B_m rely on different predicates that drive their firms' similarity perception. For example, one may believe that the critical driver of firms' similarity is the range of products they offer (e.g., Hoberg and Phillips' (2016) text-based classification or our industries built with image), and another believes it is a production process (e.g., SIC or NAICS classifications). Furthermore, even if investors A_n and B_m share common views on the dominant role of the products offered in the context of the companies' similarities, their final judgments related to the industry classification may differ because of the distinct cognitive processes. As an example, one can search for peers to tractor producers. One respect will be related to a machine being produced, matching this firm with all other tractor producers. This perspective will also enter tractors close to excavators, which may be treated as similar devices. However, another respect can be related to the unique features of tractors—which makes their applications materially different. One tractor can be designed to assist agriculture, while the other works in open-pit mines. These vast machines place tractors close to bucket-wheel excavators. Thus, different respects of investors An and Bm similarity define two distinct classification schemes: X and Y.

[29] According to Tversky's featural theory of similarity (e.g., Gati and Tversky (1984) and Tversky (1977)), Murphy and Medin (1985) noted that "the relative weighting of a feature (as well as the relative importance of common and distinctive features) varies with the stimulus context and task, so there is no unique answer to the question on how similar is one object to another" (p. 296). Even more bluntly in this regard is a philosopher Goodman (1972, p. 437), who called similarity "invidious, insidious, a pretender, an imposter, a quack." Goodman says that similarity of A to B is an ill-defined, meaningless notion unless one can say "in what respects" A is similar to B.

6.2 Respects of firms similarity

As similarities can be ambiguous, representing market industries has no clear solution. The solution depends on the relative perception of similarity as a concept that links peers into groups and on the desired application. From this perspective, each industry classification can only be one approximation of how different investors define similar companies. Meanwhile, the best industry classification is the one that is the most useful in a particular application. Referring to our earlier relationship, if a demand/supply related to industry X is higher than to Y, then industry X is better than Y for applications that benefit from it. Furthermore, if all existing industry schemes (e.g., NAICS, SIC, GICS, or our image industry classification) only approximate classifications defined by the market participants, we should consider the aspects that drive market participants to define industries.

Indeed, the wide availability of popular industry classifications (such as SIC, NAICS, or GICS) builds a two-sided relationship—where investors that know these classifications treat them as a dimension of firm similarity. This results in similar (different) demand/supply profiles for stocks from the same (different) SIC, NAICS, or GICS industries. However, given that these systems often define stocks inconsistently and each delivers several granularities, investors who represent peers based on them can become confused.[30] From the cognitive psychology perspective, well-known classifications may serve as some predicates determining peers' definitions, but certainly, they are not the only determinant.

Not surprisingly, alternative aspects of firm similarity may have a dominant role in determining investment behavior. Hoberg and Phillips (2016) show that a significant market participant compares firms' product offerings to identify peers. We agree that considering products to measure firms' similarity is intuitive. After all, a company's cognitive process is knowing what products it offers to customers. However, determining similarities is not a

[30]It is challenging to define peers for a stock based on existing classifications. When an investor searches for company A's peers and reviews peers indicated by SIC, NAICS, or GICS, then—depending on the classification—they will get different peers. Worse, each classification has several granularities. Peers based on SIC2 or SIC4 are very different. So, an investor must decide by themself what their peers are.

universal task; evaluating the products' similarities may differ depending on the cognitive sources. Hoberg and Phillips (2016) argue that a valuable source of such information is the description of the company activity defined by the management and is available in the business description section of annual 10K reports. We agree that reading this description may be the basis for some investors to learn about the product offering. However, we argue that these descriptions are not a precise and common source of information to define the companies' product offerings. First, text descriptions in the 10K report usually have a broader context than a company's product offering. It also relates to, e.g., company history, size, competitive advantages, or suppliers; this way, it is a mix of information describing a company. Therefore, drawing universal conclusions about product offerings based solely on this description is tricky. Second, a subjective management description often underlines only the company's advantages. Third, many unprofessional market participants are unaware of a section like that. After all, a business description of, e.g., Apple, Ford, or Coca-Cola, is not a common source of information about their product offerings.

We also argue that illustrations of offered products build a more familiar, objective, and precise picture of the companies' product offerings than management descriptions. First, photos and movies have a dominant impact on human cognitive processes. This is related to the unquestionable widespread use of image-based communication based on television, the internet, and mobile devices (Branthwaite, 2002; Dewan, 2015). Back to the Apple, Ford, and Coca-Cola examples, we believe that most investors see plenty of photos representing their products. Second, there are different sources of images that represent companies' product offerings. Google aggregates most of them. Therefore, the representation of photos available at Google is objective; it draws a holistic picture of the products offered by the company. Third, scientists discovered that our brain has excellent abilities regarding image processing. We process images up to 60,000 times faster than text (Potter et al., 2014). In an age where an overload of information is delivered to us, using the natural abilities of our mind increases our efficiency by building a clear image of companies' product offerings. Consequently, all these features make the image an important respect of firm similarity and explain the mechanism influencing the high application value of our image industry classification.

Finally, we demonstrate the empirical evidence supporting our theory about images' importance in determining firm similarities. First, since the R^2 of FORECASTED MONTHLY RETURN from Table 3 is by far the highest relative to other classifications with similar granularity, image industries show the highest agreement about investors' expectation of future stock returns. This means there is a high demand/supply from investors who share common expectations about firms' peers aggregated with the image. This relation is the foundation for the high benefits of image-based portfolio diversification that was evaluated in Section 5.1.

Second, the volatility of volume shocks demonstrates the rate of investors' agreement. The lower the volatility of volume shocks, the higher the investor agreement about stocks in the industry. Figure 7 visualizes that the rate of agreement about stocks classified into a single industry is the highest for four-digit GICS and image-based classification. It is consistent with results observed for industry diversification benefits and industry momentum, where these two classification deliver the highest Sharpe ratios.

7 Conclusion

In this study, we apply machine learning approaches to classify sectors by associating businesses with their picture representation. We employ machine vision and unsupervised clustering to determine the relationship between businesses based on their customer-facing product offerings. We present an industry categorization that finds peers in a manner analogous to the human brain by comparing picture similarities across companies. We demonstrate that sectors grouped with the image are valuable for applications that capitalize on investor overreaction.

First, we show that image industries are suitable for portfolio diversification. Second, the image-based industry momentum technique outperforms most competing industry categorization methods. Finally, we demonstrate that the economic homogeneity of enterprises grouped with the image is high; this indicates substantial relationships between firms' economic status and images that might define their product offering.

The picture enables the construction of industry classifications with distinctive characteristics. It is dynamic, allowing for a rapid reassignment of enterprises across sectors in response to changes in their product offers. It may also be non-unique, meaning that a company offering many items may be classified into multiple industries. It also has certain drawbacks as it does not adequately categorize items that are difficult to communicate visually—such as services, high-tech, finance, and multi-product conglomerates.

Future studies should expand the dimensions of photos utilized to construct industries. Researchers should attempt to develop technology to identify photos presenting distinctive goods and services. One can also examine the usefulness of implementing our classifications in conjunction with other classifications.

References

Barroso, P., & Santa-Clara, P. (2015). Momentum has its moments. *Journal of Financial Economics*, *116*(1), 111–120.

Bhojraj, S., Lee, C. M., & Oler, D. K. (2003). What's my line? a comparison of industry classification schemes for capital market research. *Journal of Accounting Research*, *41*(5), 745–774.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools.*

Branthwaite, A. (2002). Investigating the power of imagery in marketing communication: evidence-based techniques. *Qualitative Market Research: An International Journal*, *5*(3), 164–171. doi:10.1108/13522750210432977

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh international world-wide web conference (www 1998)*

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. *the Journal of Finance*, *53*(6), 1839–1885.

Dewan, P. (2015). Words Versus Pictures: Leveraging the Research on Visual Communication. *Partnership: The Canadian Journal of Library and Information Practice and Research*, *10*(1). doi:10.21083/partnership.v10i1.3137

Diermeier, J. J., Ibbotson, R. G., & Siegel, L. B. (1984). The Supply of Capital Market Returns. *Financial Analysts Journal*, *40*(2), 74–80. doi:10.2469/faj.v40.n2.74

Fama, E. F., & French, K. R. (1997). Industry costs of equity. *Journal of financial economics*, *43*(2), 153–193.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. doi:10.1016/0010-0285(84)90013-6

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Incorporated

Goodman, N. (1972). Seven Strictures on Similarity. In *Problems and projects*. Bobs-Merril.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. doi:10.48550/ARXIV.1502.01852

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. doi:10.48550/ARXIV.1207.0580

Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, *124*(5), 1423–1465.

Ibbotson, R. G., Diermeier, J. J., & Siegel, L. B. (1984). The Demand for Capital Market Returns: A New Equilibrium Theory. *Financial Analysts Journal*, *40*(1), 22–33. doi:10.2469/faj.v40.n1.22

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. doi:10.1126/science.aaf7894

Jiang, J., Kelly, B. T., & Xiu, D. (2020). (re-) imag (in) ing price trends. *forthcoming, Journal of Finance*, (21-01).

Jing, Y., & Baluja, S. (2008). VisualRank: Applying PageRank to Large-Scale Image Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(11), 1877–1890. doi:10.1109/TPAMI.2008.121

Kahle, K. M., & Walkling, R. A. (1996). The impact of industry classifications on financial research. *Journal of financial and quantitative analysis*, *31*(3), 309–335.

Kaustia, M., & Rantala, V. (2015). Social learning and corporate peer effects. *Journal of Financial Economics*, *117*(3), 653–669.

Kaustia, M., & Rantala, V. (2021). Common analysts: Method for defining peer firms. *Journal of Financial and Quantitative Analysis*, *56*(5), 1505–1536.

Krishnan, J., & Press, E. (2003). The north american industry classification system and its implications for accounting research. *Contemporary Accounting Research*, *20*(4), 685–717.

Kumar, A., Rai, P., & Daume, H. (2011). Co-regularized Multi-view Spectral Clustering. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 24), Curran Associates, Inc.

Lewellen, S. (2012). Firm-specific industries.

Merton, R. C. (1973). An Intertemporal Capital Asset Pricing Model. *Econometrica*, *41*(5), 867–887

Moskowitz, T. J., & Grinblatt, M. (1999). Do industries explain momentum? *The Journal of finance*, *54*(4), 1249–1290.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. doi:10.1037/0033-295X.92.3.289

Ng, A., Jordan, M., & Weiss, Y. (2001). On Spectral Clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (Vol. 14), MIT Press

Obaid, K., & Pukthuanthong, K. (2022). A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. *Journal of Financial Economics*, *144*(1), 273–297.

Perry, R., Mischler, G., Guo, R., Lee, T., Chang, A., Koul, A., ... Vogelstein, J. T. (2021). mvlearn: Multiview Machine Learning in Python. *Journal of Machine Learning Research*, *22*(109), 1–7.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in rsvp at 13 ms per picture. *Attention, Perception, & Psychophysics*, *76*(2), 270–279.

Pukthuanthong, K. (2006). Underwriter learning about unfamiliar firms: Evidence from the history of biotech ipos. *Journal of Financial Markets*, *9*(4), 366–407.

Ramnath, S. (2002). Investor and analyst reactions to earnings announcements of related firms: An empirical analysis. *Journal of Accounting Research*, *40*(5), 1351–1376.

Rauh, J. D., & Sufi, A. (2012). Explaining corporate capital structure: Product markets, leases, and asset similarity. *Review of Finance*, *16*(1), 115–155.

Scoles, S. (2019). Researchers spy signs of slavery from space. *Science*, *363*(6429), 804. doi:10.1126/science.363.6429.804

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS computational biology*, *15*(1), e1006633.

Shi, J., & Malik, J. (1997). Normalized cuts and image segmentation. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 731–737.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. doi:10.1038/nature16961

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. doi:10.48550/ARXIV.1409.1556

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. doi:10.48550/ARXIV.1905.11946

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352. doi:10.1037/0033-295X.84.4.327

von Luxburg, U. (2007). A Tutorial on Spectral Clustering. doi:10.48550/ARXIV.0711.0189

Yu, J., Wang, Z., Majumdar, A., & Rajagopal, R. (2018). DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States. *Joule*, *2*(12), 2605–2617. doi:https://doi.org/10.1016/j.joule.2018.11.021

Table 1: Description and Summary Statistics of Image Industries 25 & 50 Unique

The table presents an overview of industries formed with firms' photos for 25 (Panel A) and 50 (Panel B) classes in a unique setup. Our sample covers NYSE, AMEX, and NASDAQ stocks—excluding financial and service companies (SIC codes 6,000-8,900). Image Industries are formed every second year in 2016 to 2021, allowing time-variation in industrial classification. We report the average number of stocks assigned to each industry (No. of Stocks) and the minimum monthly count (in parentheses). We download images describing the companies from Google, which allows us to gather good-quality images for most companies in our research sample. We report the average monthly capitalization of stocks in each industry (Market Cap. (bn USD)) and share of stocks classified with photos in the whole market capitalization (Avg. % of Market Cap.) as well as in our research sample (in parentheses). Finally, we show the average monthly return of the stock in each industry (Avg. Month. Excess. Ret.), the inter-industry correlation between all stocks in the industry (Inter. Corr.), and the relation between our industries to two-digit SIC codes in the form of an indication of the five most common SIC codes among the companies assigned to each of our industries. We calculate inter-industry correlations for sectors that consist of at least five stocks.

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Inter. Corr.	TOP5 SIC codes
PANEL A: Image Industries 25 Unique						
0	85.5 (59)	995.4	3.6 (5.2)	0.0004	0.282	('13', '49', '29', '35', '28')
1	24.9 (14)	260.5	0.8 (1.2)	0.0055	0.201	('35', '28', '99', '36', '50')
2	20.2 (17)	95.1	0.3 (0.5)	0.0057	0.213	('38', '35', '36', '37', '99')
3	28.1 (13)	207.7	0.7 (1.0)	0.0080	0.237	('35', '38', '36', '37', '28')
4	34.5 (24)	243.0	0.8 (1.2)	0.0097	0.314	('42', '37', '49', '45', '47')
5	83.0 (54)	1,184.3	4.0 (5.9)	0.0081	0.200	('28', '49', '99', '36', '51')
6	10.0 (6)	72.5	0.2 (0.3)	0.0178	0.245	('38', '36', '35', '33', '49')
7	22.3 (16)	121.4	0.4 (0.6)	0.0080	0.336	('35', '33', '37', '30', '50')
8	48.0 (27)	301.1	1.0 (1.5)	0.0024	0.245	('15', '99', '25', '58', '57')
9	231.2 (159)	3,445.5	11.5 (16.8)	0.0053	0.180	('99', '28', '56', '38', '59')
10	60.5 (32)	769.6	2.7 (3.9)	0.0096	0.188	('36', '35', '38', '99', '50')
11	60.4 (35)	722.9	2.6 (3.7)	0.0078	0.183	('49', '28', '10', '13', '99')
12	14.7 (9)	306.9	0.9 (1.3)	0.0039	0.225	('37', '55', '99', '36', '50')
13	35.3 (26)	520.9	1.8 (2.6)	0.0029	0.279	('37', '45', '99', '38', '42')
14	23.0 (16)	290.8	0.9 (1.3)	0.0139	0.295	('36', '99', '35', '50', '38')
15	12.6 (9)	173.1	0.6 (0.9)	0.0088	0.328	('35', '37', '50', '36', '39')
16	13.2 (7)	209.7	0.7 (1.0)	0.0157	0.153	('36', '20', '35', '38', '58')
17	67.5 (36)	497.7	1.5 (2.2)	0.0091	0.254	('35', '50', '33', '36', '37')
18	6.1 (2)	151.8	0.5 (0.7)	0.0164	0.534	('36', '99')
19	8.4 (6)	184.5	0.6 (0.9)	0.0079	0.460	('40', '37', '73', '34', '36')
20	69.8 (57)	1,225.4	3.6 (5.4)	0.0042	0.164	('28', '38', '99', '35', '36')
21	75.8 (58)	353.4	1.3 (1.9)	0.0095	0.179	('38', '36', '50', '28', '35')
22	20.0 (14)	228.0	0.7 (1.1)	0.0087	0.247	('38', '35', '36', '99', '28')
23	17.7 (9)	164.3	0.6 (0.9)	0.0024	0.248	('25', '28', '50', '30', '99')
24	10.0 (6)	115.6	0.4 (0.6)	0.0078	0.282	('35', '37', '38', '34', '36')
Sum	1,082.7 (711)	12,840.8	42.8 (62.6)			

Table 1: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Inter. Corr.	TOP5 SIC codes
PANEL B: Image Industries 50 Unique						
0	11.0 (6)	92.5	0.4 (0.5)	0.0013	0.332	('13', '10', '35', '49', '33')
1	5.5 (4)	11.2	0.0 (0.1)	-0.0120	0.250	('38', '35', '36', '99', '37')
2	70.4 (40)	1,152.0	4.0 (5.9)	0.0088	0.219	('28', '99', '49', '50', '58')
3	19.0 (9)	158.9	0.6 (0.8)	0.0107	0.370	('42', '37', '47', '49', '51')
4	19.3 (15)	310.4	1.0 (1.4)	0.0106	0.303	('36', '35', '99', '38', '56')
5	18.1 (14)	153.4	0.5 (0.7)	0.0116	0.262	('37', '42', '55', '49', '45')
6	14.1 (12)	133.4	0.5 (0.7)	0.0116	0.372	('35', '36', '37', '38', '34')
7	10.9 (6)	319.6	0.9 (1.3)	0.0210	0.299	('36', '35', '38', '50', '59')
8	25.9 (15)	239.9	0.8 (1.2)	0.0071	0.252	('36', '38', '35', '50', '48')
9	22.9 (20)	389.0	1.3 (2.0)	0.0039	0.310	('37', '45', '99', '38', '47')
10	30.4 (27)	453.9	1.5 (2.3)	-0.0020	0.317	('13', '29', '49', '28', '44')
11	4.2 (3)	16.0	0.1 (0.1)	-0.0100	-	('28', '22', '31', '34', '59')
12	51.4 (36)	1,266.8	3.5 (5.2)	0.0058	0.225	('28', '35', '50', '99', '38')
13	33.7 (17)	338.4	1.3 (1.9)	0.0170	0.215	('38', '35', '36', '99', '49')
14	15.4 (2)	133.0	0.5 (0.7)	0.0124	0.153	('49', '99', '28', '36', '40')
16	8.4 (2)	133.8	0.4 (0.6)	-0.0016	0.359	('37', '40', '34', '36', '73')
17	41.2 (36)	347.2	1.2 (1.8)	0.0076	0.190	('36', '35', '38', '73', '99')
18	55.3 (25)	406.7	1.4 (2.0)	0.0061	0.192	('38', '36', '35', '33', '28')
19	5.8 (2)	166.6	0.6 (0.9)	0.0141	0.294	('36', '10', '33')
22	16.6 (11)	126.5	0.5 (0.7)	0.0004	0.239	('25', '38', '50', '57', '35')
23	13.3 (9)	213.2	0.7 (1.0)	0.0099	0.239	('38', '35', '28', '99', '36')
24	17.8 (12)	156.2	0.5 (0.7)	0.0154	0.300	('35', '37', '99', '49', '73')
25	15.8 (9)	280.2	0.8 (1.2)	0.0026	0.224	('37', '55', '99', '50', '36')
26	53.5 (43)	819.9	2.7 (3.9)	0.0070	0.155	('20', '28', '99', '38', '35')
27	35.9 (15)	669.8	1.9 (2.8)	0.0051	0.219	('59', '56', '53', '99', '54')
28	17.5 (16)	58.7	0.2 (0.3)	0.0051	0.266	('56', '31', '30', '23', '39')
30	4.4 (2)	29.7	0.1 (0.1)	0.0156	0.066	('38', '99', '36', '35', '87')
31	28.4 (12)	145.0	0.4 (0.6)	0.0080	0.243	('13', '38', '99', '49', '28')
32	9.2 (5)	111.4	0.5 (0.6)	-0.0055	0.503	('13', '29', '44', '73', '16')
33	3.0 (3)	22.6	0.1 (0.1)	0.0192	-	('36', '50')
34	35.7 (19)	374.2	1.3 (1.9)	0.0032	0.168	('49', '10', '99', '28', '14')
35	1.9 (1)	0.8	0.0 (0.0)	-0.0140	-	('28', '35', '38', '36')
36	13.1 (7)	69.2	0.2 (0.4)	0.0096	0.302	('35', '37', '38', '36', '59')
37	38.6 (25)	351.9	1.1 (1.7)	0.0104	0.168	('99', '28', '38', '48', '50')
38	16.8 (14)	329.1	1.2 (1.8)	0.0044	0.198	('28', '35', '26', '36', '50')
40	14.0 (5)	47.1	0.2 (0.2)	0.0041	0.218	('35', '38', '36', '34', '37')
41	1.0 (1)	2.3	0.0 (0.0)	0.0110	-	('87',)
45	11.0 (6)	257.2	0.9 (1.3)	0.0103	0.230	('49', '28', '99', '36', '34')
47	60.4 (38)	723.7	2.2 (3.3)	0.0036	0.260	('37', '35', '33', '50', '36')
48	41.1 (25)	215.1	0.8 (1.1)	0.0045	0.242	('15', '99', '25', '58', '17')
49	11.6 (10)	148.8	0.5 (0.7)	0.0027	0.339	('35', '30', '36', '37', '50')
Sum	923.6 (579)	11,375.0	37.4 (54.7)			

Table 2: Peer Groups' Inter-industry Correlations - Comparison

We present a comparison of different industry classification techniques by inter-industry correlations (Correlations %), the average number of industries (Number of Industries), and the average monthly number of classified stocks (Number of Stocks). We compare two classifications based on Image Industries: non-unique with 25 classes and unique with 25 classes, with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 and 48 classes, NAICS industry classification with 20 classes, three digits NAICS codes, four and six digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 and 50 classes. The sample covers stocks classified with Image Industries from 2016 to 2021. We use sectors with at least five stocks to calculate the average inter-industry correlation.

Industry Classification Name	Correlation (%)	Number of Industries	Number of Stocks
PANEL A: Image Industries 25 - comparison			
Image Industries 25	0.260	25.0	1,124.0
Image Industries 25 Unique	0.255	25.0	1,124.0
Industry_MG	0.280	20.0	1,124.0
Fama-French 30 Industries	0.276	30.0	1,119.9
NAICS Industries	0.260	17.8	1,122.8
4-digit GICS	0.255	22.0	1,115.3
Icode 25 Industries	0.242	24.3	1,113.2
PANEL B: Image Industries 50 - comparison			
Image Industries 50	0.267	39.0	959.3
Image Industries 50 Unique	0.260	39.0	959.3
2-digit SIC	0.304	56.6	1,124.0
Fama-French 48 Industries	0.276	45.8	1,119.9
3-digit NAICS	0.303	70.7	1,122.8
6-digit GICS	0.316	57.8	1,115.3
Icode 50 Industries	0.253	44.5	1,085.0

Table 3: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Market Information

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable $vble$ represents 10 ratios using market information: 1) MARKET to BOOK; 2) PRICE to BOOK; 3) MONTHLY RET; 4) FORECASTED MONTHLY RET; 5) MARKET LEVG; 6) EV to SALES; 7) PE; 8) BETA; 9) MARKET CAP; and 10) TOBIN'S Q for each firm i at month t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry ind excluding firm i at month t. We regress FORECASTED MONTHLY RET on the industry average from MONTHLY RET. In Panel A we compare two classifications based on Image Industries with 25 classes (non-unique and unique) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes. In Panel B we compare two classifications based on Image Industries with 50 classes (non-unique and unique) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes. In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers stocks classified with Image Industries from 2016-2021. We calculate regression for industries that have at least five members. Table A1 in Appendix shows details of ratios calculation.

Industry Classification Name	MARKET to BOOK	PRICE to BOOK	MONTHLY RET	FORECASTED MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q
PANEL A: Image Industries 25 - comparison										
Image Industries 25	29.3	20.8	25.2	2.3	33.8	24.5	9.6	33.5	38.1	27.8
Image Industries 25 Unique	30.2	18.5	24.3	2.4	35.9	24.2	10.0	34.2	36.9	28.2
Industry_MG	27.8	18.3	26.5	1.9	31.0	24.9	9.6	37.1	35.6	27.1
Fama-French 30 Industries	28.7	18.2	27.0	1.8	31.4	27.5	8.2	35.2	35.2	27.7
NAICS Industries	28.5	21.5	26.5	2.3	30.6	27.0	9.5	35.4	37.9	27.3
4-digit GICS	27.4	17.0	28.8	1.5	30.7	25.8	7.0	34.1	38.3	26.7
Icode 25 Industries	26.3	17.3	26.3	1.9	30.2	24.2	8.4	34.1	34.0	25.9
PANEL B: Image Industries 50 - comparison										
Image Industries 50	31.0	21.4	26.2	2.3	36.6	25.8	10.9	34.1	38.6	29.4
Image Industries 50 Unique	31.5	20.6	24.8	2.3	38.1	25.6	10.8	31.6	39.6	29.8
2-digit SIC	30.4	18.8	28.4	1.9	33.7	27.8	9.3	35.4	36.2	29.0
Fama-French 48 Industries	31.0	18.7	28.4	1.9	33.3	30.6	9.0	35.4	35.0	29.5
3-digit NAICS	29.4	19.6	26.9	2.4	32.0	27.1	10.0	32.7	34.4	27.3
6-digit GICS	31.4	19.9	30.1	1.4	33.8	30.8	7.9	33.9	40.6	29.3
Icode 50 Industries	29.0	18.6	27.4	2.0	32.7	29.7	9.1	29.9	35.7	27.6

Table 4: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Accountancy Information

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable $vble$ represents 16 ratios using accountancy information: 1) TOTAL ASSETS; 2) NET SALES; 3) DIVIDEND PAYOUT; 4) PROFIT MARGIN; 5) DEBT to EQUITY; 6) SALES GROWTH; 7) R&D EXPENSE to SALES; 8) R&D GROWTH; 9) SG&A to # EMPLOYEES; 10) SG&A GROWTH; 11) FORECASTED EPS; 12) EPS GROWTH; 13) DEBT to ASSET; 14) RNOA; 15) ROE; and 16) ASSETS to SALES for each firm i at quarter t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry ind excluding firm i at month t. In Panel A we compare two classifications based on Image Industries with 25 classes (non-unique and unique) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes. In Panel B we compare two classifications based on Image Industries with 50 classes (non-unique and unique) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes. In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers stocks classified with Image Industries with 25 classes from 2016 to 2021. We calculate regression for industries that have at least five members. Table A1 in Appendix shows details of ratios calculation.

Industry Classification Name	TOTAL ASSETS	NET SALES	DIV PAYOUT	PROFIT MARGIN	DEBT to EQUITY	SALES GROWTH	R&D EXPENSE to SALES	R&D GROWTH	SG&A to # EMPLOYEES	SG&A GROWTH	FORECASTED EPS	EPS GROWTH	DEBT to ASSETS	RNOA	ROE	ASSET to SALES
PANEL A: Image Industries 25 - comparison																
Image Industries 25	44.8	44.5	3.9	24.2	15.5	44.6	22.7	27.7	21.5	42.1	46.3	16.7	28.8	17.7	16.6	20.7
Image Industries 25 Unique	43.8	44.9	5.6	24.1	16.0	43.7	24.4	30.6	23.1	42.7	47.1	16.6	28.7	19.7	18.1	20.5
Industry_MG	32.6	36.0	4.7	23.9	16.2	40.4	23.0	24.2	23.1	34.2	42.8	14.4	28.9	18.9	15.9	25.6
Fama-French 30 Industries	36.3	37.7	4.1	24.3	14.6	39.3	20.1	24.0	22.3	34.5	41.7	15.6	28.0	20.0	13.0	22.2
NAICS Industries	43.6	43.4	3.0	24.0	11.9	41.3	18.9	24.2	24.7	35.6	42.7	16.3	30.6	18.6	14.2	23.3
4-digit GICS	40.8	38.8	3.1	24.1	11.3	41.1	21.5	25.4	22.9	34.5	41.6	15.6	28.9	18.2	13.1	26.3
Icode 25 Industries	38.6	40.6	4.5	23.0	13.4	40.6	19.4	28.1	22.7	36.3	43.1	17.2	26.9	19.0	14.7	21.9
PANEL B: Image Industries 50 - comparison																
Image Industries 50	42.9	45.4	4.8	24.5	17.0	48.2	24.2	31.8	22.7	42.1	47.7	15.8	28.3	18.4	17.9	20.2
Image Industries 50 Unique	43.1	48.4	6.0	28.2	16.9	46.9	29.7	32.0	21.2	42.6	47.6	17.4	27.9	22.2	18.1	21.9
2-digit SIC	38.5	39.5	5.7	26.3	15.8	40.9	21.8	24.8	20.6	33.3	43.6	15.9	27.2	20.3	15.6	26.6
Fama-French 48 Industries	36.6	37.1	5.1	28.4	14.5	41.2	24.9	25.5	22.3	34.5	40.9	15.4	26.9	20.9	14.7	27.6
3-digit NAICS	35.4	39.8	5.0	26.4	16.2	40.4	20.8	25.9	20.9	34.9	42.2	15.9	24.7	22.2	16.4	23.2
6-digit GICS	43.8	43.4	4.4	24.6	14.5	42.1	21.1	24.6	23.3	33.1	44.7	16.7	28.2	22.1	15.7	26.3
Icode 50 Industries	35.5	39.3	5.7	25.7	17.2	42.0	22.8	26.0	23.0	35.9	45.9	17.4	27.4	20.3	18.8	25.9

Table 5: Industry diversification benefits - comparison with the Image Industry

We demonstrate the average performance of portfolios built with different industry classification techniques formed with 1) unique image industries with 25 classes (Image Industries), 2) Moskowitz and Grinblatt (1999) (Industry_MG), 3) Fama-French industries with 30 classes, 4) NAICS industry classification with 20 classes, 5) four digits GICS codes, and 6) transitive Hoberg and Phillips (2016) classifications with 25 classes (Icode 25 Industries). To create a portfolio, we randomly select one stock from each industry each month and measure the portfolio performance. We perform 500 trials per industry and report the average performance statistics for three different stock weighing techniques: 1) equal weights (Panel A); 2) market weights (Panel B); 3) mean-variance optimized portfolios to maximize Sharpe Ratio (Panel C); and 4) conditional value-at-risk (CVaR) optimized portfolios to minimize risk (Panel D). To optimize portfolios (Panel C and D) we use three years of historical monthly returns before the monthly optimization date. We use the same stock universe for every industry classification consisting of stocks with the Image Industry 25 Unique classification. In each row, the best observation is marked as dark green, the second best as light green, and the third best as beige. The best observations are the highest for Annual Ret., Sharpe Ratio, and Calmar Ratio, and the lowest for Annual Std. Dev., Max DrownDown, and Max 1m. Loss. The sample covers the years 2016 to 2021. We use sectors with at least five stocks for the Image Industry.

Industry Classification Name	Image Industries	Industry_MG	Fama-French 30 Industries	NAICS Industries	4-digit GICS	Icode 25 Industries
PANEL A: Equally weighted portfolios						
Annual Ret.	0.097	0.064	0.065	0.090	0.070	0.078
Annual Std. Dev.	0.255	0.270	0.268	0.306	0.246	0.267
Max DrownDown	0.472	0.545	0.539	0.546	0.474	0.507
Max 1m. Loss	0.283	0.310	0.324	0.343	0.273	0.314
Sharpe Ratio	0.389	0.241	0.248	0.299	0.290	0.291
Calmar Ratio	0.221	0.126	0.129	0.181	0.167	0.163
PANEL B: Value weighted portfolios						
Annual Ret.	0.159	0.121	0.110	0.121	0.109	0.111
Annual Std. Dev.	0.291	0.236	0.212	0.225	0.235	0.288
Max DrownDown	0.370	0.380	0.335	0.343	0.319	0.383
Max 1m. Loss	0.278	0.243	0.226	0.234	0.224	0.276
Sharpe Ratio	0.735	0.609	0.572	0.588	0.582	0.533
Calmar Ratio	0.542	0.430	0.394	0.425	0.461	0.398
PANEL C: Max. Sharpe Ratio portfolios						
Annual Ret.	0.080	0.049	0.064	0.081	0.065	0.036
Annual Std. Dev.	0.247	0.240	0.224	0.244	0.219	0.269
Max DrownDown	0.415	0.433	0.366	0.404	0.373	0.494
Max 1m. Loss	0.255	0.257	0.235	0.266	0.233	0.286
Sharpe Ratio	0.357	0.243	0.334	0.351	0.334	0.167
Calmar Ratio	0.258	0.180	0.230	0.273	0.252	0.117
PANEL D: Min. CVaR Ratio portfolios						
Annual Ret.	0.063	0.042	0.016	0.060	0.043	0.020
Annual Std. Dev.	0.218	0.207	0.203	0.217	0.189	0.216
Max DrownDown	0.386	0.409	0.433	0.396	0.355	0.447
Max 1m. Loss	0.227	0.237	0.245	0.248	0.194	0.226
Sharpe Ratio	0.314	0.220	0.105	0.298	0.255	0.121
Calmar Ratio	0.220	0.154	0.088	0.213	0.185	0.087

Table 6: Industry Dynamics

The table compares the statistics of industry dynamics. In Panel A we compare unique Image Industries with 25 classes with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes. In Panel B we compare unique Image Industries with 50 classes with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes. We estimate the industry dynamics as the frequency of reclassification of a given company among industries. For example, the dynamics for a company that is classified into one industry for the entire period is zero. The sample covers all stocks classified with Image Industries 25 (Panel A) Image Industries 50 (Panel B) and additionally with all other techniques, from the years 2016-2021. We additionally require that each firm has classification in year t-1 to ensure that the captured changes in the industry are exclusively due to re-classifications. The reported statistics is the proportion of firms with a new classification to all observations. They are based on annual firm observations.

Industry Classification Name	Dynamics
PANEL A: Image Industries 25 - comparison	
Image Industries 25	0.217
Industry_MG	0.027
Fama-French 30 Industries	0.027
NAICS Industries	0.039
4-digit GICS	0.000
Icode 25 Industries	0.104
PANEL B: Image Industries 50 - comparison	
Image Industries 50	0.225
2-digit SIC	0.031
Fama-French 48 Industries	0.029
3-digit NAICS	0.066
6-digit GICS	0.000
Icode 50 Industries	0.111

Table 7: Industry momentum & short-term reversal

The table compares Sharpe ratios of momentum and short-term reversal industry portfolios built with different industry classification techniques formed with 1) unique image industries with 25 classes (Image Industries), 2) Moskowitz and Grinblatt (1999) (Industry_MG), 3) Fama-French industries with 30 classes, 4) NAICS industry classification with 20 classes, 5) four digits GICS codes, and 6) transitive Hoberg and Phillips (2016) classifications with 25 classes (Icode 25 Industries). We build an industry momentum strategy in the same way as Moskowitz and Grinblatt (1999) by investing long (short) for six, nine, or twelve months in three industries with the highest (lowest) six months momentum. The short-term reversal strategy is built by investing long in three industries for six, nine, or twelve months with the lowest one-month returns. The returns of industries are calculated as market-weighted (Panel A) or equally weighted (Panel B). The table has 12 rows representing different strategies. The first phrase demonstrates the name of the strategy (Momentum or Reversal), and the second phrase denotes the investment horizon (six, nine, or twelve months). In each row, the highest Sharpe ratio is marked as dark green, the second highest as light green, and the third highest as beige. The sample covers the years 2016 to 2021. We use sectors with at least five stocks for the Image Industry.

	Image Industries	Industry_MG	Fama-French 30 Industries	NAICS Industries	4-digit GICS	Icode 25 Industries
PANEL A: Market weighted industry returns						
Momentum 6m	0.718	0.250	0.466	0.066	0.889	-0.240
Momentum 9m	0.647	0.108	0.267	-0.155	0.733	-0.246
Momentum 12m	0.690	-0.059	0.319	-0.213	0.636	0.054
Reversal 6m	1.395	1.002	0.974	1.355	1.438	1.409
Reversal 9m	1.440	1.028	0.965	1.346	1.443	1.353
Reversal 12m	1.484	1.087	0.992	1.395	1.517	1.381
PANEL B: Equally weighted industry returns						
Momentum 6m	0.511	0.342	0.346	0.690	0.453	0.184
Momentum 9m	0.332	0.191	0.233	0.369	0.140	-0.011
Momentum 12m	0.225	0.096	0.173	0.060	-0.039	0.008
Reversal 6m	1.047	0.728	0.708	0.844	0.969	1.012
Reversal 9m	1.027	0.689	0.678	0.866	0.899	0.984
Reversal 12m	1.055	0.740	0.761	0.923	0.938	1.020

Table 8: Industry momentum - "Random" Industry Portfolios

We compare Sharpe ratios of 'random' industry portfolios with Image Industries Unique 25. We build an industry momentum strategy in the same way as Moskowitz and Grinblatt (1999) by investing long (short) for six, nine, or twelve months in three industries with the highest (lowest) six months momentum. To construct a "random" industry, we replace every true stock in Image Industries with another stock with almost the same six-month return. We find similar stocks by ranking 6-month returns and picking a replacement stock that differs by "n" ranks. The table has three columns, where each column represents strategy with different "n" shifts, e.g., column Shift_-1 states replacing an actual stock with another whose 6-months momentum rank is one point lower. Column Shift_0 represents the original strategy. The returns of industries are calculated as market-weighted (Panel A) or equally weighted (Panel B). In each row, the highest Sharpe ratio is marked as dark green. The sample covers 2016-2021, where we use the first months to create stock 6-month momentum. We use sectors with at least five stocks for the Image Industry.

	Shift_-1	Shift_0 or IIC	Shift_1
PANEL A: Market weighted industry returns			
Momentum 6m	-0.128	0.718	0.320
Momentum 9m	-0.240	0.647	0.153
Momentum 12m	0.028	0.690	-0.050
PANEL B: Equally weighted industry returns			
Momentum 6m	-0.492	0.511	-0.367
Momentum 9m	-0.423	0.332	-0.449
Momentum 12m	-0.256	0.225	-0.509

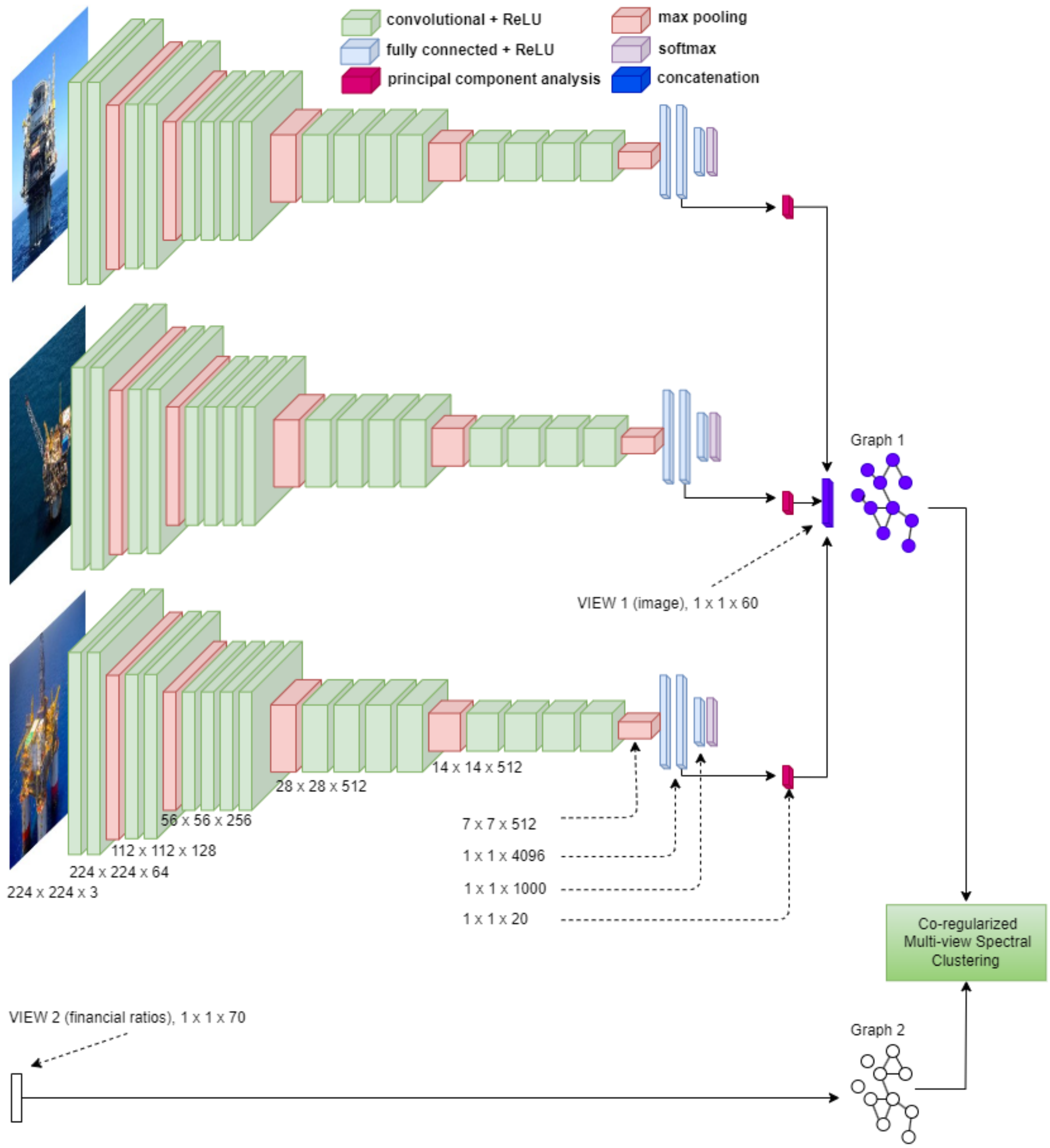

Figure 1: Multi-view clustering

Note: The figure presents the diagram of Co-regularized Multi-view Spectral Clustering (CMVSC) (Kumar, Rai, & Daume, 2011) augmented to define Image Industry Classification. Two views define the inputs for each company: 1) three photographs showing the company's business profile; and 2) a set of 70 financial indicators taken from WRDS. The photos dimensions are 224-224-3, where the first dimension shows the height, the second the width, and the third the colors. To detect objects on photos, we use a convolutional neural network VGG-19 that is 19 layers deep (Simonyan & Zisserman, 2014). The networks consist of convolutional layers that create a feature map, pooling layers that scale down the information generated by the convolutional layer, and fully connected layers that compile the data extracted by previous layers to form the final output. VGG-19 pretrained with more than 1m photos is designed to classify objects into 1,000 categories. We use the numerical representation of identified objects from the last, but one fully connected layer with dimensions 1x1x4096. Next, we apply Principal Complement Analysis (PCA) to reduce this dimension to 1x1x20 per single photo. Finally, to create the numerical representation of the first view, we join the information about identified objects from three photos by concatenating three vectors to single one with dimensions 1x1x60. The clustering procedure based on CMVSC creates one graph per each view and combine the co-regularized information from graphs to form clusters.

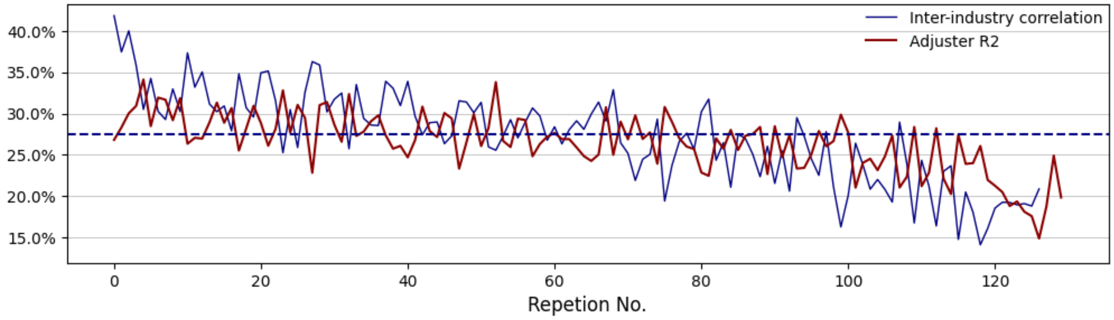

Figure 2: Inter-industry correlation and adjusted R2 in guided clustering procedure

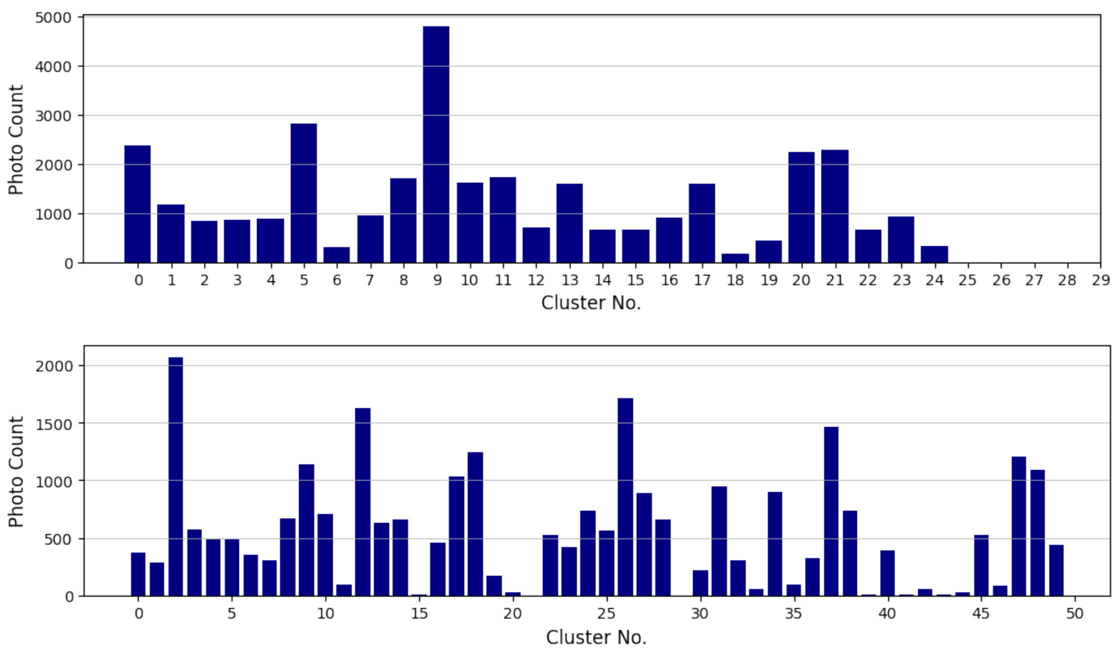

Figure 3: Average Number of Photos per Industry

The upper (lower) figure presents the average photo count for Image Industry 25 (Image Industry 50) in each of two-years periods: 2016 to 2017, 2018 to 2019, 2020 to 2021.

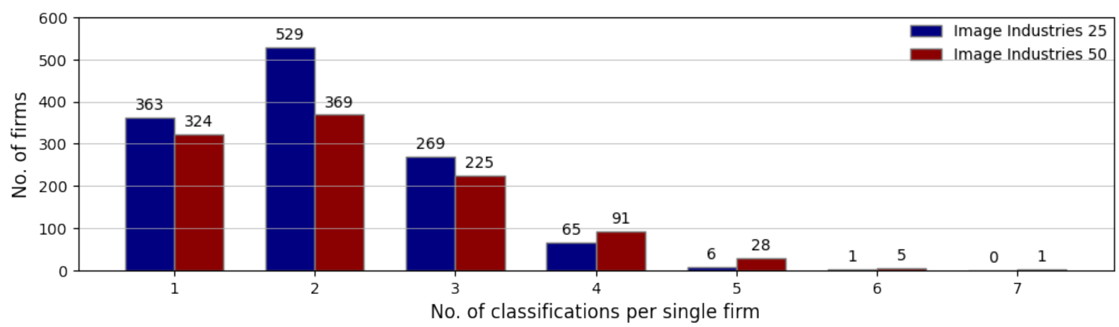

Figure 4: Number of firms with multiple classifications

This chart illustrates how often companies have been classified into one or more classes with nonunique Image Industry Classification consisting of 25 and 50 classes.

Figure 5: Photos representing industries with highest inter-industry correlations

This figure presents photos from Image Industries 50 presenting three industries with the highest inter-industry correlations. Each industry is represented with 10 randomly selected stocks and their three randomly selected photos. Photos are downloaded from Google from the years 2016 to 2017, which is in the middle of our testing sample.

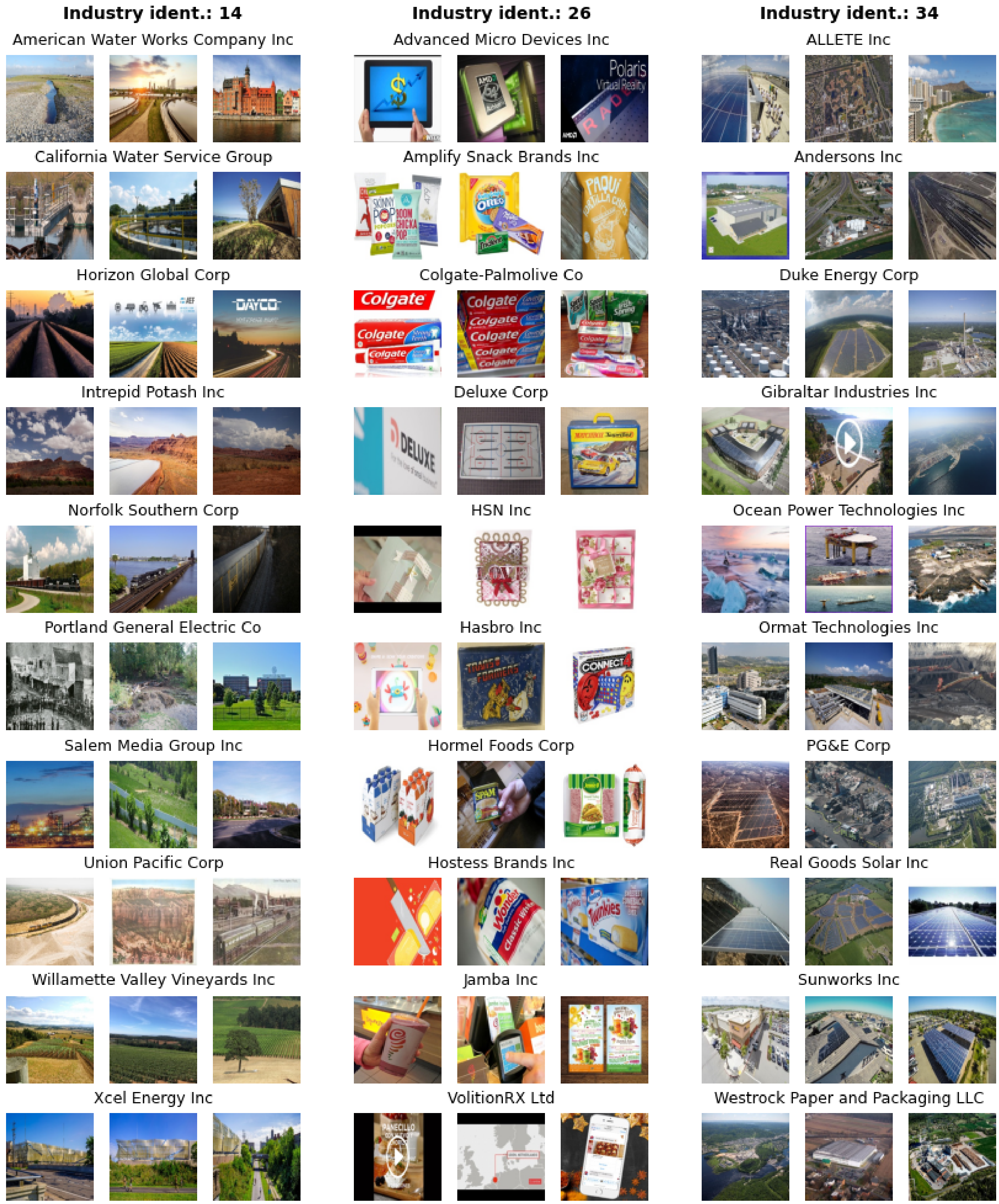

Figure 6: Photos representing industries with the lowest inter-industry correlations

This figure presents photos from Image Industries 50 presenting three industries with the lowest inter-industry correlations. Each industry is represented with 10 randomly selected stocks and their three randomly selected photos. Photos are downloaded from Google from years 2016 to 2017, which is in the middle of our testing sample. We excluded industry no. 30 that is represented with less than 10 stocks.

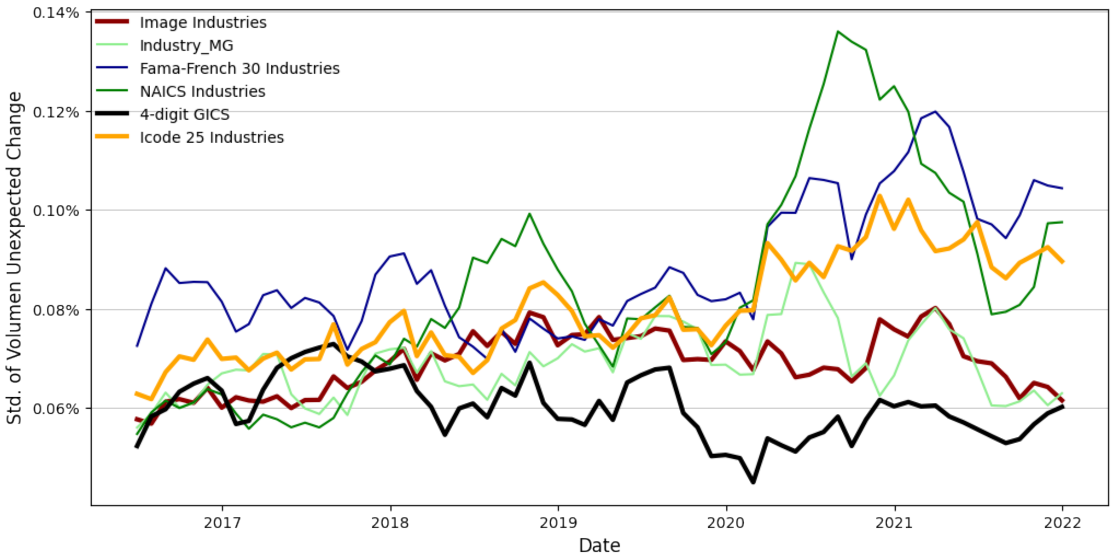

Figure 7: Investor's agreement

We show how investors agree about industries in different classifications. We measure agreement for a classification as the mean of industries' standard deviations of unusual volume changes of stocks. The unusual change of volume is the change of volume $VolCh_{t,i} = vol_{t,i}/vol_{t-1,i}$ in the month t for stock i, divided by the average monthly volume change in the previous six months $\mathrm{UnVolCh}_{t,i} = \mathrm{VolCh}_{t,i} \,/\, \left(\frac{1}{6}\sum_{t=-6}^{-1} \mathrm{VolCh}_{t,i}\right)$ The figure demonstrates the six-month moving average of the agreement.

Appendix

A.1 Photo cleaning procedure

This section demonstrates our photo-cleaning procedure. It covers two primary cleaning techniques: mechanical and contextual.

In mechanical cleaning, we start with an object recognition task on each photo with a convolutional neural network (CNN) VGG19 model (Simonyan & Zisserman, 2014). We eliminate 46,000 photos with suits (representing people) and road signs (representing a majority firm's logotypes). Second, we use Python implementation of open-source Tesseract 5.0 software to detect text on photos. Photos with text represent mostly notes, logotypes, or other unrelated objects to the firms' business. We eliminate 915k photos with any text. Third, we use Open Source Computer Vision Library (Bradski, 2000) to identify people's faces with CascadeClassifier. We remove 45,000 photos representing the front of people's faces. Fourth, we find that photos with a dominating white background typically represent graphs, and photos with extreme size proportions represent some logotypes still in the database. We remove 50,000 photos with more than 90% of white points and photos with any height-to-width proportions greater than 2.5:1 and 1:2.5. After mechanical cleaning, we stay with 900,000 photos.

Next, we introduce a technique that we call contextual cleaning. After mechanical cleaning, we have relatively clean photos; however, many represent objects unrelated to the company's business activity. We call these photos unclear. With contextual cleaning, we distill photos depicting specific objects related to the firm's business activity (clear) from unclear. This procedure involves three steps: 1) create a training dataset consisting of explicit photos that are correctly linked to several defined industries; 2) train a CNN model[31];—and 3) use CNN to distinguish clear photos from unclear ones with a 95% probability on the testing sample.

[31] We describe in detail the implemented CNN network architecture in Section 3.3

We build a testing dataset with photos based on the Fama-French 49 industries classification (FF49). Industries from FF49 are linked with SIC subsectors names, creating a dictionary of relations, where phrases from SIC subsector names are unique keys and FF49 industries are values. Our dictionary covers 560 phrases directly related to FF49 sectors. We download roughly 100 photos representing each phrase from Google, giving us 60k images. To further improve the quality of this dataset, we manually investigate each photo and eliminate pictures unrelated to the industry. After database cleaning, we have close to 12k images from 34 classes. Next, we go to step two of contextual cleaning and train the CNN model. Finally, we use the set of 900k photos as the testing sample and split photos for clear and unclear. The final collection of clear images consists of 423k photos. We use 204k photos from 2009 to 2013 as the training sample and 219k to create a time-varying industry classification.

A.2 Tables and Figures

Table A1: Ratios definitions

The table provides definitions of variables used in this study. Panel A presents the ratios that use market data which we calculate monthly. Panel B shows ratios based only on accountancy information that we update with quarterly frequency. In Panel C, we show some additional variables that are used as interim steps to calculate some ratios from Panel A or B. We download all financial information from Compustat and market data from CRSP. All lowercase variables in column Definition present a symbol in Compustat.

Variable Name	Full Ratio Name	Updates	Definition
PANEL A: Set of Ratios Using Market Information			
MARKET to BOOK	MARKET to BOOK ratio	monthly	Book assets (atq) minus BOOK_EQUITY plus MARKET_CAP all divided by TOTAL_ASSETS.
PRICE to BOOK	Price-to-Book	monthly	MARKET_CAP divided by total common equity (ceqq).
MONTHLY RET	Monthly return	monthly	Monthly return from CRSP
FORECASTED MONTHLY RET	Forecasted monthly return	monthly	Monthly return one month in the future from CRSP.
MARKETLEVG	Market Leverage	monthly	BOOK_DEBT divided by TOTAL_ASSETS minus BOOK_EQUITY plus MARKET_CAP.
EV to SALES	Enterprise Value-To-Sales	monthly	The sum of MARKET_CAP, long-term debt (lltq), and debt in current liabilities (dlcq) all divided by NET_SALES.
PE	Price-to-Earnings	monthly	MARKET_CAP divided by the sum of the latest four quarter reported net income before extraordinary items (ibq).
BETA	Beta (36 months)	monthly	Beta from the single index model based on monthly returns over the previous 36 months downloaded from WRDS.
MARKET CAP	Market capitalization	monthly	Average price times shares outstanding (SHROUT), prices as average from bid offer (BID) and ask (ASK), all divided by 1,000. Data from CRSP.
TOBIN Q	Tobin's Q	monthly	MARKET_CAP plus long-term debt total (dlttq) plus debt in current liabilities (dlcq) all divided by TOTAL_ASSETS
PANEL B: Set of Ratios Using Only Accountancy Information			
TOTAL ASSETS	Total assets	quarterly	atq
NET SALES	Net Sales	quarterly	Sum of the latest four quarterly reported net sales (saleq) from Compustat.

Table A1: (continued)

Variable Name	Full Ratio Name	Updates	Definition
DIV PAYOUT	Dividend Payout	quarterly	Sum of the latest four quarter reported dividend per share (dvpsxq) divided by sum of the latest four quarter reported earnings per share (epspxq).
PROFIT MARGIN	Profit Margin	quarterly	Sum of the latest four quarter reported net operating income after depreciation (oiadpq) divided by NET_SALES.
DEBT to EQUITY	Leverage	quarterly	Total liabilities (ltq) divided by total stockholders' equity (seqq).
SALES GROWTH	Sales Growth	quarterly	The logarithm of (net sales 1 year in the future divided by current value NET_SALES).
R&D to SALES	Scaled R&D Expense	quarterly	Sum of the latest four quarter reported research and development expense (xrdq) divided by NET_SALES.
R&D GROWTH	R&D Growth	quarterly	The logarithm of (sum of the next four quarter reported research and development expense (xrdq) divided by sum of the latest four quarter reported research and development expense (xrdq)).
SG&A to EMPLOYEES	SG&A Expansion	quarterly	Sum of the latest four quarter reported selling, general and administrative expenses (xsgaq) divided by a number of employees (emp).
SG&A GROWTH	SG&A Growth	quarterly	The logarithm of (sum of the next four quarter reported selling, general and administrative expenses (xsgaq) divided by the sum of the latest four quarter reported selling, general and administrative expenses (xsgaq)).
FORECASTED EPS	Forward EPS (next fiscal year)	quarterly	Earning per share for fiscal year 1 from IBES
EPS GROWTH	Forward to trailing EPS	quarterly	The logarithm of (earning per share for fiscal year 1 from IBES divided by the sum of the latest four quarter reported earnings per share (epspxq)).
DEBT to ASSETS	Book Leverage	quarterly	BOOK_DEBT to TOTAL_ASSETS.
RNOA	Return on Net Operating Assets	quarterly	Sum of the latest four quarterly reported net operating income after depreciation (oiadpq) divided by the sum of property, plant, and equipment (ppentq) and current assets (actq), less current liabilities (lctq).

Table A1: (continued)

Variable Name	Full Ratio Name	Updates	Definition
ROE	Return on Equity	quarterly	Sum of the latest four quarter reported net income before extraordinary items (ibq) divided by COMMON_EQUITY.
ASSETS to SALES	Asset Turnover	quarterly	TOTAL_ASSETS divided by NET_SALES.
PANEL C: Other Variables Used as Input to Calculate Ratios			
COMMON EQUITY	Common Equity	quarterly	ceqq
BOOK EQUITY	Book Equity	quarterly	Stockholders' equity (seqq) minus preferred stock liquidating value (pstkq) plus balance sheet deferred taxes and investment tax credit (txditcq).
BOOK DEBT	Book Debt	quarterly	TOTAL_ASSETS minus BOOK_EQUITY.

Table A2: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Market Information for All Stocks

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable $vble$ represents 10 ratios using market information: 1) MARKET to BOOK; 2) PRICE to BOOK; 3) MONTHLY RET; 4) FORECASTED MONTHLY RET; 5) MARKET LEVG; 6) EV to SALES; 7) PE; 8) BETA; 9) MARKET CAP; and 10) TOBIN'S Q for each firm i at month t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry ind excluding firm i at month t. We regress FORECASTED MONTHLY RET on the industry average from MONTHLY RET. In Panel A we compare two classifications based on Image Industries with 25 classes (non-unique and unique) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes. In Panel B we compare two classifications based on Image Industries with 50 classes (non-unique and unique) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes. In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers all stocks from NYSE, AMEX, and NASDAQ from 2016 to 2021. We calculate regression for industries that have at least five members. Table A1 in Appendix shows details of ratios calculation.

Industry Classification Name	MARKET to BOOK	PRICE to BOOK	MONTHLY RET	FORECASTED MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q
PANEL A: Image Industries 25 - comparison										
Image Industries 25	29.3	20.8	25.2	2.3	33.8	24.5	9.6	33.5	38.1	27.8
Image Industries Unique	30.2	18.5	24.3	2.4	35.9	24.1	10.0	34.2	36.9	28.2
Industry_MG	22.8	20.9	29.3	3.4	23.5	24.7	13.5	32.7	34.5	22.6
Fama-French 30 Industries	23.4	20.2	29.8	3.5	24.6	24.9	12.5	30.9	35.4	23.2
NAICS Industries	23.9	21.5	28.9	3.4	24.7	24.0	12.4	32.2	31.5	23.4
4-digit GICS	28.9	23.1	26.6	1.9	28.9	25.4	12.4	32.0	40.8	27.7
Icode 25 Industries	24.6	20.4	25.2	1.9	29.0	25.3	10.1	30.3	36.7	22.9
PANEL B: Image Industries 50 - comparison										
Image Industries 50	31.0	21.4	26.2	2.3	36.6	25.8	10.9	34.1	38.6	29.4
Image Industries 50 Unique	31.5	20.6	24.8	2.3	38.1	25.6	10.8	31.6	39.6	29.8
2-digit SIC	24.6	19.6	30.3	3.5	25.8	22.9	13.1	31.5	36.6	24.5
Fama-French 48 Industries	25.2	19.4	30.4	3.4	26.3	23.5	12.8	31.3	37.1	24.8
3-digit NAICS	24.5	19.6	30.1	3.6	27.2	22.4	13.0	30.9	34.2	23.4
6-digit GICS	28.8	21.7	26.6	1.9	29.2	26.1	12.4	32.1	40.1	27.4
Icode 50 Industries	24.9	20.0	24.9	2.1	28.6	24.4	11.7	29.3	36.6	24.1

Table A3: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Accountancy Information for All Stocks

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable $vble$ represents 16 ratios using accountancy information: 1) TOTAL ASSETS; 2) NET SALES; 3) DIVIDEND PAYOUT; 4) PROFIT MARGIN; 5) DEBT to EQUITY; 6) SALES GROWTH; 7) R&D EXPENSE to SALES; 8) R&D GROWTH; 9) SG&A to # EMPLOYEES; 10) SG&A GROWTH; 11) FORECASTED EPS; 12) EPS GROWTH; 13) DEBT to ASSET; 14) RNOA; 15) ROE; and 16) ASSETS to SALES for each firm i at quarter t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry ind excluding firm i at month t. In Panel A we compare two classifications based on Image Industries with 25 classes (non-unique and unique) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes. In Panel B we compare two classifications based on Image Industries with 50 classes (non-unique and unique) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes. In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers all stocks from NYSE, AMEX, and NASDAQ from 2016 to 2021. We calculate regression for industries that have at least five members. Table A1 in Appendix shows details of ratios calculation.

Industry Classification Name	TOTAL ASSETS	NET SALES	DIV PAYOUT	PROFIT MARGIN	DEBT to EQUITY	SALES GROWTH	R&D EXPENSE to SALES	R&D GROWTH	SG&A to # EMPLOYEES	SG&A GROWTH	FORECASTED EPS	EPS GROWTH	DEBT to ASSETS	RNOA	ROE	ASSET to SALES
PANEL A: Image Industries 25 - comparison																
Image Industries 25	44.8	44.5	3.9	24.2	15.5	44.6	22.7	27.7	21.5	42.1	46.3	16.7	28.8	17.7	16.6	20.7
Image Industries Unique	43.8	44.9	5.6	24.1	16.0	43.7	24.4	30.6	23.1	42.7	47.1	16.6	28.7	19.7	18.1	20.5
Industry_MG	35.0	38.8	4.7	24.5	13.8	31.5	17.1	15.8	26.9	25.3	30.3	13.8	22.4	16.5	14.2	21.7
Fama-French 30 Industries	34.5	39.5	4.1	19.4	13.3	31.2	18.2	22.7	25.5	25.4	31.3	14.0	23.3	13.4	11.9	22.4
NAICS Industries	27.8	36.9	4.8	23.4	11.3	31.5	19.1	23.0	24.5	27.4	28.9	16.0	20.4	14.0	13.7	21.8
4-digit GICS	51.5	47.2	4.1	25.2	13.4	32.2	21.5	20.7	26.9	25.1	40.1	16.7	26.1	13.7	14.1	25.3
Icode 25 Industries	43.7	39.8	4.5	22.7	13.7	33.0	24.1	25.7	25.1	27.3	43.9	18.1	23.5	15.6	14.9	26.5
PANEL B: Image Industries 50 - comparison																
Image Industries 50	42.9	45.4	4.8	24.5	17.0	48.2	24.2	31.8	22.7	42.1	47.7	15.8	28.3	18.4	17.9	20.2
Image Industries 50 Unique	43.1	48.4	6.0	28.2	16.9	46.9	29.7	32.0	21.2	42.6	47.6	17.4	27.9	22.2	18.1	21.9
2-digit SIC	32.6	39.1	4.2	21.6	12.3	31.8	19.0	21.9	24.2	24.9	31.4	14.6	21.6	13.7	12.6	21.4
Fama-French 48 Industries	32.8	40.0	3.9	21.5	12.1	32.1	19.0	22.6	24.0	25.3	30.3	14.7	21.6	13.4	12.0	21.7
3-digit NAICS	25.5	32.7	4.7	20.4	13.5	31.7	18.2	24.4	23.0	27.6	32.9	16.4	20.3	15.2	14.3	20.9
6-digit GICS	48.8	46.3	4.3	23.6	13.2	32.9	21.6	21.1	25.5	24.9	40.8	16.7	24.7	14.1	13.3	25.1
Icode 50 Industries	42.1	40.4	4.4	25.0	15.1	32.8	24.7	22.5	23.6	27.8	43.7	18.4	23.7	15.8	14.7	25.4

Table A4: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Market Information for Stocks with Prices Not Smaller than USD 5

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable $vble$ represents 10 ratios using market information: 1) MARKET to BOOK; 2) PRICE to BOOK; 3) MONTHLY RET; 4) FORECASTED MONTHLY RET; 5) MARKET LEVG; 6) EV to SALES; 7) PE; 8) BETA; 9) MARKET CAP; and 10) TOBIN'S Q for each firm i at month t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry ind excluding firm i at month t. We regress FORECASTED MONTHLY RET on the industry average from MONTHLY RET. In Panel A we compare two classifications based on Image Industries with 25 classes (non-unique and unique) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes. In Panel B we compare two classifications based on Image Industries with 50 classes (non-unique and unique) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes. In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers stocks classified with Image Industries from 2016 to 2021 with prices not smaller than USD 5. We calculate regression for industries that have at least five members. Table A1 in Appendix shows details of ratios calculation.

Industry Classification Name	MARKET to BOOK	PRICE to BOOK	MONTHLY RET	FORECASTED MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q
PANEL A: Image Industries 25 - comparison										
Image Industries 25	28.3	19.2	26.1	3.3	31.6	20.8	10.4	29.7	35.3	26.9
Image Industries Unique	28.3	17.3	25.1	3.2	32.1	23.3	10.4	31.8	33.5	26.1
Industry_MG	26.5	18.3	27.5	2.7	28.4	25.5	10.5	34.1	30.4	25.6
Fama-French 30 Industries	26.7	17.9	28.2	2.7	28.2	26.1	9.2	32.0	31.4	26.0
NAICS Industries	27.9	20.9	26.9	2.9	27.3	24.9	9.7	28.8	35.4	26.8
4-digit GICS	26.5	17.4	29.7	2.5	27.7	25.7	7.7	30.8	35.8	25.7
Icode 25 Industries	24.7	17.1	27.2	2.9	27.3	24.4	9.3	30.6	31.2	23.9
PANEL B: Image Industries 50 - comparison										
Image Industries 50	28.4	21.2	27.2	3.1	33.5	23.7	11.1	29.5	34.8	26.9
Image Industries 50 Unique	29.0	19.6	26.0	3.1	33.3	23.0	11.2	27.5	33.8	27.2
2-digit SIC	28.1	18.9	29.5	2.6	30.2	28.8	10.0	31.3	31.7	26.8
Fama-French 48 Industries	28.8	17.9	29.7	2.6	29.9	30.6	9.8	31.3	30.8	27.6
3-digit NAICS	27.7	19.9	28.2	3.2	29.3	26.9	10.6	29.8	31.7	26.0
6-digit GICS	29.9	20.5	31.2	2.0	29.8	30.5	8.9	31.6	35.6	27.9
Icode 50 Industries	26.6	18.0	28.1	3.0	29.9	28.6	10.0	27.5	31.4	25.4

Table A5: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Accountancy Information for Stocks with Prices Not Smaller than USD 5

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable $vble$ represents 16 ratios using accountancy information: 1) TOTAL ASSETS; 2) NET SALES; 3) DIVIDEND PAYOUT; 4) PROFIT MARGIN; 5) DEBT to EQUITY; 6) SALES GROWTH; 7) R&D EXPENSE to SALES; 8) R&D GROWTH; 9) SG&A to # EMPLOYEES; 10) SG&A GROWTH; 11) FORECASTED EPS; 12) EPS GROWTH; 13) DEBT to ASSET; 14) RNOA; 15) ROE; and 16) ASSETS to SALES for each firm i at quarter t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry ind excluding firm i at month t. In Panel A we compare two classifications based on Image Industries with 25 classes (non-unique and unique) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes. In Panel B we compare two classifications based on Image Industries with 50 classes (non-unique and unique) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes. In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers stocks classified with Image Industries from 2016 to 2021 with prices not smaller than USD 5. We calculate regression for industries that have at least five members. Table A1 in Appendix shows details of ratios calculation.

Industry Classification Name	TOTAL ASSETS	NET SALES	DIV PAYOUT	PROFIT MARGIN	DEBT to EQUITY	SALES GROWTH	R&D EXPENSE to SALES	R&D GROWTH	SG&A to # EMPLOYEES	SG&A GROWTH	FORECASTED EPS	EPS GROWTH	DEBT to ASSETS	RNOA	ROE	ASSET to SALES
PANEL A: Image Industries 25 - comparison																
Image Industries 25	33.8	36.5	4.5	20.3	15.1	44.2	17.1	28.3	21.8	40.5	45.7	16.0	25.6	16.7	15.2	18.7
Image Industries Unique	32.6	35.2	6.2	22.0	16.8	45.1	20.4	29.0	21.2	41.5	45.3	16.6	25.2	20.2	16.0	20.6
Industry_MG	30.9	32.7	5.3	21.6	16.8	40.3	19.2	21.8	22.2	32.8	43.4	14.7	27.9	20.3	16.4	23.1
Fama-French 30 Industries	32.4	32.7	4.9	23.6	15.2	38.0	21.5	19.4	19.9	33.2	41.4	15.2	26.4	21.1	14.5	22.5
NAICS Industries	39.5	38.0	3.7	17.6	13.4	40.8	16.2	26.8	23.3	33.7	43.6	16.1	29.1	22.2	14.2	18.2
4-digit GICS	35.1	31.9	3.9	18.7	13.4	40.3	18.2	23.3	22.0	32.9	42.0	15.4	27.2	20.2	13.6	23.0
Icode 25 Industries	31.0	34.1	5.1	22.7	14.9	40.3	20.9	23.9	23.8	34.6	42.8	17.0	26.3	21.0	15.4	24.1
PANEL B: Image Industries 50 - comparison																
Image Industries 50	33.2	36.2	5.7	22.7	16.3	47.0	21.8	28.8	21.4	40.7	45.8	15.8	24.9	18.0	16.4	19.9
Image Industries 50 Unique	32.8	36.2	6.6	24.5	18.4	45.9	25.2	28.8	20.1	39.9	43.7	16.7	23.9	19.3	17.5	20.9
2-digit SIC	34.7	36.3	5.9	25.7	16.0	39.7	22.5	21.8	21.2	31.0	43.1	15.8	25.2	22.3	15.5	27.6
Fama-French 48 Industries	35.5	34.5	6.1	25.8	15.4	40.4	22.6	23.5	21.1	32.8	41.2	15.1	25.5	23.0	15.4	28.3
3-digit NAICS	32.7	34.2	5.7	24.5	17.5	40.6	25.5	22.9	21.6	33.4	41.5	15.6	25.6	22.2	18.6	23.7
6-digit GICS	38.5	37.7	4.7	23.1	14.5	41.5	21.7	21.9	21.8	32.3	44.8	16.2	25.3	24.0	15.8	26.9
Icode 50 Industries	30.6	33.2	6.2	24.4	15.9	41.4	21.3	23.2	22.1	35.1	44.0	17.4	27.2	22.1	18.2	25.3

Table A6: Industry Momentum - Volatility Targeting

The table compares Sharpe ratios of momentum industry portfolios built with different industry classification techniques formed with 1) unique image industries with 25 classes (Image Industries), 2) Moskowitz and Grinblatt (1999) (Industry_MG), 3) Fama-French industries with 30 classes, 4) NAICS industry classification with 20 classes, 5) four digits GICS codes, and 6) transitive Hoberg and Phillips (2016) classifications with 25 classes (Icode 25 Industries). We build an industry momentum strategy in the same way as Moskowitz and Grinblatt (1999) by investing long (short) in three industries with the highest (lowest) six, nine, or twelve months momentum. We extend industry momentum with the volatility targeting procedure proposed by Barroso and Santa-Clara (2015). The returns of industries are calculated as market-weighted (Panel A) or equally weighted (Panel B). In each row, the highest Sharpe ratio is marked as dark green, the second highest as light green, and the third highest as beige. The sample covers the years 2016-2021. We use sectors with at least five stocks for the Image Industry.

	Image Industries	Industry_MG	Fama-French 30 Industries	NAICS Industries	4-digit GICS	Icode 25 Industries
PANEL A: Market weighted industry returns						
Momentum 6m 6m	0.513	0.093	0.045	-0.086	0.996	-0.189
Momentum 6m 9m	0.439	-0.068	0.002	-0.185	0.807	-0.211
Momentum 6m 12m	0.488	-0.228	0.110	-0.511	0.787	-0.029
PANEL B: Equally weighted industry returns						
Momentum 6m 6m	0.445	0.535	0.505	0.137	0.092	0.4
Momentum 6m 9m	0.213	0.345	0.330	-0.075	-0.119	0.07
Momentum 6m 12m	0.101	0.210	0.171	-0.185	-0.139	0.085