Just Look: Knowing Peers with Image Representation*

Tomasz Kaczmarek† Kuntara Pukthuanthong‡

April 8, 2024

Abstract

What an industry looks like? 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 Firm Similarities (IFS). IFS 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 efficiency of pair trading strategies, diversification benefits, and industry momentum profits. This success is attributed to high investor agreement within an industry, leading to substantial aggregated demand and supply effects on stock prices. IFS excels in industries characterized by expected growth and intangibility.

Keywords: Images, firm similarities, Diversification, Industry momentum

JEL Codes: G00, G11, G12

*Code and data are available upon request. We will post and update the data of firm similarities on our websites. We thank Fred Bereskin, Kate Holland, Dhagash Mehta, Mike O'Doherty, and seminar participants at the University of Missouri Columbia, Missouri State University, University of Missouri St. Louis, the Fields Institute for Research in Mathematical Sciences at the University of Toronto, Poznan University, and Blackrock

†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 the 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

What does an industry *look* like? How does the image-based industry classification diverge from conventional methods? Capital market research often necessitates grouping enterprises into more homogeneous categories, with firm similarities as a key criterion. Researchers utilize industry groups to narrow their analysis, select control firms, benchmark performance, and provide descriptive statistics. Meanwhile, investors seek peers for portfolio diversification, strategy formulation, and valuation. Traditional industry classification methods, such as SIC, NAICS, and GICS, segment firms based on production processes, supply chains, or financial metrics. Alternatives explore business descriptions in 10-K reports or analyst coverage. However, these methods may not capture the full spectrum of similarities, leading to diverse groupings of similar companies. Our study introduces an innovative approach: classifying industries based on visual data, offering a fresh perspective on firm similarities and industry dynamics.[1]

Existing classifications do not utilize images to identify similarities among firms, even though images are the most efficient and intuitive medium for human cognition. In an era of information overload, leveraging our natural ability to process visual information can enhance efficiency. Research by Potter, Wyble, Hagmann, and McCourt (2014) reveals that the human brain can process images in as little as 13 milliseconds, with 90% of information transmitted to the brain being visual. This results in images being processed 60,000 times faster than text, a phenomenon known as the picture superiority effect, which underscores the greater memorability of images compared to words.

[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).

Furthermore, images can convey narratives more effectively than text or numbers alone, capturing attention and engaging viewers more. They encapsulate intricate details, emotions, and scenes in a single snapshot, conveying information more effectively and efficiently than text or numbers. Obaid and Pukthuanthong (2022) found that news photos have a higher persuasive power and predictive ability than text. Visuals provide refined information that may be overlooked or too burdensome to process in lengthy texts or dull figures, and they can freeze moments in time, capturing fleeting expressions or gestures that may not be fully captured by other data forms. A photo captures multi-dimensional information efficiently and effectively, making it useful for clustering industries.

In this study, we introduce a novel approach for measuring firm similarity, utilizing graphical objects that encapsulate the essence of their business operations. We term this approach Image-based Firm Similarity (IFS).[2] These objects often represent a company's products but can also depict elements like final products in the supply chain, raw materials, or other business-related visuals. We validate the efficacy of IFS through the R^2 values obtained from financial ratio regressions of individual firms against their respective industries. IFS stands out due to its unparalleled adaptability and responsiveness, surpassing traditional industry classifications. Images, being inherently more engaging and requiring less cognitive processing than text or numbers, enhance the appeal and effectiveness of IFS, particularly for firms with distinctive products. This characteristic makes IFS especially suited for capturing innovation through new product offerings or significant product enhancements.

Building on the innovative framework of IFS, we create firm similarities with varying levels of detail, specifically through classifications of 45 and 73 classes.[3] The effectiveness of IFS is rooted in its reflection of the human brain's natural propensity to group visually similar objects (Shen, Horikawa, Majima, & Kamitani, 2019). This cognitive process is supported by research from Branthwaite (2002) and Dewan (2015), highlighting visual communication's

[2]Throughout this study, the terms photos, images, photographs, and pictures are used interchangeably. Our IFS encompasses all these visual representations.

[3]In the classification with 45 (73) classes, there are 25 (50) industries with at least 5 firms during the formation period (2009-2013).

significant impact on human decision-making.

In financial markets, where investment strategies are often influenced by short-term fluctuations, stock categorization based on visual similarities can provide investors with a more intuitive and relatable framework. When evaluating a firm and its competitors, investors are likely to draw on the visual representation of products, which is increasingly relevant in today's image-centric digital landscape. Our analysis reveals that stock categorization based on visual similarities effectively captures this innate human response, with industry momentum strategies as prime examples.

Furthermore, we build upon Hoberg and Phillips (2016), who identify competitors through textual analysis by employing the more direct and unambiguous language of images (Branthwaite, 2002; Dewan, 2015). IFS is predicated on the notion that photographs can capture complex relationships and showcase product offerings in ways that text and numbers cannot. Our methodology yields R^2 values that are competitive with other classification techniques, affirming the robustness of IFS. Additionally, the relatively low cross-correlation of stock returns across industries, as evidenced by superior Sharpe and Calmar ratios, underscores the enhanced diversification benefits offered by our image-based approach.

Practical industry classifications should ideally satisfy several criteria, as outlined by (Kaustia & Rantala, 2021)(henceforth, KR). These include a balance of observable and unobservable traits, genuine economic linkages within groups, adaptability to changes in business and economic structures, simplicity in comprehension and implementation, and reasonable implementation costs. Our Image-based Firm Similarity (IFS) aligns with these requirements. Firstly, IFS captures the fundamentals, such as product resemblance, and intangibles, such as color, taste, and sentiment. Secondly, it emphasizes economic linkages through product similarity rather than financial ratios. Thirdly, IFS's dynamic nature, powered by the human brain's rapid processing of images, allows it to adapt to changes in product offerings quickly. This makes it a valuable tool for investors seeking to leverage pair trading strategies, diversify portfolio risk, or maximize the Sharpe ratio while capitalizing

on industry momentum. We provide the data on our website to facilitate broader usage, ensuring accessibility and ongoing updates for academics and practitioners.

We primarily use Google Images for image collection due to its extensive repository of firm-level photos.[4] Although company annual reports could be a source, they typically contain a limited number of high-quality images, often supplemented by graphs and tables rather than visuals.

Furthermore, images from annual reports are often biased, presenting firms in a favorable light and resembling advertisements, making them unsuitable for sector grouping. Our machine learning-based approach leverages Google's extensive image archive, allowing for more accurate outcomes and deeper insights as the algorithm accesses a broader range of images.

We introduce a novel algorithm that captures similarities between entities represented by diverse images, particularly focusing on companies depicted through varied and numerous visuals. We employ machine learning algorithms to extract characteristic features from over two million photos. We utilize the state-of-the-art Deep Convolutional Neural Network (CNN), specifically the VGG19 model (Simonyan & Zisserman, 2014), along with other machine learning techniques such as transfer learning and object recognition. This comprehensive analysis of image-based similarities between companies allows us to establish a statistically significant measure of similarity based on visual data, enriching our understanding of corporate relationships through image analysis.

Defining similarities between companies based on imagery presents a formidable challenge, primarily due to the diverse nature of the input photos. A significant portion of the images downloaded from Google, despite the total exceeding two million, are tangentially related to the enterprises' product offerings. We find that more than 60% of these images require removal—ranging from logos and faces to contextually irrelevant pictures—before we can begin forming firm similarities. This cleaning process is crucial to accurately represent-

[4]While news sources are another option, they often have limitations, such as a focus on prominent corporations and a predominance of logos over meaningful images.

ing a company's products through images. The task is complicated by the varied nature of the imagery associated with each company.

For instance, Coca-Cola's images might showcase its iconic delivery trucks, misleadingly suggesting a categorization within transportation, while the prevalence of iPads in corporate imagery could erroneously imply that a vast swath of NYSE-listed companies belongs to the electronics manufacturing sector. Such scenarios highlight the limitations of unsupervised algorithms, which might inadvertently focus on random objects and the inadequacies of semi-supervised or supervised learning when used alone to align companies with industries based on non-imagery criteria. Our method incorporates several targeted steps to mitigate the risk of misclassification when using images for industry classification. Initially, we pre-process images to filter out common but non-indicative objects, such as delivery trucks, in company photos. We then employ feature extraction techniques to distill essential visual elements that better represent a company's primary business.

Dimensionality reduction is applied to simplify the image data, emphasizing meaningful visual patterns. Subsequent clustering groups companies based on these refined visual features, with clusters continually adjusted for accuracy. Finally, we rigorously validate our IFS against traditional classifications and, through out-of-sample testing, use historical images to predict future classifications. This ensures our approach is precise, economically relevant, and free from look-ahead bias.

Out-of-sample (OOS) testing is a key element of our study, emphasizing our commitment to addressing concerns of look-ahead bias and data mining in finance. This procedure ensures the predictive accuracy and economic relevance of Image Firm Similarities (IFS) across various market conditions and timeframes. By forecasting IFS for future periods using historical image data, such as predicting IFS for 2018-2019 with images from 2015-2017 and, similarly, for 2020 to 2021 using images from 2017 to 2019, we demonstrate the reliability of our image-based classifications. The effectiveness of our method in capturing the dynamic nature of firms and their economic environments is validated by comparing our results with established

industry classifications like SIC, NAICS, and GICS. This comparison highlights the ability of IFS to detect subtle yet economically significant similarities among firms through visual data.

We validate our measure using R^2 of the firm's financial ratio regression on the same ratio of the firm's corresponding industry. We apply the same financial ratios used by (Kaustia & Rantala, 2021). Our benchmarks include SIC, NAICS, and GICS-based classification with granularity closest to IFS. 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 IFS, is created based on quantitative methods and uses the companies' product offerings as the main similarity measure. IFS exceeds both HP and GICs for most accountancy-based ratios, and, among 16 financial rations, it is the strongest in nine ratios, including net sales, dividend payout ratio, profit margin, debt to equity ratio, sales growth, R&D expense to sales, R&D growth, SG&A growth, and EPS growth. It performs worse in the market-based ratios group, where, among ten ratios, it achieves the best results for beta, forecasted returns, or PE ratio, and for others, it performs on the level of SIC-based classifications, but its performance is worse than HP and GICS.

Notably, IFS performs well in ratios that capture expectation and growth. High R^2 from forecasted market returns and forecasted EPS prove the ability to cluster firms with similar market expectations. Amazingly, IFS has R^2 of R&D growth and sales growth, even 3-5 percent points higher than that of HPs and GICs; therefore, IFS also captures growth.

In addition to the conventional market and accounting indicators, we present ratios that capture intangibles, including innovation and human capital, such as R&D growth, R&D per unit of sales,[5] and SG&A per employee. IFS surpasses HPs and GICS in all intangibles ratios except SG&A to employees.

[5]the greater its R&D-to-sales ratio, the earlier a firm invests in R&D in its existence.

Our R^2 results corroborate the premise that our image-based firm similarities capture 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 industry classifications. A total of 17%-22% of IFS-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. IFS 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 several examples. SIC (NAICS) reclassified Exxon Mobil Corp from petroleum & coal products to oil & gas extraction in 2017. Our images demonstrate the same reclassification but two years earlier, based on photos from 2013 to 2015. SIC reclassified Hornbeck Offshore Services Inc. from water transportation to oil & gas extraction in 2018, while image-based similarity demonstrated a comparable change in peer structure from 2014 to 2016. Other examples of image agility in reclassification include, e.g., FMC Corp reclassified from industrial machinery & equipment to chemical & allied products, Oshkosh Corp from transportation equipment to industrial machinery & equipment, or SEACOR Holdings Inc. from water transportation to oil & gas extraction. The dexterity of firm similarities is crucial for real-time investment.

Next, we present the three applications that utilize image-based similarities: pair trading strategy based on sales and EPS growth, diversification, and momentum. The pair trading strategy invests in companies with similar profiles defined through imagery, text (HP), and common analysts (KR) and ranks firms according to their growth exposure, adopting long positions in high-growth companies and short positions in those experiencing low growth. The pair trading strategy exploits temporary mispricings between pairs of stocks that have the same similarities. The strategy's effectiveness is assessed using Sharpe and Calmar

ratios, which unequivocally indicate the highest performance when the similarity between companies is based on our image-based similarity metric.

For diversification purposes, our image firm similarities with 45 classes have the top three Sharpe and Calmar ratios 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 perform particularly strongly for equal- and value-weighted portfolios but do not work well with optimized portfolios. The outcome validates that the industries identified by IFS are distinctive and potentially provide diversification advantages.

We show that IFS presents the fastest dynamics compared to other industry classification approaches. Specifically, IFS has over 200% faster dynamics than HP. There are several reasons to explain this.

First, the core of IFS's dynamism lies in its ability to rapidly adapt to and reflect the changes in companies' activities, strategies, and market positions through visual cues. Unlike traditional systems like SIC, NAICS, and GICS, which primarily analyze textual descriptions and numerical data that may lag behind real-world changes, IFS leverages the latest in machine learning and image processing technologies to analyze visual content. This method captures real-time shifts in a company's product offerings, branding, and consumer engagement strategies.

Second, more immediate and intuitive visual data provides a direct window into a company's current state, bypassing the delays inherent in updating textual descriptions or waiting for new financial reports.

Third, the human brain processes images much faster than text, mirroring the IFS's swift adaptability in recognizing and categorizing industry dynamics based on visual similarities. This approach enables a more responsive classification system and taps into the unstructured data realm, offering a richer, more nuanced understanding of industry affiliations and evolutions. Consequently, by harnessing the untapped potential of visual data, the IFS methodology establishes itself as uniquely dynamic and capable of keeping pace with

the rapid and multifaceted changes characteristic of today's business environments.

Turning to investment strategies based on firm similarities, we adhere to the techniques proposed by the industry momentum of Moskowitz and Grinblatt (1999) and the volatility-adjusted momentum of Barroso and Santa-Clara (2015), and short-term reversal. We demonstrate that our 45 classes deliver the highest Sharpe ratios for equal- and value-weighted industries. In addition, strategies built on IFS 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 IFS is directly due to industry momentum rather than the momentum of individual companies.

What explains the functionality of IFS? 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 IFS provides reasonable agreement within an industry. This characterization describes the benefit of adopting IFS in pair trading, diversification, and momentum strategies.

This project aims to develop a new categorization metric for industries. We do not assert that we are the best industry classification. Our IFS has some limitations as follows. First, our IFS does not perform well with a business that images, such as consulting firms and abstract technologies, cannot represent. Our IFS thrives in the industries that manufacture products or services, such as oil drilling, waste management, etc. Second, we cannot achieve high granularity, as is available in SIC codes. For example, 3 and 4-digit SIC codes have as many as 300 to 400 clusters. The IFS does not go beyond 73 clusters. Third, IFS can only go back a few years due to image availability.

Our contributions are fourfold. First, we contribute to the industry classification liter-

ature. In the next section, we explain the distinction between methods in detail. 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 the neuroscience literature. We show that the human brain can cluster photos better than text and numbers. We also contribute to 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. Industry classification is based on business activities, while firm similarities can be based on characteristics. We thus call our technique as image firm similarities. 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. On the other hand, IFS 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. Finally, while other approaches provide qualitative classifications, our output presents the score or the rank of how each firm is similar to others within an industry.

Both text- and analyst-based groups, described below, have an advantage over SIC, NAICS, and GICS as they are non-transitive and subject to yearly revision. On the other hand, SIC, NAICS, and GICS are transitive—making them appropriate for many applications such as industry momentum and diversification, while the non-transitive approach cannot.[6] The applications of the non-transitive approach include pair trading and comparable firm valuation.

The text-based groups were pioneered by Lewellen (2012) and Rauh and Sufi (2012), and it was formalized and extended by Hoberg and Phillips (2016) (HP). Ibriyamova, Kogan, Salganik-Shoshan, and Stolin (2019) extend the application to companies' brief descriptions. In this study, we compare our IFS 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

[6]Transitivity implies that for any firm, A and B, in the same industry, a firm C in A's industry is also in B's. HP provides two types of classifications: transitive and non-transitive. 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.

variable is a dummy variable that provides only two possible states.[7]

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. Thus, we use approximately 100 images to categorize a stock. Unlike HP, however, each of our variables (images) is a three-dimensional 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.[8]

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. Therefore, we only include their peer groups as our benchmark in this study for the pair trading simulation. Notably, their high granularity of industry groups likely accounts for their excellent R^2 scores.

The other non-transitive approaches including clustering by characteristics (He, Wang, and Yu, 2021), technologies (Lee, Sun, Wang, and Zhang, 2019), and Edgar search (Lee, Ma, and Wang, 2015).

[7]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.

[8]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 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 we use as input are incomparable. For further explanation about pixels, see https://levelup.gitconnected.com/pixels-arrays-and-images-ef3f03638fe7

3 Methodologies and Data

Our study presents a refined approach to industry classification, leveraging the visual data inherent in company product images. This methodology unfolds through a sequence of stages, initially identifying similarities between companies based on image analysis. This is followed by creating industry clusters and concludes with periodic updates based on new data. This approach provides a dynamic and nuanced understanding of industry relationships, offering a contemporary lens through which to examine firm connectedness in the digital age.

The core of our method begins with calculating similarities between firms, distilled from comparing their respective image sets. This process is quantified by the following equation, representing the similarity measure between two firms, A and B, based on their image sets:

$$S(A, B) = 1 - \text{distance}(A, B) \tag{1}$$

Here, $S(A, B)$ denotes the similarity score, and $distance(A, B)$ encapsulates the aggregated measure of similarity between the image sets of firms A and B, refined through a dimensionality reduction and pairing process that ensures a focus on substantive visual correlations.

Following identifying firm similarities, we use a clustering procedure to delineate industries, grounded in the premise that firms with higher mutual similarities suggest a shared industry space. The clustering can be formalized as:

$$C = \text{cluster}(S(A, B)) \tag{2}$$

In this equation, C represents the resulting set of industry clusters derived from applying a clustering algorithm to the matrix of similarity scores $S(A, B)$ across all firm pairs. This methodological step is pivotal, as it translates the complex web of pairwise firm similarities into a coherent structure of industries, each defined by the collective visual narrative of its constituent firms.

The dynamic aspect of our methodology allows for the reevaluation and update of these clusters as new images emerge, ensuring that the industry classification remains reflective of the latest market developments. This process underscores the adaptability and relevance of our approach in capturing the evolving landscape of industry affiliations.

Through these streamlined steps, our study advances a novel framework for industry classification, harnessing the untapped potential of image data to reveal deeper insights into firm relatedness and industry dynamics. This methodology challenges traditional classification schemes and enriches our understanding of the visual dimensions that underpin industry relationships in the contemporary business environment.

In the next part of this section, we start by discussing the data we use. Then, we discuss the technology used to define similarities between companies. Furthermore, we demonstrate how we construct Image Industries. Finally, we explain the methodology to update firms assigned to each industry that sets a backbone for the out-of-sample testing.

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. We collect all photos representing firms directly from the Google search engine via Python API. The API allows the collection of 100 photos for a single query. 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.[9] To achieve time-varying classification, we download photos for each year separately. We use the following

[9]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, 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.

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. We retrieve photos for each year from 2009 to 2021. Finally, we collect close to 2 million photos and group our data sample into 3-year rolling windows. 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 always 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 demonstrate a detailed procedure of photo cleaning in Appendix A.1.

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 based on the upload time.

In addition, we build a collection of 19 stock-level financial ratios based on Kaustia and Rantala (2021) and group them into ratios using market information and ratios based on accounting data only. Table A1 in the Internet Appendix demonstrates a detailed calculation methodology for each ratio. We download the SIC and NAICS codes from CRSP, GICS from Compustat, similarity scores with industries based on text classification from Hoberg and Phillips (2016), and similarity scores on common analysts from Kaustia and Rantala (2021) 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 three-year rolling windows.

We build image-based industries in four significant steps. First, we start with the fea-

[10]Compustat delivers static classifications. CRSP gives dynamic but only for SIC and NAICS. Therefore, we have dynamic data from CRSP (SIC, NAICS) and static data from Compustat (GICS). See Table 8

ture extraction procedure from each image. This step is important to define the numerical representation of objects from photos. The process is described in Section 3.2. Second, we move to identify yearly updated firms' similarities based on their image representation from all 3-year rolling window periods. Our procedure of peer definition is demonstrated in Section 3.3. Third, we use data from the image definition training sample to create IFS. The established industries are constant for the entire out-of-sample testing period.[11] Lastly, as companies dynamically change their product offerings depending on market conditions, we update the composition of each industry on a biennial basis. We describe industries' definition and their updates in Section 3.4.

3.2 Extracting content from image

Defining image industries is meticulously designed to pinpoint objects emblematic of each industry, enabling the grouping of companies based on visual similarities in their associated images. This identification is crucial in distinguishing industries by aligning companies with shared visual characteristics, facilitating a more intuitive and visually coherent industry classification system.

The foundational step in this procedure involves the comparison of images between pairs of companies to uncover similarities. Figure 1 demonstrates that the comparison is grounded in the extraction of numerical representations from each image, a task accomplished through the application of the VGG19 model (Simonyan & Zisserman, 2014). The VGG19, a convolutional neural network (CNN) model pre-trained on the ImageNet dataset, is renowned for its effectiveness in image recognition tasks. Its architecture, designed for deep image processing, consists of 19 layers, including 16 convolutional layers, 3 fully connected layers, 5 max-pooling layers, and a softmax output layer to classify images into 1000 categories.

The process of feature extraction with VGG19 unfolds as follows:

[11] Given that access to the photos' history is limited and the testing period lasts eight 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.

- Input Processing: Each image is resized to match the VGG19's input dimensions of 224x224 pixels. The input layer adjusts the image to this uniform size to ensure consistency across the dataset.

- Convolutional Layers: The model employs multiple convolutional layers, each applying filters to the image to capture a range of features, from basic edges and textures in the initial layers to more complex patterns and objects in deeper layers. These layers are represented mathematically as:

$$F_{l+1} = \text{ReLU}\left(W_l * F_l + b_l\right) \tag{3}$$

  where F_{l+1} is the feature map obtained from layer $l+1$, W_l and b_l are the weights and biases of layer l, F_l is the input feature map to layer l, and ReLU denotes the Rectified Linear Unit activation function that introduces non-linearity, enhancing the network's learning capability.

- Feature Vector Extraction: After processing through the convolutional and pooling layers, the image's representation is flattened and passed through fully connected layers, culminating in a 4096-dimensional feature vector. This vector encapsulates the image's essence, encoding the visual information necessary for identifying similarities across images.

With its substantial dimensionality, this feature vector serves as the basis for comparing images between companies, enabling the identification of visually similar photos that suggest shared industry characteristics. By analyzing these vectors across the dataset, our methodology discerns patterns and commonalities that signal industry affiliations, laying the groundwork for clustering companies into image-based industries.

Through this intricate process, we harness the power of deep learning and the rich visual data encapsulated in company-related images to define firm similarities in a manner that

transcends traditional classification approaches. This offers a novel perspective on company grouping based on the visual domain.

3.3 Identification of Firms' Similarity

Building upon the foundation of image comparison through feature extraction via the VGG19 model, our methodology for defining image industries involves refining the comparison process to identify meaningful similarities between firms based on their photos.

The cornerstone of our image comparison process is cosine similarity, a metric adept at capturing the orientation differences between high-dimensional vectors. This measure is crucial for assessing the degree of similarity between images, providing a range that signifies complete similarity to total dissimilarity. The formula for cosine similarity between two vectors, a and b, is given by:

$$\text{cosine\_similarity}(a, b) = \frac{a \cdot b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{4}$$

However, applying cosine similarity directly to the 4096-dimensional feature vectors extracted by the VGG19 model introduces challenges, primarily due to the "curse of dimensionality." This phenomenon, where the space becomes sparse in high dimensions, can obscure meaningful distinctions between images. To mitigate this, we implement Principal Component Analysis (PCA) to reduce the dimensionality of our feature vectors, thereby retaining the most significant aspects of the data while eliminating redundancy. PCA achieves this by transforming the original data into a set of linearly uncorrelated variables (principal components), with the transformation defined as:

$$Z = XP \tag{5}$$

where X represents the matrix of original feature vectors, P denotes the matrix of principal components, and Z is the transformed data in reduced dimensions. We require at least 70%

of variability to be preserved, focusing the analysis on the most informative features of the images.

Upon reducing the dimensionality, the similarity between images is reassessed using the refined feature vectors. This creates a distance matrix for each pair of companies, reflecting the cosine similarities between their images. To systematically pair images from different companies based on these similarities, we solve the linear sum assignment problem, aiming to maximize the overall similarity (or minimize the total distance). The assignment problem can be formulated as:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} \text{cost}_{ij}\, x_{ij} \tag{6}$$

where $cost_{ij}$ is the distance between the i photo from one company and the j photo from another, and x_{ij} is a binary variable indicating whether these photos are matched (Crouse, 2016).

Recognizing that not all photos contribute equally to identifying industry similarities—particularly those with higher distances—we set a threshold to distinguish closely related photos. By adjusting this threshold based on empirical observations and robustness tests, we ensure that only the most relevant matches influence our similarity scores.[12]

Given the potential for random similarities in such a vast dataset, we employ a hypothesis testing framework, where H0 says there's a random connection between companies. We test H0 with Weighted Least Squares (WLS) regression (Strutz, 2016). This approach allows us to account for heteroscedasticity in our data, offering a more accurate model. The WLS regression is expressed as:

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \Sigma) \tag{7}$$

where Y is the vector of observed similarities, X is the design matrix, β represents the regression coefficients, and Σ is the diagonal matrix of weights. By comparing the observed similarity against the 99% confidence interval derived from the WLS model, we can determine

[12]We use the 0.4 threshold in the study. We verified that the results with thresholds 0.35 and 0.45 are comparable.

whether H0 is rejected or, in other words, if the similarity between companies exceeds what would be expected by chance.

To address type I errors in our hypothesis testing, we adopt a strategy that evaluates similarity scores for each year based on photo representations from the latest three years ($t-2$, $t-1$, t). We then analyze similarity ratios between each pair of companies based on their three latest observations. We define similarity in period t as requiring a company to be similar in at least two of the three latest years. This ensures that any single-period similarity is not seen as nonrandom. A nonrandom connection is established if the null hypothesis is rejected in at least two of the three latest years. Furthermore, to reduce type II errors, we require a pair of companies to maintain this similarity for the next period once we establish a nonrandom similarity.

Let's illustrate this with an example. Consider the similarity evaluation between companies A and B using their photos from 2009-2013. To establish a similarity in a given year (denoted as period t), we require photos from the preceding two years ($t-2$ and $t-1$) in addition to the current year (t), allowing us to compute similarities for the periods 2011, 2012, and 2013. Let's suppose that the similarities for 2011 and 2013 are determined to be nonrandom, while that of 2012 is random. In this scenario, two of the three similarities within 2011-2013 are deemed nonrandom. Consequently, we consider companies A and B as peers in 2013. To mitigate type II errors and ensure robustness, we extend this peer relationship to the subsequent year, 2014. This approach necessitates five years to define peer groups, prompting us to utilize photos from 2009 to 2013 to establish the earliest similarities in our dataset. The peer groups defined for 2013 are subjected to an out-of-sample testing approach in the subsequent year, 2014. Furthermore, our testing procedure is designed to update similarities once every two years, aligning with a one-year rewriting period to minimize type II errors.

Analyzing similarities among firms generates a similarity matrix, illustrating the connections between each pair of firms at a given period t. Given that our methodology for

similarity identification focuses on capturing statistically significant similarities while rejecting random ones, the resulting similarity matrix is highly sparse. In aggregate comparison, just over 1% of all possible connections are deemed statistically significant, indicating that nearly 99% of entries in the matrix receive a value of zero.

This integrative methodology, from initial feature extraction to the final statistical validation, forms the bedrock of our approach to defining image industries. By carefully navigating the complexities of high-dimensional image data and incorporating rigorous hypothesis testing, we ensure that our firm similarities are grounded in visual similarities and statistically significant, providing a compelling framework for understanding industry relatedness through the lens of image data.

3.4 Image industry definition

The definition of image firm similarities (IFS) is a critical phase in our methodology, designed to cluster firms into industries based on visual similarities. Given the unique characteristics of our similarity matrix, which is highly sparse and composed only of statistically significant similarities, we adopt a two-step approach to address the challenges inherent in clustering with such data.

Scaling Similarity Scores

The first step involves a nuanced scaling procedure for similarity scores. Traditional clustering methodologies often scale all similarities to fall within a 0-1 range. However, given that our similarities are already filtered for statistical significance, a standard scaling would obscure the distinction between firms with the lowest statistically significant similarity and those without statistical similarity. To circumvent this issue, we adopt a more tailored scaling approach, where all significant similarities are scaled to range from 0.5 to 1, while pairs deemed not statistically similar are assigned a zero similarity. This scaling preserves the granularity of our similarity scores, ensuring that the clustering process reflects genuine, statistically validated similarities.

Clustering Methodology

The second step addresses the challenge posed by our sparse similarity matrix. Clustering involves identifying and assigning similar instances to clusters or groups of similar cases. Traditional clustering methods, such as k-neighbors or spectral clustering, falter in the face of such sparsity (Géron, 2019). Inspired by Hoberg and Phillips (2016), we implement a dedicated clustering methodology. Initially, each company is placed in a singular cluster. We then iteratively measure the distance between clusters, merging the two with the highest similarity at each step. This process continues until we achieve the desired number of clusters. Additionally, after each merging step, we conduct a single reclassification to ensure each company is assigned to its most appropriate cluster, acknowledging that cluster mergers may displace some firms from their optimal grouping.

Dynamic Classification Update

We introduce a biannual updating process to accommodate the dynamic nature of firm similarities in response to new images. This process relies on a similarity matrix predominantly composed of unseen photos, given that it is based on a rolling 3-year observation period, with approximately 66% of the photos being new in each update. Starting with an individual firm, we identify the cluster to which it and its peers were previously assigned as defined by the new similarity matrix. We calculate the mean similarity for each potential cluster from the previous period and select the one with the highest mean similarity for reassignment. This procedure is replicated for all firms, followed by a reclassification step. Here, we assess whether the current cluster assignment remains optimal for each firm by calculating the mean distance to each cluster. If a more suitable cluster is identified, we reclassify the firm accordingly. This iterative process ensures that each firm is associated with the cluster that best reflects its current visual representation.

3.5 Out-of-sample setting

To affirm the economic relevance of the Image Industry Classification (IIC) and the efficacy of its associated applications, we conduct a rigorous examination in an out-of-the-sample (OOS) setting. Central to our OOS approach is the precise timestamping of all photos utilized, a feature ensured by the Google search algorithm employed for image downloads. This process allows for the collection of photos within specified upload periods, such as querying images of Noble Energy Inc. from 2016 with the criteria 'Noble Energy Inc. products after 2016-01-01 before 2016-12-31.' This ensures that the images retrieved could not have been uploaded later than in 2016, safeguarding against any look-ahead bias. However, it's noteworthy that images taken before 2016 and uploaded within the year do not compromise our analysis due to the absence of look-ahead bias.

Each image in our database is tagged with the year it was made available online. As detailed in the Section 3.3, the similarity between any two companies for year t is determined based on images from years $t-2$, $t-1$, and t. Moreover, to mitigate Type I and II errors, this assessment indirectly incorporates images from years $t-3$ and $t-4$. Therefore, we employ a classification delayed by one period for conducting our economic tests. For instance, when evaluating our classification against market or financial data from 2014 to 2015, we utilize firm similarities and IIC classification for the year 2013, which was formulated based on images from 2009 to 2013. Consequently, photos uploaded from 2014 to 2015 are not used to test IIC performance for that period. This OOS procedure is methodically repeated across four cycles: we predict IIC for 2016 to 2017 using photos from 2011 to 2015, 2018 to 2019 with images from 2013 to 2017, and 2020 to 2021 with photos from 2015 to 2019. This comprehensive OOS framework underpins the results presented in the subsequent sections, ensuring a robust validation of our findings free from predictive bias.

This methodological framework culminates in defining a time-varying IFS capable of adapting to new visual data and preserving the relevance of firm similarities over time. With the clusters defined, the subsequent section of our study will delve into the economic

characterization of these image-based industries, examining their economic homogeneity and exploring potential applications of our visual similarity-based clustering approach. This analysis will not only underscore the practical value of the IFS but also highlight its potential to offer fresh insights into the dynamics of industry classification and economic behavior.

4 Results

This section describes the relatedness of firms clustered with photos that illustrate their business activity. We start by reporting the characteristics of firms' similarities captured with image and Image Industries formed with those similarities. Then, we examine IFS'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

In exploring firm relatedness through visual data, our approach harnesses the power of an image to form firm similarities. This methodology effectively captures the essence of companies' business activities, as illustrated by the visual representations of peers from notable large-cap companies: Exxon Mobil Corp, Walmart Inc., Citigroup Inc., and Johnson & Johnson. These examples highlight the diverse nature of objects that can underscore similarities within and across industries and are visualized in Figures 2, 3, 4, and 5.

For Exxon Mobil Corp, the range of images spans refineries, petroleum stations, drilling machines, and tankers, reflecting the broad spectrum of its operations. Walmart Inc. is visually represented through internal and external store views, grocery counters, consumer products, and delivery trucks bearing its logo, showcasing its retail dominance. Citigroup Inc. presents a more abstract connection, where images of glass skyscrapers, prominent

building logotypes, and scenes of individuals engaging in financial activities, such as discussing in TV studios or using banking apps, depict the financial services industry's essence. Johnson & Johnson's image similarities are more straightforward, with visuals of cosmetics and standardized packaged chemicals highlighting its product offerings. This nuanced identification of similarities reveals how diverse objects and scenes can bind companies within the same industry, showcasing the sophistication of our image-based analysis.

However, it's noteworthy that, despite the overall success of this method, some instances of misclassification occur, such as the association of US Bancorp with Exxon Mobil Corp demonstrated at Figure 2. This is attributed to the visual overlap in the presentation of logotypes and the architectural resemblance between refineries and skyscrapers. Such examples underscore the challenges and complexities of defining industries solely based on visual data.

Building on the foundation of visual similarities, the next phase of our analysis delves into the visual features of Image Industries. We build the Image Industries with a large set of photos. Table 1 shows the final average number of photos used to link firms to industries in a single period is above 200 thousand. The typical industry is represented by at least several hundred images per period. 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 6 and Figure 7 illustrate the random representations of six sectors from Image Industries 73 characterized by high inter-industry correlation. The convergence of objects describing randomly selected companies is apparent. For example, most photos representing Industry 40 show cables and small electronic devices. However, in detail, objects recognized within this industry include special electronics and smartphones. We can see the same tendency in Industry 15 (bottles, food in plastic packaging, cosmetics, chemicals) or Industry 1 (trucks, trains, or cars). It demonstrates the first characteristic of Image Industries is selective similarity aggregation, where one class collects similar objects but not identical

ones. If the clustering procedure 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 six Industries signified in Figure 8, and Figure 9 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, but it may also be 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 2 demonstrates the characteristics of Image Industries. First, our Image Industries classify 50.8% to 53.3% of our stock sample, translating into a monthly average of 1,084 to 1,110 firms. Photo representation of unclassified firms is of non-distinguishing quality to indicate what class it should be assigned to.[13] Second, many industries contain a small number of companies. This phenomenon is attributed to the fact that the similarity matrix used to construct industries is very sparse. Some companies have only one peer or at most a few. In clustering, such companies remain their sole peers, forming industries with few members. Therefore, our study employs two terms to describe the number of industries in a given classification. The first pertains to the total number of industries, including those composed of a single company. The second, however, encompasses the number of industries with a minimum of five companies at the time of the first industry definition, i.e., in 2013.[14] Consequently, a classification that has 25 (50) industries comprising at least five companies totals 45 (73) clusters. Using industries with a minimum of five companies is especially crucial for economic tests and proposed applications, where calculations require a minimum of five companies per industry.

[13]Image Industries classify 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 from 1983 to 2013.

[14]Table A2 demonstrates descriptive statistics of image industries defined in 2013. Industries marked in bold have at least five stocks.

Table 3 presents the number of industries and the number of companies for classifications used as benchmarks in our tests. The number of companies in other classifications is constrained by the image industries to standardize the research sample.

4.2 Economic homogeneity of firms clustered with image

To verify economic relations between companies 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 to explain contemporaneous firm-level indicators ($vble_t$). We estimate the average adjusted R^2 for all firms within each cluster. R^2 is estimated from a regression below for each firm i:

$$\text{v}ble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t, \tag{8}$$

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. 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.

We use the same set of ratios as Kaustia and Rantala (2021) extended by FORE-

CASTED MONTHLY RET, TOBIN'S Q, R&D GROWTH, SG&A to # EMPLOYEES, SG&A GROWTH, FORECASTED EPS, and EPS GROWTH. 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 significantly impacts economic homogeneity tests. The higher the number of industries in the classification, the higher the R^2. Therefore, we compare the IFS of 45 and 73 industries to schemes that have the most similar numbers of industries.[15] The 2-digit SIC has over 50 classes, which is too high compared to IFS with 45 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 IFS 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 3 demonstrates the average yearly number of industries with at least five stocks per classification.

To be comparable, Tables 4 and 5 demonstrate a pair-wise results comparison that stems from the *same* set of firms; i.e., those classified with Image Industries —as well as with all other classification techniques. Table 4 compares Image Industries with variables based on market information. The average rank for two classifications based on image varies from the first to the last. The homogeneity of image classifications is extreme when evaluating three variables: FORECASTED MONTHLY RET, MARKET LEVG, and PE. Overall, the economic homogeneity based on market-based ratios is average and similar to industries

[15]The 45 industries contain 25 industries with five firms for each industry in 2013; the 73 industries contain 50 industries with five firms for each in 2013.

based on SIC, including Fama-French Industries.

Next, we extend our comparison with ratios based only on accountancy information. Table 5 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, R&D EXPENSE to SALES, R&D GROWTH, SG&A GROWTH, EPS GROWTH, DIV PAYOUT, PROFIT MARGIN, DEBT TO EQUITY, and the worst for ASSET to SALES. The image and text-based comparison show the relative overperformance of similar classifications in 60% of observations. The Image Industries performs better than the GICS in 63% of cases.

Finally, to maximize the sample size, we change the approach to build the stock universe and match our classification to competitors with all available firms. Tables A3 in the Appendix reports the results for market-based indicators. The overall rankings are more favorable for image industries. In this comparison, image-based classification is equal to that of the GICS industries. The most significant upgrades come from MARKET to BOOK, MONTHLY RET, MARKET LEVG, and EV to SALES, with no downgrades. Accountancy-based ratios are available in Table A4 in the Appendix—ratios. Image industries are the strongest classification in this comparison. They rank at first or second place in 75% of observations. We also test how the elimination of microcaps impacts results. Tables A5 and A6 in the Appendix demonstrate the exclusion of micro-caps does not change the relative position of image industries in rankings.

Overall, the homogeneity of firms clustered with images depends on the character of data used to create ratios. It reaches the top-performing GICS classification for accountancy-based ratios and is equal to SIC-based industries for market-based ratios. Furthermore, the high variation in R^2 rankings suggests that image-based industries differ significantly from competitors. The following example illustrates the advantage of IFS as a classification based on product similarities. Ameren Corp is classified with IFS 73 to industry 45, with an exceptionally high inter-industry correlation of 0.335. Such a high correlation level means

that companies classified in Industry 45 are characterized by intense homogeneity. Ameren Corp is an American power company.

The IFS classified this company into industry 45 through photos of electric power plants. Close-up images also depict products of, e.g., Kosmos Energy Ltd, Mexco Energy Corp, or Noble Energy Inc—so these firms are peers along with IFS. Instead, two-digit SIC classifies Ameren Corp to electric, gas, sanitary services, three-digit NAICS to utilities, and six-digit GICS to multi-utilities. Because of this—in the case of classifications not based on product similarities— Ameren Corp is not a peer of Kosmos Energy Ltd, Mexco Energy Corp, or Noble Energy Inc.

In contrast, HP classifications that are also product-related agree with IFS about the Ameren Corp peers. This example shows that product-oriented classifications categorize companies that need power plants to deliver electricity to the same industry as electricity 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 three methods for using image-based similarities or industry classifications to build investment portfolios. The first method is inspired by the strong abilities of the images to capture similarities in growth that are demonstrated with high economic homogeneity on growth-related ratios. Second method presents the relative benefits of using the image to achieve portfolio diversification; meanwhile, the third one shows the possibilities of using the IFS 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 45 meets this requirement. Table 2 demonstrates the number of stocks per industry for IFS with 45 (Panel A) and 73 classes (Panel B). IFS 73 is too granular because the average number of stocks per industry is below 20, and at the formation period 23 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 IFS with 45 classes. This classification has 25 classies with at least 5 firms, so on average, have 43 stocks per industry with at least 5 firms.[16] Analogous to the comparative analysis in the previous section, we compare the results of applications based on IFS 45 to five classifications with a similar number of industries.

5.1 Pair trading based on growth

Building on the findings from Section 4, where Image Industries showcases significant economic homogeneity in growth-related ratios, we are motivated to develop an investment strategy capitalizing on the high similarity among peers regarding growth potential. Our proposed pair trading strategy seeks to leverage the anticipated performance disparity between firms with high and low growth profiles. Our pair trading strategy focuses on peer firms to capitalize on their relative growth prospects while mitigating market-wide risks. By pairing similar firms within the same industry, we isolate our investment on their compara-

[16]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.

tive growth potential, reducing the impact of industry-wide trends. This approach allows for a more precise comparison, enhancing our ability to identify mispricings between firms with high and low growth. Consequently, our strategy is not merely about selecting high-growth firms but about exploiting the relative valuation discrepancies among peers, leading to a more targeted and risk-adjusted investment approach. This approach involves constructing a portfolio by pairing peers identified through image-based similarities, then investing long in firms showing high growth and short in those with low growth, as observed in the previous month. Growth is assessed using two standard metrics, SALES GROWTH, and EPS GROWTH, with peers, including the focal company, sorted into quintiles to facilitate the high-low strategy evaluation.

To enrich our analysis and provide a benchmark, we also identify peers using textual similarities, as delineated by Hoberg and Phillips (2016) and common analyst coverage, following Kaustia and Rantala (2021). This comparative framework spans 2016 to 2021, offering a broad temporal lens to assess the strategy's effectiveness.

The results, detailed in Table 6, underscore the superior performance of the pair trading strategy rooted in image similarities, particularly evident through its high Sharpe ratio for both SALES GROWTH and EPS GROWTH. Notably, the strategy achieves an exceptional Sharpe ratio above 3 when predicated on SALES GROWTH, coupled with the max of 1 million loss of merely 2.2%, propelling the Calmar ratio to an extraordinary level well above 10. This is nearly double the ratios observed in strategies based on textual similarities and common analyst coverage.

These findings affirm the exceptional prowess of image-based similarities in capturing firms' growth potential and translate this statistical significance into a viable trading strategy with remarkable investment performance. Moving forward, we will explore the applications of image-based industries, particularly focusing on their diversification capabilities and momentum within the industry.

5.2 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. [17].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 7 reports the results for stocks classified with Image Industry 45. The random portfolios built with Image Industries 45 perform the best in the Sharpe ratio for portfolios optimized to minimize risk, take the second position when portfolios are equally weighted or optimized to maximize the Sharpe ratio, and the third position with the value-weighted

[17]An alternative to industry classification for portfolio diversification might involve utilizing market data to examine company correlations, optimize the balance between expected returns and risk, or directly minimize risk exposure. However, this approach bears a significant limitation: the relationships between stock returns can be transient and arbitrary, not necessarily reflecting the underlying economic foundations tied to the companies' operational profiles. Diversification based solely on historical return correlations may result in a superficial risk mitigation strategy, as these correlations may not persist over time and could lead to misinformed portfolio decisions.

In light of these considerations, diversification that relies exclusively on return correlations without the insights provided by industry classification may only offer an illusion of reduced risk. Industry classification emerges as a vital alternative or complementary strategy to portfolio optimization. For instance, implementing maximum investment limits per industry can provide a structured approach to diversification, ensuring a portfolio is not overly exposed to sector-specific risks and is better aligned with broader economic cycles and industry-specific developments.

By integrating industry classification into portfolio construction, investors can achieve a more nuanced and economically grounded diversification strategy. This method not only counters the temporal and random nature of return correlations but also enriches the investment process with a deeper understanding of the industries that shape different market segments. Consequently, the employment of industry classifications facilitates a more stable and informed diversification strategy and serves as a crucial dimension in the pursuit of optimized portfolio performance and risk management.

approach. The relative overperformance of image-based schemes is definitive, and only the four-digit GICS achieves better results for not-optimized portfolios. All in all, classifications based on an image deliver substantial diversification benefits for equally weighted, value-weighted, maximum Sharpe, and minimum risk portfolios. Table 8 shows that image-based classifications are the most dynamic compared to competing classifications. The high dynamics of classifications correspond well with what is 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.

Table 7 presents the performance outcomes for stocks grouped under Image Industry 45. When constructing random portfolios, those formed using Image Industry 45 excel in terms of the Sharpe ratio for risk-minimized portfolios, rank second for equally weighted and Sharpe ratio-maximized portfolios, and take third place in value-weighted portfolios. The superior performance of image-based classification is clear, with only the four-digit GICS outperforming in non-optimized portfolios. IFS provides significant diversification advantages across equally weighted, value-weighted, maximum Sharpe, and minimum risk portfolios. Table 8 demonstrates that FTS is the most dynamic among the compared classifications, aligning with market observations where companies frequently adjust their product offerings to stay competitive. This dynamism in industry assignment reinforces the diversification benefits of employing images in portfolio construction.

The high dynamics of IFS allow for a more responsive adaptation to market changes, ensuring that portfolios remain well-diversified even as the market landscape evolves. By quickly reflecting shifts in company strategies or industry trends, IFS helps maintain a balanced portfolio composition, reducing the risk of overconcentration in outdated or less relevant sectors. This agility in adjusting to market dynamics is a key factor in maximizing diversification benefits, as it enables investors to spread their risk across a more accurately represented range of industries and firms. Consequently, the use of IFS in portfolio construc-

tion not only captures the current state of the market but also proactively adapts to future changes, thereby enhancing the overall resilience and performance of the portfolio.

5.3 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 the performance of industry momentum strategies 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.[18]

Table 9 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 45 ranks eight five times and second once. The momentum

[18]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).

strategies built with IFS dominate both value and equal-weighted portfolios. In the case of reversal strategies, IFS delivers sound results for value-weighted portfolios, but it gets a bit worse Sharpe ratios in equal-weighted settings. In addition, IFS demonstrates the most robust results. Its Sharpe ratio is solid for the momentum, reversal, value, and equally weighted portfolios.[19]

Next, we create "random" industry portfolios and compare their Sharpe ratios with Image Industry 45 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 10 demonstrates the simulation results, where Sharpe ratios of Image Industry 45 are higher than those of any random portfolio.

6 Underlying mechanism

The industry categorization significantly influences stock trading, contingent upon how market players discern their peers. An optimal industry classification should manifest a pronounced consensus (or lack thereof) concerning stocks within the same (or different) industries. The more cohesive the agreement within an industry, the more pronounced the influence of aggregated demand and supply on stock prices, amplifying the efficacy of any market-based industry categorization application. This principle underpins the application of industry categorization based on market pricing (Diermeier, Ibbotson, & Siegel, 1984; Ibbotson, Diermeier, & Siegel, 1984; Merton, 1973).

[19]Table A7 reports the results of industry momentum strategy extended with volatility targeting mechanism as proposed by Barroso and Santa-Clara (2015). Moreover, IFS has high application benefits for this application.

Section 5.1 illustrates that image-based similarities facilitate the construction of highly profitable pair trading strategies. Meanwhile, Sections 5.2 and 5.3 demonstrate that the image industry classification surpasses all other schemes in delivering the most substantial portfolio diversification benefits and profitability from the industry momentum strategy. This section delves into the underlying theories and presents empirical evidence supporting these assertions.

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 firms 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 A's 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, 5.2 and 5.3. 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.[20] 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 arbi-

[20] 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.

trary 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.

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 that of 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 who 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.[21] 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 customers. However, determining similarities is not universal; 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

[21]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.

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 widespread use of image-based communication based on television, the internet, and mobile devices (Branthwaite, 2002; Dewan, 2015). Back to Apple, Ford, and Coca-Cola examples, we believe most investors see plenty of photos representing their products. Second, different sources of images 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. The R^2 of FORECASTED MONTHLY RETURN from Table 4 is very high relative to other classifications with similar granularity, suggesting that 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 benefits of image-based portfolio diversification evaluated in Section 5.2.

Second, the dispersion of analysts' EPS estimates across stocks demonstrates the rate of

investors' agreement. The lower the dispersion of analysts' estimates, the higher the investor agreement about stocks in the industry. Figure 10 visualizes that the highest agreement rate for stocks classified into a single industry is for Image Industries 45. It is consistent with results observed for pair trading, industry diversification benefits, and industry momentum, where Image Industries 45 delivers 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 identification of peers with image improve performance of a pair trading strategy based on growth. Second, image industries are suitable for portfolio diversification. Third, 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 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 all firms listed at major stock exchanges. 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)*

Crouse, D. F. (2016). On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, *52*(4), 1679–1696. doi:10.1109/TAES.2016.140952

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

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *The Journal of Finance*, *57*(5), 2113–2141. doi:https://doi.org/10.1111/0022-1082.00490. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/0022-1082.00490

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, W., Wang, Y., & Yu, J. (2021). Similar stocks. *Available at SSRN 3815595*.

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

Ibriyamova, F., Kogan, S., Salganik-Shoshan, G., & Stolin, D. (2019). Predicting stock return correlations with brief company descriptions. *Applied Economics*, *51*(1), 88–102.

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.

Lee, C. M., Ma, P., & Wang, C. C. (2015). Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, *116*(2), 410–431.

Lee, C. M., Sun, S. T., Wang, R., & Zhang, R. (2019). Technological links and predictable returns. *Journal of Financial Economics*, *132*(3), 76–96.

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

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.

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.

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.

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

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

Strutz, T. (2016, January). *Data fitting and uncertainty (2nd edition)*.

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

Table 1: Image Data Sample

The table provides a succinct overview of the image dataset after cleaning, organized into three-year rolling windows to ensure a comprehensive representation of each firm's business activity through visual data. Key metrics detailed in the table include: 1) Period: each row corresponds to a distinct three-year window during which images were aggregated, 2) #Firms: the total number of unique firms represented within each period, 3) #Photos: the total count of photos collected per period, illustrating the dataset's visual depth, 4) #Pairs: the number of analyzed pairs of firms for each period, indicating the comparative analysis breadth, 5) #Photos / #Firms: the average number of photos per firm, reflecting the visual data's richness per company, 6) #Pairs / #Firms: the average number of analyzed pairs per firm, showing the extent of inter-firm visual comparisons. All photos were sourced from Google using a Python API, highlighting the dataset's reliance on publicly available, online visual representations of firms' activities.

Period	#Firms	#Photos	#Pairs	#Photos / #Firms	#Pairs / #Firms
2009-2011	2,951.0	209,898.0	4,352,725.0	71.1	1,475.0
2010-2012	2,894.0	209,459.0	4,186,171.0	72.4	1,446.5
2011-2013	2,954.0	218,636.0	4,361,581.0	74.0	1,476.5
2012-2014	2,963.0	223,169.0	4,388,203.0	75.3	1,481.0
2013-2015	2,989.0	227,050.0	4,465,566.0	76.0	1,494.0
2014-2016	3,065.0	233,812.0	4,695,580.0	76.3	1,532.0
2015-2017	3,145.0	241,272.0	4,943,940.0	76.7	1,572.0
2016-2018	3,218.0	247,398.0	5,176,153.0	76.9	1,608.5
2017-2019	3,285.0	251,882.0	5,393,970.0	76.7	1,642.0
2018-2020	3,313.0	253,926.0	5,486,328.0	76.6	1,656.0
2019-2021	3,384.0	254,889.0	5,724,036.0	75.3	1,691.5

Table 2: Description and Summary Statistics of Image Industries 45 & 73

The table presents an overview of industries formed with firms' photos for 45 (Panel A) and 73 (Panel B) classes. Classification with 45 (73) classes has 25 (50) industries with at least 5 firms in the forming period (2009-2013). Our sample covers NYSE, AMEX, and NASDAQ stocks. Image Industries are updated every second year from 2014 to 2021, allowing time-variation in industrial classification. We report the average number of stocks assigned to each industry (No. of Stocks), the average monthly capitalization of stocks in each industry (Market Cap. (bn USD)), and the share of stocks classified with photos in the whole market capitalization (Avg. % of Market Cap.). 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 45 (25)						
0	29.4	641.2	2.3	0.0051	0.286	('45', '37', '38', '73', '48')
1	60.1	677.5	2.4	0.0049	0.294	('35', '37', '42', '40', '49')
2	15.4	566.6	2.2	0.0059	0.167	('35', '48', '38', '50', '36')
3	36.0	430.9	1.9	0.0006	0.202	('36', '56', '38', '48', '73')
4	45.1	962.8	3.3	-0.0017	0.155	('20', '99', '56', '28', '51')
5	47.5	1,027.5	4.0	0.0042	0.268	('60', '20', '28', '38', '67')
6	51.5	782.2	2.6	0.0115	0.248	('36', '35', '73', '50', '99')
7	36.4	296.4	1.2	0.0097	0.362	('36', '35', '38', '50', '73')
8	39.1	92.5	0.4	0.0037	0.171	('60', '36', '38', '35', '28')
9	21.6	48.7	0.2	-0.0008	0.291	('25', '36', '57', '50', '33')
10	45.0	934.7	3.1	0.0037	0.182	('28', '99', '25', '38', '59')
11	16.5	265.4	1.0	0.0002	0.165	('28', '38', '87', '59', '99')
12	11.6	68.5	0.2	0.0108	0.149	('73', '28', '99', '38', '26')
13	13.8	264.7	0.9	0.0065	0.240	('49', '99', '36', '73', '46')
14	63.3	1,207.5	4.0	-0.0021	0.201	('58', '99', '59', '53', '54')
15	26.6	650.4	2.3	-0.0039	0.260	('60', '99', '73', '62', '63')
16	14.4	122.7	0.4	0.0038	0.189	('38', '73', '28', '99', '35')
17	35.8	319.7	1.3	0.0063	0.221	('58', '60', '20', '28', '70')
18	44.0	379.0	1.5	0.0050	0.249	('60', '15', '70', '79', '73')
19	16.3	80.2	0.3	0.0103	0.245	('10', '14', '99', '60', '28')
20	12.6	52.8	0.2	0.0060	0.220	('60', '99', '20', '49', '28')
21	24.0	223.3	0.7	0.0078	0.215	('38', '35', '36', '99', '26')
22	32.5	278.0	0.8	0.0015	0.170	('55', '37', '60', '50', '99')
23	34.9	386.0	1.3	0.0051	0.276	('60', '73', '70', '65', '63')
24	3.3	2.5	0.0	-0.0005	-0.006	('10', '32', '49', '20', '14')
25	45.7	717.1	2.5	-0.0044	0.248	('13', '29', '49', '28', '44')
26	28.0	143.7	0.5	0.0010	0.203	('38', '15', '35', '50', '99')
27	34.0	137.0	0.5	0.0038	0.242	('35', '36', '37', '38', '30')
28	29.0	220.4	0.8	0.0048	0.292	('60', '63', '49', '28', '80')
29	49.1	555.2	1.9	0.0007	0.251	('36', '38', '73', '35', '99')
30	15.1	123.4	0.4	-0.0038	0.252	('56', '31', '30', '99', '37')
31	2.0	5.1	0.0	-0.0004	-	('34', '37', '38')
32	2.0	2.2	0.0	-0.0239	-	('10', '28')
33	38.4	920.6	4.2	-0.0255	0.278	('13', '29', '44', '49', '16')
34	13.3	56.3	0.3	-0.0056	0.182	('16', '20', '22', '25', '28')

Table 2: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Inter. Corr.	TOP5 SIC codes
35	2.0	16.0	0.1	-0.0223	-	('37',)
36	2.0	0.8	0.0	-0.0310	-	('35', '36')
37	1.9	8.9	0.0	0.0090	-	('59', '50')
38	1.0	0.0	0.0	0.0138	-	('60', '67')
39	2.0	8.9	0.0	-0.0159	-	('61', '63')
40	23.1	38.0	0.2	-0.0003	0.132	('60', '79', '73', '26', '28')
41	2.0	4.1	0.0	-0.0180	-	('13', '30')
42	13.0	101.5	0.5	-0.0072	0.212	('30', '31', '56', '60')
43	2.0	1.8	0.0	0.0060	-	('34',)
44	2.0	22.8	0.1	0.0152	-	('34',)
Sum	1,084.2	13,845.8	50.8			
PANEL B: Image Industries 73 (50)						
0	26.4	564.9	2.0	0.0054	0.282	('45', '37', '38', '73', '48')
1	41.2	477.6	1.7	0.0058	0.321	('42', '37', '40', '49', '99')
2	9.6	27.5	0.1	0.0086	0.208	('34', '38', '37', '99', '36')
3	10.3	44.8	0.2	0.0030	0.211	('56', '23', '51', '38', '48')
4	7.4	36.3	0.1	0.0041	0.140	('38', '36', '56', '73', '48')
5	10.3	126.3	0.5	-0.0017	0.250	('28', '60', '99', '73', '67')
6	38.3	724.1	2.4	0.0102	0.237	('36', '35', '73', '50', '99')
7	12.7	138.8	0.5	0.0035	0.190	('38', '35', '28', '73', '36')
8	27.4	251.1	1.0	0.0060	0.288	('36', '38', '73', '50', '35')
9	20.1	65.1	0.3	0.0009	0.144	('60', '36', '28', '38', '35')
10	15.3	170.9	0.6	0.0072	0.337	('60', '36', '73', '35', '65')
11	23.3	247.2	0.9	0.0062	0.170	('28', '60', '59', '99', '56')
12	11.2	42.8	0.2	0.0112	0.177	('38', '35', '36', '73', '99')
13	31.9	1,181.7	4.2	0.0032	0.178	('20', '28', '38', '59', '99')
14	8.5	29.0	0.1	0.0054	0.304	('35', '36', '30', '37', '38')
15	23.1	276.3	1.2	0.0020	0.311	('20', '25', '57', '28', '50')
16	10.4	41.6	0.2	0.0047	0.241	('20', '37', '35', '33', '36')
17	15.6	244.8	1.0	0.0022	0.186	('28', '25', '38', '49', '57')
18	8.6	54.3	0.2	0.0090	0.162	('73', '99', '28', '20', '48')
19	11.5	31.0	0.1	0.0081	0.186	('36', '28', '73', '99', '13')
20	9.5	135.3	0.4	0.0084	0.272	('49', '99', '36', '73', '46')
21	19.4	83.7	0.3	0.0065	0.202	('36', '38', '35', '48', '73')
22	24.5	229.8	0.8	-0.0017	0.227	('58', '99', '20', '36', '28')
23	14.3	420.3	1.6	-0.0055	0.318	('60', '62', '73', '38', '99')
24	21.0	105.2	0.4	0.0053	0.176	('99', '36', '60', '28', '38')
25	9.8	64.8	0.2	0.0115	0.372	('60', '70', '79', '99', '24')
26	26.2	448.0	1.5	0.0054	0.155	('28', '58', '99', '38', '20')
27	14.3	511.7	1.7	0.0048	0.390	('60', '62', '67', '63', '73')
28	11.1	42.6	0.2	0.0058	0.257	('36', '73', '35', '50', '99')
29	12.9	113.5	0.4	-0.0054	0.211	('60', '78', '10', '23', '49')
30	12.5	20.5	0.1	-0.0024	0.178	('60', '99', '70', '67', '24')
31	9.2	29.6	0.1	0.0018	0.208	('60', '22', '24', '63', '50')
32	9.9	61.5	0.2	0.0102	0.251	('10', '14', '28', '38', '73')
33	5.7	17.6	0.1	-0.0099	0.180	('56', '60', '53', '48', '28')
34	10.6	60.3	0.2	0.0043	0.252	('38', '36', '33', '50', '26')

Table 2: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Inter. Corr.	TOP5 SIC codes
35	13.0	149.0	0.5	0.0062	0.248	('38', '35', '56', '23', '50')
36	17.7	221.6	0.7	-0.0021	0.289	('35', '38', '37', '50', '73')
37	27.1	273.4	0.8	0.0053	0.164	('55', '37', '50', '79', '99')
38	7.4	159.3	0.5	-0.0071	0.353	('49', '28', '10', '13', '15')
39	29.4	129.3	0.4	-0.0039	0.219	('60', '15', '63', '70', '65')
40	8.9	115.5	0.4	0.0096	0.449	('35', '36', '38', '33', '50')
41	8.8	18.0	0.1	0.0063	0.218	('38', '35', '60', '20', '36')
42	30.9	779.7	2.4	0.0064	0.182	('59', '53', '54', '55', '35')
43	23.0	403.6	1.6	-0.0111	0.191	('20', '59', '99', '53', '54')
44	17.8	147.6	0.6	0.0042	0.170	('20', '55', '37', '73', '36')
45	30.3	491.0	1.6	0.0041	0.335	('13', '29', '49', '44', '99')
46	32.5	322.4	1.3	0.0040	0.294	('15', '13', '70', '29', '25')
47	9.8	88.1	0.3	0.0076	0.227	('35', '36', '37', '73', '48')
48	10.9	157.9	0.5	-0.0009	0.300	('35', '63', '60', '36', '62')
49	17.3	317.0	1.0	0.0045	0.247	('36', '35', '73', '99', '60')
50	21.7	152.6	0.6	-0.0037	0.251	('60', '36', '35', '73', '63')
51	12.9	33.9	0.1	0.0104	0.236	('60', '38', '36', '35', '99')
52	13.2	268.9	1.0	0.0087	0.111	('60', '36', '38', '35', '99')
53	15.6	476.6	2.1	0.0033	0.293	('58', '99', '53', '54', '59')
54	15.1	142.6	0.5	-0.0083	0.184	('56', '31', '30', '99', '37')
55	5.7	30.7	0.1	-0.0005	0.149	('30', '31', '34', '56', '99')
56	20.9	417.6	1.6	-0.0060	0.292	('60', '13', '29', '79', '73')
57	17.3	170.9	0.7	0.0048	0.300	('70', '79', '60', '73', '63')
58	2.0	16.0	0.1	-0.0223	-	('37',)
59	2.0	0.8	0.0	-0.0310	-	('35', '36')
60	4.0	20.8	0.1	-0.0059	-	('15', '28', '30', '99')
61	45.9	405.7	1.9	-0.0050	0.157	('35', '36', '38', '73', '48')
62	1.9	8.9	0.0	0.0090	-	('59', '50')
63	1.0	0.0	0.0	0.0138	-	('60', '67')
64	2.0	8.9	0.0	-0.0159	-	('61', '63')
65	5.2	21.3	0.1	0.0084	0.501	('60', '73', '59')
66	2.0	4.1	0.0	-0.0180	-	('13', '30')
67	29.3	1,155.5	5.3	-0.0038	0.183	('48', '73', '35', '36', '34')
68	8.0	24.2	0.1	-0.0018	0.054	('58', '63', '99')
69	13.0	101.5	0.5	-0.0072	0.212	('30', '31', '56', '60')
70	2.0	1.8	0.0	0.0060	-	('34',)
71	2.0	22.8	0.1	0.0152	-	('34',)
72	7.8	15.6	0.1	0.0046	0.051	('60', '37', '49', '73', '79')
Sum	1,109.7	14,395.7	53.3			

Table 3: Peer Groups' - Comparison

We present a comparison of different industry classification techniques by 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 that have 45 (Panel A) and 73 classes (Panel B), 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 2014 to 2021.

Industry Classification Name	Number of Industries	Number of Stocks
PANEL A: Image Industries 45 (25) - comparison		
Image Industries	34.9	1,011.0
Industry_MG	20.0	1,011.0
Fama-French 30 Industries	29.8	1,009.0
NAICS Industries	19.0	1,010.2
4-digit GICS	24.6	1,007.4
Icode 25 Industries	25.3	990.0
PANEL B: Image Industries 73 (50) - comparison		
Image Industries	60.9	1,011.0
2-digit SIC	61.4	1,011.0
Fama-French 48 Industries	46.8	1,009.0
3-digit NAICS	77.7	1,010.2
6-digit GICS	66.3	1,007.4
Icode 50 Industries	45.3	990.0

Table 4: 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 classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) 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 classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) 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 2014-2021. We calculate regression for industries that have at least five members. Table A1 in the 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 45 (25) - comparison										
Image Industries	25.2	18.9	23.9	2.4	31.7	23.8	10.0	31.5	36.0	22.8
Industry_MG	26.1	21.7	26.7	2.1	31.7	25.7	10.1	31.8	36.5	24.5
Fama-French 30 Industries	25.4	20.5	26.8	2.1	31.1	25.9	8.3	30.5	37.6	23.8
NAICS Industries	26.2	21.6	26.3	2.2	31.5	24.8	9.3	30.7	36.7	25.0
4-digit GICS	27.5	20.7	28.0	2.1	32.2	26.2	10.5	31.2	39.8	25.7
Icode 25 Industries	23.7	21.9	26.0	2.7	31.6	26.7	9.1	29.4	34.2	21.3
PANEL B: Image Industries 73 (50) - comparison										
Image Industries	26.5	18.7	23.1	2.6	31.5	25.1	10.5	32.1	34.6	23.6
2-digit SIC	28.1	22.4	26.7	2.2	32.5	24.5	11.2	30.9	35.0	26.3
Fama-French 48 Industries	27.4	21.3	26.2	2.3	31.3	25.9	10.5	31.0	35.2	25.6
3-digit NAICS	27.3	22.0	25.7	2.1	31.3	25.1	10.8	30.2	35.3	24.6
6-digit GICS	27.9	21.7	27.1	2.2	32.0	27.1	10.1	30.7	38.5	25.6
Icode 50 Industries	24.5	20.2	26.3	2.9	32.3	25.7	10.7	30.1	32.9	22.0

Table 5: 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 classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) 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 classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) 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 2014 to 2021. We calculate regression for industries that have at least five members. Table A1 in the 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 45 (25) - comparison																
Image Industries	41.9	43.4	6.0	27.9	16.6	39.1	22.2	22.9	23.5	33.4	42.5	15.8	26.1	16.7	15.6	22.9
Industry_MG	40.9	41.3	4.5	24.4	17.3	37.8	19.6	16.8	27.0	30.7	43.7	17.2	31.1	18.0	13.4	26.6
Fama-French 30 Industries	39.9	40.4	4.4	24.3	16.0	35.4	21.2	17.8	25.1	31.6	42.1	16.9	29.1	17.4	12.9	25.4
NAICS Industries	43.7	41.6	4.9	24.6	14.2	36.4	19.2	9.4	27.9	30.3	41.7	17.2	31.6	16.7	15.4	25.2
4-digit GICS	42.8	41.3	5.7	24.6	14.5	36.7	19.3	17.4	25.7	30.1	42.5	17.3	31.1	15.1	13.6	26.8
Icode 25 Industries	37.7	38.2	5.8	26.7	16.5	35.9	22.1	18.5	26.0	29.9	40.5	18.0	30.0	17.7	16.1	27.6
PANEL B: Image Industries 73 (50) - comparison																
Image Industries	39.2	45.1	7.6	28.5	17.6	38.3	27.4	26.3	21.9	33.6	41.4	16.9	24.9	19.8	16.7	23.4
2-digit SIC	37.2	39.6	7.1	25.8	14.3	35.2	20.1	18.9	25.9	29.5	43.2	17.3	28.1	18.6	15.9	25.6
Fama-French 48 Industries	38.9	40.1	6.6	26.0	15.5	34.1	22.9	19.7	24.8	29.4	41.5	17.2	28.1	17.8	15.2	25.8
3-digit NAICS	35.3	40.1	6.0	27.6	15.9	34.0	19.8	21.4	26.1	28.2	42.8	17.8	27.8	19.1	18.3	25.6
6-digit GICS	40.7	39.7	6.3	26.1	15.7	35.6	20.5	19.0	27.7	28.8	41.8	17.7	28.2	18.3	16.4	26.5
Icode 50 Industries	39.4	37.1	6.1	28.7	17.6	37.7	25.1	20.8	25.3	31.1	40.6	18.4	28.5	18.0	17.3	26.7

Table 6: Pair Trading Strategy on Growth

The table showcases the outcomes of pair trading strategies rooted in firm growth metrics, specifically SALES GROWTH and EPS GROWTH. Firms and their peers, identified through similarities in images (IIC), texts (HP), and shared analysts (KR), are sorted into quintiles based on growth observed in the preceding month. Investment positions are then held for one month, favoring firms with high growth (long) over those with low growth (short). The analysis, spanning 2014 to 2021, considers firms with a minimum of five peers. Text similarity data is derived from Hoberg and Phillips (2016) (HP), while common analyst data is from Kaustia and Rantala (2021) (KR). This table elucidates the performance differential between the top and bottom growth quintiles, reflecting the efficacy of growth-based pair trading. Table A1 in the Appendix shows details of ratios calculation.

	SALES GROWTH			**EPS GROWTH**		
	IIC	HP	KR	IIC	HP	KR
Annual Ret.	0.260	0.219	0.192	0.136	0.114	0.039
Annual Std. Dev.	0.083	0.084	0.086	0.080	0.070	0.053
Max DrownDown	0.026	0.040	0.039	0.081	0.058	0.070
Max 1m. Loss	0.022	0.040	0.039	0.073	0.057	0.043
Sharpe Ratio	3.126	2.627	2.237	1.691	1.629	0.733
Calmar Ratio	10.107	5.511	4.886	1.671	1.960	0.555

Table 7: Industry diversification benefits - comparison with the Image Industry

We demonstrate the average performance of portfolios built with different industry classification techniques formed with 1) image industries with 45 classes (25 classes with at least five firms in 2013) (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 monthly 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 (Panels 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 45 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 sample covers the years 2014 to 2021. We use sectors with at least five stocks.

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.	0.046	0.036	0.043	0.044	0.048	0.047
Annual Std. Dev.	0.231	0.243	0.237	0.265	0.214	0.241
Max DrownDown	0.485	0.521	0.501	0.550	0.444	0.497
Sharpe Ratio	0.199	0.149	0.185	0.167	0.226	0.197
Calmar Ratio	0.104	0.077	0.092	0.089	0.118	0.103
PANEL B: Value weighted portfolios						
Annual Ret.	0.106	0.086	0.090	0.092	0.100	0.107
Annual Std. Dev.	0.205	0.220	0.193	0.223	0.198	0.216
Max DrownDown	0.332	0.395	0.328	0.390	0.297	0.326
Sharpe Ratio	0.559	0.466	0.514	0.459	0.598	0.586
Calmar Ratio	0.382	0.284	0.323	0.295	0.426	0.409
PANEL C: Max. Sharpe Ratio portfolios						
Annual Ret.	0.128	0.117	0.138	0.125	0.116	0.120
Annual Std. Dev.	0.227	0.215	0.221	0.224	0.229	0.225
Max DrownDown	0.330	0.323	0.311	0.349	0.335	0.335
Sharpe Ratio	0.578	0.564	0.643	0.570	0.529	0.549
Calmar Ratio	0.441	0.428	0.497	0.403	0.420	0.415
PANEL D: Min. CVaR Ratio portfolios						
Annual Ret.	0.132	0.109	0.119	0.119	0.108	0.107
Annual Std. Dev.	0.213	0.203	0.212	0.205	0.214	0.211
Max DrownDown	0.313	0.321	0.317	0.312	0.327	0.330
Sharpe Ratio	0.634	0.571	0.587	0.585	0.532	0.526
Calmar Ratio	0.487	0.409	0.445	0.447	0.403	0.381

Table 8: Industry Dynamics

The table compares the statistics of industry dynamics. In Panel A we compare Image Industries with 45 classes (25 classes with at least five firms in 2013) 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 Image Industries with 73 classes (50 classes with at least five firms in 2013) 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 45 (Panel A), Image Industries 73 (Panel B), and additionally with all other techniques from the years 2014-2021. 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 are 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 45 (25) - comparison	
Image Industries	0.167
Industry_MG	0.020
Fama-French 30 Industries	0.021
NAICS Industries	0.033
4-digit GICS	0.000
Icode 25 Industries	0.084
PANEL B: Image Industries 73 (50) - comparison	
Image Industries	0.217
2-digit SIC	0.027
Fama-French 48 Industries	0.025
3-digit NAICS	0.053
6-digit GICS	0.000
Icode 50 Industries	0.092

Table 9: 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) image industries with 45 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 2014 to 2021. We use sectors with at least five stocks.

	Image Industries	Industry_MG	Fama-French 30 Industries	NAICS Industries	4-digit GICS	Icode 25 Industries
PANEL A: Market weighted industry returns						
Momentum 6m	0.606	0.189	0.786	0.060	0.398	0.436
Momentum 9m	0.683	0.117	0.590	-0.109	0.215	0.183
Momentum 12m	0.766	0.104	0.495	-0.192	0.119	0.029
Reversal 6m	1.178	0.941	0.814	0.990	1.144	0.911
Reversal 9m	1.178	0.974	0.823	0.941	1.153	0.923
Reversal 12m	1.206	0.995	0.836	0.967	1.190	0.964
PANEL B: Equally weighted industry returns						
Momentum 6m	0.650	0.088	0.283	0.282	0.235	0.093
Momentum 9m	0.524	-0.075	0.140	0.108	0.029	-0.059
Momentum 12m	0.383	-0.200	-0.059	0.072	-0.051	-0.147
Reversal 6m	0.498	0.567	0.577	0.596	0.668	0.616
Reversal 9m	0.500	0.583	0.567	0.624	0.694	0.611
Reversal 12m	0.533	0.607	0.611	0.664	0.766	0.709

Table 10: Industry momentum - "Random" Industry Portfolios

We compare Sharpe ratios of 'random' industry portfolios with Image Industries with 45 classes (25 classes with at least five firms in 2013). 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-month 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 2014-2021, where we use the first months to create stock 6-month momentum. We use sectors with at least five stocks.

	Shift_-1	Shift_0 or IIC	Shift_1
PANEL A: Market weighted industry returns			
Momentum 6m 6m	0.086	0.606	-0.042
Momentum 6m 9m	0.181	0.682	-0.074
Momentum 6m 12m	0.126	0.766	-0.127
PANEL B: Equally weighted industry returns			
Momentum 6m 6m	0.026	0.650	0.117
Momentum 6m 9m	0.192	0.524	0.115
Momentum 6m 12m	0.121	0.383	0.102

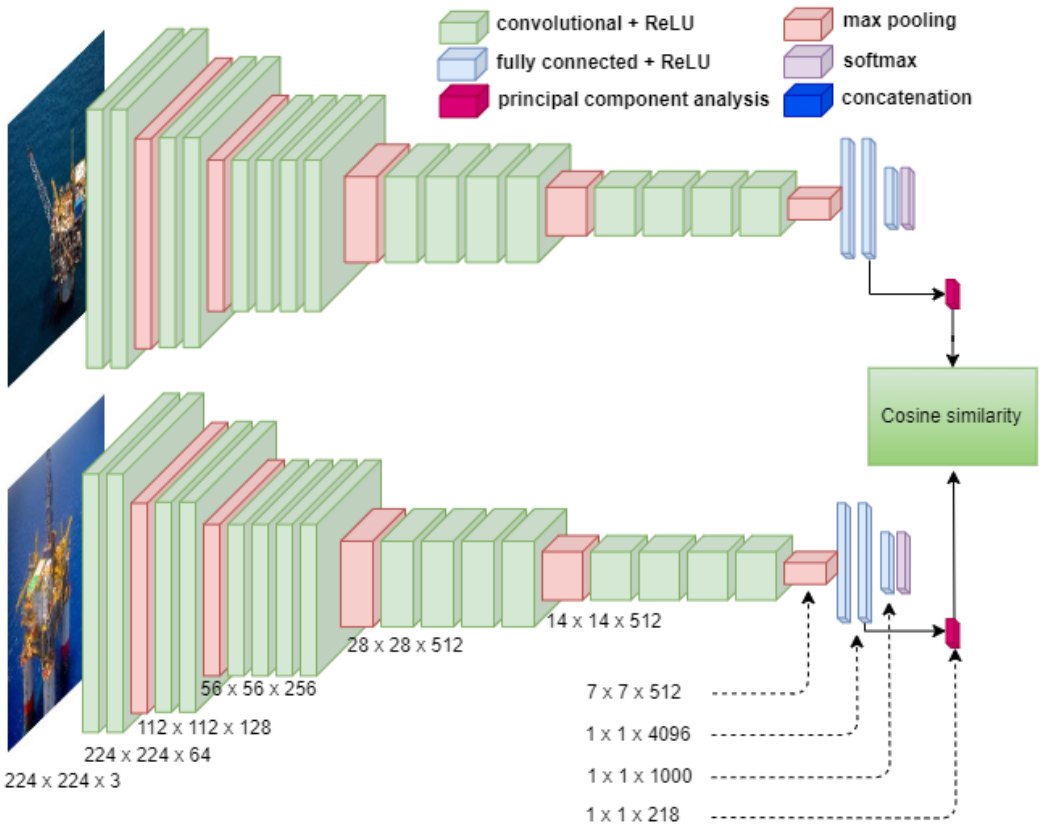

Figure 1: Image similarity

Note: The figure presents the architecture of image comparison. In the first step, images are standardized to dimensions 224x224x3, where the first dimension shows the height, the second the width, and the third the colors. Second, 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. Third, we apply Principal Complement Analysis (PCA) to reduce this dimension and represent at least 70% of the variation. The reduced vector has a dimension of 1x1x218. Finally, feature vectors are the input to define cosine similarity between two photos.

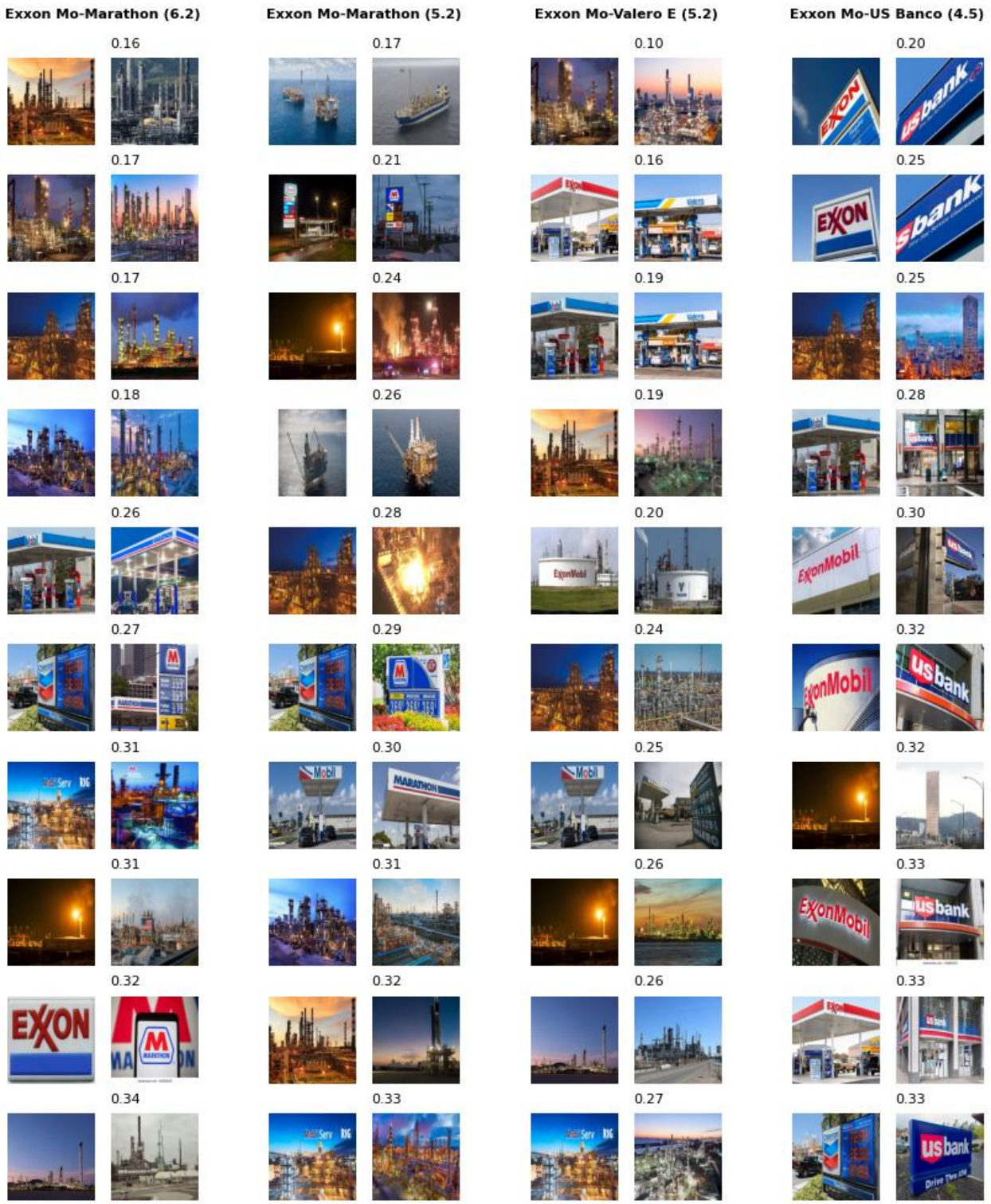

Figure 2: Photos representing four peers of Exxon Mobil Corp with the highest similarity score

This figure presents photos related with Exxon Mobil Corp and its four peers with the highest similarity score - from the left: Marathon Petroleum Corp, Marathon Oil Corp, Valero Energy Corp, and US Bancorp. Each set of photos is titled with peers names and similarity score in the brackets. Each pair of photos demonstrate the cosine similarity distance measure.

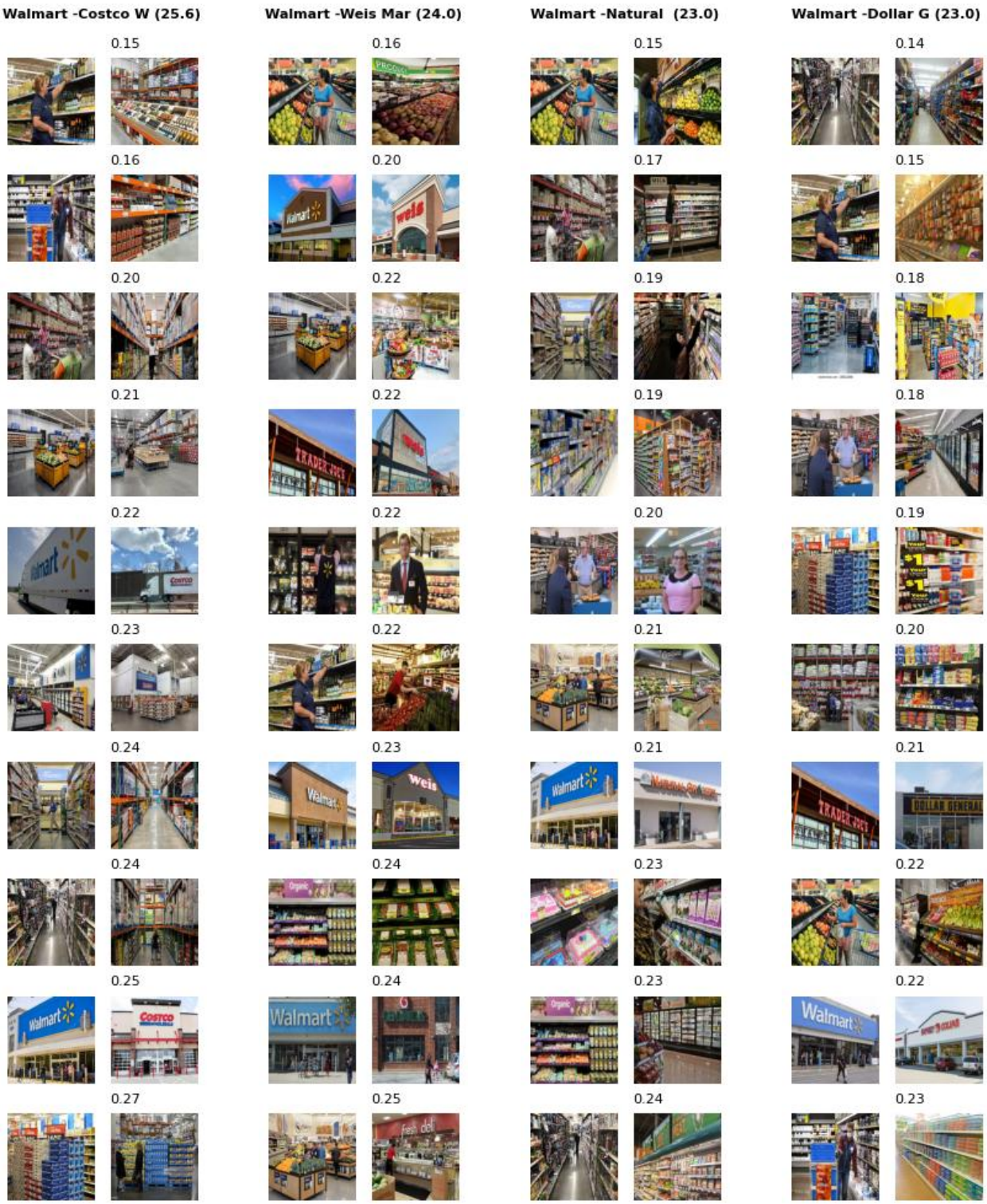

Figure 3: Photos representing four peers of Walmart Inc with the highest similarity score

This figure presents photos related with Walmart Inc and its four peers with the highest similarity score - from the left: Costco Wholesale Corp, Weis Markets, Natural Grocers By Vitamin Cottage Inc, and Dollar General Corp. Each set of photos is titled with peers names and similarity score in the brackets. Each pair of photos demonstrate the cosine similarity distance measure.

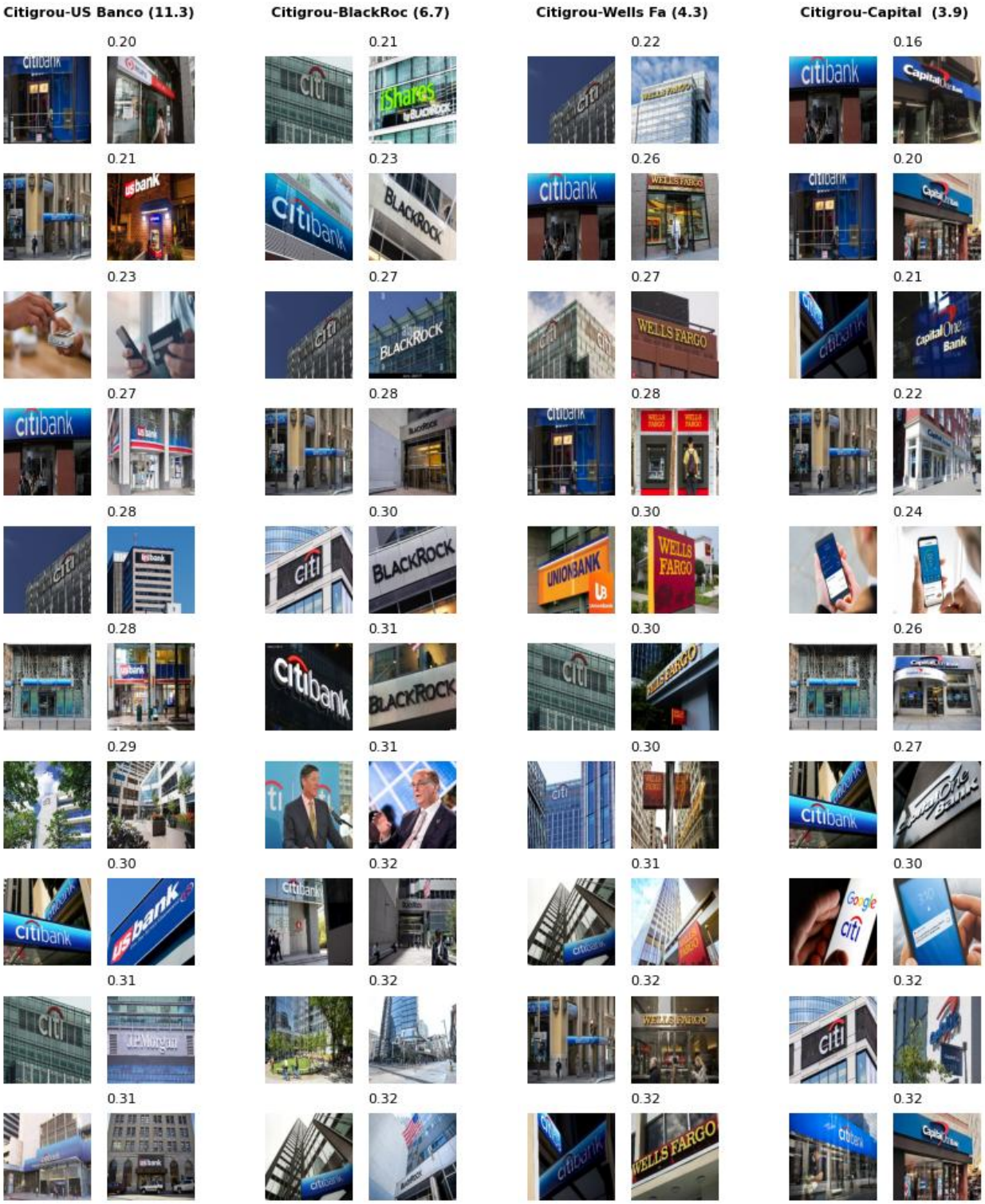

Figure 4: Photos representing four peers of Citigroup Inc with the highest similarity score

This figure presents photos related with Citigroup Inc and its four peers with the highest similarity score - from the left: US Bancorp, BlackRock, Wells Fargo & Co, and Capital One Financial Corp. Each set of photos is titled with peers names and similarity score in the brackets. Each pair of photos demonstrate the cosine similarity distance measure.

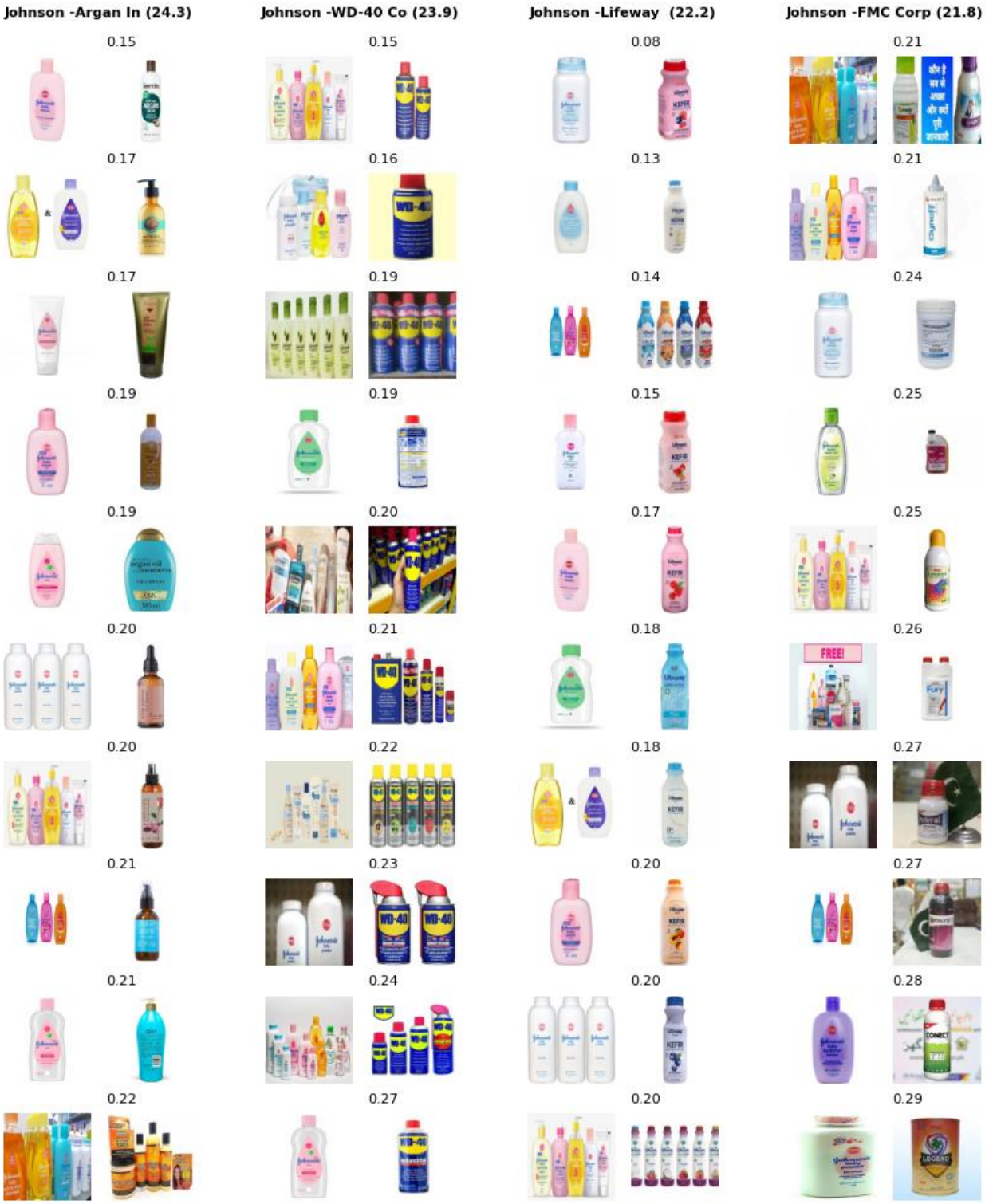

Figure 5: Photos representing four peers of Johnson & Johnson with the highest similarity score

This figure presents photos related with Johnson & Johnson and its four peers with the highest similarity score - from the left: Argan Inc, WD-40 Co, Lifeway Foods Inc, and FMC Corp. Each set of photos is titled with peers names and similarity score in the brackets. Each pair of photos demonstrate the cosine similarity distance measure.

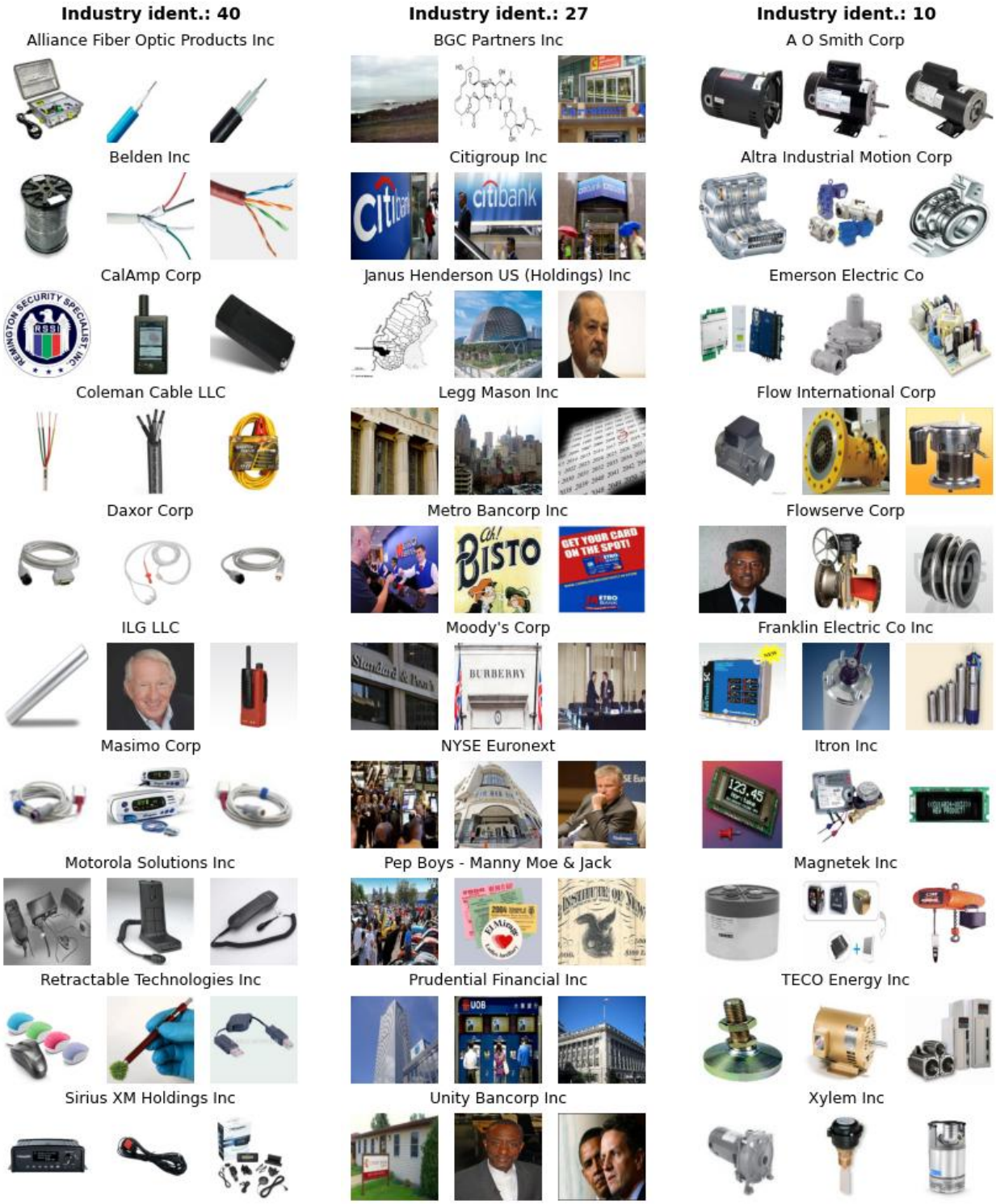

Figure 6: Photos representing industries with highest inter-industry correlations (1)

This figure presents photos from Image Industries 73 presenting three industries with high 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 2009 to 2013 and are used to form industries. We excluded industries represented with less than 10 stocks.

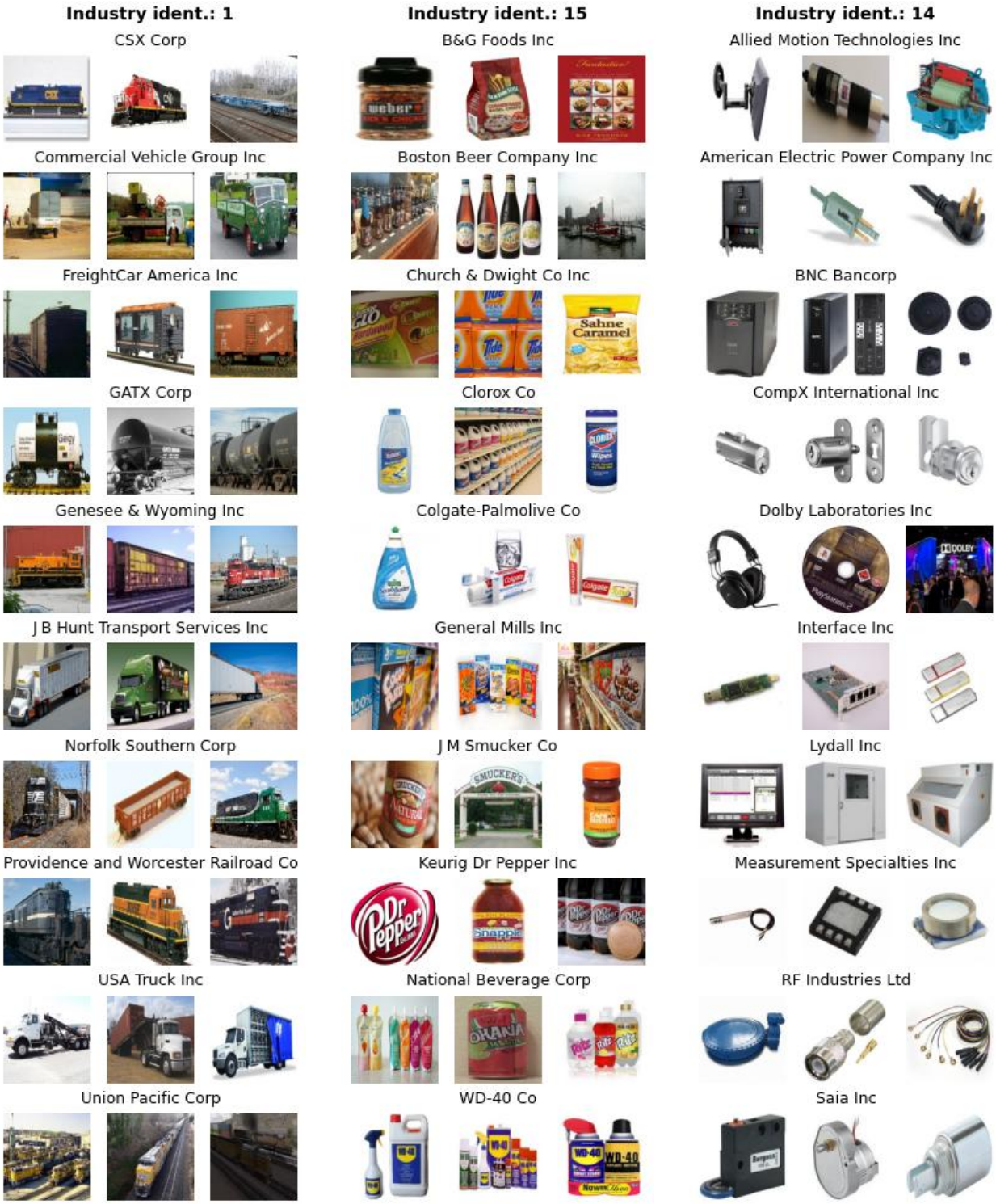

Figure 7: Photos representing industries with highest inter-industry correlations (1)

This figure presents photos from Image Industries 73 presenting three industries with high 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 2009 to 2013 and are used to form industries. We excluded industries represented with less than 10 stocks.

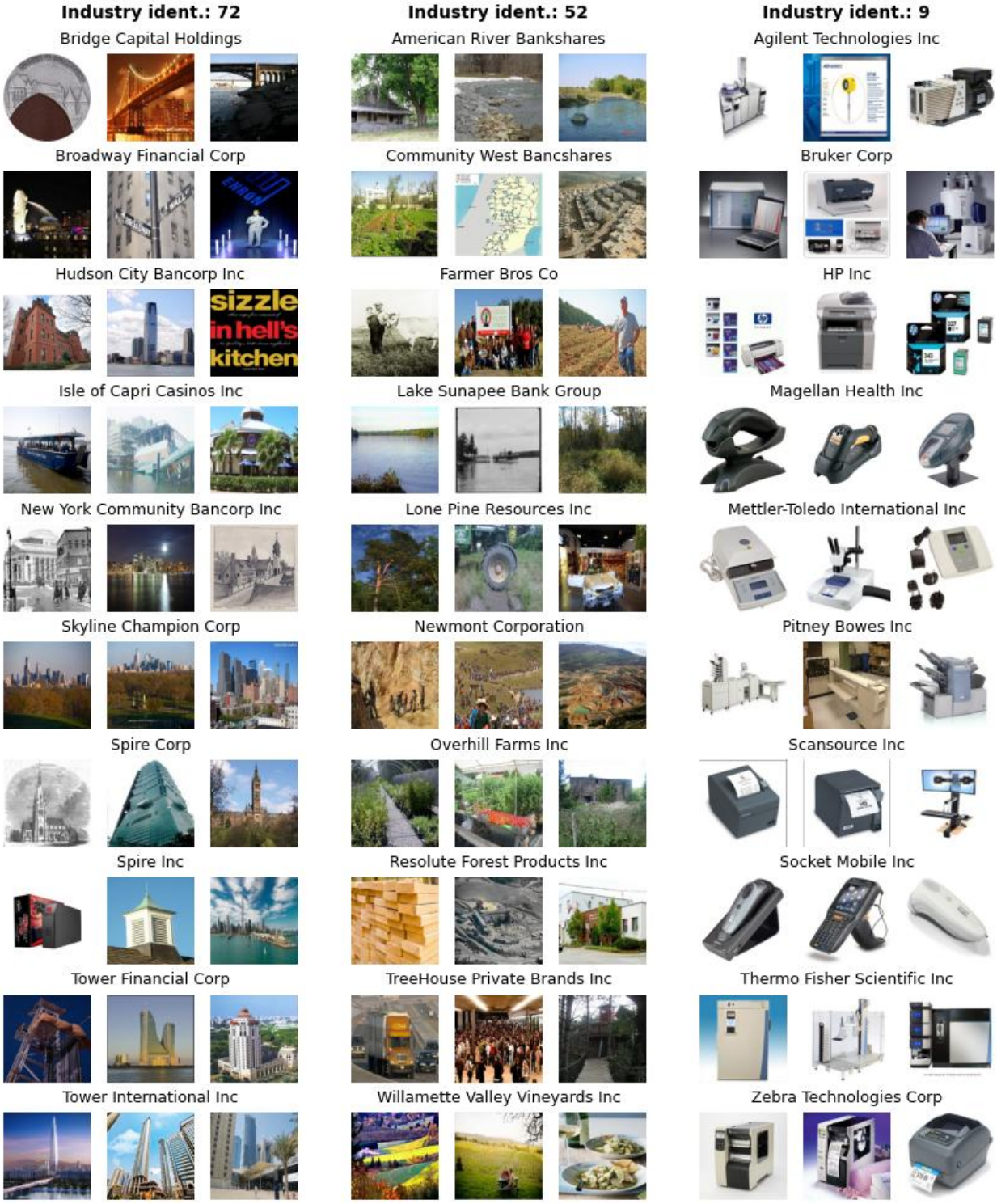

Figure 8: Photos representing industries with the lowest inter-industry correlations (1)

This figure presents photos from Image Industries 73 presenting three industries with low 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 2009 to 2013 and are used to form industries. We excluded industries represented with less than 10 stocks.

Figure 9: Photos representing industries with the lowest inter-industry correlations (2)

This figure presents photos from Image Industries 73 presenting three industries with low 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 2009 to 2013 and are used to form industries. We excluded industries represented with less than 10 stocks.

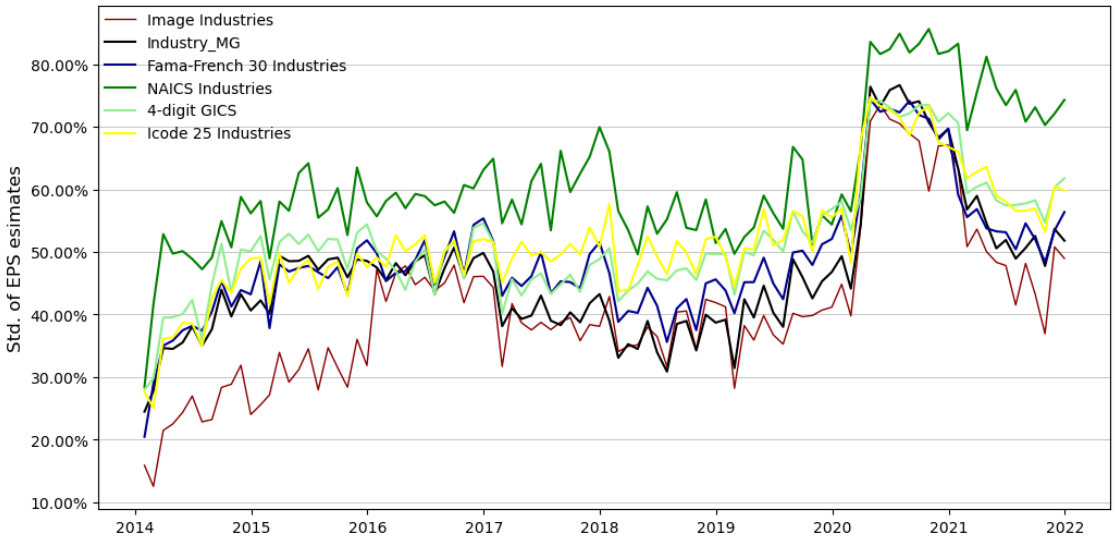

Figure 10: Investor's agreement

The chart illustrates the level of investor agreement regarding the behavior of companies within the same industries. We adopt the measure of agreement proposed by Diether, Malloy, and Scherbina (2002). Investor opinion diversity regarding a company i at time t is computed as the analysts' forecasts in the month prior to the fiscal period end date divided by the absolute value of the mean forecast. Subsequently, investor opinion diversity regarding all companies within a given industry j at time t is expressed as the standard deviation of the estimated ratios, while investor opinion divergence regarding industry classification at time t is the mean of these standard deviations. Forecast data is sourced from I/B/E/S detail files. The dataset covers the period from 2014 to 2021 and pertains to industries comprising at least 5 companies.

Appendix

A.1 Photo cleaning procedure

This section demonstrates our photo-cleaning procedure. Our database, upon downloading images from Google, includes 4.2 million images. Many of these are unsuitable for processing in our study. This section elucidates the methodology employed to cleanse the dataset by removing extraneous images.

The cleaning process begins with identifying and removing images containing text. Photos with text are predominantly comprised of notes, logotypes, or other elements unrelated to the firm's business operations. This step is pivotal in the data-cleaning process and accounts for the largest reduction in image count. We utilize the Python implementation of the open-source Tesseract 5.0 software for text detection on photos. The algorithm identifies text in 2.2 million images, representing over half of the downloaded images.

Subsequently, we eliminate images dominated by human faces. While some photos featuring human faces can convey certain business characteristics of firms, images overwhelmed by faces can occur across any company, irrespective of its industry. These often include photos of company employees, frequently showcased for marketing purposes. We employ the Open Source Computer Vision Library (OpenCV) (Bradski, 2000) to detect images dominated by faces, specifically utilizing the CascadeClassifier. The algorithm flags 0.2 million images as being dominated by human faces.

Further analysis reveals that photos with a predominantly white background typically depict graphs, and those with extreme size proportions often represent logos still present in the database. We remove images dominated by a white background (where 90%

After the comprehensive image-cleaning process, we are left with a sample of 1,629,175 images, averaging between 120,000 and 150,000 images per year. This refined dataset forms the basis for our subsequent analysis, ensuring that the visual data accurately reflects the business activities and characteristics of the firms.

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 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: Description and Summary Statistics of Image Industries 45 & 73 in industry formation period

The table presents a static overview of industries at the time of the first industry definition (year 2013) formed with firms' photos. Classification with 45 (73) classes has 25 (50) industries with at least five firms. Our sample covers NYSE, AMEX, and NASDAQ stocks. We report the average number of stocks assigned to each industry (No. of Stocks), the average monthly capitalization of stocks in each industry (Market Cap. (bn USD)), and the share of stocks classified with photos in the whole market capitalization (Avg. % of Market Cap.). Finally, we report the relationship between our industries to two-digit SIC codes by indicating the five most common SIC codes among the companies assigned to each of our industries. Industries marked in bold consist of at least five stocks.

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	TOP5 SIC codes
PANEL A: Image Industries 45 (25)				
0	30	463.5	2.3	('45', '37', '38', '35', '36')
1	63	406.1	2	('42', '35', '37', '40', '38')
2	2	2.4	0	('30', '36')
3	107	1,254.8	6.3	('36', '38', '48', '73', '35')
4	15	80.1	0.4	('56', '23', '51', '38', '60')
5	97	1,864.1	9.4	('20', '28', '38', '51', '59')
6	21	66.9	0.3	('60', '63', '67', '73', '82')
7	103	574.8	2.9	('36', '35', '50', '73', '38')
8	36	165.7	0.8	('35', '36', '38', '37', '34')
9	35	94.2	0.5	('36', '33', '35', '37', '30')
10	21	35.4	0.2	('25', '57', '24', '37', '52')
11	8	32.8	0.2	('99', '20', '53', '54', '59')
12	17	60.4	0.3	('28', '38', '26', '39', '73')
13	3	15.7	0.1	('73',)
14	6	35.8	0.2	('36', '49', '87', '96')
15	2	1.8	0	('20', '99')
16	2	41.3	0.2	('49', '73')
17	71	855.2	4.3	('58', '20', '53', '54', '99')
18	29	604.6	3	('73', '60', '62', '63', '99')
19	2	3.8	0	('73', '79')
20	4	7.4	0	('17', '73', '80', '99')
21	12	37.7	0.2	('26', '15', '27', '28', '30')
22	30	29.0	0.1	('60', '10', '13', '49', '67')
23	59	298.7	1.5	('60', '70', '15', '49', '63')
24	2	2.0	0	('10', '49')
25	21	31.2	0.2	('60', '20', '28', '49', '73')
26	36	202.0	1	('38', '35', '73', '36', '50')
27	39	180.9	0.9	('55', '37', '36', '50', '73')
28	2	74.7	0.4	('49',)
29	2	2.5	0	('64', '73')
30	1	0.1	0	('99',)
31	2	13.5	0.1	('37', '38')
32	2	1.8	0	('10', '28')
33	39	968.6	4.9	('13', '29', '44', '49', '16')
34	14	49.9	0.3	('16', '20', '22', '21', '25')
35	2	17.2	0.1	('37',)
36	2	0.9	0	('35', '36')
37	2	8.7	0	('50', '59')
38	1	0.0	0	('67',)

Table A2: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	TOP5 SIC codes
39	2	8.7	0	('61', '63')
40	24	33.9	0.2	('60', '79', '73', '26', '28')
41	2	4.4	0	('13', '30')
42	13	78.7	0.4	('30', '31', '56', '60')
43	2	2.2	0	('34',)
44	2	19.5	0.1	('34',)
Sum	987	8,733.6	43.9	
PANEL B: Image Industries 73 (50)				
0	24	328.1	1.6	('45', '37', '38', '35', '47')
1	43	320.9	1.6	('42', '37', '40', '35', '47')
2	2	2.4	0	('30', '36')
3	9	29.4	0.1	('38', '48', '34', '37', '59')
4	8	22.9	0.1	('56', '51', '23', '38')
5	2	0.3	0	('28',)
6	9	22.3	0.1	('73', '35', '38', '36', '37')
7	8	56.1	0.3	('63', '67', '73', '82')
8	76	373.7	1.9	('36', '35', '50', '73', '10')
9	12	145.8	0.7	('35', '38', '36', '50', '80')
10	14	95.9	0.5	('35', '36', '38', '49', '50')
11	22	35.4	0.2	('60', '22', '13', '34', '49')
12	10	16.1	0.1	('73', '79', '26', '28', '35')
13	43	809.9	4.1	('28', '38', '59', '51', '36')
14	16	48.0	0.2	('36', '38', '73', '22', '35')
15	46	918.9	4.6	('20', '28', '54', '26', '29')
16	21	78.3	0.4	('35', '37', '33', '36', '30')
17	21	35.4	0.2	('25', '57', '24', '37', '52')
18	5	5.4	0	('20', '59', '67', '73', '99')
19	11	54.9	0.3	('28', '26', '38', '99')
20	3	15.7	0.1	('73',)
21	10	76.0	0.4	('36', '37', '38', '13', '23')
22	6	35.8	0.2	('36', '49', '87', '96')
23	2	1.8	0	('20', '99')
24	31	22.2	0.1	('36', '38', '99', '50', '73')
25	2	41.3	0.2	('49', '73')
26	37	243.0	1.2	('58', '20', '99', '25', '28')
27	14	427.1	2.1	('60', '62', '73', '55', '63')
28	2	3.8	0	('73', '79')
29	15	44.0	0.2	('10', '60', '49', '12', '61')
30	4	7.4	0	('17', '73', '80', '99')
31	15	41.4	0.2	('60', '63', '28', '36', '78')
32	11	11.4	0.1	('28', '38', '73', '26', '36')
33	6	6.8	0	('60', '28', '49', '56', '70')
34	8	18.9	0.1	('26', '27', '35', '51', '59')
35	7	57.1	0.3	('56', '23', '57', '60')
36	8	7.9	0	('60', '13', '38', '65', '67')
37	7	6.2	0	('20', '16', '21', '22', '60')
38	2	2.0	0	('10', '49')
39	7	2.3	0	('28', '32', '56', '59', '73')

Table A2: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	TOP5 SIC codes
40	15	50.0	0.3	('36', '33', '38', '99', '48')
41	14	34.1	0.2	('38', '35', '36', '50')
42	24	214.2	1.1	('35', '37', '38', '36', '50')
43	3	7.0	0	('73', '99')
44	38	180.6	0.9	('55', '37', '36', '50', '73')
45	2	74.7	0.4	('49',)
46	40	228.9	1.2	('70', '15', '65', '79', '24')
47	13	170.7	0.9	('73', '60', '63', '35', '47')
48	2	2.5	0	('64', '73')
49	5	10.5	0.1	('36', '25', '60', '73')
50	1	0.1	0	('99',)
51	2	13.5	0.1	('37', '38')
52	9	3.0	0	('60', '20', '10', '24')
53	29	639.0	3.2	('53', '54', '58', '59', '55')
54	8	135.3	0.7	('39', '34', '48', '51', '53')
55	2	1.8	0	('10', '28')
56	38	935.0	4.7	('13', '29', '44', '49', '16')
57	7	43.7	0.2	('16', '22', '25', '28', '32')
58	2	17.2	0.1	('37',)
59	2	0.9	0	('35', '36')
60	4	18.8	0.1	('15', '28', '30', '99')
61	49	379.1	1.9	('35', '36', '38', '73', '48')
62	2	8.7	0	('50', '59')
63	1	0.0	0	('67',)
64	2	8.7	0	('61', '63')
65	6	17.7	0.1	('60', '73', '59')
66	2	4.4	0	('13', '30')
67	32	921.2	4.6	('48', '35', '36', '73', '34')
68	8	25.0	0.1	('58', '63', '99')
69	13	78.7	0.4	('30', '31', '56', '60')
70	2	2.2	0	('34',)
71	2	19.5	0.1	('34',)
72	9	14.8	0.1	('60', '37', '49', '73', '79')
Sum	987	8,733.6	43.9	

Table A3: 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 classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) 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 classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) 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 NYSE, AMEX, and NASDAQ stocks from 2014 to 2021. We calculate regression for industries that have at least five members. Table A1 in the 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 45 (25) - comparison										
Image Industries	25.2	18.9	23.9	2.4	31.7	23.8	10.0	31.5	36.1	22.8
Industry_MG	22.6	19.4	21.3	1.8	26.9	24.0	12.1	28.5	35.7	22.3
Fama-French 30 Industries	22.3	19.2	21.7	1.8	27.2	22.6	10.9	26.7	36.7	22.3
NAICS Industries	23.3	19.4	21.5	1.9	27.9	22.2	10.8	25.8	35.7	23.0
4-digit GICS	26.3	21.5	23.9	1.8	28.1	24.8	10.9	28.0	38.3	25.3
Icode 25 Industries	22.5	19.7	22.6	2.1	28.8	23.6	10.7	26.6	33.9	21.4
PANEL B: Image Industries 73 (50) - comparison										
Image Industries	26.5	18.7	23.1	2.6	31.5	25.1	10.5	32.1	34.6	23.6
2-digit SIC	24.0	19.2	21.7	1.8	27.7	22.9	11.5	28.4	35.9	23.8
Fama-French 48 Industries	24.3	19.1	21.8	1.8	27.4	22.0	11.2	27.4	36.7	23.9
3-digit NAICS	24.9	19.7	21.8	2.1	28.7	21.5	11.1	27.9	34.9	23.7
6-digit GICS	26.9	20.9	23.8	1.7	28.6	25.4	10.5	27.7	37.9	25.9
Icode 50 Industries	23.3	18.9	22.4	2.3	28.9	23.5	11.8	26.2	33.1	22.3

Table A4: 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 classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) 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 classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) 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 NYSE, AMEX, and NASDAQ stocks from 2014 to 2021. We calculate regression for industries that have at least five members. Table A1 in the 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 45 (25) - comparison																
Image Industries	41.9	43.4	6.0	27.9	16.6	39.1	22.2	22.9	23.5	33.4	42.5	15.8	26.1	16.7	15.6	22.9
Industry_MG	48.5	37.1	4.2	23.9	16.7	28.3	18.7	3.7	27.7	23.3	36.3	14.0	27.7	16.5	14.5	25.0
Fama-French 30 Industries	47.3	43.4	3.3	19.6	14.6	27.8	18.7	8.6	26.6	22.9	39.9	13.9	27.1	13.9	11.6	21.9
NAICS Industries	45.7	40.4	4.1	21.0	12.7	28.2	19.6	5.0	28.0	24.2	35.7	16.0	30.3	15.7	14.5	23.6
4-digit GICS	47.9	42.9	3.9	23.1	13.3	28.6	21.1	7.0	27.6	21.3	38.7	14.9	28.2	13.0	14.0	23.8
Icode 25 Industries	41.7	36.9	4.4	20.5	14.9	30.2	18.6	9.2	25.0	25.9	40.5	16.6	27.1	14.7	14.1	22.7
PANEL B: Image Industries 73 (50) - comparison																
Image Industries	39.2	45.1	7.6	28.5	17.6	38.3	27.4	26.3	21.9	33.6	41.4	16.9	24.9	19.8	16.7	23.4
2-digit SIC	45.5	41.6	3.9	21.1	13.4	28.2	18.6	9.4	25.6	22.3	38.2	14.2	27.0	13.8	13.1	23.4
Fama-French 48 Industries	45.8	42.6	3.5	21.7	13.8	28.4	18.9	10.6	25.6	22.9	38.4	14.4	28.1	14.1	11.9	22.0
3-digit NAICS	39.3	39.2	4.4	22.6	14.7	28.5	16.6	10.3	24.9	23.7	39.3	15.1	26.6	14.6	14.0	21.8
6-digit GICS	46.6	43.4	3.9	23.2	12.9	29.5	20.2	13.9	26.0	21.8	40.0	15.6	26.8	13.7	12.6	23.6
Icode 50 Industries	40.3	37.7	4.5	21.2	14.6	31.5	18.8	12.2	25.2	26.5	40.2	15.6	26.6	14.8	14.0	22.6

Table A5: 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 classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) 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 classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) 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 2014 to 2021 with prices not smaller than USD 5. We calculate regression for industries that have at least five members. Table A1 in the 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 45 (25) - comparison										
Image Industries	24.8	19.7	25.2	3.1	27.8	24.0	9.9	30.2	32.5	21.8
Industry_MG	24.1	21.8	27.6	2.7	28.7	26.7	10.4	30.7	34.1	23.1
Fama-French 30 Industries	23.8	21.1	27.6	2.5	27.5	27.3	9.0	28.7	34.6	22.4
NAICS Industries	25.5	22.9	27.1	3.0	28.2	22.6	9.7	27.3	34.8	24.2
4-digit GICS	26.7	21.9	29.4	2.5	29.5	26.9	10.7	29.5	37.2	24.6
Icode 25 Industries	22.9	22.4	27.2	3.0	28.7	26.1	9.7	27.2	31.1	20.5
PANEL B: Image Industries 73 (50) - comparison										
Image Industries	25.1	18.5	24.3	3.2	26.8	24.6	10.5	30.1	30.5	21.9
2-digit SIC	27.1	22.9	27.8	2.8	29.1	26.8	11.9	29.8	31.8	25.3
Fama-French 48 Industries	26.5	22.2	27.2	2.8	28.0	27.2	11.1	29.4	32.5	24.5
3-digit NAICS	25.9	23.0	26.7	3.4	28.4	25.2	12.0	30.2	32.0	23.2
6-digit GICS	27.3	23.4	28.1	2.5	30.0	28.9	11.4	30.4	35.2	24.3
Icode 50 Industries	22.6	20.2	27.4	3.4	29.7	25.1	11.4	27.8	29.8	20.0

Table A6: 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 classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) 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 classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) 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 2014 to 2021 with prices not smaller than USD 5. We calculate regression for industries that have at least five members. Table A1 in the 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 45 (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 73 (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 A7: Industry Momentum - Volatility Targeting

The table compares Sharpe ratios of momentum industry portfolios built with different industry classification techniques formed with 1) image industries with 45 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 2014-2021. We use sectors with at least five stocks.

	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	0.457	0.002	0.667	-0.068	0.021	0.190
Momentum 6m 9m	0.472	-0.078	0.427	-0.275	-0.162	-0.057
Momentum 6m 12m	0.508	-0.121	0.327	-0.403	-0.196	-0.219
PANEL B: Equally weighted industry returns						
Momentum 6m 6m	0.653	0.129	0.334	0.202	0.02	0.107
Momentum 6m 9m	0.483	-0.064	0.123	0.027	-0.196	-0.040
Momentum 6m 12m	0.259	-0.251	-0.138	-0.070	-0.235	-0.181