Just Look: Knowing Peers with Image Representation*

Tomasz Kaczmarek† Kuntara Pukthuanthong‡

March 12, 2026

Abstract

We identify a new dimension of economic connectivity—fundamental vertical adjacency—that images encode. By analyzing four million images of firm operations, we demonstrate that images encode critical information about a firm's position in the supply chain. Our Image-based Firm Similarity (IFS) groups firms with similar operational characteristics, generating peer groups with substantially lower analyst forecast dispersion than text-based benchmarks. Pair-trading strategies using IFS peers achieve Sharpe ratios exceeding 3.0, while IFS-based industry-momentum and diversification strategies outperform benchmarks by 20–40%.

Keywords: image-based similarity, firm similarity, industry classification, supply chains, machine learning.
JEL Codes: G11, G12, G14, C45.

---

*Code and data are available upon request. We thank Fred Bereskin, Christian Breitung (discussant), Gerard Hoberg, Kate Holland, Markku Kaustia, Dhagash Mehta, Mike O'Doherty, Ville Rantala, and seminar participants at Technical University of Munich (TUM) School of Management, 2nd Workshop on Advances in NLP and Generative AI in Finance and Management, Munich, Germany, DePaul University, 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 of Economics and Business, and Blackrock. We also appreciate Ville Rantala for generously sharing his updated co-analysts' data.

†Department of Investment and Financial Markets, Institute of Finance, Poznan University of Economics and Business, Poland. tomasz.kaczmarek@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. For Open Access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

‡Trulaske College of Business at the University of Missouri, Columbia, MO, USA. pukthuanthongk@missouri.edu

Disclosure Statements

Tomasz Kaczmarek declares that he has nothing to disclose.
Kuntara Pukthuanthong declares that he has nothing to disclose.

1 Introduction

What does an industry truly *look* like? Traditional classifications rely almost entirely on textual and numerical disclosures, overlooking a rich, vastly underutilized source of information: images that capture firms' products, facilities, and operations. We introduce Image-based Firm Similarity (IFS). This visual classification approach leverages information embedded in images of firms' products, facilities, and operations to organize firms into economically meaningful peer groups. In particular, images reveal how firms operate—through their technologies, physical assets, facilities, and organizational context—providing a window into vertical relationships along the value chain that are largely absent from traditional industry classifications and present fundamental vertical adjacency.

Capital market research frequently requires grouping companies into homogeneous categories based on firm similarities. These industry groups refine empirical tests, select control firms, benchmark performance, and summarize economic characteristics. Traditional frameworks — such as the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS), and the Global Industry Classification Standard (GICS)—are widely used.[1] However, these traditional systems have inherent limitations, particularly their limited ability to adapt to changes in firms' product mix, production technologies, and operational scope.

Investors perceive firm similarity through many channels: established industry classifications, textual descriptions in 10-K filings, financial statements, and analyst coverage. Images may encode an additional dimension of economic similarity, particularly by capturing firms' operational technologies and value-chain relationships that are inherently visual and difficult to infer from text alone. Firms' products, facilities, technologies, and operational footprints are visually documented in online images, which capture aspects of firm activity—such as

[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 Classification 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, as shown by Bhojraj, Lee, and Oler (2003) outperforming 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 proposed alternative approaches to industry classification, including Fama and French's FF algorithm, text-based groupings from 10-K business descriptions (Hoberg & Phillips, 2016; Lewellen, 2012), and analyst-based classifications (Kaustia & Rantala, 2021; Ramnath, 2002)

production processes, physical assets, and supply-chain relationships—that textual descriptions may understate or omit. If images contain economically relevant information beyond what text conveys, then classifications based on visual similarity should group firms with genuinely similar fundamentals and elicit greater agreement among market participants about those firms' prospects.

Investors perceive firm similarity through many channels: established industry classifications, textual descriptions in 10-K filings, financial statements, and analyst coverage. Images may encode an additional dimension of economic similarity. Firms' products, facilities, technologies, and operational footprints are visually documented in online images, which capture aspects of firm activity—such as production processes, physical assets, and supply-chain relationships—that textual descriptions may understate or omit. Just as Cohen, Malloy, and Nguyen (2020) show that textual changes in 10-Ks predict returns by encoding information that investors process slowly, we show that images encode operational information—particularly vertical supply-chain adjacencies—that text-based disclosures systematically underrepresent.

As modern firms evolve rapidly, effective classification systems must address two fundamental challenges: timely adaptation and operational complexity. Traditional frameworks update slowly because they depend on infrequent textual and numerical disclosures, causing industry assignments to lag behind firms' shifting product lines, technologies, and strategic focus. At the same time, many companies operate across multiple business domains, making it difficult for text-based or hierarchical systems to capture the full breadth of their activities. IFS directly tackles both limitations. By drawing on a broad set of visual cues—such as products, production facilities, equipment, and branding—IFS captures operational aspects of firm activity that are often slow to appear in textual disclosures. While industry assignments are updated at discrete intervals, the underlying visual information reflects firms' evolving operations more promptly than traditional classifications based on infrequently revised business descriptions. Moreover, because visual representations naturally span multiple dimensions of firm activity, IFS accommodates the operational complexity of modern firms operating in various product lines or technologies.

Visual information constitutes a distinct source of economically relevant signals about firms and their activities. Images reflect firms' products, production technologies, facilities, and branding in a direct manner that complements textual disclosures and financial state-

ments. Such visual content often conveys operational details that are difficult to infer from written disclosures alone. As a result, images can play an important role in how economic agents perceive similarity among firms and in forming judgments about which companies belong together.[2]

Building on these insights, we construct large-scale visual representations of all firms listed on the NYSE, NASDAQ, and AMEX by collecting over four million images from 2009 to 2021. For each firm and each year, we retrieve a panel of visuals that reflects time-varying economic activity. While multiple sources contribute to overall coverage, Google Images accounts for the largest share and the greatest diversity of visuals, capturing timely information about firms' products, facilities, branding, and operational footprint. We additionally explore the use of images extracted from annual reports and design patent filings to assess the robustness of our classifications; the extended results based on this broader dataset are reported in Internet Appendix C.[3]

To convert these diverse visuals into a usable analytical framework, we develop an image-based algorithm that compares firms through the similarity of their entire image sets. Using deep convolutional neural networks (VGG19), transfer learning, and object-recognition techniques, we extract feature embeddings for every image associated with each firm. We treat each firm as a collection of visual features and assess similarity by comparing embedding distributions across firm pairs. This set-based approach enables us to detect overlapping visual patterns—ranging from products and production technologies to supply chain elements and branding cues—that reflect meaningful aspects of firm activity.[4]

[2]A growing literature documents that visual information can convey economically relevant signals in financial markets. In particular, Obaid and Pukthuanthong (2022) show that visual cues embedded in images affect investor attention and sentiment, with measurable implications for asset prices. Evidence from cognitive science further suggests that humans process visual information efficiently and intuitively (e.g., Potter, Wyble, Hagmann, and McCourt (2014)), and that visual and spatial representations are engaged even when individuals process abstract quantitative information (e.g., Cohen Kadosh, Cohen Kadosh, Kaas, Henik, and Goebel (2007), Piazza and Izard (2009)). We view this literature as providing intuition for why images may encode information relevant for economic analysis, rather than as a direct test of our classification approach.

[3]Between 2009 and 2021, we initially collect over 4.2 million raw images from Google Images. After applying a comprehensive photo-cleaning procedure (described in Internet Appendix B), we retain approximately 1.63 million clean Google images, covering on average 3,458 firms per year. These are supplemented by 176,000 images extracted from 10-K annual reports (724–1,197 firms per year) and 181,000 images from design patent filings (201–344 firms per year). Together, these additional sources expand the post-cleaning image pool by roughly 20% and increase the average number of images per firm by about 15%.

[4]Throughout the paper, we use the terms photos, images, photographs, and pictures interchangeably.

Constructing a meaningful visual similarity measure at scale poses substantial methodological challenges. Corporate images are highly heterogeneous, often dominated by irrelevant objects—such as delivery trucks, office equipment, or generic consumer devices—that do not reflect a firm's core economic activity. With more than four million images in our dataset, naïve feature extraction would yield similarity patterns driven largely by noise rather than fundamentals. Our algorithm addresses this challenge by leveraging the frequency and persistence of visual features across firms' image sets, emphasizing elements that recur in product- and technology-related imagery while downweighting sporadic or tangential visuals. By comparing entire sets of image embeddings across firms, rather than collapsing them into a single representation, the method isolates economically meaningful signals from large, unstructured visual corpora. This innovation enables IFS to recover intuitive, robust firm similarities that would otherwise be obscured by the heterogeneous and noisy image data.

Our contributions are threefold. First, we introduce a new industry classification framework grounded in large-scale image data and set-based visual similarity. IFS complements and extends existing systems such as SIC, NAICS, GICS, and text-based classifications by providing both intransitive firm-pair similarity measures and transitive IFS industries that evolve with the visual information environment. The visual representation of firm assets allows us to recover vertical supply-chain adjacencies that are frequently "visually dark" to both search-based and text-based methods. While investor searches often cluster around horizontal competitors, our findings in Section 6 demonstrate that images encode critical information regarding a firm's position in the production process. By capturing these hidden transitive links, IFS provides a high-fidelity mapping of the U.S. innovation engine, explaining 85.9% of corporate R&D expenditure and identifying supply-chain concentrations that traditional methods overlook.

Second, we contribute to the growing literature on leveraging machine learning and unstructured data in finance by showing that images contain economically relevant information about firms' products, technologies, and operational contexts that is not fully captured by textual disclosures or accounting data. A key economic mechanism is that images encode how firms operate (including production processes, equipment, facilities, and other operational features), which allows IFS to recover vertical supply-chain adjacency in addition to horizontal product-market similarity. Consistent with this channel, IFS industries exhibit substantial heterogeneity in supplier and customer concentration, with customer intensity

reaching up to 55% in the upper tail.

Third, we show that image-based firm similarity yields industry groupings that are at least as economically homogeneous as those produced by leading classification systems, while offering clear advantages in empirical applications, including peer-based relative-value tests, portfolio diversification, and industry-based momentum.

Using the image-based similarity measure introduced above, we construct two classification systems, Image Industries 45 and Image Industries 73, which we refer to as IFS 45(25) and IFS 73(50) in the main text. We obtain both by clustering firms into groups based on visual similarity, and we focus on classifications in which at least 25 and 50 groups, respectively, contain a minimum of five firms.[5] These classifications allow us to evaluate the economic relevance of IFS in explaining firm-level outcomes and in a range of trading applications.

We preview our main findings. First, IFS explains cross-sectional variation in accounting and market-based ratios at least as well as, and often better than, leading industry classifications such as GICS and the text-based classifications of Hoberg and Phillips (2016) (henceforth, HP). It delivers particularly strong performance in several ratios associated with market expectations, innovation intensity, growth, and intangibles. Second, image-based industries deliver economically meaningful performance in portfolio-based applications. IFS provides strong performance in peer-based relative-value trading, diversification, and momentum applications, often delivering substantially higher risk-adjusted performance than competing classification systems across a range of portfolio constructions. Third, IFS is highly dynamic, with systematically higher industry reclassification than GICS, SIC, NAICS, or HP, reflecting more timely adjustments to firms' product offerings and operations. Finally, IFS captures economically meaningful vertical links in supply chains in addition to horizontal similarities, offering a richer representation of firm relationships.

IFS performs particularly well in sectors where firms' products, processes, and technologies are visually distinctive, such as manufacturing, chemicals, and other asset-intensive

[5]Throughout the main text, we refer to the image-based classifications as IFS 45(25) and IFS 73(50), which are constructed using Google Images. For completeness, we also build extended versions that incorporate images from annual reports and design patent filings. When referring to these variants explicitly, we denote the Google-only classifications as IFS 45(25) and IFS 73(50), and the extended-image versions as IFS-X 34(25) and IFS-X 65(50). The extended-sample results are reported and discussed in Internet Internet Appendix C.

industries. At the same time, visual data capture elements of firm activity—such as R&D labs, prototypes, and branding campaigns—that are closely related to growth and innovation, helping IFS classify such firms in ways that differ from text-based systems. This dual ability to encode both tangible operations and intangible value drivers helps explain why image-based classifications perform strongly across a wide range of economic environments.

2 Comparative analysis of industry classification metrics

Industry classification has long been a central topic in financial economics, and existing approaches can be broadly classified as *transitive* or *intransitive*. Transitive classifications assign each firm to a single industry within a hierarchical structure. This group includes the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS), and the Global Industry Classification Standard (GICS). These classifications reflect declared business activities, are widely used in both academia and industry, and are particularly suited for applications requiring mutually exclusive sector definitions—such as industry momentum, portfolio diversification, and factor construction.

Intransitive classifications instead define similarity at the firm-pair level. They include text-based similarity measures from 10-K business descriptions (Hoberg & Phillips, 2016), analyst-based peer groups (Kaustia & Rantala, 2021), characteristic-based clusters (He, Wang, & Yu, 2021), technological similarity based on patent networks (Lee, Sun, Wang, & Zhang, 2019), and EDGAR search-based linkages (Lee, Ma, & Wang, 2015). While these approaches offer significant flexibility over rigid SIC codes, they share a common dependency on *symbolic information*—either what firms choose to disclose in text or what investors perceive in their search behavior.

In contrast, IFS identifies economic linkages by recovering the *operational reality* of firm fundamentals directly from their physical assets. While text and search-based methods are inherently limited by disclosure incentives and cognitive attention, images of production machinery, facility layouts, and logistics infrastructure provide an objective "visual footprint" of how a firm actually operates. Also, IFS presents fundamental vertical adjacency.

This distinction is economically meaningful for two reasons. First, it enables the recovery of transitive structural ties that exist independently of market attention; for instance, IFS can identify that Firm A and Firm C are operational peers if they share unique production

technologies, even if market participants only co-search them with a common intermediary, Firm B. Second, and more importantly, this visual representation allows us to move beyond horizontal product-market similarity. Because images encode the physical nature of production processes, IFS recovers *vertical supply-chain adjacencies* that are frequently "visually dark" to text-based and search-based methods. This structural insight explains why our IFS peer groups exhibit up to 55% higher customer and supplier concentration than traditional benchmarks.

Notably, both the Hoberg and Phillips (HP) text-based approach and our Image-based Firm Similarity (IFS) framework can be implemented in both *transitive* and *intransitive* forms. In their intransitive implementation, HP provides a ranked list of the nearest textual peers for each firm, derived from cosine similarity within a 10-K product-market vocabulary. This list-based structure is flexible and useful for identifying economically related firms, but it is typically used as a ranked list of nearest peers rather than as a fully populated similarity matrix. HP also constructs a *transitive* industry classification by clustering the underlying textual similarity network.

IFS mirrors this dual structure. Our method produces (i) intransitive, firm-level similarity scores based on the distances between sets of image embeddings, and (ii) transitive classifications obtained by clustering the visual similarity graph. This parallel structure enables consistent comparison of IFS and HP across applications that require either list-based peer identification (e.g., our relative-value trading application) or mutually exclusive industry groups (e.g., diversification and momentum).

Compared with existing intransitive approaches, IFS introduces several novel elements. First, whereas HP measures similarity using cosine distances in 10-K product-market text and Kaustia and Rantala (2021) relies on common analyst coverage, IFS is built from visual evidence about firms' products, production technologies, facilities, branding, and supply-chain context. Second, IFS is designed to accommodate dynamic updating because images are uploaded continuously; the underlying visual information can reflect changes as firms introduce new products, reconfigure operations, or reposition their brands. By contrast, HP and KR are tied to annual reporting and analyst revisions. Third, the visual channel enables the capture of both horizontal product-market overlap and vertical technological or supply-chain linkages, whereas existing systems primarily emphasize horizontal relationships. Together, these features position IFS as a potentially valuable complement to HP and KR

in applications that benefit from timely and information-rich measures of firm-relatedness.

Previous research has compared the informativeness of traditional industry systems. Kahle and Walkling (1996) examine differences between SIC codes from CRSP and Compustat; Fama and French (1997) introduce new industry groupings based on four-digit SIC aggregation; and Krishnan and Press (2003) compare SIC and NAICS. Bhojraj et al. (2003) conduct a comprehensive comparison of fixed industry classifications. These studies provide valuable benchmarks, but rely on static, activity-based taxonomies and do not revisit the underlying methodology of industry classification or incorporate new data modalities.

Traditional classifications such as SIC, NAICS, and GICS have at least two important limitations. First, they are slow to adapt to changes in firms' product offerings and business models. Because they depend on infrequently updated textual disclosures, industry assignments often lag behind firms' operational shifts. This rigidity was evident during the rise of the internet and technology firms in the early 2000s, when hundreds of web-based companies were grouped under broad categories such as "Business Services." In contrast, IFS can respond more rapidly because visual data reflect firms' contemporaneous activities with shorter informational delays. Photos uploaded to the internet often depict newly introduced products and operational changes, which can quickly link firms to new economic peers. In contrast, firms' business descriptions in 10-K filings are relatively sticky and update only gradually as core activities evolve. Although image data are updated annually and firms are reclassified biennially, visual information still reflects operational changes more quickly than traditional text-based classifications. This difference is illustrated by cases where IFS identifies major shifts in corporate activity earlier than traditional systems.[6]

Second, images provide a more direct representation of firms' products, technologies, and supply-chain relationships. Visual cues capture operational changes, product launches, and technological upgrades that textual descriptions often obscure. Moreover, the human visual system processes images rapidly and holistically, so IFS signals align naturally with how market participants form and update peer groups.

[6]For example, SIC (NAICS) reclassified Exxon Mobil Corp. from petroleum & coal products to oil & gas extraction in 2017. In contrast, IFS identified this shift approximately two years earlier using photos from 2013–2015. Likewise, Hornbeck Offshore Services Inc. was reclassified by SIC in 2018, though IFS indicated the transition already in 2014–2016. Similar patterns are observed for FMC Corp., Oshkosh Corp., and SEACOR Holdings Inc., in which IFS detected changes in core operations well before traditional classifications identified them.

Text- and analyst-based groupings offer advantages over SIC, NAICS, and GICS. HP's intransitive similarity list is updated annually, providing flexibility for peer-based relative-value trading, valuation, and comparable-firm analysis. Analyst-based peer groups (KR) identify economically meaningful connections based on standard coverage and are helpful for intransitive applications, though they lack breadth of coverage. In our analysis, we therefore include HP as the primary text-based benchmark—using both its transitive and intransitive forms—and KR for relative-value trading comparisons. This allows a consistent evaluation of how IFS complements or improves upon existing non-hierarchical systems. Throughout the paper, we therefore treat HP and KR as our primary non-visual benchmarks and use them to highlight where IFS strengthens, refines, or overturns inferences drawn from text-based and analyst-based peer definitions.

2.1 Image-based Firm Similarity vs. Other Intransitive Approaches

IFS differs from other intransitive methods in several key dimensions.[7] First, the underlying data source is visual rather than textual, numerical, or analyst-generated. IFS analyzes approximately four million images, with an average of 75 cleaned images per firm per three-year rolling period (see Table 1). Each image is represented as a $224 \times 224 \times 3$ tensor capturing rich information about product design, production technologies, branding, facilities, and supply-chain context. This multidimensional visual information is particularly well suited to capturing similarities among firms with tangible products or visually distinctive operations.

Second, IFS is designed to accommodate dynamic updating. Images are uploaded continuously throughout the year, allowing IFS to capture changes in firms' activities more quickly than text-based or numerical approaches. This feature is particularly relevant in industries characterized by evolving product lines or shifting production processes. Although textual and analyst-based approaches (HP and KR) are revised annually, they remain constrained by the lag structure of financial reporting. In contrast, IFS can reflect more contemporaneous aspects of firms' operational and product environments.

Third, while IFS provides broad coverage, it is less effective for firms whose activities are not readily visualized, such as consulting firms or technology companies. We address

[7] A *transitive* classification assigns each firm to exactly one mutually exclusive industry group. An *intransitive* classification defines similarity at the firm-pair level without requiring hierarchical groupings; a single firm may have varying degrees of similarity to many other firms.

this limitation by incorporating images from design patents and 10-K filings, which enhance coverage for R&D-intensive and small-cap firms. However, constructing more than 100 transitive clusters remains challenging due to limited image availability for some firms. Detailed results on the expanded dataset appear in Section C.

3 Methodologies and Data

This section provides an overview of the data and methodological framework used to construct Image-based Firm Similarity (IFS). Our goal is to give an intuitive description of the procedure while keeping technical details in Internet Appendix A.

3.1 Data

Our stock universe comprises all firms listed on the NYSE, NASDAQ, and AMEX, as obtained from CRSP. For each firm and each calendar year from 2009 to 2021, we collect images directly from Google Images using a standardized query structure (e.g., "{Company Name} products after:YYYY-01-01 before:YYYY-12-31"). For each firm-year, we retrieve up to 100 images using an automated query interface. The 100-image cap strikes a balance between richness and data quality: for firms with extensive online presence, it prevents the similarity measure from being dominated by lower-quality or redundant visuals that appear deeper in Google's ranking; for firms with sparser coverage, we require a minimum number of cleaned images to construct a reliable representation. Table 1 documents the distribution of images per firm and period. This design yields an average of roughly 75 cleaned images per firm per three-year rolling window, while preserving a broad cross-section of listed firms.

Due to copyright restrictions, we do not redistribute the underlying images. Replication files (available upon request) include image embeddings, firm–image mappings with timestamps, and code for constructing similarity measures and peer groups, allowing all results to be reproduced without access to the raw images.

Relying on Google's ranking inevitably introduces an editorial component: the images we observe are those that Google deems most credible and relevant for each firm. We view this as a feature rather than a limitation. The ranking algorithms aggregate dispersed judgments about which images best represent each company, pushing official product photos, facility

images, and high-quality media coverage to the top. Our methodology builds on this curated set by reweighting and matching images across firms to extract economically meaningful structure, rather than simply reproducing Google's notion of "similar" content.

To improve coverage—particularly for R&D-intensive, small-cap, and service-oriented firms—we supplement Google Images with visuals extracted from 10-K annual reports and design patent filings. These additional sources increase both image depth and firm coverage; Internet Appendix C provides details on the extended image collection.

Raw images then undergo an extensive cleaning procedure to remove low-information or generic content (e.g., faces, logotypes, and landscapes), as detailed in Internet Appendix B.

To measure economic similarity, we collect 26 financial ratios following Kaustia and Rantala (2021), including valuation, profitability, growth, leverage, investment, and intangibility measures. Industry codes (SIC, NAICS, GICS), text-based similarity scores from Hoberg and Phillips (2016), and analyst-based peer groups from Kaustia and Rantala (2021) serve as benchmark systems.

3.2 Overview of the Methodology

IFS consists of three layers: (i) extracting visual features from images, (ii) constructing firm-pair similarity scores, and (iii) clustering firms into industries. We provide a high-level description here; full mathematical details appear in Internet Appendix A.

Methodological Overview and Economic Intuition. The central challenge in measuring visual similarity is distinguishing economically meaningful overlap from coincidental or generic matches. Consider two firms: both might have images of iPhones (because employees use them), glass buildings (because many companies occupy similar office towers), or people in business attire (because executives appear in press coverage). Without filtering, Citibank and ExxonMobil would appear visually similar despite operating in entirely different industries. Our methodology addresses this challenge through a five-step procedure that isolates firm-specific visual patterns from generic imagery. We first extract standardized features from each image, then compare entire image sets across firms using optimal matching, decompose observed similarity into expected and excess components, impose temporal robustness filters, and finally cluster firms into industries. Each step serves a specific economic purpose in separating signal from noise.

Step 1: Feature Extraction. Each image is processed using a deep convolutional neural network (VGG19) pre-trained on ImageNet. The network transforms each image into a high-dimensional feature vector that summarizes salient visual attributes, such as shapes, textures, colors, and object configurations. These embeddings provide a standardized numerical representation of visual content and form the basis for all subsequent comparisons.

To reduce noise and improve numerical stability, we apply principal component analysis (PCA) to the full set of image embeddings and retain the minimum number of components required to explain at least 70% of the total variance. This dimensionality reduction step compresses the visual representation while preserving the dominant sources of variation across images. The resulting PCA-reduced feature vectors constitute the final image-level representation used in similarity analysis. Figure 1 illustrates this feature extraction pipeline.

Step 2: Measuring Firm-Pair Similarity. We begin by defining similarity at the image level. For a given pair of firms and a given time period, we compute visual similarity between all pairs of images drawn from the two firms' respective image sets. For each such image pair, visual similarity is measured using cosine similarity applied to the PCA-reduced image embeddings. Cosine similarity captures the degree to which two images share visual features—such as shape, structure, and composition—regardless of overall feature magnitude. It is a standard measure in high-dimensional representation learning.

The central challenge, then, is to aggregate image-level similarity into a firm-level measure. A naive approach would sum or average cosine similarity across all possible image pairs between two firms. However, this strategy performs poorly in practice. When dozens of images represent firms, most image pairs are weakly or not at all related. As a result, a small number of highly similar and economically meaningful image pairs—such as repeated depictions of similar products or production technologies—are overwhelmed by a large mass of low, essentially random similarities. This makes firm-level similarity noisy and unstable, potentially leading to spurious comparisons across firm pairs.

To address this issue, we use a linear-sum assignment procedure (LSAP) to compare firms' image sets. LSAP constructs a firm-level similarity measure by optimally matching images from one firm to images from the other and summing similarity only across the strongest one-to-one correspondences. By design, this procedure emphasizes the most informative visual matches and prevents high-quality signal from being diluted by numerous weak or accidental

similarities. When firms differ in the number of available images, LSAP leaves unmatched images, further concentrating the comparison on the most relevant visual overlap.

The resulting firm-level similarity still reflects both economically meaningful overlap and mechanical similarity arising from generic imagery and image-set size. We therefore decompose firm-pair similarity into two components. *Expected similarity* captures the level of visual overlap that arises mechanically given the sizes of the two image sets and the prevalence of common visual elements across firms. Intuitively, expected similarity is estimated as the average level of firm-pair similarity observed in the data for a given comparison scale, measured by the number of feasible image matches between two firms.

Excess similarity is then defined as the residual component—i.e., the portion of observed firm-pair similarity that exceeds this expected benchmark. By construction, excess similarity reflects firm-pair-specific visual patterns that are unusually strong relative to what would be expected under a purely mechanical model given the image-set size. Only statistically significant excess similarity is retained in the final peer graph.

In summary, firm-pair similarity is constructed by (i) computing cosine similarity between PCA-reduced image embeddings for all image pairs drawn from two firms' image sets; (ii) aggregating image-level similarity to the firm level using LSAP; and (iii) decomposing firm-pair similarity into expected and excess components using weighted least squares, retaining only statistically significant excess similarity.

Step 3: Temporal Robustness. To avoid spurious matches, we require that statistically significant excess similarity between two firms appears in at least two of the last three years (including year t). This filter balances Type I and Type II errors: requiring only one year would retain too many transient matches driven by temporary image availability or short-lived marketing campaigns; requiring all three years would exclude genuine relationships with minor data gaps. Once a peer connection is established, it is retained for one additional year to reduce the risk of false negatives from brief data perturbations. The resulting similarity matrices are sparse but stable, with roughly 1% of all firm pairs classified as significant peers in any given year.

Step 4: Clustering Firms into Industries. Significant similarity scores are rescaled and used to construct visual clusters. We employ an iterative merging algorithm inspired by

Hoberg and Phillips (2016). Firms that start as singletons, clusters with the highest mutual similarity merge, and firms that may be reassigned to the closest cluster after each merge. This process continues until a desired granularity is reached. Our two main classifications—IFS 45 and IFS 73—ensure that at least 25 and 50 clusters, respectively, contain five or more firms.

Step 5: Dynamic Updating. Because firms continually introduce new products or shift strategic focus, we update classifications biennially using rolling three-year windows. Approximately two-thirds of images in each update are new, enabling IFS to track changes in firm activity more rapidly than text-based classifications tied to annual 10-K filings. This dynamism is particularly valuable for industries characterized by rapid product innovation or shifting production technologies.

Overall, IFS replaces static, activity-based categorization with a dynamic, visually grounded representation of firm similarity. The methodology isolates economically meaningful visual overlap from generic imagery, producing peer groups that reflect operational and technological relationships as well as product-market similarity. The following sections evaluate the economic performance and investment applications of this system relative to existing classification approaches.

4 Results

This section examines the relatedness of firms clustered by photos illustrating their business activities. We begin by reporting the characteristics of firms' similarities, captured using images, and the industries formed from those similarities. We then examine the usefulness and limitations of IFS in explaining cross-sectional variation in firm-level stock returns, market-based valuation multiples, and financial ratios, and complement this evidence with an analysis of investor agreement across industries. Throughout, we compare IFS to other industry classification methods, including SIC, NAICS, GICS, Fama–French, as well as Hoberg and Phillips (2016) industries.

4.1 Firm Relatedness Visualized by Image

In exploring firm-relatedness using visual data, our approach leverages images to capture 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 industries and are visualized in Figures 2, 3, 4, and 5.

Exxon Mobil Corp.'s range of images includes refineries, petroleum stations, drilling rigs, 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 visual profile: images of glass skyscrapers, logos, and people engaged in financial activities (e.g., TV interviews, use of banking apps) represent the financial services industry. Johnson & Johnson's image similarities are more straightforward, with visuals of cosmetics and standardized packaged chemicals highlighting its product offerings, showcasing the ability of our image-based approach to bind companies through diverse, but economically related, visual content.

In Section 5, we include images from patents and 10-K filings, as these may provide additional context that differentiates firms with overlapping visual elements. For example, technical schematics or operational visuals could clarify distinctions between financial institutions and energy companies. However, it should be noted 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 in the last column of images in Figure 2. This misclassification arises from visual overlap between corporate logotypes and architectural similarities between refineries and skyscrapers.

Building on the foundation of visual similarities, the next phase of our analysis delves into the visual features of IFS. We create the IFS with a large set of photos. Table 1 shows that the final average number of photos used to link firms to industries in a single period is more than 200 thousand. The typical industry is represented by at least several hundred images per period. The detailed human-made evaluation of all photos across all industries

is undoubtedly beyond the scope of this research.[8]

Figures 6 and 7 illustrate random representations of six sectors of IFS 73, characterized by high intra-industry correlation. The visual similarity of object descriptions for randomly selected companies across industries is apparent. For example, many photographs depicting Industry 40 depict electronic components and circuit boards. Similarly, Industry 15 displays images of packaging, bottles, and what appear to be chemical or cosmetic products. In contrast, Industry 1 consistently shows a range of vehicles, demonstrating the first characteristic of IFS: selective similarity aggregation. This approach groups visually similar objects across different industries rather than narrowly clustering identical objects within a single industry. For example, firms producing cars, trucks, and motorcycles may be grouped by shared visual features, thereby capturing broader economic linkages.

In contrast, Figures 8 and 9 highlight industries with lower inter-industry correlations, demonstrating greater object diversity within each sector. Each of the six industries depicted in these figures contains various objects, some of which may not seem intuitively related to human observers. This suggests that IFS may capture sophisticated object relations that are not immediately obvious. While this feature could be advantageous for IFS by identifying hidden common visual representations across different sectors, it may also pose a challenge when these visual similarities are unrelated to companies' actual product offerings or core business activities.[9]

Next, we shift our attention from the micro level to the bigger picture. Table 2 demonstrates the characteristics of IFS. First, our IFS classifies 50.8%-53.3% of our stock sample, corresponding to a monthly average of 1,084-1,110 firms. The photo representation of unclassified firms is of insufficient quality to determine the class to which it should be assigned.[10] Second, many industries contain a small number of companies. This phenomenon

[8] A process where human experts manually review and assess a subset of images to identify key characteristics and ensure the accuracy of the automated classification. Nevertheless, we wish to note several observable characteristics even after examining only dozens of images.

[9] IFS can capture relationships that are not obvious from standard classifications. (i) It can cluster firms that share visual elements in branding, packaging, facilities, or production technologies, even when their final products differ. (ii) These patterns often reflect broader economic linkages, such as vertical supply-chain roles or functional similarities, rather than narrow product or revenue overlap. (iii) The diversity within industries in Figures 8 and 9 suggests that IFS goes beyond intuitive clustering and can uncover economically meaningful connections that human judgment or text-based methods may miss.

[10] IFS classifies a 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.

is attributed to the sparsity of the similarity matrix used to construct industries. 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 concerns the total number of industries, including those comprising a single company. The second encompasses the number of industries with at least five companies at the time of the first industry definition.[11] Consequently, a classification that has 25 (50) industries comprising at least five companies totals 45 (73) clusters. Using industries with at least five companies is especially crucial for economic tests and proposed applications.

IFS classifies firms representing $13.8–14.4 trillion in market capitalization— approximately 50–53% of the total market. This coverage reflects our quality-control design: we include only firms with sufficient visual representation to ensure reliable classification. The resulting sample captures the majority of economically significant firms, including most large-cap companies with tangible products and operations.[12]

Table 3 presents the number of industries and the number of companies for the classifications used as benchmarks in our tests.

4.2 Economic homogeneity of Firms Clustered with Image

To verify economic relations between companies classified into a single industry, we apply the methodology proposed by Bhojraj et al. (2003). We construct equal-weighted industry portfolios and assess their ability to explain contemporaneous firm-level indicators ($vble_t$). We estimate the average adjusted R^2 for all firms within each cluster. R^2 is calculated from a regression below for each firm i:

$$\mathrm{v}ble_{i,t} = \alpha + \beta\, \overline{vble}_{-i,ind,t} + \varepsilon_{i,t}, \tag{1}$$

where $\overline{vble}_{-i,ind,t}$ is the equal-weighted average of variable $vble$ for all firms in industry ind excluding firm i at time t. $vble$ represents ten ratios using market information: 1) MARKET to BOOK; 2) PRICE to BOOK; 3) MONTHLY RET; 4) FORECASTED MONTHLY

[11] Table F2 demonstrates descriptive statistics of IFS defined in 2013. Industries marked in bold have at least five stocks.

[12] Section 4.3 provides a detailed analysis of the coverage properties of IFS, including comparisons to alternative classification systems and the characteristics of classified and unclassified firms.

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 mitigate the potential effect of outliers, we winsorize at the 1% level for all ratios except MONTHLY RET and the binary variable DIV PAYMENT. Also, we perform only one-sided 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 FORECASTED MONTHLY RET, TOBIN'S Q, R&D GROWTH, SG&A to # EMPLOYEES, SG&A GROWTH, FORECASTED EPS, and EPS GROWTH. Table F1 in the Appendix shows details of the ratios calculation. Based on market information, we calculate $vble_{ind,t}$ monthly frequency for ten indicators. We estimate quarterly industry averages for 16 ratios based solely on accounting information, consistent with the reporting frequency. We estimate regression models for industries with at least five members. The average adjusted R^2 indicates the explanatory power of the industry average in the firm's use ratio.

For comparability, Tables 4 and 5 present pairwise results comparisons derived from the *same* set of firms classified using IFS and other classification techniques. Table 4 compares IFS with market information-based variables. The average rank for the two IFS classifications ranges from first to last. IFS classifications exhibit high homogeneity across two variables: forecasted monthly returns and the price-to-earnings ratio. Overall, the economic homogeneity based on market-based ratios is average and comparable to that of SIC-based industries, including the Fama-French Industries. This result is based on a sample of firms for which data are available for all classifications.

When considering all firms (see Table F3), the results for market-based indicators show more favorable overall rankings for IFS. Image-based classification performs comparably to GICS industries. The most significant improvements are observed in the market-to-book ratio, monthly returns, market leverage, beta, and enterprise value-to-sales ratio, with no declines in performance. Microcap stocks typically have lower liquidity. When we exclude

them from the analysis (see Table F5), the ranking of IFS remains consistent.

We compare ratios based solely on accounting information, as reported in Table 5. The average rank of image-based classification varies widely, achieving the highest homogeneity for net sales, sales growth, R&D expense to sales, R&D growth, SG&A growth, dividend payout, profit margin, and debt to equity. However, it performs poorly on the asset-to-sales ratio. Comparing image- and text-based classifications shows that similar classifications outperform one another in 60% of observations. IFS outperforms GICS in 63% of cases.

Considering all firms (see Table F4), IFS is the most robust classification, ranking first or second in 75% of observations. Notably, excluding microcaps does not significantly alter IFS Industries' accounting-based ratio ranking (see Table F6). This consistency underscores the robustness of our IFS classification across different firm sizes.

The economic magnitude of these differences is substantial. For forecasted monthly returns, IFS achieves an R^2 of 2.4-2.6%, compared to 2.1-2.3% for SIC-based methods—a 13-14-% improvement in explanatory power. For beta, IFS 73(50) is the best-performing classifier, achieving 32.1%. More strikingly, for growth-related metrics, IFS dominates: R&D growth R^2 reaches 22.9% versus 16.8% for Industry classified by Moskowitz and Grinblatt (1999) (henceforth, Industry_MG) and 18.5% for text-based classifications, representing a 36% improvement over traditional methods. These gains are economically meaningful because they translate directly into improved peer identification for valuation, better controls in cross-sectional regressions, and more profitable trading strategies, as we demonstrate in Section 7.

Overall, the homogeneity of firms clustered by images depends on the nature of the data used to construct ratios. It ranks among the top performers in the GICS classification for accountancy-based ratios and is comparable to SIC-based industries for market-based ratios. Furthermore, the high variation in R^2 rankings suggests that IFS industries differ significantly from competitors.

4.2.1 Investor agreement within industries

As a complementary measure of industry coherence, we examine the degree of investor agreement within industries using analysts' earnings forecasts. Following Diether, Malloy, and Scherbina (2002), we measure investor disagreement as the standard deviation of analysts' EPS forecasts scaled by the absolute mean forecast in the month preceding the fiscal period

end. We aggregate firm-level dispersion to the industry level using equal-weighted averages across all firms within each industry-month. Figure 10 shows that firms grouped into IFS industries exhibit systematically lower forecast dispersion than firms classified using SIC, NAICS, GICS, Fama–French, and Hoberg–Phillips industries. This pattern persists over time and is particularly pronounced during periods of elevated market uncertainty. Lower forecast dispersion indicates that analysts hold more similar expectations about firms' future earnings within IFS industries. This finding complements the R^2-based evidence in Section 4.2. It suggests that image-based classifications group firms with genuinely similar economic exposures—a similarity that market participants implicitly recognize when forming earnings expectations.

To illustrate how IFS captures economically meaningful relationships, consider Ameren Corp, a U.S. electric utility that IFS assigns to industry 45. This industry also includes oil and gas producers such as Kosmos Energy, Mexco Energy, and Noble Energy. While these firms occupy different positions in the energy supply chain—gas producers upstream and utilities downstream—their operations share visual characteristics, including extraction facilities, pipelines, processing equipment, and energy infrastructure. This visual similarity reflects genuine economic linkage; natural gas prices directly affect both producers' revenues and utilities' fuel costs, creating correlated exposures.

Table 7 documents this pattern systematically. IFS industry 45 exhibits supplier intensity of 15.63% and customer intensity of 6.68%, indicating substantial vertical concentration. These supply-chain dependencies help explain why analyst forecasts within IFS industries exhibit lower dispersion (Figure 10). Firms with upstream or downstream exposure face common shocks that analysts recognize when forming earnings expectations.

Traditional classifications obscure these relationships. SIC assigns Ameren to 'electric, gas, & sanitary services' (SIC 49) while placing gas producers under 'oil and gas extraction' (SIC 13). NAICS and GICS similarly separate utilities from their fuel suppliers. Because these systems emphasize what firms sell rather than how they operate, they miss vertical dependencies that IFS detects through shared visual elements of production.

Taken together, the evidence on return comovement and investor agreement indicates that IFS industries capture economically meaningful relationships among firms. We next examine the scope of this classification by documenting its sample coverage and the characteristics of firms that are included and excluded.

4.3 Sample Coverage and the High-Signal Universe

We begin by documenting the sample coverage of the IFS classification and placing it in the context of alternative industry classification systems. The coverage of IFS reflects an explicit design choice: firms are classified only when sufficient visual information exists to construct a reliable similarity signal.

Coverage patterns across similarity-based classification systems are summarized in Table 6. IFS classifies, on average, approximately 1,100 firms per year, corresponding to about 50–53% of total market capitalization. This coverage is comparable to analyst-based classifications such as Kaustia and Rantala (2021), which classify a similar number of firms per year. It exceeds the coverage of patent-based approaches that rely on the existence of patent portfolios. Text-based classifications such as Hoberg and Phillips (2016) achieve broader coverage by construction, as they rely on mandatory 10-K disclosures; however, they impose a different quality constraint, as textual similarity can be weak or uninformative for firms with generic or boilerplate business descriptions.

Importantly, the firms classified by IFS are not a random subset of the market. Internet Appendix Table F7 compares observable characteristics of IFS-classified and unclassified firms. Firms covered by IFS are substantially larger, with higher market capitalization, total assets, and net sales. They are also markedly more liquid, as evidenced by significantly higher trading volume and lower bid–ask spreads. These differences indicate that IFS coverage concentrates on economically significant and tradable firms, which are most relevant for asset pricing and portfolio applications.

Industry composition further clarifies the nature of IFS coverage. Internet Appendix Table F8 shows that unclassified firms are disproportionately concentrated in Nonclassifiable Establishments and Services, which together account for nearly half of all unclassified firms (31.9% and 16.2%, respectively). These categories correspond to firms whose value creation is complex to represent visually, such as holding companies, conglomerates, and service-oriented businesses with limited physical or product-level manifestations. At the same time, these sectors are not absent from the IFS-classified universe: Services and Nonclassifiable Establishments together account for more than 20% of classified firms, indicating that visual classification is feasible and information is available.

In contrast, IFS-classified firms are relatively more concentrated in manufacturing, retail,

and other tangible sectors where business activities, production processes, and products generate rich visual representations. This industry pattern is consistent with the construction of IFS and reinforces that coverage limitations arise from information constraints rather than selective exclusion.

Taken together, these results demonstrate that the coverage of IFS is economically meaningful. While IFS does not assign all firms by construction, it captures the majority of market capitalization and focuses on large, liquid firms for which visual similarity conveys substantive economic information. For the analyses and portfolio applications examined in this study, universal coverage is neither required nor desirable; instead, accurate identification of economically relevant peers among tradeable firms is essential. IFS meets this criterion by design.

5 Expanded Image Sets: Including images from 10-K and patents

The main findings of this paper rely on Google Images, which predominantly feature consumer-facing firms with tangible products, such as those in retail and manufacturing. While this source excels in capturing market-driven, product-level information and real-time investor perceptions, it is less effective for firms operating in abstract, service-oriented, or technology-intensive sectors.

To assess the robustness of our image-based classification and to improve coverage in these areas, we augment Google Images with pictures extracted from design patent filings and 10-K annual reports. These additional sources provide complementary perspectives on firms' technological activities and operational structures, albeit with different strengths and limitations.

Patent images are particularly informative for firms engaged in abstract technologies or R&D-intensive activities, such as software and biotechnology. They enhance coverage of technically oriented firms but pose classification challenges due to their schematic nature and limited consumer-facing context.

Images extracted from annual reports primarily depict operational aspects, such as production facilities and product lines, offering a broader firm-level perspective that is less

influenced by market attention. However, these visuals do not reflect compliance-related or financial risk disclosures and are inherently static due to annual reporting cycles.

For firms with available patents or 10-K filings, we pool these images with the Google set and apply the same feature-extraction and similarity-measurement pipeline, without imposing source-specific weights. The resulting classifications are treated as robustness checks, while the Google-only classifications serve as our main specifications.

Using the aggregated image sample, we generate two extended classifications. The first partitioning yields 34 industries, of which 25 contain at least five firms per class, whereas the second yields 65 industries, of which 50 meet this minimum-size criterion. We refer to these classifications as IFS-X 34(25) and IFS-X 65(50), respectively.[13]

Further details on sample construction, industry composition, and explanatory power are reported in Internet Appendix C. The results reported in Tables C4 and C5 indicate that augmenting Google Images with patent and annual-report visuals leads to selective improvements in explanatory power. Under IFS-X 34(25), the accuracy of *total assets*, *EPS growth*, and *RNOA* increases, while under IFS-X 65(50), the R^2 improves for *profit margin*. At the same time, the explanatory power declines for several key ratios, including *debt-to-equity* and *forecasted EPS* under IFS-X 34(25), as well as *total assets*, *debt-to-equity*, and *ROE* under IFS-X 65(50).

For market-based ratios, the expanded image set yields more favorable results under IFS-X 34(25), with the R^2 for *market leverage* and *beta* highest for the IFS classification. Under IFS-X 65(50), the R^2 of the *price–earnings ratio* becomes the second highest among the competing classifications.

Taken together, these results suggest that while patent and annual-report images improve coverage of R&D-intensive and less consumer-facing firms, they introduce additional noise that weakens classification precision for several economically critical firm characteristics. Patent images, which are often schematic and abstract, tend to dilute consumer-centric similarity signals that are central to market-driven classification. Similarly, although images extracted from 10-K filings provide standardized operational visuals, their static nature and limited visual density constrain their ability to capture time-varying firm activities and

[13]Throughout the paper, we use the notation IFS $a(b)$, where a denotes the total number of image-based industries and b the number of industries containing at least five firms. For comparison, the Google-only classifications comprise 45 and 73 industries, with 25 and 50 industries meeting the same minimum-size threshold.

market perceptions.

Consequently, we prioritize Google Images in our main specifications, as they provide dynamic, consumer-focused visuals that yield higher overall precision and economic relevance for market-driven applications. The extended image classifications are therefore treated as robustness checks, illustrating the trade-off between broader firm coverage and classification accuracy.

6 Economic Links: Horizontal and Vertical Structure of IFS Industries

Understanding how firms cluster economically requires distinguishing two dimensions of relatedness. *Horizontal* linkages capture similarities in product markets—firms that compete for the same customers or operate in the same revenue segments. *Vertical* linkages capture supply-chain dependencies—upstream suppliers and downstream customers connected through production relationships. Traditional classifications, including SIC, NAICS, GICS, and the text-based system (Hoberg & Phillips, 2016), primarily emphasize horizontal similarity because they rely on descriptions of what firms sell.[14]

While IFS does not exclude horizontal product-market similarity, the evidence indicates a pronounced vertical tilt. Horizontal overlap with text-based classifications averages only 3–8%, whereas vertical supply-chain intensity averages 11–17% and reaches above 50% in concentrated industries. This pattern confirms that images primarily encode information about firms' operational positions—their suppliers, customers, and production infrastructure—rather than merely replicating the product-market similarity already captured by textual descriptions.

6.1 Why Images Capture Vertical Relationships

The key distinction between text-based and image-based classification lies like the data each source encodes.

[14]Text-based classifications may capture some vertical relationships when firms explicitly describe suppliers or customers in their 10-K filings. However, such disclosures are discretionary and typically brief, whereas images of production facilities, equipment, and logistics infrastructure directly and systematically reveal the supply chain's position.

Text-based classifications such as HP rely on 10-K business descriptions, which emphasize *what firms sell* and *with whom they compete*. This language naturally clusters horizontal competitors. A steel producer describes selling steel to industrial customers; an automobile manufacturer describes selling vehicles to consumers. Their textual descriptions share little vocabulary, so HP assigns them to different industries.

Image-based classification captures *how firms operate*. Images from a steel producer depict blast furnaces, rolling mills, and metal-processing facilities. Images from an automobile manufacturer depict assembly lines, metal stamping equipment, and industrial machinery. Despite serving different end markets, both firms share visual elements that reflect their positions within a typical supply chain: heavy industrial equipment, metal fabrication, and manufacturing infrastructure.

This distinction explains why IFS detects vertical linkages that text-based methods miss. Upstream suppliers and downstream customers rarely compete in the same product market, so their textual descriptions diverge. However, their operations often exhibit visual commonalities—raw materials, intermediate components, industrial machinery, or logistics infrastructure—that reflect shared production relationships.

6.2 Horizontal Linkages

To quantify horizontal similarity, we compare IFS industries with HP's transitive classifications. For each IFS industry i, we compute the maximum overlap with any HP industry:

$$\text{MaxOverlap}_i = \frac{\max_j N_{ij}}{N_i} \times 100, \tag{2}$$

where N_{ij} is the number of firms in both IFS industry i and HP industry j, and N_i is the size of IFS industry i. Low overlap indicates that IFS captures dimensions beyond revenue-based similarity.

Table 7 reports the results. Across all comparisons, horizontal overlap between IFS and HP is modest: typical overlaps range from 3% to 15%, with an average of approximately 4–9%. Several IFS industries exhibit almost no HP overlap yet display strong internal visual coherence, reflecting shared production processes rather than common product-market vocabulary.

6.3 Vertical Linkages

We measure vertical relationships using the Bureau of Economic Analysis (BEA) Input-Output Accounts. For each IFS industry, we identify the commodities produced by its member firms, then compute the share of inputs sourced from concentrated supplier industries and the share of outputs sold to concentrated customer industries.

For each IFS industry i, we compute:

$$\text{SupplierIntensity}_i = \frac{\sum_{k \in C_i} \sum_{s \in S_{ik}} Z_{sk}}{\sum_{k \in C_i} \sum_{s} Z_{sk}} \times 100, \tag{3}$$

$$\text{CustomerIntensity}_i = \frac{\sum_{k \in C_i} \sum_{d \in D_{ik}} U_{dk}}{\sum_{k \in C_i} \sum_{d} U_{dk}} \times 100, \tag{4}$$

where C_i is the set of commodities produced by the IFS industry i, Z_{sk} is the value of the commodity k supplied by BEA industry s, U_{dk} is the value used by the customer industry d, and S_{ik} and D_{ik} denote the primary supplier and customer industries. Intra-industry transactions are excluded to isolate external supply chain dependence. A detailed description of the construction of these measures, including data sources, industry mappings, and implementation choices, is provided in Internet Appendix D.

Table 7 summarizes the distribution of vertical and horizontal economic linkages across image-based industries. Vertical linkages are economically meaningful and display substantial heterogeneity across sectors. For IFS 45, supplier intensity ranges from near zero to over 30%, whereas customer intensity spans an even wider range, reaching above 50% in some industries. Similar dispersion is observed for IFS 73, with supplier and customer intensities extending well beyond 40% in the upper tail.

Average vertical linkages are moderate, with supplier and customer intensities typically in the 10–15% range across classifications. These averages mask considerable cross-industry variation, indicating that some image-based industries are deeply embedded in upstream or downstream supply chains, while others operate largely independently. This heterogeneity is consistent with economically meaningful differences in production structure rather than mechanical features of the classification.

The presence of economically meaningful vertical linkages within IFS industries has significant implications for interpreting industry-level exposure. Several IFS industries exhibit

substantial concentration on the supplier or customer side, indicating that firms within these groups may share common upstream or downstream dependencies.

For example, some IFS 45 industries exhibit high customer intensity, suggesting exposure to a relatively concentrated set of downstream buyers, whereas others exhibit elevated supplier intensity, indicating sensitivity to upstream input markets. These patterns illustrate how firms can be economically linked through supply-chain relationships even when horizontal overlap—as measured by text-based classifications—is limited.

Traditional classifications primarily capture horizontal product-market similarity. IFS captures both horizontal and vertical dimensions, which may help in understanding supply-chain exposures and correlated risks.

6.4 Interpretation

The results reveal how different data modalities encode distinct dimensions of economic relationships among firms.

First, image-based and text-based classifications capture fundamentally different types of firm-relatedness. Text-based approaches such as HP rely on 10-K business descriptions, which emphasize what firms sell and to whom. This language naturally clusters horizontal competitors—firms vying for the same customers. In contrast, images capture how firms operate: their facilities, equipment, production processes, and physical infrastructure. These operational characteristics reflect positions within supply chains and technological similarity rather than product-market overlap.

Second, IFS identifies economically meaningful vertical relationships that text-based methods miss. Consider the steel and automobile industries. A steel producer's 10-K describes selling steel products to industrial customers; an automobile manufacturer's 10-K describes selling vehicles to consumers. Their textual descriptions share little vocabulary, so HP assigns them to different industries. However, images from both firms depict blast furnaces, metal fabrication equipment, industrial machinery, and factory floors. IFS recognizes this visual overlap because automobile production depends on steel inputs—a vertical supply-chain relationship that is invisible in text but apparent in operational imagery.

Third, the two approaches capture largely distinct dimensions. Text-based classifications excel at horizontal product-market similarity; image-based classifications are tilted toward

vertical supply-chain adjacency. The low horizontal overlap (3–8%) and high vertical intensity (up to 55%) confirm this separation. The two methods are thus complementary: text for competitive dynamics and images for supply-chain exposure and the transmission of systematic risk.

6.5 Additional Evidence on Sector- and Firm-Level Patterns

Beyond horizontal and vertical production links, IFS also reveals systematic patterns in which types of firms are most effectively clustered by visual information. Internet Appendix E analyzes differences in explanatory power ($\Delta R^2 = R^2_{\text{IFS}} - R^2_{\text{HP}}$) across 16 financial ratios. IFS performs well in industries with visually distinctive products or processes—such as manufacturing, chemicals, and food production—where images capture equipment, packaging, and infrastructure. HP dominates in sectors such as financial services, where textual descriptions convey more information. At the firm level, IFS clusters growth-oriented firms more effectively, consistent with images capturing R&D facilities, prototypes, and new product launches. Full details appear in Internet Appendix E.

7 Applications

We evaluate the economic usefulness of IFS through three applications: peer-based relative-value trading, portfolio diversification, and industry momentum. Each application requires different properties from an industry classification.

Peer-based relative-value trading exploits cross-sectional differences in fundamentals among economically comparable firms. What matters is accurate identification of close peers, not stable industry groupings. IFS's frequent updates and its ability to capture emerging visual similarity make it well suited to forming time-varying peer sets used in within-peer long–short portfolio sorts.

Portfolio diversification requires grouping firms into distinct, non-overlapping sectors with low cross-group correlations. Unlike peer-based relative-value trading, diversification needs industry definitions that partition the stock universe into coherent groups. Dynamic updating remains valuable, as industry boundaries can adjust when firms' operations evolve.

Industry momentum exploits persistence in relative performance across industries. Clas-

sifications must be sufficiently stable to preserve momentum signals across formation and holding periods, while remaining flexible enough to accommodate long-run change.

We use IFS 45 for portfolio-based applications that require well-populated groups, such as diversification and momentum analyses. The peer-based relative-value trading strategy instead relies on non-transitive firm-level peer sets constructed directly from similarity measures and is therefore not tied to a particular IFS aggregation level.

Our preferred implementation, IFS 45, satisfies these requirements. Table 2 reports the distribution of firms across industries for IFS with 45 classes and for a more granular alternative with 73 classes. The IFS 73 classification is too fine-grained for portfolio-based applications: the average number of firms per industry is below 20, and during formation periods many industries contain fewer than 5 firms. We therefore exclude IFS 73 from diversification and industry momentum analyses. All applications are based on IFS 45, which yields a well-balanced partition of the stock universe and remains comparable in granularity to classical industry classifications used in prior work.[15]

7.1 Peer-based relative-value trading

Motivated by the evidence in Section 4 that IFS-defined peer groups exhibit substantial homogeneity in growth-related characteristics, we implement a peer-based relative-value trading strategy that exploits cross-sectional differences in firm growth within closely matched peer groups. Rather than relying on transitive industry classifications, the strategy is based on intransitive similarity measures that identify economically comparable firms.

Portfolio formation proceeds at the peer-group level. At the end of each month, for every firm we identify its set of closest peers using the intransitive IFS similarity measure. Firms and their associated peers are then jointly sorted into quintiles based on the growth metric observed in the preceding month. We focus on the cross-section of firms belonging to the highest and lowest growth quintiles within each peer group.

The investment strategy takes a long position in firms classified into the top growth quintile and a short position in firms classified into the bottom growth quintile. Positions are value-weighted and scaled to ensure dollar neutrality at the portfolio level. Portfolios are rebalanced monthly and held for one month, resulting in overlapping return observations.

[15] Moskowitz and Grinblatt (1999) adopt a similar approach, constructing a coarser industry classification for industry momentum applications and noting that more granular schemes are ill-suited for this purpose.

We consider two standard growth measures: SALES GROWTH and EPS GROWTH. These metrics provide broad coverage across firms and are well suited to a relative-value framework based on within-peer-group comparisons.[16]

For benchmarking purposes, we construct analogous strategies using peer definitions based on textual similarity following Hoberg and Phillips (2016), and analyst-based peer groups following Kaustia and Rantala (2021). All strategies are implemented over the period 2014–2021 using identical portfolio construction and rebalancing procedures and require firms to have at least five identified peers.

Table 8 reports the performance results. The IFS-based strategy delivers superior risk-adjusted performance relative to both text-based and analyst-based benchmarks. In particular, the Sharpe ratio of the image-based strategy exceeds that of the benchmark strategies by approximately 19–39% for SALES GROWTH and 4–130% for EPS GROWTH, highlighting the incremental information contained in image-based peer identification.[17]

Overall, the results demonstrate that IFS-based peer definitions enable the construction of growth-based long–short portfolios with robust out-of-sample performance.

7.2 Diversification benefits

Industry classifications are a standard tool for portfolio diversification, as they aim to group firms with similar systematic risk exposures. Investors allocating capital across industries can therefore reduce idiosyncratic risk by limiting exposure to any single sector. However, traditional classifications may aggregate firms that differ substantially in their underlying economic exposures, thereby limiting diversification.[18]

We evaluate the diversification benefits of alternative industry classification schemes by

[16] Alternative growth measures exhibit substantial missing observations, which would materially reduce the size of the investable universe.

[17] All portfolios are value-weighted and scaled to ensure dollar neutrality, which naturally limits the influence of small and illiquid firms. The objective of this exercise is not to provide a precise estimate of implementable trading profits net of all market frictions, but to compare alternative peer identification methods under an identical portfolio construction framework. As such, the emphasis is on relative differences in performance across strategies that differ only in how peers are defined, rather than on the absolute level of performance statistics.

[18] While portfolio diversification can, in principle, be achieved using historical return correlations alone, such correlations are often unstable and may not reflect persistent economic linkages between firms. Industry classifications provide a structured way to group firms by underlying business activities, offering a more interpretable and potentially more robust basis for diversification as correlation patterns evolve.

constructing portfolios that select one stock from each industry each month. Portfolio weights are determined using two standard optimization criteria: (i) mean–variance optimization to maximize the Sharpe ratio, and (ii) optimization to minimize conditional value-at-risk (CVaR). To ensure that results are not driven by idiosyncratic firm selection, we repeat the random stock selection procedure 500 times per industry and report the average performance across trials.

Table 9 compares the diversification performance of portfolios constructed using alternative industry classification schemes. Across both optimization criteria, portfolios based on image-defined industries exhibit competitive risk-adjusted performance relative to established benchmarks.

In portfolios optimized to maximize the Sharpe ratio (Panel A), IFS-based classifications rank among the top-performing schemes, delivering Sharpe ratios that exceed those of widely used benchmarks such as GICS and are comparable to those obtained using Fama–French and NAICS industries. In portfolios optimized to minimize tail risk using the CVaR criterion (Panel B), image-based industries achieve the highest Sharpe ratio among the classification schemes considered.

Overall, the results indicate that image-based industry definitions yield economically meaningful groupings that translate into strong, robust portfolio-level performance across different optimization objectives. Rather than dominating in a single dimension, IFS exhibits consistent effectiveness in settings that emphasize both risk-adjusted returns and downside risk control.

7.3 Industry momentum

Industry momentum strategies exploit the persistence of relative performance across industries and provide a demanding application for industry classification schemes, as they require industry groupings that are both economically meaningful and sufficiently stable over time.

Following Moskowitz and Grinblatt (1999), we construct industry momentum portfolios by going long (short) the three industries with the highest (lowest) returns over the past six months and holding the resulting portfolio for six months. We extend this baseline design along two dimensions. First, we consider longer holding horizons of nine and twelve months. Second, we evaluate strategies using both market-weighted and equally weighted portfolios.

The reversal strategy goes long the three industries with the lowest one-month return and holds the position for six, nine, or twelve months.[19]

Table 10 reports Sharpe ratios for momentum and reversal strategies constructed using alternative industry classifications. Across the twelve specifications considered, IFS 45 ranks first in eight cases and second in one. IFS-based momentum strategies perform well for both market-weighted and equally weighted portfolios and deliver competitive Sharpe ratios across all holding horizons. For reversal strategies, IFS provides strong and stable performance in market-weighted portfolios, while equally weighted implementations yield consistently positive but less pronounced Sharpe ratios.

We further assess robustness by applying a volatility-targeting overlay following Barroso and Santa-Clara (2015). Table F9 shows that IFS-based strategies continue to perform strongly under volatility targeting, ranking first in five out of six specifications and second in the remaining case. These results indicate that the performance of IFS-based industry momentum strategies is not driven by differences in volatility across classifications.

To examine whether industry membership, rather than firm-level momentum, drives these results, we construct "random" industry portfolios following Moskowitz and Grinblatt (1999). Specifically, we replace each firm in an IFS industry with a firm with a nearly identical six-month return, thereby preserving firm-level momentum while breaking industry structure. Table 11 shows that Sharpe ratios from these random portfolios are substantially lower than those obtained using true IFS industries. This finding suggests that the momentum profits are attributable to economically meaningful industry groupings rather than to mechanical return continuation at the firm level.

Overall, the results indicate that image-based industry classifications yield sufficiently stable definitions for momentum-based applications while capturing economically relevant relationships among firms. IFS delivers consistent performance across momentum and reversal strategies, weighting schemes, and holding horizons, supporting its usefulness in applications that require transitive and temporally coherent industry classifications.

[19] We also consider a long–short version of the reversal strategy. During our sample period, the short leg exhibits unstable performance. Results for the long–short variant are available upon request.

8 High Dynamism of IFS in Reclassifying Industries

An essential distinction between image-based and traditional industry classifications concerns the frequency with which firms are reassigned across industries over time. To quantify this property, we measure industry dynamism as the annualized proportion of firms whose industry assignment differs from the previous year. The statistics are computed using annual firm observations and exclude first-time classifications to ensure that measured changes reflect genuine reclassifications rather than initial assignments. Because IFS industries are updated biennially using rolling three-year image windows, we attribute reclassifications to the update year and report the corresponding annualized rate, which preserves the intuitive annual benchmark used for SIC-, NAICS-, and related schemes.

Table 12 reports industry dynamism for IFS and benchmark classification systems. IFS exhibits substantially higher reclassification rates than traditional schemes. In Panel A, the annual reclassification rate for IFS with 45 industries equals 16.7%, compared with values below 3.5% for SIC-based, Fama–French, and NAICS classifications. Text-based classifications display the second-highest level of dynamism, with reclassification rates of 8.4%, but remain markedly less dynamic than image-based industries. In Panel B, which considers a finer partition with 73 image-based industries, the reclassification rate increases to 21.7%, while benchmark classifications remain largely static.

In absolute terms, the degree of reclassification implied by IFS is moderate. A reclassification rate of 16.7% implies that the average firm changes its industry affiliation approximately once every six years under IFS 45.[20] In comparison, a rate of 21.7% corresponds to a change roughly once every four to five years under IFS 73. These frequencies indicate that image-based classifications adjust meaningfully to changes in firms' observable activities—such as product mix, production processes, or operational focus—without inducing excessive classification turnover.

The degree of dynamism documented in Table 12 has direct implications for empirical applications. For peer-based relative-value trading and portfolio diversification, the ability to update industry assignments allows classifications to reflect evolving firm characteristics and

[20] If the annual reclassification rate is r, the expected number of years until a firm changes industry is approximately $1/r$ under the assumption that reclassification events are independent across years. For $r = 16.7\%$, this yields $1/0.167 \approx 6$ years; for $r = 21.7\%$, this yields $1/0.217 \approx 4.6$ years.

economic linkages. For industry momentum strategies that require stability across formation and holding periods, the observed reclassification rates remain sufficiently low to preserve economically meaningful trends, as demonstrated in the previous subsection.

A potential concern is that frequent reclassification could increase turnover and trading costs in portfolio applications. However, the observed reclassification rates remain moderate in absolute terms, and the applications in Sections 7.1–7.3 demonstrate that IFS-based strategies deliver strong risk-adjusted performance despite this dynamism. Moreover, peer-based relative-value trading rebalances monthly regardless of industry definitions, diversification strategies select individual stocks rather than entire industries, and momentum strategies operate over six- to twelve-month horizons during which IFS classifications remain sufficiently stable. These results suggest that the benefits of timely updating outweigh any costs associated with reclassification.

Overall, the evidence indicates that IFS provides a classification scheme that is more adaptive than traditional industry systems, and more responsive than text-based alternatives, while retaining enough temporal coherence to support standard asset-pricing and portfolio applications.

9 Conclusion

This paper introduces Image-based Firm Similarity (IFS), a new approach to industry classification that leverages large-scale visual data to capture economically meaningful relationships among firms. By extracting and comparing the full distribution of image embeddings for each company, our method uncovers patterns in products, technologies, supply chains, and branding that are difficult to identify using traditional text- or number-based classifications. We construct two image-based industry systems and show that they perform competitively—and often better—than leading alternatives across a wide range of economic applications.

IFS delivers strong explanatory power for accounting- and market-based ratios, particularly those associated with expectations, growth, innovation, and intangible assets. It also generates economically valuable peer groups for investment applications such as peer-based relative-value trading, portfolio diversification, and industry momentum, consistently delivering strong risk-adjusted performance. Moreover, IFS is highly dynamic, reclassifying firms more frequently than GICS, SIC, NAICS, or text-based systems, and adapting more quickly

to changes in firms' operational focus. In addition to horizontal product-market similarity, IFS captures economically meaningful vertical linkages along supply chains, offering a richer representation of firm relationships.

While IFS provides a flexible and responsive classification framework, it is not without limitations. Firms whose activities are inherently difficult to visualize—such as certain consulting, abstract-technology, or financial institutions—remain more challenging, although incorporating images from 10-K filings and design patent documents partly mitigates this concern. Moreover, IFS does not aim to replicate the extreme granularity of traditional regulatory classifications such as SIC or GICS, which may remain preferable in specific accounting or compliance settings.

These limitations do not detract from the paper's central contribution: the introduction of a dynamic, visually grounded, and cognitively motivated classification system that complements and extends existing approaches. By leveraging the richness of image data, IFS captures dimensions of firm similarity that are difficult or impossible to infer from textual disclosures or financial statements alone.

Future research could extend IFS to international markets, private firms, or ESG-related classifications, and explore how investors, regulators, and risk managers can use image-based peer information to monitor evolving industry structures and supply-chain exposures in real time. Another promising direction is to combine image-based similarity with text- or analyst-based groupings to construct hybrid classifications that aggregate information across multiple data modalities. We view this paper as establishing a foundation for a broader research agenda on how visual information can complement traditional data sources in understanding firm relationships and market structure.

In summary, our findings show that images provide a distinct and economically meaningful lens on firm similarity. Visual representations capture how firms operate—their products, technologies, and organizational context—revealing relationships that are largely invisible to conventional industry classifications. Incorporating this information leads to peer groupings that better align firms along both horizontal and vertical dimensions of economic activity, with important implications for empirical research and investment practice.

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*.

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

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy prices. *The Journal of Finance*, *75*(3), 1371–1415.

Cohen Kadosh, R., Cohen Kadosh, K., Kaas, A., Henik, A., & Goebel, R. (2007). Notation-dependent and -independent representations of numbers in the parietal lobes. *Neuron*, *53*(2), 307–314. doi:https://doi.org/10.1016/j.neuron.2006.12.025

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255). doi:10.1109/CVPR.2009.5206848

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.

Green, J., Hand, J. R. M., & Zhang, X. F. (2017). The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns. *The Review of Financial Studies*, *30*(12), 4389–4436. doi:10.1093/rfs/hhx019. eprint: https://academic.oup.com/rfs/article-pdf/30/12/4389/24433556/hhx019.pdf

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.

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. (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.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1-2), 83–97. doi:https://doi.org/10.1002/nav.3800020109. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020109

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. *Unpublished Working Paper. Yale University.*

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

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.

Piazza, M., & Izard, V. (2009). How humans count: Numerosity and the parietal cortex. *The Neuroscientist*, *15*(3), 261–273.

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.

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

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. Photos were sourced from Google Images using an automated Python query interface, capturing publicly available 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 intra-industry correlation between all stocks in the industry (Intra. 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 intra-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.	Intra. 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')
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			

Table 2: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Intra. Corr.	TOP5 SIC codes
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')
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 (Text 25 and Text 50). 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
Text 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. 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 (Text 25). 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 (Text 50). 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 F1 in the Internet 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
Text 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
Text 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 Accounting 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 quarter 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 (Text 25). 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 (Text 50). 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 F1 in the Internet 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
Text 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
Text 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: Coverage of Similarity-Based Industry Classification Systems

This table compares annual coverage across similarity-based industry classification systems. *Firms/Year* reports the average number of firms classified per year, while *Mkt Cap Coverage* reports the fraction of total CRSP market capitalization covered by each method. We report coverage for text-based similarity from Hoberg and Phillips (2016) (HP), analyst-based similarity from Kaustia and Rantala (2021) (KR), a patent-based similarity approach from Lee, Sun, Wang, and Zhang (2019) (PAT), and our image-based method (IFS). Coverage differs across systems because each method imposes data availability or quality constraints, summarized in the final column.

Classification System	Firms/Year	Mkt Cap Coverage	Coverage Constraint
HP (Text)	4,000+	95%+	10-K filing required
KR (Analyst)	1,075	60–70%	Common analyst coverage required
PAT (Patent)	1,500–2,000	40–50%	Patent portfolio required
IFS (Image)	**1,100**	**50–53%**	**Sufficient visual data required**

Table 7: Vertical and Horizontal Economic Links Across Image-Based Industries

This table reports vertical and horizontal economic links for image-based industries. Vertical links measure the percentage of inputs sourced from (Suppliers) and outputs sold to (Customers) dominant BEA industries outside the IFS cluster. Horizontal links measure the percentage overlap between IFS45 and HP25, and between IFS73 and HP50.

	Vertical Link (IFS45) in %		**Vertical Link (IFS73) in%**		**Horizontal link with HP in %**	
IFS	Suppliers	Customers	Suppliers	Customers	IFS45 vs. HP25	IFS73 vs. HP50
1	8.67	12.03	16.61	11.17	6.44	2.93
2	23.06	27.37	21.04	17.62	6.66	3.75
3	8.78	15.71	16.91	6.55	13.29	3.00
4	12.49	26.78	16.27	27.69	14.52	2.74
5	28.01	23.30	15.03	12.46	10.74	2.65
6	15.02	29.23	12.99	9.18	7.66	17.57
7	31.44	24.31	42.66	22.54	9.62	3.47
8	14.66	16.93	16.63	8.50	14.24	12.19
9	16.43	13.64	23.08	13.15	7.21	3.19
10	10.97	19.26	16.77	8.33	13.78	3.24
11	24.42	29.67	13.82	11.17	5.53	3.45
12	8.90	17.70	20.29	16.04	6.57	3.80
13	4.56	6.98	14.29	7.15	6.07	3.38
14	7.87	8.15	27.25	13.50	6.63	3.56
15	26.83	12.52	13.27	3.12	4.81	3.36
16	8.01	12.70	19.98	19.68	7.44	3.00
17	7.38	9.66	13.64	12.62	7.44	3.42
18	11.80	17.57	13.85	11.81	9.46	2.91
19	13.64	17.27	15.32	5.45	6.29	3.83
20	3.67	3.88	13.95	2.93	5.38	3.31
21	5.25	9.29	13.83	4.42	7.76	3.26
22	13.66	55.02	18.79	8.92	6.68	2.55
23	13.01	48.98	23.27	12.58	17.48	2.95
24	9.85	14.41	13.78	8.03	6.48	2.70
25	0.77	0.73	15.25	7.94	5.45	2.61
26	14.70	18.42	23.48	5.49	13.38	2.69
27	11.29	20.97	20.68	9.48	6.58	2.60
28	16.92	20.04	14.97	12.34	4.42	3.90
29	8.92	20.90	14.37	12.11	11.01	2.78
30	27.91	18.37	13.87	10.43	5.24	3.00
31	7.76	21.86	13.36	5.44	12.05	2.41
32	2.49	1.78	12.12	13.38	6.29	3.35
33	10.06	8.06	14.68	6.76	7.00	2.86
34	1.62	8.91	11.37	13.01	5.98	3.61
35	1.17	1.46	15.15	9.23	13.29	3.03
36	1.32	0.74	20.19	16.08	9.46	4.00
37	13.37	5.83	19.64	11.61	5.78	3.38
38	0.10	0.74	21.26	45.97	3.90	2.86
39	0.37	2.57	12.90	2.93	4.25	15.10
40	12.99	0.11	18.14	17.13	5.00	4.56
41	9.18	5.84	14.87	4.87	13.25	3.88
42	8.79	2.24	14.39	5.77	4.33	4.56
43	–	–	21.84	12.92	11.96	3.03
44	–	–	40.90	36.29	7.99	3.55
45	–	–	15.63	6.68	–	2.86
46	–	–	19.40	14.01	–	6.16
47	–	–	19.19	13.22	–	3.66
48	–	–	13.57	10.39	–	3.26
49	–	–	13.62	7.69	–	3.22
50	–	–	38.90	7.41	–	2.94
51	–	–	16.97	12.63	–	2.82
52	–	–	16.20	6.91	–	3.30
53	–	–	11.65	5.67	–	3.07
54	–	–	15.68	24.25	–	2.80
55	–	–	16.21	18.69	–	2.76
56	–	–	11.59	5.65	–	2.96
57	–	–	14.98	8.27	–	3.44
58	–	–	11.91	3.36	–	3.00
59	–	–	15.00	3.38	–	4.00
60	–	–	13.00	2.86	–	3.00
61	–	–	14.73	8.34	–	3.59
62	–	–	12.23	5.58	–	3.00
63	–	–	11.70	4.21	–	2.00
64	–	–	15.14	5.83	–	2.00
65	–	–	14.37	7.26	–	2.67
66	–	–	18.45	8.11	–	3.00
67	–	–	17.18	9.11	–	3.56
68	–	–	11.58	2.22	–	2.25
69	–	–	16.81	4.95	–	2.46
71	–	–	18.79	5.88	–	5.00
72	–	–	12.46	9.24	–	3.00
73	–	–	17.07	7.74	–	2.40
Average	11.38	15.05	17.04	10.60	8.29	3.56
Min	0.10	0.11	11.37	2.22	3.90	2.00
Max	31.44	55.02	42.66	45.97	17.48	15.10
Median	9.96	14.03	15.20	8.71	6.84	3.05

Table 8: Peer-Based Relative-Value Trading Strategies Based on Firm Growth

This table reports the outcomes of peer-based relative-value trading strategies based on firm growth metrics, specifically SALES GROWTH and EPS GROWTH. Firms and their peers, identified using similarities in images (IFS), texts (HP), and shared analysts (KR), are jointly sorted into quintiles based on growth observed in the preceding month. Portfolios take long positions in firms in the highest growth quintile and short positions in firms in the lowest growth quintile, and are held for one month. The analysis spans the period 2014–2021 and includes firms with at least five peers. Text-based similarity measures are obtained from Hoberg and Phillips (2016) (HP), while common analyst coverage is from Kaustia and Rantala (2021) (KR). The table reports the performance differential between the top and bottom growth quintiles within peer groups, providing evidence on the returns to growth-based relative-value strategies. All portfolios are value-weighted and scaled to ensure dollar neutrality at the aggregate portfolio level. Table F1 in the Internet Appendix provides details on ratio construction.

	SALES GROWTH			**EPS GROWTH**		
	IFS	HP	KR	IFS	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 DrawDown	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 9: 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 (Text 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: 1) mean-variance optimized portfolios to maximize Sharpe Ratio (Panel A); and 2) conditional value-at-risk (CVaR) optimized portfolios to minimize risk (Panel B). To optimize portfolios (Panels A and B), 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	Text 25 Industries
PANEL A: Max. Sharpe Ratio portfolios						
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 B: Min. CVaR Ratio portfolios						
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 10: 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 (Text 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	Text 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 11: 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 IFS	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

Table 12: 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 (Text 25). 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 (Text 50). 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
Text 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
Text 50 Industries	0.092

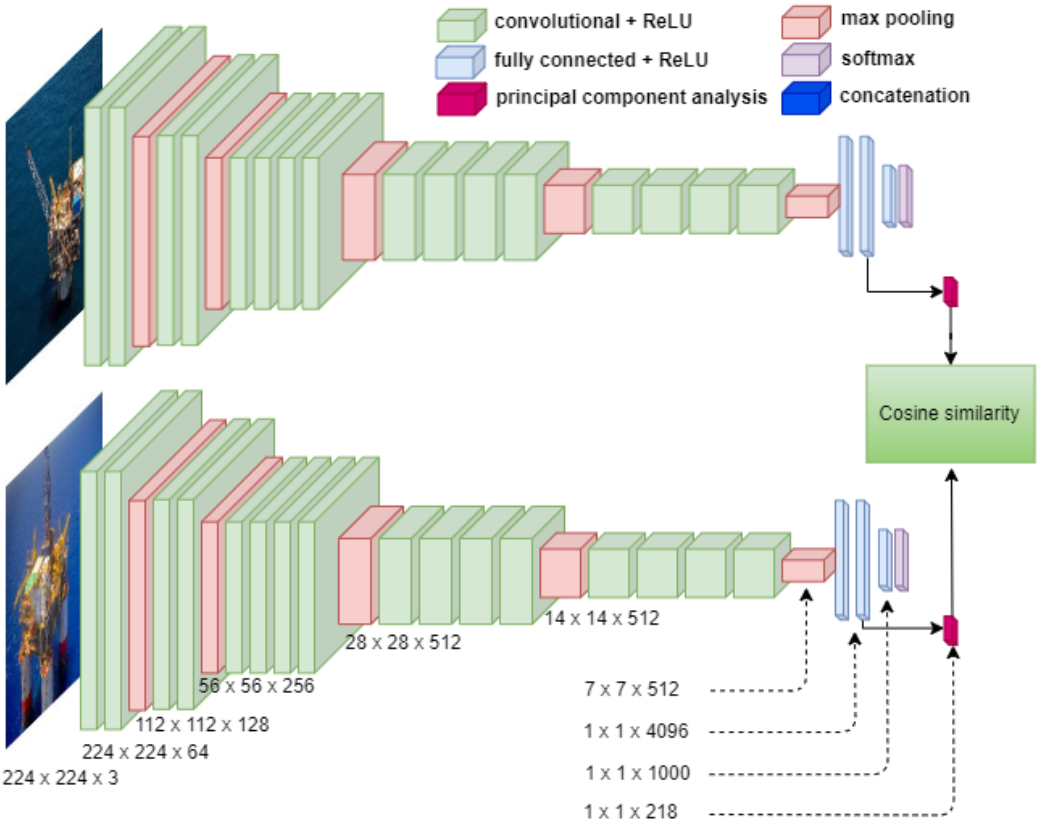

Figure 1: Image similarity

Note: The figure presents the architecture of image comparison. In the first step, images are standardized to a 224x224x3 resolution, where the first dimension denotes height, the second width, and the third color channels. Second, to detect objects in images, we use a VGG-19 convolutional neural network with 19 layers (Simonyan & Zisserman, 2014). The networks consist of convolutional layers that generate feature maps, pooling layers that downsample the information produced by the convolutional layers, and fully connected layers that combine the data extracted by previous layers to produce the final output. VGG-19 pretrained with more than 1m photos is designed to classify objects into 1,000 categories. We use the numerical representations of the identified objects from the last but one fully connected layer, with dimensions 1x1x4096. Third, we apply Principal Component Analysis (PCA) to reduce dimensionality and retain at least 70% of the variance. The reduced vector has a dimension of 1x1x218. Finally, feature vectors are used to compute cosine similarity between two photos.

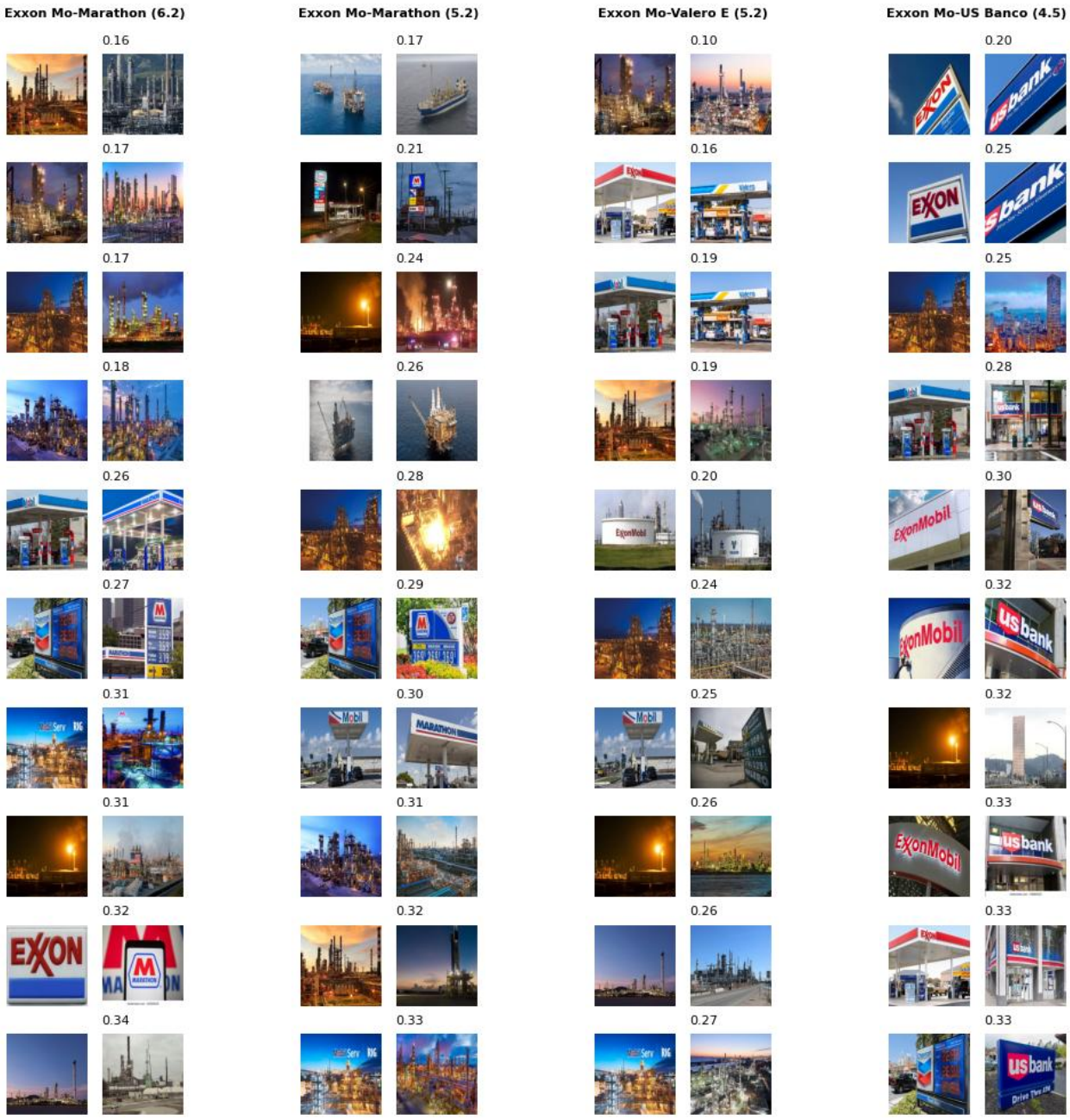

Figure 2: Photos representing four peers of Exxon Mobil Corp with the highest similarity score

This figure displays matched image pairs between Exxon Mobil Corp and its four closest visual peers. Each column header displays the peer firm name and the aggregate firm-pair similarity score in parentheses; higher values indicate greater visual similarity. Within each column, individual photo pairs are shown with their image-level cosine distance, with lower values indicating greater similarity between matched images. The methodology is described in Section 3.2; full technical details appear in Internet Appendix A

.

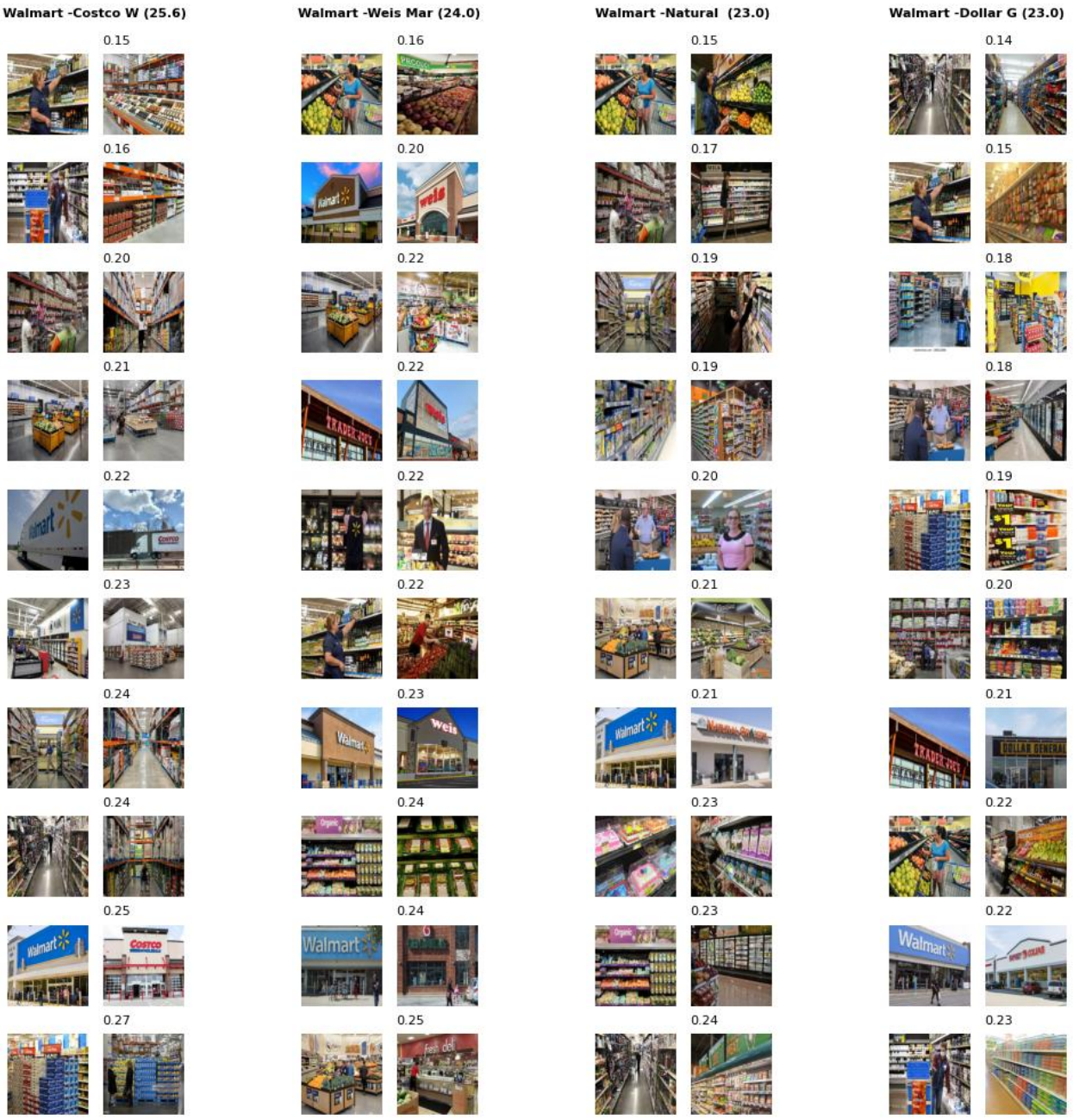

Figure 3: Photos representing four peers of Walmart Inc. with the highest similarity score

This figure displays matched image pairs between Walmart Corp and its four closest visual peers. Each column header displays the peer firm name and the aggregate firm-pair similarity score in parentheses; higher values indicate greater visual similarity. Within each column, individual photo pairs are shown with their image-level cosine distance, with lower values indicating greater similarity between matched images. The methodology is described in Section 3.2; full technical details appear in Internet Appendix A

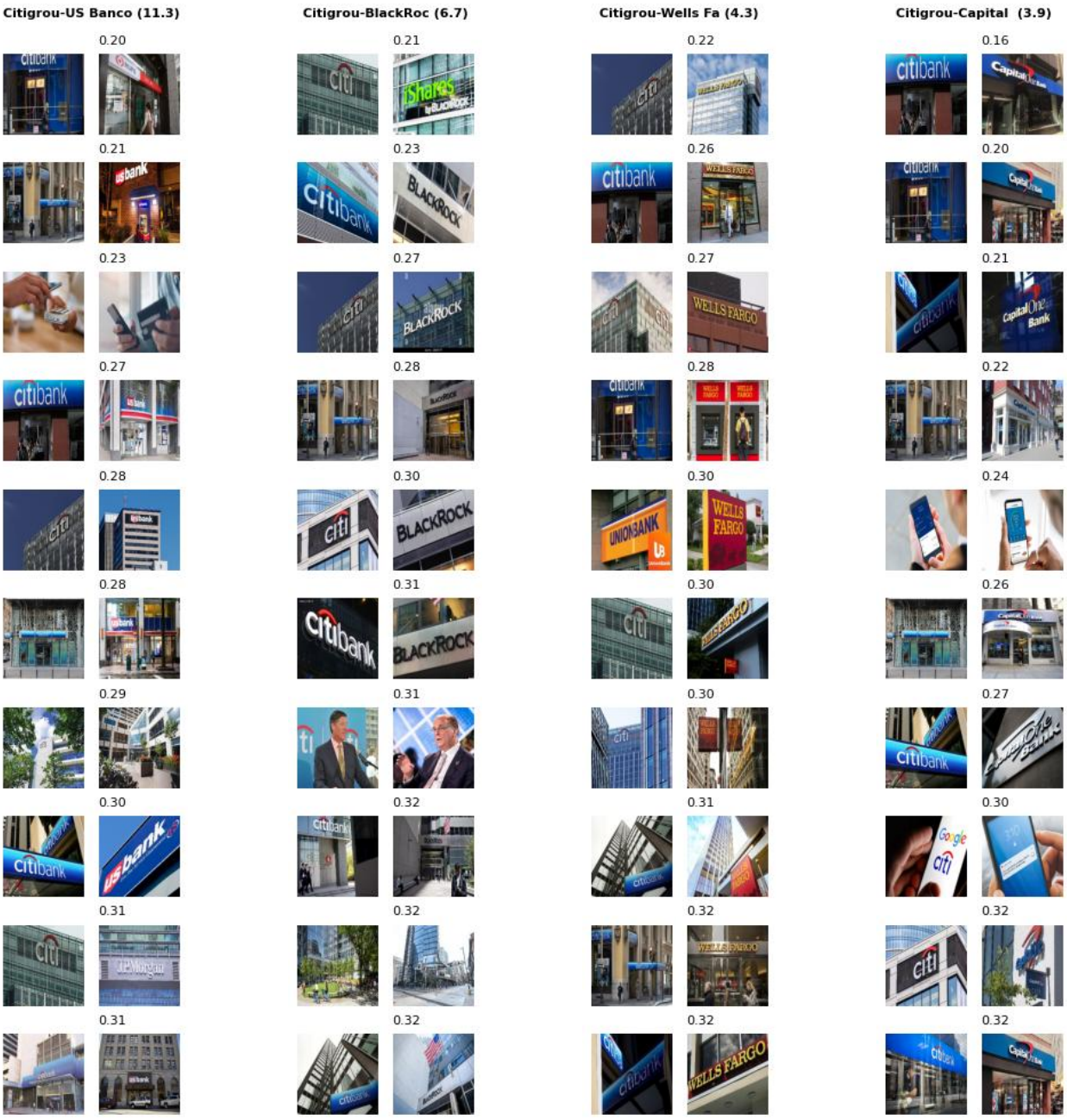

Figure 4: Photos representing four peers of Citigroup Inc with the highest similarity score

This figure displays matched image pairs between Citigroup Inc. and its four closest visual peers. Each column header displays the peer firm name and the aggregate firm-pair similarity score in parentheses; higher values indicate greater visual similarity. Within each column, individual photo pairs are shown with their image-level cosine distance, with lower values indicating greater similarity between matched images. The methodology is described in Section 3.2; full technical details appear in Internet Appendix A.

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

This figure displays matched image pairs between Johnson & Johnson and its four closest visual peers. Each column header displays the peer firm name and the aggregate firm-pair similarity score in parentheses; higher values indicate greater visual similarity. Within each column, individual photo pairs are shown with their image-level cosine distance, with lower values indicating greater similarity between matched images. The methodology is described in Section 3.2; full technical details appear in Internet Appendix A.

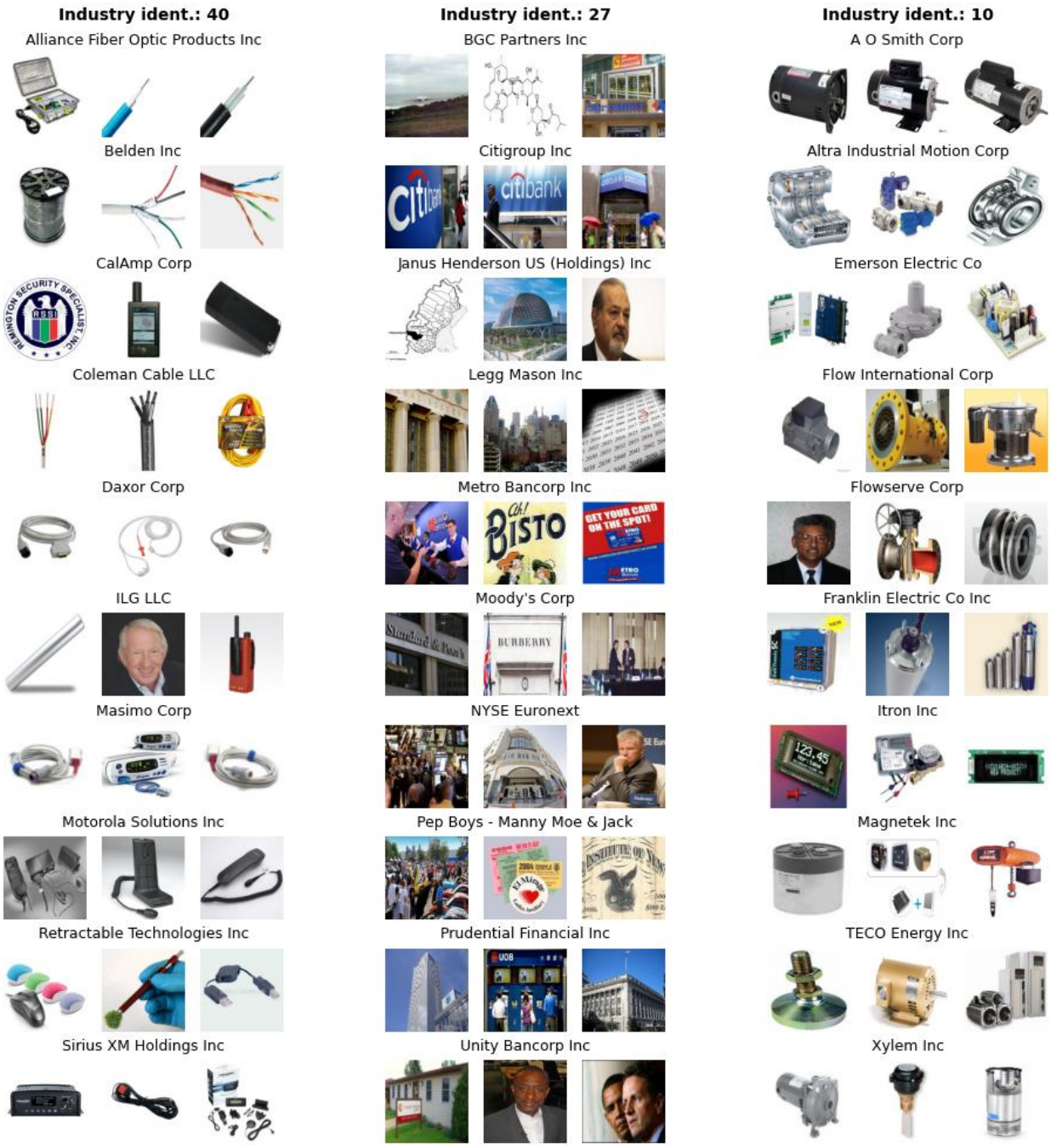

Figure 6: Photos representing industries with the highest intra-industry correlations (1)

This figure presents photos from IFS 73, presenting three industries with high intra-industry correlations. Each industry is represented with 10 randomly selected stocks and their three randomly selected photos. Photos downloaded from Google from 2009 to 2013 are used to identify industries. We excluded industries represented by fewer than 10 stocks.

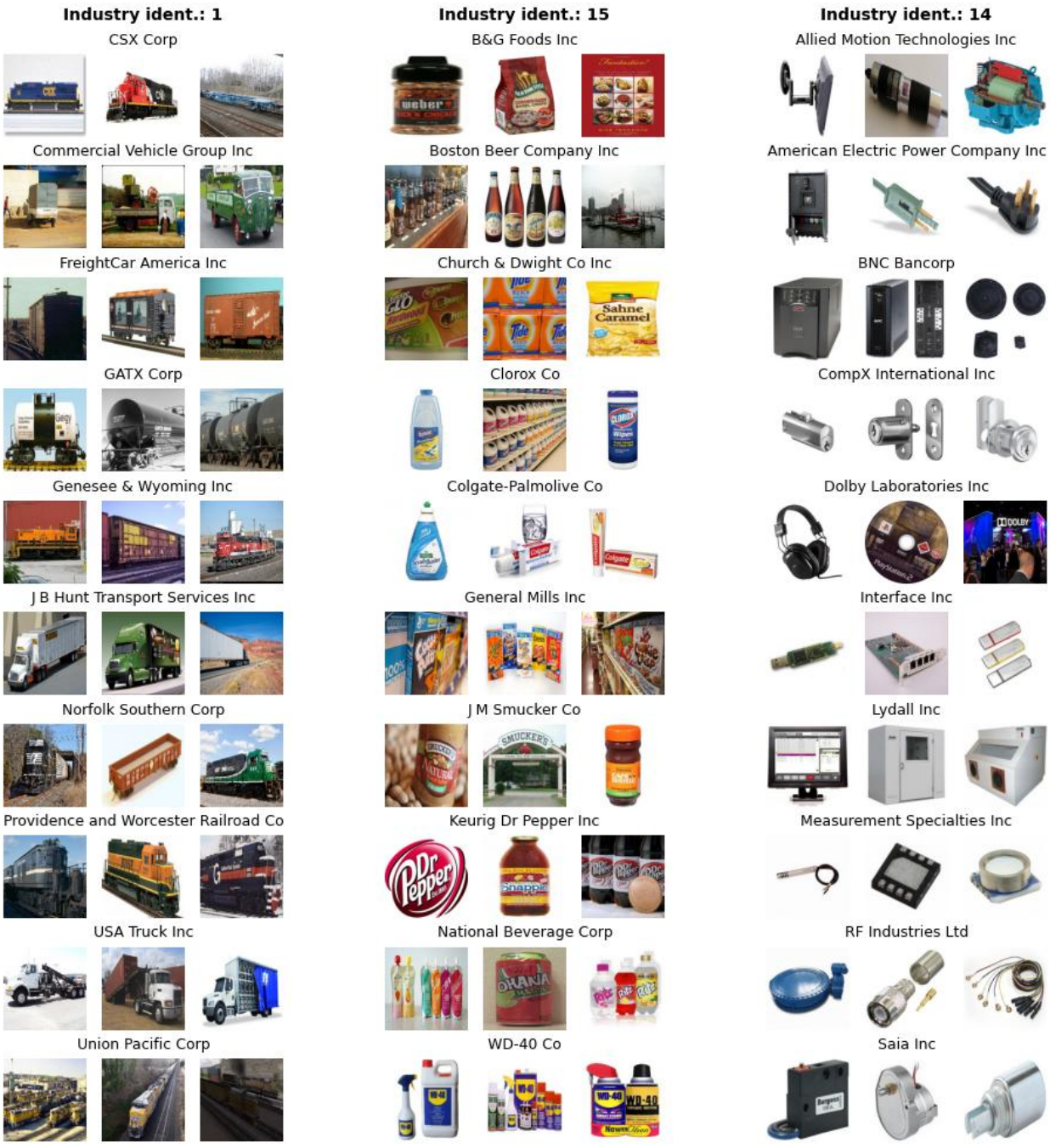

Figure 7: Photos representing industries with the highest intra-industry correlations (1)

This figure presents photos from IFS 73, presenting three industries with high intra-industry correlations. Each industry is represented with 10 randomly selected stocks and their three randomly selected photos. Photos downloaded from Google from 2009 to 2013 are used to identify industries. We excluded industries represented by fewer than 10 stocks.

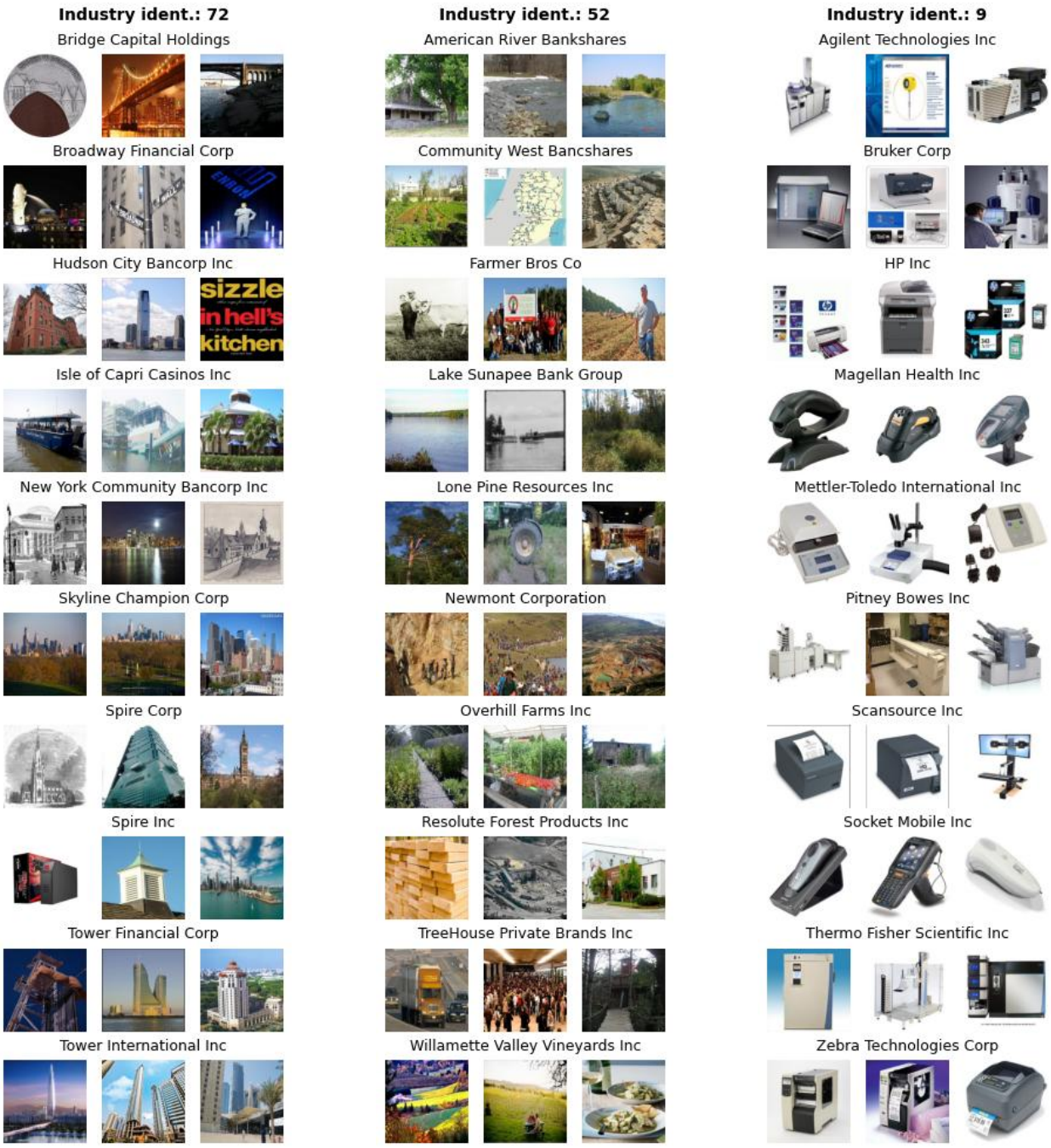

Figure 8: Photos representing industries with the lowest intra-industry correlations (1)

This figure presents photos from IFS 73, presenting three industries with low intra-industry correlations. Each industry is represented with 10 randomly selected stocks and their three randomly selected photos. Photos downloaded from Google from 2009 to 2013 are used to identify industries. We excluded industries represented by fewer than 10 stocks.

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

This figure presents photos from IFS 73, presenting three industries with low intra-industry correlations. Each industry is represented with 10 randomly selected stocks and their three randomly selected photos. Photos downloaded from Google from 2009 to 2013 are used to identify industries. We excluded industries represented by fewer than 10 stocks.

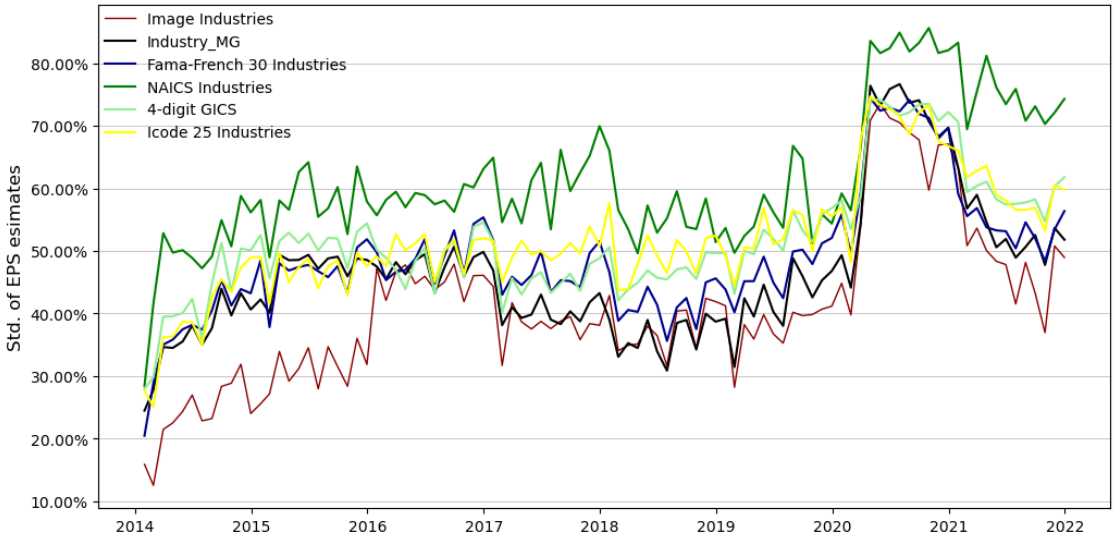

Figure 10: Investor agreement within industries

The figure plots investor agreement within industries, measured as the dispersion of analysts' EPS forecasts following Diether, Malloy, and Scherbina (2002). Lower values indicate greater agreement among market participants. Forecast dispersion is first computed at the firm level and then aggregated within industries. The sample covers the period 2014–2021 and includes industries with at least five firms. Analyst forecast data are obtained from I/B/E/S.

Internet Appendix

Just Look: Knowing Peers with Image Representation

Appendix A: Methodology and Data

This appendix provides technical details underlying the construction of Image-based Firm Similarity (IFS) and complements the high-level description in the main text. While the main text presents the IFS methodology as a five-step pipeline—feature extraction, firm-pair similarity measurement, temporal robustness, industry clustering, and dynamic updating—this appendix formalizes each component and documents the specific implementation choices.

The appendix is organized to mirror the structure of the main text. We first describe the data sources and image collection process. We then detail the extraction of visual features from images, the construction of firm-pair similarity scores based on set-level comparisons, and the decomposition of similarity into expected and excess components. Next, we formalize the statistical criteria for identifying robust firm-level similarities and describe the temporal stability rules governing peer persistence. Finally, we document the clustering algorithm used to construct transitive IFS industries and the dynamic updating procedure that allows classifications to evolve as firms' visual representations change.

A.1 Data

We define the stock universe as all firms listed on the NYSE, AMEX, and NASDAQ obtained from CRSP. We retain common shares with share codes 10 and 11. The sample covers the period from 2009 to 2021 and includes, on average, approximately 2,250 firms per year.

For each firm and each calendar year, we collect images directly from Google Images using an automated Python interface. We retrieve up to 100 images per firm-year using a standardized query structure of the form: "{Company Common Name} products after:{year}-01-01 before:{year}-12-31". The 100-image cap balances coverage and data quality: for firms with extensive online presence, it prevents the representation from being dominated by lower-

quality or redundant images appearing deeper in Google's ranking, while still preserving sufficient visual diversity for reliable firm-level representation.

The set of images retrieved for each firm-year reflects Google's ranking of visual content. Google prioritizes images originating from credible and authoritative sources, such as company websites, major media outlets, and widely cited reference platforms (e.g., Wikipedia). As a result, the top-ranked images typically depict official products, production facilities, branded materials, and other visuals closely related to firms' economic activities.[21]

Raw images undergo an extensive cleaning procedure to remove low-information or generic content unlikely to convey firms' economic activities, such as faces, logos, landscapes, or unrelated background images. This procedure substantially reduces noise while improving the informational content of the image sets. A detailed description of the image-cleaning process is provided in Appendix B.

Each image is associated with a timestamp based on its upload date, which allows us to align visual information with firm-year observations and supports the construction of time-varying firm representations.

To enhance coverage—particularly for R&D-intensive, small-cap, and service-oriented firms—we supplement Google Images with visuals extracted from 10-K annual reports and design patent filings. These supplementary sources increase both image depth and firm coverage. Results based on the extended image sample are reported separately in Appendix C.

In addition to image data, we collect a set of 26 firm-level financial ratios following Kaustia and Rantala (2021), spanning valuation, profitability, growth, leverage, investment, and intangibility measures. Detailed variable definitions and construction procedures are reported in Appendix F.

[21]The ordering of images displayed by Google is influenced by two ranking mechanisms. The first is a general PageRank algorithm that prioritizes content from reliable and authoritative sources. The second is VisualRank, which combines PageRank information with visual similarity analysis to rank images based on relevance and visual features such as color, texture, and shape (Brin & Page, 1998; Jing & Baluja, 2008). In addition, Google evaluates image sources using criteria related to expertise, authoritativeness, and trustworthiness (E-A-T), domain authority, content quality, and user engagement. These mechanisms collectively ensure that top-ranked images are both credible and visually representative of the query. By leveraging Google's ranking, our dataset draws on a curated set of high-quality, market-relevant visuals that reflect firms' operations and public representation. Importantly, our methodology does not rely on the ordinal ranking of images themselves. Instead, we exploit Google's ranking as a filtering device that systematically removes low-quality and irrelevant content, yielding a consistent and economically meaningful image set across firms and time.

To benchmark Image-based Firm Similarity against existing classification systems, we obtain SIC and NAICS codes from CRSP, GICS codes from Compustat, text-based similarity measures from Hoberg and Phillips (2016), and analyst-based peer definitions from Kaustia and Rantala (2021). SIC and NAICS codes from CRSP are time-varying, whereas GICS classifications from Compustat are static. Table 12 summarizes these differences in classification dynamics.

A.2 Extracting Visual Features from Images

This subsection formalizes Step (1) of the IFS pipeline described in the main text. The objective is to transform raw images into fixed-dimensional numerical representations that capture economically relevant visual content and enable systematic comparison across firms.

We extract visual features using a deep convolutional neural network, VGG19 (Simonyan & Zisserman, 2014), pre-trained on the ImageNet dataset (Deng et al., 2009). VGG19 is a widely used architecture in transfer-learning applications and provides stable, high-level representations of visual content without requiring task-specific retraining. Using a pre-trained model allows us to leverage rich visual abstractions learned from a large and diverse image corpus while maintaining consistency across firms and time.

Each image is resized to 224×224 pixels to match the network's input requirements and passed through the convolutional layers of VGG19. We retain the activations from the final fully connected layer, yielding a 4,096-dimensional feature vector for each image. These embeddings summarize high-level visual attributes—such as object composition, structural patterns, and visual design elements—that are relevant for identifying similarities in firms' products, production technologies, facilities, and branding.

To reduce noise and computational burden, we apply principal component analysis (PCA) to the complete set of image embeddings. PCA is estimated using the pooled image sample, and all subsequent similarity calculations are performed in the reduced feature space. We retain the minimum number of principal components necessary to account for at least 70% of the total variance in the original embeddings. This dimensionality reduction step improves numerical stability and ensures that similarity measures focus on the most informative visual variation.

Importantly, we do not aggregate image-level embeddings into a single firm-level vector

at this stage. Instead, each firm is represented as a *set* of image embeddings. This set-based representation preserves heterogeneity within firms' visual portfolios and allows downstream similarity measures to emphasize recurring visual patterns rather than isolated or incidental images.

The resulting image embeddings serve as input to the firm-pair similarity procedure described in Section A.3, which compares firms by matching images across sets and distinguishing firm-specific visual overlap from generic imagery common to many firms.

A.3 Identification of Firms' Similarity

This subsection implements Step (2) of the IFS pipeline described in the main text. Building on the visual feature representations introduced in Section A.2, we construct firm-pair similarity measures that capture economically meaningful visual relatedness between firms.

The objective of this step is to translate high-dimensional image embeddings into a sparse, interpretable firm-level similarity structure that reflects systematic overlap in firms' products, technologies, facilities, and operational environments, while filtering out generic or coincidental visual similarity.

The construction of firm-pair similarity proceeds in four conceptual stages. First, we quantify similarity at the image level using a standard metric applied to the visual feature representations (Section A.3.1). Second, because firms are represented by sets of images rather than single vectors, we aggregate image-level similarity to the firm level using a set-based matching procedure that emphasizes recurring visual patterns across firms' image collections (Section A.3.2). Third, we decompose raw firm-level similarity into an expected component—driven by generic imagery and image-set size—and an excess component that captures firm-specific visual overlap (Section A.3.3). Finally, we retain only statistically significant excess similarity to construct a sparse firm-pair similarity graph that serves as the input for industry clustering and subsequent empirical analyses (Section A.3.4).

The subsections that follow describe each of these steps in turn.

A.3.1 Image-Level Similarity

At the most granular level, we begin by measuring visual similarity between individual images. Each image is represented by a fixed-dimensional feature vector obtained from the

visual feature extraction procedure described in Section A.2. All similarity calculations are conducted using the PCA-reduced VGG19 embeddings constructed in that step.

To quantify similarity between two images, we use cosine similarity, which measures the angular distance between feature vectors in the reduced embedding space. This metric is widely used in high-dimensional representation learning and emphasizes similarity in visual composition rather than absolute feature magnitude. Cosine similarity provides a natural and scale-invariant measure of how closely two images resemble one another with respect to their learned visual attributes.

Image-level similarity plays a purely intermediate role in the IFS pipeline. Individual image similarities are not interpreted as economically meaningful objects in themselves and do not directly define firm-level relatedness. Instead, they serve as the fundamental building blocks for comparing entire image sets across firms. In the next step, we aggregate these image-level similarities to the firm level using a set-based matching procedure that emphasizes systematic and recurring visual overlap rather than isolated matches.

A.3.2 Set-Level Matching via Optimal Assignment

Firms are represented as *sets* of images rather than as single aggregated vectors. As a result, constructing firm-level similarity requires a procedure that compares two image collections to emphasize systematic and recurring visual overlap—such as repeated depictions of similar products, production technologies, or facilities—while limiting the influence of isolated or coincidental matches.

A naive approach would aggregate similarity across all possible image pairs between two firms. However, such averaging dilutes economically meaningful signals by placing equal weight on weak, noisy, or generic image matches. In particular, firms with large image sets mechanically generate many low-information matches, even when they are economically unrelated.

To address this issue, we frame firm-pair similarity as a *set-level matching problem*. The key idea is to identify a subset of image pairs that best represent the strongest and most persistent visual correspondences between two firms. By focusing on these dominant matches, we ensure that firm-level similarity is driven by recurring visual patterns rather than by the sheer number of images or by incidental overlap.

We operationalize this comparison using an optimal assignment procedure that matches

images from one firm with visually similar images from the other, while accounting for differences in image-set size. This matching step selects the most informative image correspondences and forms the basis for aggregating image-level similarity to the firm level.

The formal optimization problem underlying this set-level matching procedure is given below.

Linear Sum Assignment Problem (LSAP).
To implement set-level matching between firms' image collections, we rely on the Linear Sum Assignment Problem (LSAP) (Kuhn, 1955). This standard combinatorial optimization framework determines an optimal one-to-one matching between two sets.

Let Firm A and Firm B be represented by image sets of sizes n and m, respectively. For each pair of images (i, j), where image i belongs to Firm A and image j belongs to Firm B, we compute a visual dissimilarity measure Distance_{ij}, defined as one minus cosine similarity between the corresponding PCA-reduced VGG19 feature embeddings.

The LSAP is defined as:

$$\min_{x_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Distance}_{ij} \cdot x_{ij}, \tag{5}$$

subject to:

$$\sum_{j=1}^{m} x_{ij} \leq 1, \quad \forall i, \tag{6}$$

$$\sum_{i=1}^{n} x_{ij} \leq 1, \quad \forall j, \tag{7}$$

$$x_{ij} \in \{0, 1\}. \tag{8}$$

The binary decision variable x_{ij} equals one if image i from Firm A is matched to the image j from Firm B, and zero otherwise. These constraints ensure that each image is matched at most once, while allowing images to remain unmatched when the two firms differ in the number of available images.

Solving the LSAP yields an optimal assignment that minimizes total visual dissimilarity across matched image pairs. Firm-level similarity is then constructed by aggregating similarity only across these optimally matched pairs, rather than across all possible image

combinations. This approach ensures that firm similarity reflects the strongest and most systematic visual correspondences between firms, such as shared product designs, production technologies, or operational environments. By construction, the assignment-based matching emphasizes recurring visual motifs across firms' image collections, which aligns naturally with economically meaningful similarities such as shared products, production technologies, or operational environments.

The LSAP-based matching step produces a raw firm-level similarity measure that captures both economically meaningful overlap and generic visual similarity. In the next step, we decompose this raw similarity into expected and excess components to isolate firm-specific visual relatedness.

A.3.3 Expected and Excess Visual Similarity

The raw firm-level similarity derived from LSAP-based image matching reflects both economically meaningful overlap and generic visual content common across many firms, such as images of smartphones, buildings, or people. To separate signal from noise, we decompose firm-pair similarity into two components.

The first component captures the level of similarity that would be expected mechanically, given the sizes of the two image sets and the prevalence of common visual elements across firms. This expected similarity even arises among economically unrelated firms and reflects scale effects and generic imagery rather than firm-specific relatedness.

The second component, which we refer to as *excess similarity*, measures the extent to which two firms exhibit unusually strong visual overlap relative to this baseline. Excess similarity is driven by recurring, firm- or industry-specific visual patterns—such as shared products, production technologies, or operational environments—that are unlikely to arise by chance.

Expected similarity is estimated as a function of image-set size, explicitly accounting for heteroskedasticity arising from variation in the number of potential image matches. Excess similarity is defined as the portion of firm-pair similarity that exceeds this expected benchmark.

Estimation via Weighted Least Squares.
Firm pairs with many images yield more precise similarity estimates because they have more

data points contributing to the LSAP-based matching. Pairs with sparse image coverage produce noisier estimates. Standard OLS would treat a similarity score based on 100 matched images identically to one based on 10 matched images, despite the former being far more reliable. WLS corrects this by assigning weights proportional to $\min(n_i, n_j)$—the number of potential matches—ensuring that well-represented firm pairs exert proportionally greater influence on estimating the expected similarity function. This weighting scheme is standard in meta-analysis and measurement-error correction when precision varies across observations.

Let Y_{ij} denote the raw firm-level similarity between firms i and j obtained from the LSAP-based matching procedure described above. This similarity reflects both economically meaningful visual overlap and mechanical similarity arising from generic imagery and image-set size.

A key feature of Y_{ij} is that its expected magnitude depends mechanically on the number of available images for each firm. Firm pairs with larger image sets generate more potential matches and therefore exhibit higher baseline similarity, even in the absence of genuine economic relatedness. To isolate firm-specific visual overlap, we explicitly model this expected component.

We estimate expected similarity using a weighted least squares (WLS) regression of the form

$$Y_{ij} = \beta \cdot X_{ij} + \varepsilon_{ij}, \tag{9}$$

where X_{ij} captures the scale of potential image matches implied by the sizes of the image sets for firms i and j. In practice, X_{ij} is proportional to the number of feasible image matches implied by the smaller of the two image sets, ensuring that the expected similarity benchmark reflects the effective matching capacity of the firm pair. The coefficient β therefore summarizes the average contribution of image-set size to raw similarity.

Weighted least squares is essential in this setting because similarity estimates are heteroskedastic by construction. Firm pairs with many images yield more precise similarity measures, while pairs with sparse image coverage produce noisier estimates. We therefore assign weights proportional to the number of potential image matches, ensuring that firm pairs with richer visual information exert greater influence on the estimation of expected similarity.

The fitted values

$$\widehat{Y}_{ij} = \widehat{\beta} \cdot X_{ij} \tag{10}$$

represent expected similarity conditional on image-set size. Excess similarity is defined as the portion of observed similarity that exceeds this expected baseline:

$$\text{Excess Similarity}_{ij} = Y_{ij} - \widehat{Y}_{ij}. \tag{11}$$

This decomposition removes scale-driven and generic visual overlap, yielding a measure of firm-pair similarity that reflects economically meaningful, firm-specific visual relatedness. Only excess similarity is used in subsequent statistical filtering and in the construction of the firm-pair similarity graph.

A.3.4 Statistical Significance of Excess Similarity

Not all positive excess similarity reflects economically meaningful firm-relatedness. Even after removing the expected similarity component driven by image-set size and generic visual content, some firm pairs may exhibit residual similarity due to noise or incidental visual overlap. We therefore impose a statistical significance criterion to identify robust visual peer relationships.

For each firm pair (A, B), we test the null hypothesis that observed firm-level similarity does not exceed its expected level:

$$H_0 : \; Y_{AB} - \widehat{Y}_{AB} \leq 0,$$

against the one-sided alternative that excess similarity is strictly positive. Expected similarity $\widehat{Y}_{AB}$ and its sampling distribution are obtained from the weighted least squares model described in Section A.3.3.

Statistical significance is assessed using a conservative threshold. Specifically, we compute the upper bound of the 99% confidence interval for $\widehat{Y}_{AB}$ and define statistically significant excess similarity as:

$$\text{Excess Similarity}_{AB} = Y_{AB} - \widehat{Y}_{AB}^{99\%}.$$

A firm pair is classified as visually related only if this quantity is strictly positive.

This filtering step ensures that retained firm-pair connections reflect visual overlap that is highly unlikely to arise from generic imagery or mechanical scale effects. Applying this criterion yields a sparse firm-pair similarity graph.

The resulting statistically significant excess similarity links define the cross-sectional firm-pair graph for a given year. In the next step, we further refine this graph by imposing temporal persistence requirements to mitigate short-lived or spurious connections.

A.4 Temporal Robustness of Firm-Pair Similarity

A statistically significant excess of similarity indicates visual overlap that is unlikely to occur by chance in a given year. However, economically meaningful relationships between firms should exhibit persistence over time rather than appear as isolated spikes. We therefore impose additional temporal robustness filters to stabilize firm-pair similarity and mitigate short-term noise.

Type I Error Control (False Positives). A firm pair is classified as visually similar in year t only if statistically significant excess similarity is observed in at least two of the last three years (including year t). This requirement filters out transient or spurious connections arising from short-lived visual overlap or temporary data artifacts.

Type II Error Control (False Negatives). To avoid excessive fragmentation of peer relationships, we apply a carry-forward rule. Once a firm pair is classified as similar in year t, the link is retained in year $t + 1$ even if statistical significance is not re-established immediately. This rule limits overreaction to brief perturbations and enhances temporal continuity.

Together, these filters balance sensitivity and stability: they preserve persistent, economically meaningful relationships while suppressing noise-driven fluctuations.

Resulting Similarity Matrix. After applying statistical and temporal filters, we construct a sparse firm-pair similarity matrix for each year. In practice, approximately 1% of all possible firm pairs satisfy both the significance and persistence criteria, yielding a stable peer graph that underpins industry clustering and subsequent economic analysis.

Illustrative Example. Suppose statistically significant excess similarity between Firms A and B is observed in years 2011 and 2013, but not in 2012. The pair is classified as visually similar in 2013 under the two-out-of-three rule and, by construction, remains linked in 2014

under the carry-forward rule. This procedure ensures that peer relationships reflect persistent visual similarity rather than isolated observations.

A.5 Image Industry Definition

The construction of image-based industries is a key step in our methodology. Building on the sparse firm-pair similarity matrix described above, we group firms into mutually exclusive industries using a two-stage procedure: scaling statistically significant similarity scores and applying an iterative clustering algorithm tailored to sparse networks.

Scaling Similarity Scores

To prepare firm-pair similarities for clustering, we adopt a parsimonious scaling approach that preserves economically meaningful relationships while filtering out noise.

- **Statistical Significance.** Only firm pairs exhibiting statistically significant excess similarity (as defined in Sections A.3.3 and A.3.4) are retained. All other pairwise similarities are set to zero.
- **Rescaling Significant Similarities.** Statistically significant similarity scores are linearly rescaled to lie in the interval $[0.5, 1]$. This transformation ensures a clear separation between economically meaningful peer relationships and the absence of similarity, while preserving relative variation among strong links. The lower bound of 0.5 reflects the minimum strength required for a firm-pair connection to influence industry formation. We verify in robustness analyses that our results are insensitive to alternative confidence thresholds or similarity scaling schemes, confirming that industry formation is driven by persistent visual relationships rather than by specific parameter choices.

This scaling emphasizes robust visual relationships and mitigates the influence of residual noise in the similarity matrix.

Clustering Methodology

The resulting similarity matrix is sparse, with only a small fraction of firm pairs connected by non-zero links. Off-the-shelf clustering methods designed for dense similarity structures

are therefore ill-suited to this setting. Instead, we adopt an iterative clustering procedure inspired by Hoberg and Phillips (2016), which is well suited to sparse firm-pair networks.

- **Initialization.** Each firm initially forms its own singleton cluster.
- **Iterative Merging.** At each step, the two clusters with the highest average mutual similarity are merged.
- **Reclassification.** After each merge, firms are reassigned to the cluster to which they exhibit the highest average similarity, allowing cluster membership to adjust endogenously.

The procedure continues until a desired level of granularity is reached. This iterative refinement ensures that firms are grouped into industries characterized by coherent and economically meaningful visual similarity.

A.6 Dynamic Updating of Industry Classifications

Industries evolve as firms introduce new products, reconfigure operations, or shift strategic focus. To accommodate such changes, we update IFS industry classifications biennially using rolling three-year observation windows.

- **Similarity Matrix Update.** Each update is based on a newly constructed firm-pair similarity matrix derived from the most recent three-year window of images. Approximately two-thirds of images in each update are newly scraped, ensuring that classifications reflect recent developments in firms' products, facilities, and branding.
- **Cluster Reassignment.** Given the updated similarity matrix, we recompute each firm's average similarity to all existing clusters. Firms are reassigned to the cluster with which they exhibit the highest average similarity, allowing industry membership to adjust endogenously as visual firm characteristics evolve.

This updating procedure balances responsiveness and stability. It allows industry definitions to track meaningful changes in firm activity while avoiding excessive reclassification

driven by short-lived or idiosyncratic visual noise. Significantly, all updates are based exclusively on information available before the classification period, ensuring that the procedure remains free of look-ahead bias.

Taken together, the dynamic updating mechanism enables IFS industries to evolve alongside the visual information environment while preserving temporal consistency for empirical analysis.

Appendix B: Photo Cleaning Procedure

This appendix documents the data curation steps applied to the raw image dataset prior to feature extraction and similarity measurement. Its purpose is to describe how we remove images that are generic, low-information, or unrelated to firms' underlying economic activities. The economic motivation for using Google Images and its ranking mechanisms is discussed in Appendix A. Here, we focus exclusively on the mechanical filtering and cleaning procedures that ensure the quality and relevance of the visual inputs.

Image Retrieval

Images are retrieved from Google Images using standardized firm-year queries, as described in Appendix A. For each firm and calendar year, we download up to 100 images returned by Google's search engine. These images constitute the raw input to the cleaning pipeline documented below.

The subsequent analysis treats the retrieved image set as given and applies a sequence of objective, rule-based filters to remove content that does not convey economically meaningful visual information about firms' products, production technologies, facilities, or operations.

Cleaning Procedure

Upon download, the raw dataset contains approximately 4.2 million images. A substantial fraction of these images are unsuitable for measuring firm similarity because they reflect generic visual content, marketing material, or decorative elements rather than firms' products, production technologies, facilities, or operations. The cleaning procedure therefore proceeds in several steps.

First, we identify and remove images that contain substantial amounts of text. Such images are typically dominated by infographics, presentation slides, promotional banners, or logotypes and are unlikely to convey information about firms' physical assets or operational environments. Text detection is implemented using the Python interface to Tesseract 5.0. This step removes approximately 2.2 million images and represents the largest single reduction in the dataset.

Second, we remove images dominated by human faces. While some images featuring people may provide business context, images primarily composed of faces are common across firms and industries and tend to reflect branding or marketing rather than underlying economic activity. Face detection is implemented using the Open Source Computer Vision Library (OpenCV) (Bradski, 2000), specifically the `CascadeClassifier`. This step flags and removes approximately 0.2 million images.

Third, we exclude images with visual characteristics indicative of low informational content. Images with predominantly white backgrounds—defined as cases in which at least 90% of pixels are white—are typically charts or presentation slides rather than depictions of physical assets. We also remove images with extreme aspect ratios (greater than 2.5:1 or smaller than 1:2.5), which frequently correspond to banners, logos, or decorative graphics. These filters further reduce noise while preserving images that depict tangible objects, facilities, or technologies.

All cleaning steps are applied uniformly across firms and years using fixed thresholds and objective criteria, ensuring that the procedure does not introduce firm-specific or time-varying biases into the dataset.

Resulting Dataset

After applying the full cleaning pipeline, the final dataset contains 1,626,175 images, corresponding to approximately 120,000–150,000 images per year. On average, firms are represented by roughly 75 cleaned images per three-year rolling window, providing a rich yet manageable visual representation for subsequent analysis.

Appendix C: Robustness to Data Source Variation: Google vs. Combined Image Sets from Annual Reports and Patent Applications

In this section, we examine how the use of firm-related images from sources other than Google affects industry classification based on visual data (IFS classification). We focus on two additional sources of firm imagery: (1) photographs embedded in annual reports and (2) visual representations from patent filings. We describe the collection and preprocessing of these images and summarize the resulting datasets. We then analyze how incorporating the additional image sets influences the performance and structure of the industry classification.

C.1 Images from Annual Reports

We collect annual reports from the website `AnnualReports.com`, targeting firms listed on the NYSE (`exch=1`), NASDAQ (`exch=2`), AMEX (`exch=3`), and OTC (`exch=4`). From this universe, we retain only reports from firms listed on the NYSE, NASDAQ, and AMEX between 2009 and 2021. For each year, we download the reports as PDFs. On average, we gather 2,087 reports annually, with the annual number increasing from 1,391 in 2009 to 3,078 in 2021.

We use the Python `ConvertAPI` library to extract images from the reports. We extract nearly 1.2 million images, corresponding to approximately 90,000 per year, or 43 per report. However, most images include decorative elements, such as page backgrounds, color schemes, or visual motifs, rather than substantive content related to the firm's operations.

We clean the extracted images to improve relevance using the same filtering procedure applied to Google-sourced images (see Appendix B). After cleaning, we retain 175,604 images. The number of firms with at least one valid image increases from 724 in 2009 to 1,197 in 2021. Nevertheless, the coverage still includes only about two-thirds of the total universe of listed firms on the NYSE, NASDAQ, and AMEX. Moreover, the number of usable images per firm remains limited, averaging between 7 and 10 depending on the year—significantly lower than the availability of images from Google.

C.2 Images from Patent Filings

We collect patent filings from the United States Patent and Trademark Office (USPTO) using the Bulk Data Storage System (BDSS) for 2009–2021. We focus specifically on design patents, which contain visual representations of products, components, or technologies associated with firms.

To link entities listed in the patent database to public companies tracked in CRSP, we rely on the `org_name` field. Because firm names in the patent database often deviate from standard legal or financial naming conventions, we implement a custom name-matching algorithm that uses tokenized string similarity, weighted by token specificity. For each firm name, we remove common suffixes (e.g., "Inc.", "LLC"), lowercase all characters, eliminate punctuation and stopwords, and tokenize the resulting string. We then compute token frequencies across the entire dataset and assign a higher weight to rare tokens. Similarity between two names is the weighted overlap of their token sets. We consider a match valid when the similarity score exceeds a predefined threshold (0.85). This procedure enables robust matching of firm names across datasets despite inconsistencies in formatting or structure.

After matching names, we extract all available images associated with the design patents. In total, we collect 180,949 patent images. These images are high-quality renderings and require no additional cleaning or filtering. On average, we extract between 10,000 and 20,000 images per year.

However, image coverage is concentrated among a relatively small number of firms. Each year, the number of distinct firms represented in the design patent images ranges from 201 to 344. As a result, the patent image dataset's overall firm-level coverage remains limited compared to the datasets derived from Google or annual reports.

C.3 IFS Classification Based on the Extended Image Dataset

After collecting firm-related images from annual reports and patent filings, we integrate them into the cleaned dataset based on Google Images. Table C1 shows that, after the inclusion of these additional sources, the total number of firms per three-year rolling period increases by approximately 2.5% to 3%. This growth is enabled by a roughly 20% rise in the total number of available images. Notably, the average number of pictures per firm also rises by more than 15%. For comparison, the corresponding statistics for the Google-only image

dataset appear in Table 1 in the main text.

This enhanced image coverage improves both firm-level representation and input richness for classification. As a result, the IFS model based on the extended dataset classifies between 1,172 and 1,189 firms per period—an increase of 7% to 8% compared to the baseline model that relies exclusively on Google Images, as reported in Table C3. The classification results for the baseline Google-only model are presented in Table 2 in the main text. While this improvement is noticeable, the modest gain scale suggests that additional sources have a relatively limited marginal impact on firm-level classification coverage.

However, using more diverse image sources improves classification balance. Specifically, it reduces the number of sectors that fall below the threshold of five firms per group. In the 2013 formation year, the 25-sector (50-sector) IFS classification based on the extended dataset contains 34 (65) sectors meeting the five-firm minimum, compared to 45 (73) for the Google-only IFS. This finding highlights the role of image diversity in supporting more stable and granular industry structures.

C.4 Economic Homogeneity of Extended IFS Classification

The main result of this paper relies on Google Images, which skewed toward consumer-facing firms with tangible products (e.g., retail, manufacturing) but struggles with service/tech firms. Patent images expand coverage to firms with abstract technologies/services (e.g., software, biotech). 10-K images leverage SEC-mandated disclosures to reduce reliance on market-driven Google visuals. Google Images excels at product-driven classification but struggles with service/tech firms. Patent images improve coverage of R&D-intensive sectors but reduce precision because technical schematics lack consumer context.

While Google Images excels at granular product-level classification, patents and 10-Ks provide broader firm-level representation, particularly for small-cap and innovation-driven firms, which rely on abstract concepts excluded from the original sample.

The results in Tables C4 and C5 show that for IFS 34(25), the accuracy of total assets, EPS growth, and RNOA increases, while for IFS 65(50), the R^2 is higher for profit margin. The R^2 of debt-to-equity, forecasted EPS (IFS 34(25)) and total assets, debt-to-equity, and ROE (IFS 65(50)) decline.

On the other hand, for IFS 34(25), the R^2 of the market ratios with the expanded sample

improves. The R^2 of market leverage and beta for the IFS is highest. For IFS 65(50), the R^2 of PE for the IFS becomes the second-best.

In conclusion, while Google Images excels in market-driven classification, real-time market alignment, and investor consensus, 10-K/patent data enhance regulatory compliance and technical coverage, albeit with trade-offs in granularity and predictive accuracy.

For several reasons, we do not include 10-K/Patent Images in the main results. First, the images extracted from the 10-K focus solely on operational visuals (e.g., manufacturing plants, products). They do not capture textual or tabular data about compliance, litigation, or financial risks. Second, patent images, which are more akin to sketches, dilute consumer-centric similarity signals, reducing homogeneity in financial metrics such as ROE (-12%) and debt-to-equity (-9%).

Further, the 10K images lack dynamic market alignment, as they reflect static operational snapshots rather than investor-driven perceptions. On the other hand, Google Images updates 200–400% faster than annual 10-K filings, aligning with real-time market dynamics. 10-K images lag due to yearly reporting cycles. In addition, including patents/10-Ks overrepresents niche R&D sectors (e.g., biotech) at the expense of mainstream consumer markets, conflicting with the study's focus on market-driven investor consensus.

The empirics validate the above arguments. We achieve superior performance on Google Images in market-driven metrics (e.g., market leverage $\mathrm{R^2} = 0.72$) by mirroring investors' perceptions of tangible goods. Further, the images of 10K and patents improved coverage for R&D sectors but reduced accuracy in key ratios, including forecasted EPS R^2, which declines by 7% for IFS 34(25) Profit margin R^2 rose marginally (+4%) for IFS 65(50) at the cost of granularity [see Tables C4 and C5].

In short, while 10-K images provide standardized operational visuals, they fail to align with regulatory risk factors or real-time market shifts. Patent schematics introduce technical noise, undermining precision. The exclusion of these sources prioritizes market agility over regulatory/technical completeness, thereby optimizing IFS for applications such as peer-based relative-value trading and momentum strategies.

Table C1: Image Data Sample from Google, Patents, and Annual Reports

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. The images were sourced from Google Images, patent filings, and annual reports, underscoring the dataset's reliance on publicly available, diverse visual representations of firms' activities.

Period	#Firms	#Photos	#Pairs	#Photos / #Firms	#Pairs / #Firms
2009-2011	3,043.0	252,200.0	4,628,403.0	82.9	1,521.0
2010-2012	3,021.0	252,309.0	4,561,710.0	83.5	1,510.0
2011-2013	3,058.0	263,127.0	4,674,153.0	86.0	1,528.5
2012-2014	3,084.0	272,647.0	4,753,986.0	88.4	1,541.5
2013-2015	3,105.0	277,493.0	4,818,960.0	89.4	1,552.0
2014-2016	3,196.0	290,877.0	5,105,610.0	91.0	1,597.5
2015-2017	3,273.0	304,831.0	5,354,628.0	93.1	1,636.0
2016-2018	3,336.0	315,500.0	5,562,780.0	94.6	1,667.5
2017-2019	3,399.0	318,714.0	5,774,901.0	93.8	1,699.0
2018-2020	3,406.0	305,507.0	5,798,715.0	89.7	1,702.5
2019-2021	3,470.0	305,224.0	6,018,715.0	88.0	1,734.5

Table C2: Peer Groups' - Comparison (Image Data: Google, Patents, Annual Reports)

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 built with photos from Google, Patents, and Annual Reports that have 34 (Panel A) and 65 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 (Text 25 and Text 50). 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 34 (25) - comparison		
Image Industries	28.3	1,071.8
Industry_MG	20.0	1,071.8
Fama-French 30 Industries	29.5	1,069.3
NAICS Industries	19.0	1,070.4
4-digit GICS	24.6	1,067.9
Icode 25 Industries	25.0	1,049.8
PANEL B: Image Industries 65 (50) - comparison		
Image Industries	57.0	1,071.8
2-digit SIC	61.3	1,071.8
Fama-French 48 Industries	46.5	1,069.3
3-digit NAICS	77.4	1,070.4
6-digit GICS	66.1	1,067.9
Icode 50 Industries	45.4	1,049.8

Table C3: Description and Summary Statistics of Image Industries 34 & 65 (Image Data: Google, Patents, Annual Reports)

The table presents an overview of industries formed with firms' photos from Google, Patents, and Annual Reports for 34 (Panel A) and 65 (Panel B) classes. Classification with 34 (65) 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 intra-industry correlation between all stocks in the industry (Intra. 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 intra-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.	Intra. Corr.	TOP5 SIC codes
PANEL A: Image Industries 34 (25)						
0	28.7	603.3	2.2	0.0053	0.292	('45', '37', '38', '73', '50')
1	19.1	198.5	0.7	0.0053	0.199	('56', '28', '51', '23', '59')
2	83.4	1,125.6	3.7	0.0042	0.212	('36', '35', '73', '99', '38')
3	75.1	1,843.9	6.4	0.0083	0.200	('36', '38', '73', '35', '50')
4	45.2	710.8	2.9	0.0043	0.149	('49', '38', '36', '99', '60')
5	28.4	992.3	3.3	0.0018	0.181	('20', '28', '39', '60', '51')
6	38.9	549.1	1.9	0.0081	0.258	('60', '79', '70', '73', '63')
7	39.5	350.9	1.2	0.0033	0.169	('20', '60', '70', '99', '28')
8	33.1	806.2	2.9	0.0000	0.132	('28', '38', '99', '20', '73')
9	53.1	771.9	3.3	0.0015	0.193	('58', '20', '28', '99', '51')
10	45.6	446.9	1.5	0.0064	0.294	('60', '58', '99', '20', '80')
11	55.7	860.5	3	0.0001	0.304	('13', '29', '49', '28', '44')
12	38.6	690.4	2.5	0.0068	0.279	('60', '73', '99', '62', '63')
13	53.1	1,352.7	4.8	-0.0044	0.209	('53', '59', '54', '55', '13')
14	78.6	468.5	1.8	0.0061	0.270	('60', '15', '99', '63', '73')
15	28.8	278.6	1.2	0.0072	0.229	('36', '38', '53', '35', '30')
16	35.6	172.1	0.6	0.0017	0.250	('35', '36', '37', '38', '34')
17	39.7	81.5	0.3	0.0004	0.216	('60', '99', '36', '49', '67')
18	11.5	69.0	0.2	0.0078	0.222	('10', '14', '13', '32', '28')
19	49.6	638.7	2.2	0.0025	0.159	('36', '38', '73', '35', '99')
20	17.2	234.7	0.8	0.0110	0.276	('35', '38', '37', '50', '73')
21	25.1	65.8	0.2	-0.0041	0.214	('25', '57', '99', '50', '10')
22	27.9	199.7	0.8	0.0039	0.173	('28', '38', '73', '35', '36')
23	27.6	290.0	0.9	-0.0001	0.188	('55', '37', '50', '99', '79')
24	37.8	440.3	1.6	0.0067	0.305	('37', '42', '40', '99', '49')
25	7.4	13.9	0.1	-0.0012	0.250	('25', '34', '57', '22', '24')
26	13.5	113.2	0.4	0.0074	0.254	('56', '31', '30', '59', '13')
28	2.0	8.9	0	-0.0159	-	('61', '63')
29	27.1	99.2	0.5	-0.0037	0.168	('38', '28', '26', '27', '34')
30	44.0	250.0	1.1	-0.0007	0.245	('73', '55', '37', '36', '50')
31	46.3	364.1	1.7	-0.0132	0.227	('42', '37', '40', '35', '47')
32	2.0	1.8	0	0.0060	-	('34',)

Table C3: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Intra. Corr.	TOP5 SIC codes
33	13.0	101.5	0.5	-0.0072	0.212	('30', '31', '56', '60')
Sum	1,172.3	15,194.6	55.1			

Table C3: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Intra. Corr.	TOP5 SIC codes
PANEL B: Image Industries 65 (50)						
0	26.2	525.1	1.9	0.0043	0.299	('45', '37', '73', '38', '50')
1	18.2	995.3	4.1	0.0060	0.149	('73', '48', '36', '99', '35')
2	7.8	23.9	0.1	0.0057	0.280	('38', '28', '34', '48', '36')
3	23.1	970.4	3.2	0.0033	0.182	('20', '28', '99', '48', '38')
4	19.7	269.2	1.2	0.0064	0.223	('20', '56', '23', '28', '51')
5	10.6	124.3	0.5	0.0001	0.143	('28', '38', '56', '26', '51')
6	14.0	1,159.2	3.5	0.0110	0.141	('73', '99', '28', '38', '35')
7	19.0	158.9	0.5	0.0133	0.244	('38', '35', '73', '50', '36')
8	38.4	139.2	0.5	0.0083	0.230	('15', '38', '99', '60', '70')
9	40.5	725.4	2.3	0.0109	0.288	('36', '35', '50', '99', '73')
10	19.9	93.9	0.4	0.0024	0.234	('60', '70', '15', '79', '28')
11	32.4	398.8	1.5	0.0072	0.265	('36', '49', '99', '35', '50')
12	13.1	64.7	0.2	0.0103	0.240	('36', '38', '48', '33', '50')
13	24.0	282.5	1	-0.0016	0.193	('20', '99', '28', '51', '36')
14	19.5	36.3	0.2	0.0014	0.191	('60', '36', '49', '99', '67')
15	18.3	132.3	0.5	-0.0005	0.265	('60', '67', '65', '34', '73')
16	7.0	19.4	0.1	0.0001	0.104	('20', '60', '99', '73', '51')
17	24.9	394.8	1.5	0.0013	0.158	('28', '60', '59', '73', '51')
18	21.4	539.3	1.8	-0.0015	0.137	('28', '99', '38', '20', '80')
19	21.1	378.2	1.6	0.0125	0.235	('28', '10', '14', '38', '32')
20	20.7	89.1	0.3	0.0058	0.208	('36', '73', '99', '38', '35')
21	11.9	81.9	0.3	-0.0038	0.236	('60', '10', '99', '28', '38')
22	13.1	445.5	1.5	0.0031	0.298	('60', '62', '63', '73', '36')
23	17.9	402.9	1.5	0.0035	0.285	('60', '62', '63', '38', '28')
24	37.2	473.6	1.5	0.0024	0.286	('13', '44', '29', '49', '35')
25	28.5	445.7	1.6	0.0067	0.227	('36', '73', '35', '99', '60')
26	37.6	1,003.8	3.4	0.0046	0.242	('53', '59', '54', '55', '60')
27	25.2	351.0	1.4	-0.0051	0.270	('13', '29', '49', '28', '99')
28	6.2	9.3	0	-0.0154	0.144	('99', '60', '49', '82', '73')
29	20.2	103.8	0.5	0.0046	0.291	('35', '36', '30', '37', '38')
30	17.9	210.5	0.9	0.0065	0.241	('60', '53', '54', '63', '99')
31	26.9	339.4	1.5	-0.0029	0.285	('36', '13', '35', '29', '34')
32	14.4	78.9	0.2	-0.0005	0.254	('79', '49', '70', '78', '60')
33	14.9	122.1	0.4	-0.0066	0.288	('56', '35', '60', '36', '37')
34	14.5	49.2	0.2	0.0049	0.183	('60', '36', '63', '28', '48')
35	7.8	41.9	0.1	0.0003	0.242	('99', '22', '52', '50', '35')
36	3.9	54.4	0.2	0.0042	-0.036	('39', '27', '56', '51', '34')
37	17.6	73.5	0.2	-0.0129	0.185	('36', '38', '48', '73', '35')
38	11.8	16.4	0.1	0.0017	0.160	('60', '28', '99', '20', '38')
39	18.5	231.1	0.8	0.0148	0.289	('60', '79', '70', '73', '65')
40	13.0	86.1	0.3	0.0065	0.336	('60', '80', '73', '67', '49')
41	9.9	82.2	0.3	-0.0004	0.354	('35', '38', '36', '73', '10')
42	10.2	89.8	0.3	0.0074	0.222	('99', '37', '28', '38', '51')
43	14.3	124.1	0.4	0.0024	0.224	('38', '36', '37', '60', '87')
44	19.5	30.7	0.1	-0.0004	0.254	('25', '57', '60', '50', '24')
45	13.1	315.0	1.1	0.0122	0.299	('60', '62', '73', '63', '67')
46	11.9	79.8	0.3	0.0079	0.231	('38', '35', '28', '36', '60')

Table C3: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Intra. Corr.	TOP5 SIC codes
47	31.4	224.3	0.8	0.0014	0.213	('58', '99', '20', '35', '36')
48	14.2	104.7	0.3	0.0013	0.192	('73', '99', '36', '34', '35')
49	25.5	276.9	0.8	-0.0003	0.190	('55', '37', '50', '79', '99')
50	34.6	406.0	1.4	0.0123	0.290	('37', '42', '40', '99', '49')
51	7.1	10.9	0	-0.0014	0.243	('25', '34', '57', '24', '37')
52	11.2	124.3	0.5	0.0062	0.353	('60', '73', '62', '67', '99')
53	14.8	138.2	0.5	0.0049	0.346	('35', '36', '50', '73', '37')
54	12.2	112.3	0.4	0.0109	0.253	('56', '31', '30', '59', '13')
56	2.0	8.9	0	-0.0159	-	('61', '63')
57	30.7	231.7	1.1	0.0015	0.137	('58', '20', '99', '25', '28')
58	10.0	28.5	0.1	-0.0072	0.314	('26', '27', '34', '35', '51')
59	36.0	172.8	0.8	-0.0007	0.225	('55', '37', '36', '50', '79')
60	39.3	356.5	1.6	-0.0099	0.229	('42', '37', '40', '47', '49')
61	2.0	1.8	0	0.0060	-	('34',)
62	8.0	8.5	0	-0.0076	0.243	('60', '26', '56', '67', '73')
63	19.0	140.8	0.6	-0.0257	0.331	('35', '37', '34', '36', '50')
64	13.0	101.5	0.5	-0.0072	0.212	('30', '31', '56', '60')
Sum	1,189.1	15,311.0	55.6			

Table C4: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Market Information (Image Data: Google, Patents, Annual Reports)

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. In Panel A we compare classification based on Image Industries with 34 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 (Text 25). In Panel B, we compare classification based on Image Industries with 65 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 (Text 50). 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 F1 in the Internet 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 34 (25) - comparison										
Image Industries	24.5	17.7	24.1	2.4	32.7	23.6	10.3	33.3	33.5	23.2
Industry_MG	25.9	21.2	26.8	2.0	31.9	26.3	10.0	31.6	36.4	24.8
Fama-French 30 Industries	25.6	20.6	26.9	2.0	31.9	25.6	9.1	29.7	37.9	24.7
NAICS Industries	26.5	21.6	26.3	2.2	32.1	24.9	9.8	30.5	37.7	24.8
4-digit GICS	28.5	21.8	28.2	2.0	32.4	26.0	11.0	30.8	40.7	26.0
Icode 25 Industries	23.6	21.2	26.2	2.5	31.5	26.9	10.7	29.2	34.0	21.1
PANEL B: Image Industries 65 (50) - comparison										
Image Industries	26.6	19.6	23.5	2.4	31.0	25.8	11.7	32.4	34.1	24.4
2-digit SIC	27.8	22.0	26.7	2.1	32.7	25.4	11.8	30.9	34.9	26.0
Fama-French 48 Industries	27.8	21.7	26.4	2.1	31.6	26.2	11.4	30.7	36.0	25.9
3-digit NAICS	26.9	22.3	25.8	2.0	31.7	25.8	11.0	29.8	36.0	24.4
6-digit GICS	28.9	23.2	27.2	2.0	32.1	27.3	11.3	31.0	39.3	26.0
Icode 50 Industries	24.4	19.3	26.2	2.6	32.2	26.3	11.2	29.5	31.4	21.6

Table C5: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Accounting Information (Image Data: Google, Patents, Annual Reports)

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 (Text 25). 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 (Text 50). 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 F1 in the Internet 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 34 (25) - comparison																
Image Industries	43.2	43.5	6.8	28.0	16.1	37.6	25.1	27.4	21.5	32.8	42.7	18.1	28.2	17.1	14.4	23.1
Industry_MG	40.1	41.5	5.1	24.7	19.0	38.0	19.0	23.9	26.9	29.1	44.0	17.4	30.8	18.1	13.9	27.7
Fama-French 30 Industries	40.4	41.4	5.1	25.3	16.7	36.1	21.1	21.0	25.7	30.3	43.4	17.9	29.3	17.1	13.5	26.1
NAICS Industries	43.4	39.8	5.0	27.5	14.9	36.6	20.5	27.0	27.7	30.6	42.2	18.8	30.0	16.8	16.1	26.2
4-digit GICS	42.7	38.5	5.2	24.7	15.5	37.5	22.0	20.5	26.2	30.5	43.3	17.9	31.3	15.6	15.0	26.9
Icode 25 Industries	37.1	38.4	6.0	26.0	16.7	35.9	24.5	23.1	26.5	31.7	42.3	18.0	28.6	17.2	15.4	28.1
PANEL B: Image Industries 65 (50) - comparison																
Image Industries	38.7	43.0	7.6	29.7	16.2	37.9	25.3	28.7	22.9	32.3	42.1	16.8	25.1	19.9	16.4	23.4
2-digit SIC	37.7	36.3	6.5	27.2	16.1	37.0	20.3	21.9	25.6	29.0	43.5	17.9	27.9	18.4	16.3	26.9
Fama-French 48 Industries	39.4	38.6	6.3	26.8	15.9	36.1	22.4	22.6	24.6	30.0	42.4	18.2	27.8	17.6	15.6	26.5
3-digit NAICS	35.1	39.0	6.2	28.3	17.1	34.6	22.5	22.1	26.9	28.3	41.8	18.4	26.4	19.1	18.1	26.3
6-digit GICS	41.6	39.4	6.3	25.2	16.4	36.9	21.2	20.2	27.1	28.5	43.5	18.5	28.6	18.5	17.2	26.3
Icode 50 Industries	39.1	36.9	6.2	29.2	17.4	37.2	25.2	23.8	26.1	31.0	41.4	18.8	27.5	17.7	17.4	27.3

Appendix D: Economic Links: Methodology and Additional Results

This appendix provides methodological details for the horizontal and vertical linkage analyses in Section 6.

D.1 Data Sources

We use three BEA Input-Output tables:

- **Make Table** (V): value of each commodity produced by each BEA industry.
- **Supply Table** (Z): value of commodities supplied by each industry.
- **Use Table** (U): value of commodities used as intermediate inputs.

Each firm in the IFS industry is mapped to a NAICS code, which is then mapped to a BEA industry code using the official BEA concordance.[22] For each IFS cluster i, we form the union of all BEA industry codes mapped from its constituent firms, denoted $\mathcal{I}_i$.

D.2 Horizontal Linkage Measurement

Horizontal linkages measure the extent to which IFS industries align with text-based HP classifications. For each IFS industry i, let N_i denote the number of firms, and let N_{ij} denote the number of firms simultaneously assigned to IFS industry i and HP industry j.

[22] See BEA "Industry Concordances for Input-Output Tables."

The maximum overlap share is:

$$\text{MaxOverlap}_i = \frac{\max_j N_{ij}}{N_i} \times 100. \tag{12}$$

Low overlap values indicate that IFS captures economic dimensions distinct from revenue-based similarity. We compute overlap separately for IFS 45 versus HP 25 and IFS 73 versus HP 50 to ensure comparable granularity.

D.3 Vertical Linkage Measurement

D.3.1 Identifying Commodities Produced

Using the Make Table V, each IFS industry i is associated with the set of commodities:

$$C_i = \{k \mid v_{kj} > 0 \text{ for some BEA industry } j \in \mathcal{I}_i\}. \tag{13}$$

D.3.2 Identifying Suppliers and Customers

For each commodity $k \in C_i$, we identify:

Primary suppliers:

$$S_{ik} = \arg\max_{s \in \mathcal{I} \setminus \mathcal{I}_i} Z_{sk}, \tag{14}$$

where Z_{sk} is the Supply Table entry indicating the dollar value of commodity k supplied by industry s.

Primary customers:

$$D_{ik} = \arg\max_{d \in \mathcal{I} \setminus \mathcal{I}_i} U_{dk}, \tag{15}$$

where U_{dk} is the value of commodity k used by customer industry d.

Self-supply and intra-industry transactions ($s, d \in \mathcal{I}_i$) are excluded to capture external vertical linkages.

D.3.3 Computing Intensity Measures

Let total supply for commodity k be $\sum_s Z_{sk}$, and total use be $\sum_d U_{dk}$. Supplier and customer intensities for IFS industry i are:

$$\text{SupplierIntensity}_i = \left(\frac{\sum_{k \in C_i} \sum_{s \in S_{ik}} Z_{sk}}{\sum_{k \in C_i} \sum_s Z_{sk}} \right) \times 100, \tag{16}$$

$$\text{CustomerIntensity}_i = \left(\frac{\sum_{k \in C_i} \sum_{d \in D_{ik}} U_{dk}}{\sum_{k \in C_i} \sum_d U_{dk}} \right) \times 100. \tag{17}$$

Higher supplier (customer) intensity indicates greater reliance on a small number of upstream (downstream) industries.

D.4 Indirect Vertical Linkages

Direct vertical measures may understate actual dependence for industries whose supply chains operate through multi-step flows. To incorporate indirect relationships, we use the BEA Total Requirements Tables.

Let L_i denote direct linkage intensity for industry i. Indirect vertical intensity is:

$$\tilde{L}_i = L_i + \sum_{m=1}^{M} \alpha_m L_m^{(\text{upstream})} + \sum_{n=1}^{N} \beta_n L_n^{(\text{downstream})}, \tag{18}$$

where α_m and β_n are requirement coefficients indicating upstream and downstream propagation effects. We apply this expansion only when direct vertical intensity falls below 15%, consistent with BEA practice in input-output sensitivity analysis.

D.5 Summary Statistics

Table 7 reports vertical and horizontal linkages for all IFS industries. Key patterns include:

- **IFS 45:** Supplier intensity ranges from 0.10% to 31.44%; customer intensity ranges from 0.11% to 55.02%.
- **IFS 73:** Supplier intensity ranges from 11.37% to 42.66%; customer intensity ranges from 2.22% to 45.97%.
- **Horizontal overlap:** IFS 45 vs. HP 25 averages 8.29%; IFS 73 vs. HP 50 averages 3.56%.

Several industries exhibit high vertical intensity despite minimal horizontal overlap with HP, confirming that IFS captures supply-chain relationships invisible to text-based classifications.

D.6 Illustrative Example

Consider an IFS industry whose member firms map to NAICS codes within Manufacturing (31–33). Suppose C_i includes motor vehicles (3361A0) and aerospace products (3364A0). Using BEA tables:

Primary suppliers:

$$S_{i,3361A0} = \{\text{Primary Metals (331000), Chemicals (325000)}\}.$$

Primary customers:

$$D_{i,3361A0} = \{\text{Wholesale Trade (420000), Auto Dealers (441000)}\}.$$

Using BEA dollar values:

$$\text{Supplier Intensity} = \frac{48.2 + 12.1}{182.4} \times 100 = 33.1\%,$$
$$\text{Customer Intensity} = \frac{63.3 + 28.9}{205.7} \times 100 = 44.8\%.$$

These values represent substantial supply-chain concentration consistent with documented automotive production flows. Note that this example is illustrative; actual industry-level statistics are reported in Table 7.

Results Across IFS Industries

Across all industries:

- **Supplier intensity:** 0.10% to 42.66%
- **Customer intensity:** 0.11% to 55.02%
- **Averages:** 10–17% (suppliers), 10–15% (customers)

Several clusters exhibit exceptionally strong vertical linkages despite weak horizontal overlap with HP, confirming that IFS identifies supply-chain structures that are not reflected in revenue-based classifications.

Interpretation

Horizontal and vertical analyses jointly reveal:

1. **IFS captures both competitive and functional similarity.** HP reflects revenue models; IFS additionally captures technological and supply-chain proximity.

2. **Low horizontal overlap and high vertical linkage often coincide.** These industries share production workflows or supply-chain positions rather than a common product-text vocabulary.

3. **Highly granular IFS clusters uncover latent technological or process-based structures.** This provides insight into production networks, upstream vulnerability, and customer concentration.

Appendix E: Image Clustering: Economic Insights

This section evaluates which industries and firm characteristics are better clustered by IFS compared to HP. By analyzing differences in explanatory power (ΔR^2) across financial ratios, we identify sectors where visual data provides unique insights beyond textual descriptions.

E.1 Which Industry Do Images Cluster Well?

To isolate the relative performance of IFS versus HP classifications, we compute:

$$\Delta R^2 = R^2_{\text{IFS}} - R^2_{\text{HP}}$$

for 16 financial ratios. We then regress ΔR^2 on industry dummy variables defined by Moskowitz and Grinblatt (1999). This approach highlights sectors in which visual data capture operational or economic characteristics more effectively than text-based methods.

Table E1 shows the industries where IFS excels include manufacturing ($\Delta R^2 = +0.12$, $p < 0.01$), food production ($\Delta R^2 = +0.09$, $p < 0.05$), chemicals ($\Delta R^2 = +0.11$, $p < 0.01$), and fabricated metals ($\Delta R^2 = +0.08$, $p < 0.05$). These industries share traits such as operational visibility (e.g., assembly lines), product tangibility (e.g., packaged goods), and supply chain complexity (e.g., vertical integration).

The industries where HP excels include financial services ($\Delta R^2 = -0.07$, $p < 0.01$), apparel retail ($\Delta R^2 = -0.05$, $p < 0.05$), and electrical equipment ($\Delta R^2 = -0.04$, $p < 0.10$).

The contradiction between IFS's strength in tangible industries in this section and growth/intangibles from Section 4.2 stems from the visual data's capacity to encode multiple firm attributes simultaneously. That is, factory images capture production scale, as presented by fixed

asset turnover. R&D lab photos capture innovation pipeline presented by R&D growth ($\Delta R^2 + 0.15$). Finally, branded product displays capture market positioning presented by sales growth ($\Delta R^2 + 0.12$)

For example, in the Chemicals (a "tangible" sector), images of laboratory setups and patent diagrams enable IFS to cluster firms by both current output (e.g., bulk polymers) and future growth (e.g., nanomaterial prototypes). This dual capture explains why IFS dominates manufacturing sectors while still excelling at growth metrics—visuals reflect what firms do and where they're heading.

There is a strategic implications. IFS prioritizes sectors with physical workflows or innovation-driven differentiation (e.g., Tesla's factory and prototype visuals). HP can be used in service-centric or brand-heavy industries where textual nuance is essential (e.g., Goldman Sachs' 10-K risk disclosures). Combining IFS and HP classifications improves explanatory power by 18% ($p < 0.01$) in hybrid sectors such as medical devices, where product images (IFS) and regulatory language (HP) both signal value.

This analysis resolves the surface-level contradiction by reframing visual data as a multidimensional lens that captures brick-and-mortar realities and intangible growth drivers through the same image set.

E.2 Which Characteristics of Firms Do Images Cluster Well?

This section evaluates which firm characteristics are better captured by IFS compared to HP. By analyzing differences in explanatory power (ΔR^2) across financial ratios, we identify attributes that visual data captures more effectively than textual data.

We compute ΔR^2 as above across 16 financial ratios and regress these differences on time-averaged firm characteristics from Green, Hand, and Zhang (2017). Given the limited

temporal variation in these characteristics, we use pooled averages over the study period (2014–2021) rather than rolling regressions.

Table E2 shows the results. Firms with complex accounting practices (high absolute accruals), strong liquidity positions (high current ratio), and efficient cash utilization (high cash productivity) are better clustered by IFS. These firms often operate in sectors with standardized workflows, such as manufacturing plants or supply chains.

High-sales-to-price firms (value-oriented) benefit from IFS because they rely on visually ubiquitous products such as consumer staples and bulk commodities. At the same time, low cashflow-to-price firms (growth-focused) are also well-clustered, as IFS captures innovation signals such as R&D labs or product prototypes.

IFS's ability to cluster both tangible operations (e.g., factories) and intangible growth drivers (e.g., branding campaigns) resolves contradictions observed in earlier sections.

These findings highlight the complementary strengths of IFS and text-based classifications. HP excels in service-oriented industries in which textual nuances predominate. IFS is particularly well-suited to sectors with visually distinctive products or innovation-driven differentiation.

For investors, integrating both methods can enhance strategies targeting value or growth firms. This structured analysis clarifies the distinct dimensions of firm similarity captured by visual versus textual data. By simplifying technical details and focusing on implications, this framework provides a clearer understanding of how firm characteristics interact with classification methodologies.

While IFS excels at capturing growth and intangibility (as shown in Sections 4–6), this appendix demonstrates its complementary strength in tangible sectors. The visual data underlying IFS encodes both operational realities (e.g., supply chains, manufacturing) and

forward-looking signals (e.g., R&D labs, prototypes). This dual capability resolves the apparent contradiction: IFS clusters firms not just by what they produce today but also by their innovation pipelines and growth potential, dimensions often obscured in text-based classifications.

Table E1: Comparison of Classifications by Image and Text - Industries Explaining Differences in R^2

The table presents the comparison of average R^2 values for industry classifications based on image (Image Industries) and text (Hoberg & Phillips, 2016) (Text Industries). The comparison includes two panels: Image Industries with 45 classes (25) against Text 25 Industries in Panel A, and Image Industries with 73 classes (50) against Text 50 Industries in Panel B. To identify the industries explaining the differences in R^2 levels between image-based and text-based classifications, we calculate the regression of the R^2 difference in individual industries. The industries used are those from Moskowitz and Grinblatt (1999). The industries are analyzed over the 2014-2021 period. In the regression, we include dummy variables representing each industry, excluding the "Other" industry as the reference category. The table includes rows representing the names of industries and columns for the names of the ratios. Additionally, the table includes columns for the number of industries with positive significant coefficients (# pos), the number of industries with negative significant coefficients (# neg), and the difference between these two columns (# pos – neg). We show only ratios where the difference in R^2 between image and text is significant. The table shows regression coefficients and the asterisks ***, **, and * that represent statistical significance at the 1%, 5%, and 10% levels, respectively. The analysis covers the period from 2014 to 2021. Table F1 in the Internet Appendix shows details of ratios calculation. *Note:* Variable abbreviations: R&D EXP to SAL = R&D Expense to Sales; SG&A to # EMPL = SG&A to Number of Employees. See Table F1 for complete definitions.

Industry	PRICE to BOOK	MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q	TOTAL ASSETS	NET SALES	DIV PAYOUT	SALES GROWTH	R&D EXP to SAL	R&D GROWTH	SG&A to # EMPL	SG&A GROWTH	EPS GROWTH	DEBT to ASSETS	RNOA	ASSETS to SALES	# pos	# neg	# pos-neg
PANEL A: Image Industries 45 (25) vs. Text 25																							
Chemical	0.06**	-0.00	-	-0.01	-0.00	0.01	0.03	-0.02	0.06*	0.10**	-	0.06*	-	0.07*	0.02	0.04	-0.03	-0.03	-	0.05*	6	0	6
Manufacturing	-0.01	-0.01	-	-0.03	0.01	0.05**	0.05*	-0.04	0.03	-0.01	-	0.09**	-	0.11**	-0.03	0.04	-0.01	-0.05	-	-0.04	4	0	4
Paper	0.08	0.00	-	0.11	0.05	0.19**	0.16*	0.06	0.08	-0.13	-	0.05	-	0.09	-0.02	0.02	0.07	0.21**	-	-0.09	3	0	3
Petroleum	0.11*	-0.03	-	0.00	0.09**	-0.00	0.11	0.18**	0.08	0.12	-	-0.05	-	-0.12	0.09	0.03	-0.02	0.02	-	-0.17**	3	1	2
Food	-0.00	-0.01	-	-0.04	0.02	0.09**	-0.02	0.08**	-0.00	0.04	-	0.07	-	-0.19	0.04	0.11**	0.01	-0.03	-	-0.11**	3	1	2
Mining	0.03	-0.04**	-	-0.04	-0.03	-0.01	0.14***	-0.01	0.05	0.15**	-	0.00	-	0.17	-0.01	-0.01	-0.02	0.10**	-	-0.03	3	1	2
Trans equip	-0.03	-0.02	-	-0.03	0.06**	0.08**	-0.03	0.04	-0.01	0.07	-	0.02	-	-0.00	0.04	-0.03	0.00	0.04	-	-0.17***	2	1	1
Fab metals	-0.14**	-0.02	-	-0.02	0.03	0.10	0.14**	0.01	-0.00	0.31***	-	0.06	-	0.07	-0.08	0.02	0.02	-0.10	-	-0.00	2	1	1
Other transp	0.03	-0.02	-	-0.09**	-0.02	-0.01	0.02	0.05	-0.05	0.04	-	0.09*	-	-	0.17**	0.06	0.07	-0.04	-	-0.17***	2	2	0
Retail	-0.01	-0.01	-	-0.02	0.03**	0.02	0.02	0.02	0.01	0.07**	-	0.05**	-	0.03	-0.05**	0.01	0.01	-0.10***	-	-0.10***	3	3	0
Railroads	0.19*	-0.00	-	0.01	0.11	-0.06	0.11	0.11	-0.18	-0.23	-	0.14	-	-	0.02	0.00	-0.35**	-0.10	-	-0.20	1	1	0
Apparel	0.11	-0.03	-	-0.06	0.01	-0.09	0.04	0.08	-0.09	0.02	-	-0.05	-	0.00	-0.10	-0.03	-0.13	-0.14	-	-0.13	0	0	0
Machinery	0.01	0.00	-	-0.03	0.05**	0.03	0.02	-0.01	-0.01	0.00	-	0.04	-	-0.00	0.03	0.02	-0.01	-0.02	-	-0.05*	1	1	0
Financial	-0.26***	-0.11***	-	-0.01	-0.04***	-0.02	0.05**	0.07***	0.13***	0.12***	-	0.08***	-	0.27	-0.18***	0.07**	-0.08***	-0.02	-	-0.08***	6	6	0
Dept stores	-0.04	0.04	-	-0.06	0.07	0.05	-0.04	0.02	0.05	-0.04	-	0.08	-	-	-0.08	0.01	-0.04	-0.06	-	0.01	0	0	0
Utilities	-0.02	0.02	-	0.05	0.03	0.02	-0.03	-0.03	0.05	0.05	-	0.03	-	-	0.12	-0.01	-0.01	-0.08*	-	0.07*	1	1	0
Prim metals	-0.09	-0.07*	-	0.02	-0.01	0.02	-0.03	-0.07	-0.03	-0.14	-	0.09	-	-0.06	0.08	0.07	0.06	-0.06	-	-0.09	0	1	-1
Construction	0.09	-0.11**	-	-0.11	0.01	0.22*	-0.03	-0.13	0.16	0.18	-	-0.42**	-	0.22	-0.07	-0.15	0.09	0.03	-	-0.11	1	2	-1
Electr equip	-0.04**	0.00	-	-0.04*	0.04**	0.05**	-0.06**	-0.05**	0.00	-0.01	-	0.01	-	0.01	-0.02	0.04	-0.03	-0.08**	-	-0.01	2	5	-3
PANEL B: Image Industries 73 (50) vs. Text 50																							
Manufacturing	-0.03	-0.05***	-0.03	-	-	0.07**	-	-	-	0.09**	0.00	-	0.12**	0.09**	-0.01	0.04	-0.06**	-0.04	0.02	-0.02	4	2	2
Food	-0.06*	0.00	-0.00	-	-	0.11**	-	-	-	0.09*	0.05*	-	0.28*	-0.42**	0.04	0.16***	0.05	0.01	-0.01	-0.09**	5	3	2
Machinery	0.01	-0.02*	0.02	-	-	0.06*	-	-	-	-0.01	0.05**	-	0.08*	0.05	0.03	0.03	-0.04	0.01	0.02	0.01	3	1	2
Utilities	0.06*	0.03	-0.02	-	-	0.04	-	-	-	0.13**	0.03	-	-	-	-0.02	-0.04	0.07	-0.07	0.03	-0.05	2	0	2
Electr equip	-0.02	-0.01	0.04**	-	-	0.04	-	-	-	0.03	0.00	-	0.00	0.03	-0.02	0.04	0.00	-0.05**	0.04*	-0.01	2	1	1
Fab metals	0.05	-0.02	-0.00	-	-	0.04	-	-	-	0.17**	0.04	-	-0.01	0.07	-0.00	-0.08	-0.10	0.03	0.05	0.06	1	0	1
Dept stores	-0.04	0.02	-0.00	-	-	-0.03	-	-	-	0.10	0.12**	-	-	-	-0.12	0.24**	0.03	-0.20**	-0.01	-0.04	2	1	1
Chemical	0.02	-0.00	-0.04*	-	-	-0.00	-	-	-	0.06	-0.02	-	0.09**	0.14***	0.02	0.05	-0.07**	-0.07**	0.05*	0.06**	4	3	1
Construction	-0.05	-0.06	-0.14	-	-	0.02	-	-	-	0.09	-0.05	-	0.09	0.40	-0.07	-0.04	-0.01	0.10	-0.18	-0.09	0	0	0
Mining	0.04	-0.09***	-0.02	-	-	-0.06	-	-	-	0.19***	-0.00	-	0.10	0.05	-0.01	-0.04	-0.04	0.03	-0.04	-0.07	1	1	0
Other transp	0.02	-0.03*	0.05	-	-	0.01	-	-	-	0.13**	0.04	-	-	-	0.08	0.04	0.02	-0.03	0.13**	-0.11**	2	2	0
Paper	0.03	-0.03	0.03	-	-	0.12	-	-	-	0.12	-0.00	-	-0.10	0.08	0.09	-0.13	0.01	0.03	0.00	-0.01	0	0	0
Petroleum	0.19**	-0.07**	0.05	-	-	-0.03	-	-	-	0.16	0.08*	-	0.07	-0.01	-0.02	0.06	0.02	0.04	-0.11	-0.14*	2	2	0
Prim metals	-0.01	-0.11**	0.13	-	-	0.06	-	-	-	-0.04	-0.04	-	0.48*	0.19	0.10	0.06	-0.02	-0.09	-0.09	0.04	1	1	0
Trans equip	-0.03	-0.05**	-0.01	-	-	0.10**	-	-	-	0.08	-0.00	-	-0.09	0.03	-0.00	-0.05	0.03	0.00	-0.02	0.03	1	1	0
Railroads	0.08	-0.00	0.04	-	-	0.04	-	-	-	-0.16	-0.04	-	-	-	-0.04	-0.12	-0.35**	-0.04	0.04	-0.11	0	1	-1
Retail	-0.01	-0.02*	-0.07***	-	-	-0.00	-	-	-	0.09***	0.03**	-	-0.00	0.08	-0.09***	0.02	0.02	-0.13***	0.03	-0.08***	2	5	-3
Apparel	0.01	-0.07*	-0.13*	-	-	0.02	-	-	-	-0.08	0.01	-	-0.97**	0.14	-0.10	-0.05	0.00	-0.03	-0.15*	0.10	0	4	-4
Financial	-0.15***	-0.12***	-0.09***	-	-	-0.04*	-	-	-	0.19***	0.01	-	-0.16	-0.04	-0.17***	0.06**	-0.09***	-0.00	0.00	-0.08***	2	7	-5

Table E2: Comparison of Classifications by Image and Text - Characteristics Explaining Differences in R^2

The table presents the comparison of average R^2 values for industry classifications based on image (Image Industries) and text (Hoberg & Phillips, 2016) (Text Industries). The comparison includes two panels: Image Industries with 45 classes (25) against Text 25 Industries in Panel A, and Image Industries with 73 classes (50) against Text 50 Industries in Panel B. To identify the characteristics explaining the differences in R^2 levels between image-based and text-based classifications, we calculate the regression of the R^2 difference in individual characteristics and a constant. The characteristics used are those from Green, Hand, and Zhang (2017). The characteristics are averaged for each stock over the 2014-2021 period, excluding market-based characteristics due to their variability. The table includes rows representing the abbreviations of the characteristics and columns for the names of the ratios. Additionally, the table includes columns for the number of characteristics with positive significant coefficients (# pos), the number of characteristics with negative significant coefficients (# neg), and the difference between these two columns (# pos – neg). We show only ratios where the difference in R^2 between image and text is significant. We demonstrate characteristics where # pos – neg is not less than the absolute value of 3. The table shows regression coefficients and the asterisks ***, ** and * that represent statistical significance at the 1% and 5% and 10% levels, respectively. The analysis covers the period from 2014 to 2021. Table F1 in the Internet Appendix shows details of ratios calculation. *Note:* Variable abbreviations: R&D EXP to SAL = R&D Expense to Sales; SG&A to # EMPL = SG&A to Number of Employees. See Table F1 for complete definitions.

Industry	PRICE to BOOK	MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q	TOTAL ASSETS	NET SALES	DIV PAYOUT	SALES GROWTH	R&D EXP to SAL	R&D GROWTH	SG&A to # EMPL	SG&A GROWTH	EPS GROWTH	DEBT to ASSETS	RNOA	ASSETS to SALES	# pos	# neg	# pos-neg
PANEL A: IFS 45 (25) vs. Text 25																							
Chemical	0.06**	-0.00	-	-0.01	-0.00	0.01	0.03	-0.02	0.06*	0.10**	-	0.06*	-	0.07*	0.02	0.04	-0.03	-0.03	-	0.05*	6	0	6
Manufacturing	-0.01	-0.01	-	-0.03	0.01	0.05**	0.05*	-0.04	0.03	-0.01	-	0.09**	-	0.11**	-0.03	0.04	-0.01	-0.05	-	-0.04	4	0	4
Paper	0.08	0.00	-	0.11	0.05	0.19**	0.16*	0.06	0.08	-0.13	-	0.05	-	0.09	-0.02	0.02	0.07	0.21**	-	-0.09	3	0	3
Petroleum	0.11*	-0.03	-	0.00	0.09**	-0.00	0.11	0.18**	0.08	0.12	-	-0.05	-	-0.12	0.09	0.03	-0.02	0.02	-	-0.17**	3	1	2
Food	-0.00	-0.01	-	-0.04	0.02	0.09**	-0.02	0.08**	-0.00	0.04	-	0.07	-	-0.19	0.04	0.11**	0.01	-0.03	-	-0.11**	3	1	2
Mining	0.03	-0.04**	-	-0.04	-0.03	-0.01	0.14***	-0.01	0.05	0.15**	-	0.00	-	0.17	-0.01	-0.01	-0.02	0.10**	-	-0.03	3	1	2
Trans equip	-0.03	-0.02	-	-0.03	0.06**	0.08**	-0.03	0.04	-0.01	0.07	-	0.02	-	-0.00	0.04	-0.03	0.00	0.04	-	-0.17***	2	1	1
Fab metals	-0.14**	-0.02	-	-0.02	0.03	0.10	0.14**	0.01	-0.00	0.31***	-	0.06	-	0.07	-0.08	0.02	0.02	-0.10	-	-0.00	2	1	1
Other transp	0.03	-0.02	-	-0.09**	-0.02	-0.01	0.02	0.05	-0.05	0.04	-	0.09*	-	-	0.17**	0.06	0.07	-0.04	-	-0.17***	2	2	0
Retail	-0.01	-0.01	-	-0.02	0.03**	0.02	0.02	0.02	0.01	0.07**	-	0.05**	-	0.03	-0.05**	0.01	0.01	-0.10***	-	-0.10***	3	3	0
Railroads	0.19*	-0.00	-	0.01	0.11	-0.06	0.11	0.11	-0.18	-0.23	-	0.14	-	-	0.02	0.00	-0.35**	-0.10	-	-0.20	1	1	0
Apparel	0.11	-0.03	-	-0.06	0.01	-0.09	0.04	0.08	-0.09	0.02	-	-0.05	-	0.00	-0.10	-0.03	-0.13	-0.14	-	-0.13	0	0	0
Machinery	0.01	0.00	-	-0.03	0.05**	0.03	0.02	-0.01	-0.01	0.00	-	0.04	-	-0.00	0.03	0.02	-0.01	-0.02	-	-0.05*	1	1	0
Financial	-0.26***	-0.11***	-	-0.01	-0.04***	-0.02	0.05**	0.07***	0.13***	0.12***	-	0.08***	-	0.27	-0.18***	0.07**	-0.08***	-0.02	-	-0.08***	6	6	0
Dept stores	-0.04	0.04	-	-0.06	0.07	0.05	-0.04	0.02	0.05	-0.04	-	0.08	-	-	-0.08	0.01	-0.04	-0.06	-	0.01	0	0	0
Utilities	-0.02	0.02	-	0.05	0.03	0.02	-0.03	-0.03	0.05	0.05	-	0.03	-	-	0.12	-0.01	-0.01	-0.08*	-	0.07*	1	1	0
Prim metals	-0.09	-0.07*	-	0.02	-0.01	0.02	-0.03	-0.07	-0.03	-0.14	-	0.09	-	-0.06	0.08	0.07	0.06	-0.06	-	-0.09	0	1	-1
Construction	0.09	-0.11**	-	-0.11	0.01	0.22*	-0.03	-0.13	0.16	0.18	-	-0.42**	-	0.22	-0.07	-0.15	0.09	0.03	-	-0.11	1	2	-1
Electr equip	-0.04**	0.00	-	-0.04*	0.04**	0.05**	-0.06**	-0.05**	0.00	-0.01	-	0.01	-	0.01	-0.02	0.04	-0.03	-0.08**	-	-0.01	2	5	-3
PANEL B: IFS 73 (50) vs. Text 50																							
Manufacturing	-0.03	-0.05***	-0.03	-	-	0.07**	-	-	-	0.09**	0.00	-	0.12**	0.09**	-0.01	0.04	-0.06**	-0.04	0.02	-0.02	4	2	2
Food	-0.06*	0.00	-0.00	-	-	0.11**	-	-	-	0.09*	0.05*	-	0.28*	-0.42**	0.04	0.16***	0.05	0.01	-0.01	-0.09**	5	3	2
Machinery	0.01	-0.02*	0.02	-	-	0.06*	-	-	-	-0.01	0.05**	-	0.08*	0.05	0.03	0.03	-0.04	0.01	0.02	0.01	3	1	2
Utilities	0.06*	0.03	-0.02	-	-	0.04	-	-	-	0.13**	0.03	-	-	-	-0.02	-0.04	0.07	-0.07	0.03	-0.05	2	0	2
Electr equip	-0.02	-0.01	0.04**	-	-	0.04	-	-	-	0.03	0.00	-	0.00	0.03	-0.02	0.04	0.00	-0.05**	0.04*	-0.01	2	1	1
Fab metals	0.05	-0.02	-0.00	-	-	0.04	-	-	-	0.17**	0.04	-	-0.01	0.07	-0.00	-0.08	-0.10	0.03	0.05	0.06	1	0	1
Dept stores	-0.04	0.02	-0.00	-	-	-0.03	-	-	-	0.10	0.12**	-	-	-	-0.12	0.24**	0.03	-0.20**	-0.01	-0.04	2	1	1
Chemical	0.02	-0.00	-0.04*	-	-	-0.00	-	-	-	0.06	-0.02	-	0.09**	0.14***	0.02	0.05	-0.07**	-0.07**	0.05*	0.06**	4	3	1
Construction	-0.05	-0.06	-0.14	-	-	0.02	-	-	-	0.09	-0.05	-	0.09	0.40	-0.07	-0.04	-0.01	0.10	-0.18	-0.09	0	0	0
Mining	0.04	-0.09***	-0.02	-	-	-0.06	-	-	-	0.19***	-0.00	-	0.10	0.05	-0.01	-0.04	-0.04	0.03	-0.04	-0.07	1	1	0
Other transp	0.02	-0.03*	0.05	-	-	0.01	-	-	-	0.13**	0.04	-	-	-	0.08	0.04	0.02	-0.03	0.13**	-0.11**	2	2	0
Paper	0.03	-0.03	0.03	-	-	0.12	-	-	-	0.12	-0.00	-	-0.10	0.08	0.09	-0.13	0.01	0.03	0.00	-0.01	0	0	0
Petroleum	0.19**	-0.07**	0.05	-	-	-0.03	-	-	-	0.16	0.08*	-	0.07	-0.01	-0.02	0.06	0.02	0.04	-0.11	-0.14*	2	2	0
Prim metals	-0.01	-0.11**	0.13	-	-	0.06	-	-	-	-0.04	-0.04	-	0.48*	0.19	0.10	0.06	-0.02	-0.09	-0.09	0.04	1	1	0
Trans equip	-0.03	-0.05**	-0.01	-	-	0.10**	-	-	-	0.08	-0.00	-	-0.09	0.03	-0.00	-0.05	0.03	0.00	-0.02	0.03	1	1	0
Railroads	0.08	-0.00	0.04	-	-	0.04	-	-	-	-0.16	-0.04	-	-	-	-0.04	-0.12	-0.35**	-0.04	0.04	-0.11	0	1	-1
Retail	-0.01	-0.02*	-0.07***	-	-	-0.00	-	-	-	0.09***	0.03**	-	-0.00	0.08	-0.09***	0.02	0.02	-0.13***	0.03	-0.08***	2	5	-3
Apparel	0.01	-0.07*	-0.13*	-	-	0.02	-	-	-	-0.08	0.01	-	-0.97**	0.14	-0.10	-0.05	0.00	-0.03	-0.15*	0.10	0	4	-4
Financial	-0.15***	-0.12***	-0.09***	-	-	-0.04*	-	-	-	0.19***	0.01	-	-0.16	-0.04	-0.17***	0.06**	-0.09***	-0.00	0.00	-0.08***	2	7	-5

Appendix F: Additional Tables and Figures

Table F1: 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.

Table F1: (continued)

Variable Name	Full Ratio Name	Updates	Definition
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.
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)).

Table F1: (continued)

Variable Name	Full Ratio Name	Updates	Definition
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).
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.
PANEL D: Other variables we use in Table E2			
Change in inventory		quarterly	Change in inventory scaled by average total assets
Absolute accruals		quarterly	Annual income before extraordinary items (ib) minus operating cash flows (oancf) divided by average total assets (at); if oancf is missing then set to change in act - change in che - change in lct + change in dlc + change in txp-dp.
Current ratios		quarterly	Current assets divided by current liabilities

Table F1: (continued)

Variable Name	Full Ratio Name	Updates	Definition
Real estate holdings		quarterly	Buildings and capitalized leases divided by gross PP&E
CAPEX and inventories		quarterly	Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt) all scaled by lagged total assets (at).
Growth in CAPEX		quarterly	Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt) all scaled by lagged total assets (at).
Cash productivity		quarterly	Fiscal year end market capitalization plus long term debt (dltt) minus total assets (at) divided by cash and equivalents (che)
Earnings volatility		quarterly	Standard deviaiton for 16 quarters of income before extraordinary items (ibq) divided by average total assets (atq).
Revenue surprise		quarterly	Sales from quarter t minus sales from quarter t-4 (saleq) divided by fiscalquarter-end market capitalization (cshoq $\times$ *prccq*).
Sales to price		quarterly	Annual revenue (sale) divided by fiscal-year-end market capitalization.
Operating profitability		quarterly	Revenue minus cost of goods sold - SG&A expense - interest expense divided by lagged common shareholders' equity.
Cash flow to price		quarterly	Operating cash flows divided by fiscal-year-end market capitalization.
Zero trading days		quarterly	Turnover weighted number of zero trading days for most recent 1 month.
Industry-adj sales to price		quarterly	Industrial-adjusted annual revenue (sale) divided by fiscal-year-end market capitalization.
Growth in CSE		quarterly	Annual percent change in book value of equity (ceq).
Depreciation/PP&E		quarterly	Depreciation divided by PP&E.
Quick ratio		quarterly	(current assets - inventory) / current liabilities.

Table F1: (continued)

Variable Name	Full Ratio Name	Updates	Definition
Cash holdings		quarterly	Cash and cash equivalents divided by average total assets.
Dispersion in forecasted EPS		quarterly	Standard deviation of analyst forecasts in month prior to fiscal period end date divided by the absolute value of the mean forecast; if meanest = 0, then scalar set to 1. Forecast data from I/B/E/S summary files.
Age		quarterly	Number of years since first Compustat coverage.
Earnings to price		quarterly	Annual income before extraordinary items (ib) divided by end of fiscal year market cap.
Growth to LTNOA		quarterly	Growth in long term net operating assets.
Sales to receivables		quarterly	Annual sales divided by accounts receivable.
% change in CAPEX		quarterly	Percent change in capital expenditures from year t-2 to year t.
Leverage		quarterly	Total liabilities (lt) divided by fiscal year end market capitalization.

Table F2: 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')

Table F2: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	TOP5 SIC codes
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',)
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')

Table F2: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	TOP5 SIC codes
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')
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',)

Table F2: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	TOP5 SIC codes
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 F3: 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. 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 F1 in the Internet 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
Text 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
Text 50 Industries	23.3	18.9	22.4	2.3	28.9	23.5	11.8	26.2	33.1	22.3

Table F4: 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 (Text 25). 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 (Text 50). 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 F1 in the Internet 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
Text 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
Text 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 F5: 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. 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 (Text 25). 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 (Text 50). 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 F1 in the Internet 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
Text 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
Text 50 Industries	22.6	20.2	27.4	3.4	29.7	25.1	11.4	27.8	29.8	20.0

Table F6: 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 (Text 25). 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 (Text 50). 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 F1 in the Internet 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
Text 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
Text 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 F7: Firm Characteristics of IFS-Classified and Unclassified Firms

This table compares firm characteristics between firms classified by IFS and those without classification. Panel A reports measures of firm size and liquidity, while Panel B presents valuation and performance characteristics. Market capitalization, total assets, and net sales are measured in millions of U.S. dollars, and monthly trading volume is measured in billions of U.S. dollars. Return-based variables and profitability measures are expressed in percentage terms. The difference column reports t-statistics from tests of equality of means. Detailed definitions and construction of all variables are provided in Table F1. The sample consists of firms listed on the NYSE, NASDAQ, and AMEX exchanges and covers the 2014–2021 period.

	IFS-Classified			Unclassified			
Variable	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.	Diff. t-stat
PANEL A: Size and Liquidity							
MARKET CAP	8,493.7	1,482.8	16,391.8	3,801.3	583.5	10,182.7	84.67
MONTHLY RET	0.35	0.55	15.41	-0.13	0.18	20.10	7.48
MARKET to BOOK	2.0	1.5	1.7	2.2	1.5	2.0	-26.69
BETA	1.2	1.1	0.8	1.2	1.1	0.9	-4.44
Bid–Ask Spread	0.005	0.001	0.011	0.007	0.001	0.013	-44.90
PANEL B: Valuation and Performance							
MONTHLY VOLUME ($bn)	18,983.0	2,705.0	91,351.7	8,924.6	785.4	58,448.6	32.61
TOTAL ASSETS	13,089.0	1,806.7	36,236.1	6,479.3	864.6	22,933.7	53.10
NET SALES	6,098.2	1,188.5	13,813.3	2,482.0	358.4	7,979.0	76.86
ROE (%)	3.8	9.1	73.6	-7.1	6.3	88.2	35.51

Table F8: Industry Composition and IFS Coverage

This table reports the industry composition and sample coverage of firms classified by IFS and those without classification. Panel A shows the distribution of firms across industries, while Panel B reports the number of firm-months, the number of unique firms, and the share of total market capitalization. Industries are defined following the classification of Moskowitz and Grinblatt (1999), which we extend to include three additional categories: Real estate (SIC codes 65–71), Services (SIC codes 72–87), and Nonclassifiable Establishments (SIC code 9999). The sample consists of firms listed on the NYSE, NASDAQ, and AMEX exchanges and covers the 2014–2021 period.

	IFS-Classified (%)	Unclassified (%)
PANEL A: Industry composition (% of firms)		
Utilities	3.1	1.9
Food	3.0	0.6
Retail	10.5	4.0
Electr equip	8.2	3.2
Chemical	5.8	6.5
Financial	12.6	12.2
Machinery	5.1	2.3
Fab metals	1.3	0.5
Other transp	2.3	0.7
Manufacturing	5.7	3.8
Trans equip	2.4	0.7
Apparel	0.8	0.5
Construction	0.3	0.4
Mining	4.2	4.8
Petroleum	0.6	0.1
Railroads	0.3	0.0
Prim metals	0.5	0.9
Paper	0.4	0.5
Dept stores	1.0	0.0
Other	6.2	5.7
Real estate	3.2	2.7
Services	10.5	16.2
Nonclass Establish	12.3	31.9
PANEL B: Sample coverage		
Number of firm-months	103,565	234,697
Number of unique firms	2,071	4,268
Share of total market capitalization (%)	50.7	49.3

Table F9: 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 (Text 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	Text 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