

Speculative Recommendation: Reframing AI for Interpretive Practice in the Digital Humanities

River Rain <riverjoyrain_at_gmail_dot_com>, Amsterdam University College  <https://orcid.org/0009-0005-4905-782X>
Houda Lamqaddam <h_dot_lamqaddam_at_uva_dot_nl>, Institute of Logic, Language and Computation, Department of Media Studies, University of Amsterdam  <https://orcid.org/0000-0001-6738-1934>

DOI: pending

Abstract

This project investigates the potential of recommender systems as a generative tool in humanities research using the case of an aesthetic-driven recommender system focused on animated films. Leveraging still frames from trailers and a fine-tuned convolutional neural network (VGG16), we extracted aesthetic features to develop a recommender system that uses visual characteristics as the sole input. The system demonstrates high precision, identifying both expected and surprising latent connections between films on the basis of visual similarities, demonstrating how recommender systems can support the investigation of high level concepts such as influence. This approach has broader implications within DH, showing how recommender systems can facilitate exploratory, interpretive research, how the stochasticity of AI models is well suited to support the interpretative aspect of humanistic scholarship, and exploring the potential of interdisciplinary cross-pollination in DH research.

1. Introduction

The influence of recommender algorithms on current media and information access is difficult to overstate. Increasingly, these systems are embedded into streaming platforms, social media feeds, online shopping interfaces, and scholarly search engines, profoundly shaping how users encounter and engage with the world. Their capacity to mediate cultural consumption has prompted significant interest in various fields, yet the Digital Humanities (DH) have only recently begun to explore the epistemological and methodological potential of recommender systems. Scholars such as [Arnold, Leonard, and Tilton 2017] have suggested that recommender algorithms, beyond their commercial origins, can serve as tools for reimagining cultural discovery, offering new ways to navigate and interpret digital collections. Despite this promise, recommender systems remain underutilized in DH. Several barriers have limited their uptake:

- the scarcity of large, high-quality user data in many humanities domains [Arnold, Leonard, and Tilton 2017];
- small and fragmented datasets [Atzenhofer-Baumgartner et al. 2024], [Arnold, Leonard, and Tilton 2017];
- insufficient domain-specific vocabularies that would enable meaningful content filtering [Ernst 2022]; and
- the diversity of stakeholders and research goals typical of DH projects, often paired with limited technical resources [Atzenhofer-Baumgartner et al. 2024], [Pavlidis 2022].

Nonetheless, the broader turn toward computational methods in the humanities — spanning network analysis, machine vision, and distant reading — suggests that recommendation should be reconsidered not merely as an engine for efficiency or personalization, but as a generative, interpretive process.

This paper proposes aesthetic-driven recommendation as a new avenue for DH research, using animated film as a case study. The central hypothesis is that visual aesthetics, rather than metadata, can serve as meaningful features to discover connections between cultural artifacts. In this context, recommendation becomes a form of visual analysis: an

1

2

3

opportunity to explore relationships between works that share stylistic similarities, even if they diverge in theme, origin, or popularity. Aesthetic-based systems may also offer a counterpoint to mainstream algorithmic culture by supporting serendipitous discovery and scholarly exploration.

We selected animation as the domain for this investigation for two key reasons. First, animation offers a uniquely rich terrain for visual analysis. Unlike live-action film, animation is not constrained by physical reality. It is a medium of total visual authorship, where every frame is intentionally designed. As a result, the visual identity of an animated work is often more stylized, consistent, and semantically loaded than its live-action counterpart. The look and feel of an animation frequently communicate the genre, tone, cultural context, or artistic lineage of the work. As Suzanne Buchan writes, animated films construct “realms of cinematic experience... accessible to the spectator only through the techniques available in animation filmmaking” [Buchan 2007]. These techniques allow animators to render abstract emotion, political allegory, or mythological space with precision and flair, making aesthetics not merely ornamental but narratively and affectively central. Stylistic differences in animation can also be used to reflect distinct production contexts: the glossy minimalism of American 3D family films, the expressive 2D forms of Japanese anime, the textured atmospheres of Eastern European experimental studios. These aesthetics carry historical, cultural, and industrial meaning. Thus, by building a system that recommends animated films based on aesthetic similarity alone, we are not simply developing a new discovery tool, we are also surfacing a latent taxonomy of style, form, and expression in contemporary animation. In doing so, we invite further scholarly inquiry into how aesthetics function as a navigational axis in digital media collections. The second reason for focusing on animation is pragmatic: the availability of structured, accessible data. Datasets such as MovieLens and The Movie Database (TMDb) offer metadata, user ratings, and content links for thousands of films, including animated ones. These datasets make it feasible to isolate and analyze animated titles at scale. Moreover, animated trailers — unlike full films — are often legally and freely accessible via platforms like YouTube, allowing us to retrieve a meaningful sample of each film’s visual identity without access to proprietary or copyrighted materials. The modular nature of trailers also ensures that the selected frames are diverse and often crafted to showcase the film’s most iconic or representative imagery.

In this study, we present a recommender system that uses visual aesthetics as the sole input for suggesting similar animated films. The system is built in five stages:

1. We built a corpus of animated films from the MovieLens dataset.
2. We extracted still frames for each film.
3. We fine-tuned a convolutional neural network (VGG16) using a visual art dataset to better recognize stylistic qualities.
4. We converted each film’s visual content into a 512-dimensional feature vector.
5. We use an approximate nearest neighbor algorithm (ANNOY) to identify and recommend films with similar aesthetic vectors.

We tested the system against user ratings in MovieLens to evaluate its precision and recall as a limited external benchmark, and we also examined the kinds of connections it generates, expected and unexpected, between films. What emerges is a prototype that shows promise in quantitative terms, but more importantly, illustrates the potential of visual similarity as a basis for interpretative analysis. The system reliably identifies sequels and studio siblings, but also proposes surprising stylistic matches across time, region, and production style. These connections serve to identify similarities, trace influences over time, and overall provide a novel lens through which to explore cultural material. More broadly, this project contributes to DH by positioning recommendation as an interpretive rather than predictive method. It demonstrates how tools from machine learning and computer vision can be repurposed for humanities inquiry: not to automate taste categorisation, but to illuminate new pathways through cultural material. By attending to aesthetics as a vector of discovery, we propose a shift away from recommender systems designed to maximize engagement, and toward systems that support exploration, serendipity, and critical engagement. In this way, our system functions both as a proof of concept and as a call to expand the scope of what recommendation can mean in the digital humanities.

2. Related Works

This section discusses relevant literature from the fields of animation, aesthetic analysis, and recommendation system design and how they unite to contextualize this research and explain relevant techniques.

7

2.1. Recommendation Systems and the Humanities

Broadly speaking, there are two main approaches to constructing recommender systems: content-based filtering and collaborative filtering. Content-based filtering focuses on the attributes of the items themselves. It recommends items similar to those a user has already shown interest in, using features such as genre, description, or keywords. While effective for personalizing recommendations for new users, it can suffer from limited variety because it does not explore beyond the user's existing preferences. Collaborative filtering, on the other hand, relies on the behavior and preferences of other users to make recommendations. It identifies patterns among user interactions, such as ratings, purchases, or clicks, and uses them to predict what a user might like based on the preferences of similar users (user-based) or items with similar interaction patterns (item-based). It works well for diverse datasets but struggles with the “cold start” problem — the challenge of making accurate suggestions when there is insufficient data about new users, new items, or both. Many recommendation systems combine both content-based and collaborative filtering methods in a hybrid system to minimize the drawbacks of both.

8

Currently, a key focus of the field involves refining the insights into a user's taste from their watch history [Ko et al. 2022] to develop content-based filtering recommendation systems that effectively handle the cold start problem without limiting the scope of user preferences. New recommendation algorithms are even moving toward analyzing user personality and behavior to create cognition-driven systems [Beheshti et al. 2020]. Recommender algorithms that analyze aesthetic quality based on individual history and preference have been investigated thoroughly in the fashion industry [Kalantidis, Kennedy, and Li 2013] [Yu et al. 2018]. For example, [Kalantidis, Kennedy, and Li 2013] pioneered methods of recommending clothing by measuring aesthetic similarities between clothing articles. Similar strategies have been used in image recommendation algorithms. [Lv et al. 2018] developed a method to unobtrusively create aesthetic profiles for individuals and use those to measure the aesthetic quality of photographs, conveniently dodging any problems that may arise when there is an absence of user data. [Chen et al. 2019] take a collaborative filtering approach and explore how aesthetic recommendation can be applied to live-action movie recommender algorithms based on matrix factorization. They conclude that the integration of aesthetic analysis improves their collaborative recommendation system. Though the link between aesthetics and user preferences has been investigated in the context of recommender systems, it has never been applied to animation, leaving a notable gap in the literature. Investigating an aesthetics-based recommendation system for animation promises to yield interesting insights on the general applicability of aesthetic features to recommendation systems as well as inform methods of recommending animated movies. Finally, the adoption of AI-based methods within DH raises additional epistemological and ethical questions. Recommender systems are often evaluated through predictive quantitative benchmarks, which fail as indicators of usefulness in the humanities [Washington et al. 2022], where the value of a system also lies in the kinds of connections it surfaces and the assumptions it builds. In this project, we therefore approach recommendation as a methodological device for exploratory analysis rather than a predictive accuracy challenge.

9

2.3 Aesthetic Evaluation

The vast majority of machine learning (ML) aesthetic assessment and feature extraction on visual material is focused on photographs. Various video quality assessment methods have been developed to evaluate computer-generated animation from an objective video quality perspective [Darici 2023], but there is little research on methods to evaluate the subjective aesthetic qualities of animation. This leaves a significant gap in the existing literature for exploring methodologies that assess the subjective aesthetic qualities of animation using ML. This gap is both methodological and interpretive, since animation foregrounds style, visual rhetoric, and cross-cultural aesthetic influence in ways that are especially relevant to DH inquiry.

10

Though the specific question of animation has not been addressed by the literature, it offers a rich library of convolutional neural networks (CNNs) and various methods of aesthetic feature extraction to draw upon. The field of digital aesthetic evaluation began by classifying images based on art and photography principles such as color, shape,

11

and composition. However, this initial approach struggled to capture the complexity and subtlety of aesthetics and was significantly enhanced by subsequent deep learning image classification CNNs. Deep CNNs were pioneered by [Krizhevsky, Sutskever, and Hinton 2012] with the breakthrough model Alexnet. Alexnet consists of five convolutional layers followed by max-pooling layers, three fully connected layers, and a softmax output, designed to efficiently process large-scale image data. More CNNs followed, optimizing different aspects of performance for different tasks. The VGG models used a deep, uniform architecture with smaller 3x3 convolutional filters to improve learning efficiency and feature extraction while making the network design simpler and more standardized [Simonyan and Zisserman 2014]. ResNet models added the concept of “skip connections,” which let information jump over layers, making it easier to train very deep networks without the problem of losing important details during training (a challenge called the vanishing gradient problem) (he, 2016). EfficientNet models employed a compound scaling method, which balances network depth, width, and resolution to optimize accuracy and efficiency across different model sizes [Tan and Le 2019]. These CNNs have been integrated into many different aesthetic feature extraction approaches. Jin and his colleagues trained VGG16 on the widely used dataset AVA (murray, 2012) to assess both the aesthetic quality of a photograph as well as the difficulty of predicting it [Jin, Segovia, and Süssstrunk 2016]. The OWACNN method [Lu, Chang, and Chen 2016] uses four separate ResNet CNNs to analyze different aspects of the input image: scene, emotion, object, and semantics. These are combined using the OWA pooling layer to get a holistic aesthetic quality assessment that captures the relationship between objects and aesthetics. The NIMA model [Talebi and Milanfar 2018] evaluates image aesthetics by predicting a distribution of aesthetic scores rather than a single score, reflecting subjective human judgments.

This project fine-tunes several different CNNs on an aesthetic quality dataset to create models that can predict the likelihood of different aesthetic ratings and create a distribution of aesthetic opinions. It finds VGG16 to be the most effective CNN for the task. Other projects build entirely new CNN architectures to reflect specific methods of human vision and aesthetic judgment. For example, [Zhang et al. 2019] develops CNNs that replicate the dual peripheral-foveal perception of the human eye to process an object’s aesthetic qualities similarly to how a person would.

12

3. Methodology

Corpus Creation

We selected the MovieLens 32M dataset as the source for our corpus. MovieLens is a series of film-oriented datasets collected since 1997 and is widely used in recommender system research. The 32M version includes 32 million ratings and two million tag applications applied to 87,585 movies by 200,948 users, collected in October 2023. It contains both animated and live-action films, along with genre labels, TMDB IDs, and individual user ratings. While the dataset provides rich metadata, our primary use focuses on the films themselves, which we augment with visual content to train our recommender system. User reviews are reserved for later comparison with the system output. This subsection outlines the corpus creation process, beginning with the MovieLens dataset.

13

3.1.1. Filtering for animated films.

To focus specifically on animated films, we filtered the dataset to include only those with the genre label “Animation.” This reduced the corpus to 4,617 animated movies.

14

3.1.2. Video link retrieval via TMDB API.

To perform aesthetic analysis, we needed to collect video content that effectively represents the visual characteristics of each film. Using the TMDB IDs for each movie, we accessed the TMDB API to retrieve any YouTube links associated with trailers, teasers, or clips. Short clips like trailers and teasers are particularly useful for our aesthetic analysis because they are designed to showcase a film’s visual style, whereas full-length films would provide a much broader range of content that is less focused on specific aesthetic features. Additionally, processing full-length films would have been computationally intensive, making it more practical to work with shorter, curated clips that highlight key visual elements. Not all movies had valid video links, reducing the dataset size to 2,341 films.

15

3.1.3. Frame extraction and dataset pruning.

After collecting the video links, we used a custom program built with the `yt_dlp` and `cv2` libraries to download each video and extract frames. One frame was captured every ten seconds, with the first and last ten seconds excluded to avoid title cards and credits. This method ensured that frames representing the core visual content of each video were selected. The extracted frames were resized to 224×224 pixels with padding, preserving the aspect ratio to avoid distortion. Although this process produced high-quality stills, some undesirable frames (such as text-heavy, black, or transition frames) were occasionally included. To address the large volume of extracted frames — some films had more than 100 frames — a program shuffled and randomly selected up to 20 frames per film. This random sampling ensured visual diversity while maintaining a manageable dataset size.

16

3.1.4. Dataset assembly & title standardization.

The final set of curated frames was merged with the corresponding metadata from the MovieLens dataset, producing our Augmented Animated Film Corpus. This dataset combines visual content (the frames) and textual metadata (such as movie titles and TMDB IDs) for each animated film. This final dataset serves as the primary input for training and evaluating the recommender system. To improve title matching in the recommender system, we standardized the movie titles. Titles were converted to lowercase, and any release years, articles, special characters, and non-English names were removed. For example, *Princess Mononoke (Mononoke-hime) (1997)* was simplified to *princess mononoke*. This normalization ensures that user inputs, regardless of formatting differences (such as capitalization or the inclusion of years), will be matched accurately to the movie titles in the dataset. These simplified titles were added to a new column in the Movie Frames Dataset, making the recommender system more efficient in processing user queries.

17

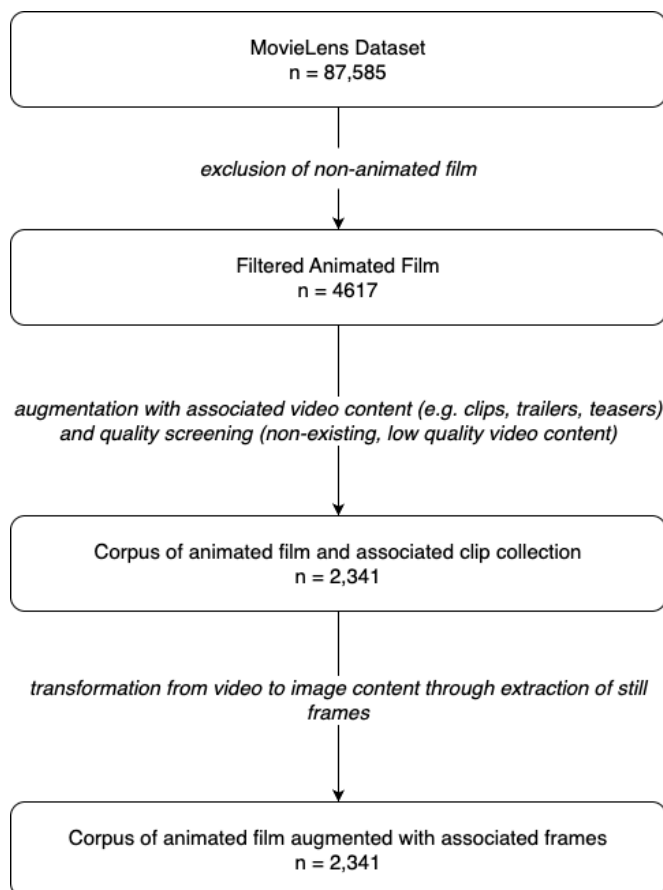


Figure 1. Corpus creation steps and count of film in our final dataset

4. Model & Processing

4.1. VGG16 as a base model.

From the variety of existing aesthetic analysis CNN models, we found VGG16 [Simonyan and Zisserman 2014] to be the most appropriate for our usage. VGG16 captures hierarchical features well, from edges in the initial layers to complex textures and patterns in deeper layers, and it is excellent at extracting spatial and visual features. The sequential design of VGG16 (stacked convolutional layers) is straightforward, making it easy to implement and modify for custom tasks (Figure 3). There are pre-trained versions available to implement, and importantly, they can be fine-tuned on other datasets. VGG16 in particular is preferable to its larger sibling, VGG19, because of the decreased computational cost. VGG16 is trained on the ImageNet dataset (deng, 2009), a large-scale dataset composed entirely of real-world photographs. As a result, when the model was applied to the animated movie dataset, it produced a number of incorrect classifications. For example, frames were frequently labeled as “television,” “website,” or “comic book,” particularly in the case of 2D animation, though some 3D frames were similarly misclassified (Figure 2). These errors reflect a mismatch between the domain of the training data (photographs) and the domain of our application (stylized or abstract animated visuals). To improve VGG16’s performance and ensure it would output more accurate results reasonable and distinct vectors, it needed to be refined on new training data. To address this gap, VGG16 must be fine-tuned not only to adapt to the 2D domain, but also to perform visual and stylistic classification specifically. This involves training the model on a curated dataset of images labeled with additional visual, semantic and stylistic attributes.

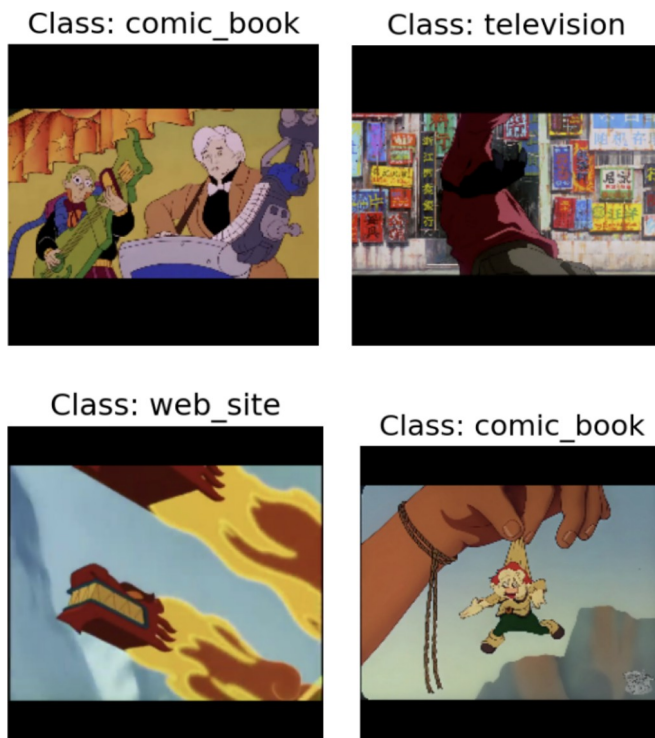


Figure 2. Unrefined VGG16 frame classifications, prior to fine-tuning.

4.2. Fine-tuning using ArtPedia.

There is currently no dataset of aesthetically classified animated frames to train or refine VGG16 with, nor was creating one in the scope of this project. To solve this problem, we reached into research in fine arts in which visual semantic alignment is more advanced than for animated film. We selected to use the Artpedia dataset as a substitute. The process diagram describing how this was accomplished is shown in Figure 3.

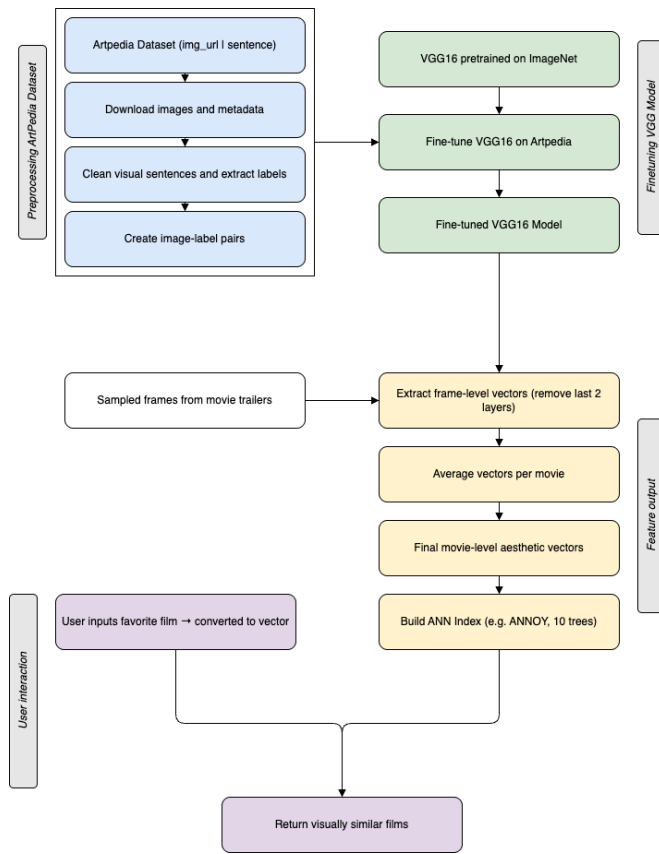


Figure 3. End-to-end pipeline of the aesthetic recommender system, from VGG16 fine-tuning to visually similar film retrieval.

The annotated Artpedia dataset [Stefanini et al. 2019] includes images of fine art, paired with visual sentences describing the content of the picture and contextual sentences containing information about the painting’s history (Figure 4). This dataset includes only 2D fine-art paintings and no photography, which is ideal because its key function is to give VGG16 further training on 2D images. Importantly, it also includes content that explicitly describes the scene in the painting. There are many large art datasets with a focus on classifying the period, artist or painting techniques, but we selected Artpedia due to its purely visual descriptions of the paintings.

To prepare this dataset for training, a program cleaned and stripped the visual sentences of their stopwords as well as common but unhelpful descriptors such as “symbol,” “foreground,” and “background” and words that contained non-standard characters or digits. It then identified the lexical category of each remaining word and removed verbs and proper nouns. Finally, it stemmed the remaining words to unify those with the same root and connotations but different conjugations. This left a set of descriptive labels corresponding to 2D fine art images. One problem remained, which is that Artpedia is a fairly small dataset, with only 2,930 total images. To artificially expand the dataset and enhance generalization, we applied data augmentation through the ImageDataGenerator during the fine-tuning process, introducing variations such as rotations, shifts, and flips. The entire process ultimately equipped VGG16 with some training on how to handle 2D images. Even if the fine art and animated frame dataset mismatch does not necessarily lead to correct classifications, it generates more robust vectors for the 2D input images, which is enough for them to appropriately differentiate themselves and be successfully grouped by the ANNOY approximate nearest neighbor algorithm.

4.3.Recommender system.

To build the recommender system, frame vectors were aggregated by movie and compared using a nearest neighbor approach. However, calculating exact distances between over 2,000 vectors is highly computationally intensive. To address this, we used ANNOY [Bernhardsson 2018], an approximate nearest neighbor algorithm that leverages random projection trees for fast and scalable similarity search. ANNOY is well-suited for high-dimensional data and allows

tuning the trade-off between speed and accuracy via the number of trees. A program constructed the ANNOY index from the VGG16-extracted movie vectors, forming the backbone of the recommender system. Given a list of animated movie titles from the Movie Frames dataset, the system returns the n most visually similar movies. A simple console interface allows users to input movie titles and receive recommendations, each accompanied by a sample image showing five randomly selected frames. This visual feedback helps users understand the aesthetic rationale behind each recommendation.

5. Results

5.1. Qualitative results and implications

The aesthetic recommender system excels at identifying both expected and surprising visual relationships between animated films, uncovering stylistic affinities that extend beyond surface similarities. Notably, the results suggest that aesthetic influence — an aspect invisible in metadata — can be traced computationally through visual features alone.

5.1.1. Capturing the visual impact of stylistic innovation.

One of the most compelling examples involves the film *Spider-Man: Into the Spider-Verse* (2018) (Figure 4). When given this input, the system recommends films such as *Puss in Boots: The Last Wish* (2022), *White Snake* (2019), and *Pil's Adventures* (2021). This finding is particularly interesting because *Into the Spider-Verse* marked a major stylistic shift in animation, blending 2D and 3D visuals to achieve a distinctive comic-book aesthetic. Its commercial and critical success signaled a broader industry move away from hyperrealistic 3D animation toward more stylized, artistically daring visual storytelling. Subsequent films, like *Puss in Boots: The Last Wish*, directly reflect this influence with their textured, hand-painted look and dynamic, frame-skipped animation style. The fact that the system recommends only post-2018 films when given *Spider-Verse* suggests it captures not just visual similarities in surface features like color or composition, but also larger stylistic movements across time. This demonstrates the system's potential for tracing aesthetic influence, a dimension of cultural production often invisible in conventional databases.

23

24

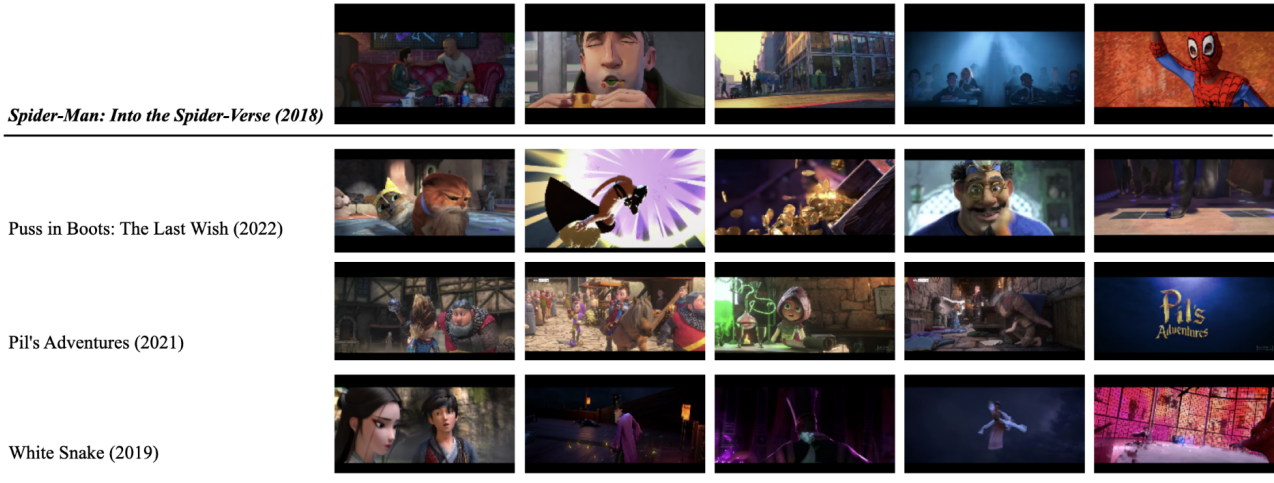


Figure 4. The system captures the visual impact of stylistic innovation, as can be seen through the recommendations based on the film *Spider-Man: Into the Spider-Verse*

5.1.2. Identifying cross-studio and cross-cultural stylistic influence.

The system also identifies more subtle visual similarities. When *Howl's Moving Castle* is used as input, the system suggests *Princess Mononoke* (1997), *From Up on Poppy Hill* (2011), and *Treasure Planet* (2002). The first two are intuitive results: Both are Studio Ghibli productions that share visual motifs such as richly layered backgrounds, delicate line work, and painterly natural environments. However, *Treasure Planet* is a more surprising inclusion. Created by Disney in the early 2000s, it blends hand-drawn and digital animation to evoke a similarly lush and fantastical aesthetic,

25

albeit in a different narrative register. This could suggest that the struggling post-renaissance Disney of 2002 was taking inspiration from Ghibli's successes at the time, such as their critically acclaimed hit *Spirited Away* (2001) which would win an Oscar that same year.



Figure 5. The painterly, hand-drawn style of *Howl's Moving Castle* (2004) is matched with other Ghibli works and with Disney's *Treasure Planet* (2002).

5.1.3. Detecting consistency in character design and visual palette.

Alongside these novel connections, the system reliably identifies more predictable stylistic continuities, such as sequels or films from the same studio (Figure 4). For example, inputting *Shrek* (2001) returns *Shrek 2* (2004), *Shrek the Third* (2007), and *Hotel Transylvania* (2012). The first two results reflect clear visual lineage within the same franchise, while the third is a more interesting suggestion: Although produced by a different studio (Sony Pictures Animation), *Hotel Transylvania* shares a similarly exaggerated character design, dark fantasy palette, and comedic tone. This demonstrates the system's capacity to identify aesthetic parallels that transcend franchise or studio boundaries.

Overall, these results highlight the significance of aesthetic similarity in interpreting cultural stylistic patterns across a corpus. By clustering films based on visual form rather than metadata, the system enables the tracing of artistic lineages, the detection of stylistic innovation, and highlighting uncovered links between animated films.



Figure 6. Querying *Shrek* (2001) returns its franchise sequels and *Hotel Transylvania*, a cross-studio match unified by exaggerated character modeling and a shared dark-fantasy register.

5.2. Quantitative comparison with viewer rating data

While the system is designed to encourage new interpretations and aesthetic connections, it is also useful to ask whether recommendations generated solely from visual similarity bear any relation to viewer preferences. To explore this question, we evaluate precision and recall using user review data from the MovieLens dataset. Because user ratings are not part of the system's input, any alignment between aesthetic similarity and viewer preference is itself a noteworthy finding. This comparison does not establish the interpretive truth of the system's outputs, but it does reveal whether visually derived recommendations also resonate with audiences in ways that are not reducible to metadata or collaborative filtering.

28

To test this, we selected the first 500,000 rows of the MovieLens ratings dataset, which includes user IDs, movie IDs, and ratings on a scale from 0 to 5. For each movie rated above a chosen threshold, we generated five recommendations ($n = 5$) and checked whether those recommended films had also been rated highly by the same user. This allowed us to calculate precision (how often a recommendation aligns with films a user also rated highly) and recall (how many user favorites the system retrieves). In fact, precision improves as the rating threshold increases, indicating the system is more effective at identifying films users genuinely liked. In fact, at higher thresholds, the system performs up to three times better than random recommendation. A t-test comparing the system's performance to a random baseline ($n = 5$, threshold = 5) yielded a t-statistic of 9.0278 and a low p-value ($p < 0.001$), confirming that the improvement is statistically significant. Recall, by contrast, remains low throughout, never exceeding 0.019. This is expected given the small number of recommendations ($n = 5$) compared to the average number of films rated per user (14.23). In this context, the precision result is more notable because it suggests that visual and aesthetic similarity may have a meaningful — if indirect — relation to viewer preference.

29

6. Discussion

6.1. Recommender systems as generative tools.

Rather than optimizing for personalization or commercial engagement, this project positions recommender systems as speculative instruments for research and interpretation. This distinction is critical for the humanities, where the goal is often to explore complexity rather than resolve it. In this study, the system surfaces aesthetic connections that invite interpretive analysis rather than asserting definitive classifications. The recommendations function not as taxonomies, but as prompts — encouraging associative discovery and uncovering latent connections among films. At the same time, this speculative use of recommendations sits in productive tension with the quantitative evaluative approach often used to assess computational systems. We initially considered quantitative comparison with viewer rating data as a limited external benchmark for assessing the validity of the matches. Through that process, however, we found this approach limited as a validation mechanism, since alignment with ratings does not establish significance nor accuracy. What it did reveal instead was an additional finding: that aesthetic similarity, even when modeled independently of user behavior, appears to bear a meaningful relation to audience preference. The quantitative results therefore point to an unexpected connection between visual form and reception. This exploratory mode aligns with core humanities methodologies, which privilege interpretation, ambiguity, and contextual nuance over prediction or ground truth. By reframing recommendation as a speculative and generative act, we highlight the potential of these systems to aid in the discovery of stylistic affinities, aesthetic influence, and cross-cultural resonance within digital cultural collections. As demonstrated here, such systems do not simply suggest what a viewer might enjoy; they provoke new questions about how visual language, artistic choice, and cultural production intersect — particularly within the context of animation.

30

6.2. Potential of AI models to identify high level concepts such as influence.

The stochasticity of AI models makes them promising tools for exploring interpretive concepts that resist fixed definitions, for instance visual similarity, aesthetic style, or cultural influence. These concepts have historically defied computational modeling due to their ambiguity, contextual variability, and resistance to objective categorization. Yet, the probabilistic nature of AI allows for precisely this type of interpretative ambiguity to be modeled rather than suppressed. In our study, by identifying visual patterns that recur across films from different studios or time periods, the system surfaces candidates for influence — not as ground truths, but as probabilistic suggestions, whose relevance must be established through contextual and historical interpretation. It is important to note, however, that to be useful, this

31

stochasticity must be made explicit. Results of probabilistic AI models must not be presented as final and accurate, but as machine estimations whose validity and relevance are left to the interpretation of the scholar.

6.3. Interdisciplinary cross-pollination in machine learning use for DH.

A central challenge in applying computational methods to the humanities is the domain-specific nature of most machine learning models and datasets. This pattern reflects the humanities' approach in favoring specificity, but also the existing silos between separate disciplines which function with different methods, under separate disciplinary roofs, and focus on different objects of study. In this study, we repurpose a dataset originally designed for fine art analysis to develop a recommender system for animated film. The successful outcome demonstrates that materials from one domain can inform and enhance computational tasks in another, when aligned by formal characteristics rather than disciplinary boundaries.

There are obvious limitations that come with this practice, for instance, the language used by scholars in a discipline, as well as the technical characteristics of the material, will undoubtedly vary among field of study. However, we argue that there is strong value in viewing interdisciplinary cross-pollination in DH not only as a hub model with digital methods feeding into and extracting from humanist frameworks, but as a set of disciplines that can interface with one another, and where digital progress in one area can be transformed and reused in another discipline in a productive and generative manner.

Within the context of Digital Humanities (DH), this project advances the case for using machine vision tools not merely to classify or retrieve objects, but to open new interpretive pathways. By recommending animated films based on visual similarity alone, the system surfaces relationships that are invisible through metadata, genre labels, or user profiling. It highlights how form — the style of animation, the texture of a background, the rhythm of color — can serve as a rich, analyzable dimension of cultural production.

7. Future Work

This project opens many avenues for future research along the lines of expanded media forms, providing multi-dimensional aspects of similarity matching, and studying the perception of such tools within the humanities.

7.1. Evaluating multi-dimensional aspects of similarity matching.

In this project, we have encoded similarity as visual distance between still frames. This definition of similarity can be expanded to incorporate dynamic visual features such as animation movement, shot composition, scene transitions, or more content-based features related to character or object characteristics. Following recent work conceptualizing similarity as nested and multi-dimensional [Lamqaddam et al. 2025] [Lamqaddam et al. 2026], incorporating a variety of similarity characteristics would provide a more holistic, personalizable view of associating films, thereby potentially identifying more interesting links among the corpus.

7.2. Application to expanded media forms.

This project looked at animated film in part due to the wide visual variability within animation. In future work, we would like to expand this research to other media such as graphic novels or video games which have similar visual characteristic, but distinct narrative structure and cultural contexts. Applying the recommender system to a cross-media corpus which includes these varied media forms would potentially allow us to trace lineage and influence across different types of visual production and test the portability of the aesthetics through time and media.

7.3. Expanding towards non-visual aesthetics.

While this project defines similarity primarily as visual characteristic, film analysis is also shaped by non-visual dimensions such as sound design [Hanson 2007], music [Brown 2023], subtitles, and captioning (zdenek, 2015). Future work could therefore expand the system to account for these features, making it possible to explore how similarity operates across multiple sensory and rhetorical registers.

7.4. Exploring scholarly perception and experience with AI and recommender systems.

Finally, although there is increasing interest in the integration of AI and recommender systems in DH, scholar resistance to such tools remains an obstacle to adoption. The perception and experience of such tools should therefore also be explored with more care to ensure that scholarly agency and needs are prioritized. Beyond simple accuracy and usability testing, the need for co-design, participatory models of curation and deployment remain central for the future of digital scholarship in the humanities.

39

8. Conclusion

In this study, we introduced an aesthetic-driven recommender system that generates similarity recommendations based solely on visual features extracted from film frames. Unlike conventional systems that rely on user profiles, collaborative filtering, or metadata, this approach foregrounds visual form as an analytical and generative framework. Through both qualitative and quantitative analyses, we show the system's capacity to identify meaningful stylistic relationships: from expected continuities within franchises or studios to more unexpected visual parallels and broader shifts in animation aesthetics.

40

The implications of these findings are twofold. First, they demonstrate that visual and aesthetic similarity are perceptible to machine learning models while also bearing a meaningful relation to viewer preferences, suggesting that aesthetic form may be more closely tied to reception than conventional metadata-driven approaches assume. Second, they highlight how recommender systems, when designed for interpretive exploration rather than optimization, can become tools for critical humanistic inquiry. Within the context of Digital Humanities, this repositions recommendation as a step in interpretive analysis, capable of surfacing patterns relevant to high-level concepts such as influence.

41

As the cultural record becomes increasingly multimodal and visually rich, the ability to computationally model aesthetic form gains renewed urgency. This project suggests that aesthetic-based recommendation can help create new research questions that emerge from style, design, and visual experience. Ultimately, this project reaffirms the relevance of computational stochasticity for DH inquiry. It shows that predictive computational approaches must not be viewed as constraining but have the potential to be generative, and expand how we view, interpret, and organise digital cultural collections.

42

Works Cited

- Arnold, Leonard, and Tilton 2017** Arnold, T., Leonard, P. and Tilton, L. (2017) "Knowledge creation through recommender systems", *Digital Scholarship in the Humanities*, 32(suppl_2), pp. ii151–ii157. Available at: <https://doi.org/10.1093/llc/fqx035>
- Atzenhofer-Baumgartner et al. 2024** Atzenhofer-Baumgartner, F., Geiger, B.C., Trattner, C., Vogeler, G. and Kowald, D. (2024) *Challenges in Implementing a Recommender System for Historical Research in the Humanities*. arXiv preprint arXiv:2410.20909. Available at: <https://doi.org/10.48550/arXiv.2410.20909>
- Beheshti et al. 2020** Beheshti, A., Yakhchi, S., Mousaeirad, S., Ghafari, S.M., Goluguri, S.R. and Edrisi, M.A. (2020) "Towards cognitive recommender systems", *Algorithms*, 13(8), p. 176. Available at: <https://doi.org/10.3390/a13080176>
- Bernhardsson 2018** Bernhardsson, E. (2018) *Approximate nearest neighbors Oh yeah (ANNOY) - AAU Social Data Science Deep Learning - 2019 portfolio*. Available at: <https://sds-aaugithub.io/M3Port19/portfolio/ann/> (Accessed: December 7, 2024).
- Brown 2023** Brown, R.S. (2023) *Overtones and Undertones: Reading Film Music*. University of California Press.
- Buchan 2007** Buchan, S. (ed.) (2007) *Animated Worlds*. Indiana University Press.
- Chen et al. 2019** Chen, X., Zhao, P., Liu, Y., Zhao, L., Fang, J., Sheng, V.S. and Cui, Z. (2019) "Exploiting aesthetic features in visual contents for movie recommendation", *IEEE Access*, 7, pp. 49813–49821. Available at: <https://doi.org/10.1109/ACCESS.2019.2910722>
- Darici 2023** Darici, F. (2023) *Analyzing video quality assessment methods on computer graphics animation videos*. BS thesis, Dept. Elect. Eng., Math. Comput. Sci. (EEMCS), Univ. Twente, Enschede, The Netherlands.

- Deng et al. 2009** Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L. (2009) "ImageNet: A large-scale hierarchical image database", in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE. Available at: <https://doi.org/10.1109/CVPR.2009.5206848>
- Ernst 2022** Ernst, F. (2022) "Towards a similarity algorithm for controlled vocabularies within the digital humanities", in *European Semantic Web Conference*. Cham: Springer International Publishing, pp. 179–188. Available at: https://doi.org/10.1007/978-3-031-11609-4_33
- Fitzgerald 2024** Fitzgerald, T. (2024) *The best streaming services for 2025*. Available at: <https://www.forbes.com/sites/entertainment/article/best-streaming-services/> (Accessed: December 12, 2024).
- Hanson 2007** Hanson, H. (2007) "Sound affects: post-production sound, soundscapes and sound design in Hollywood's studio era", *Music, Sound, and the Moving Image*, 1(1), pp. 27–49.
- Harper and Konstan 2015** Harper, F.M. and Konstan, J.A. (2015) "The MovieLens datasets: History and context", *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4), pp. 1–19. Available at: <https://doi.org/10.1145/2827872>
- He et al. 2016** He, K., Zhang, X., Ren, S. and Sun, J. (2016) "Deep residual learning for image recognition", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. Available at: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90>
- Jin, Segovia, and Sússtrunk 2016** Jin, B., Segovia, M.V.O. and Sússtrunk, S. (2016) "Image aesthetic predictors based on weighted CNNs", in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2291–2295. Available at: <https://doi.org/10.1109/ICIP.2016.7532767>
- Kahraman 2015** Kahraman, A.D. (2015) "Animation use as an educational material and animation techniques", *Online Journal of Art and Design*, 3(1), pp. 1–12. Available at: <https://www.adjournal.net/articles/31/311.pdf>
- Kalantidis, Kennedy, and Li 2013** Kalantidis, Y., Kennedy, L. and Li, L.J. (2013) "Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos", in *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, pp. 105–112. Available at: <https://doi.org/10.1145/2461466.24614>
- Ko et al. 2022** Ko, H., Lee, S., Park, Y. and Choi, A. (2022) "A survey of recommendation systems: recommendation models, techniques, and application fields", *Electronics*, 11(1), p. 141. Available at: <https://doi.org/10.3390/electronics11010141>
- Krizhevsky, Sutskever, and Hinton 2012** Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) "ImageNet classification with deep convolutional neural networks", *Advances in Neural Information Processing Systems*, 25. Available at: <https://doi.org/10.1145/3065386>
- Lamqaddam et al. 2025** Lamqaddam, H., Donoso, I., Mortier, Q., Brosens, K. and Verbert, K. (2025) "Motif-Match: Redefining similarity for digital art history through multifaceted image search", in *Digital Humanities 2025*. Available at: <https://zenodo.org/records/19682677>
- Lamqaddam et al. 2026** Lamqaddam, H., Mortier, Q., Donoso-Guzman, I., Brosens, K. and Verbert, K. (2026) "Beyond distance metrics: A nested model of similarity in visual art search", in *Proceedings of the 2026 International Conference on Advanced Visual Interfaces*. Available at: <https://dl.acm.org/doi/10.1145/3811427.3811442>
- Loukadakis, Cano, and O'Boyle 2018** Loukadakis, M., Cano, J. and O'Boyle, M. (2018) "Accelerating deep neural networks on low power heterogeneous architectures", in *11th International Workshop on Programmability and Architectures for Heterogeneous Multicores (MULTIPROG-2018)*.
- Lu, Chang, and Chen 2016** Lu, K.H., Chang, K.Y. and Chen, C.S. (2016) "Image aesthetic assessment via deep semantic aggregation", in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 232–236. IEEE. Available at: <https://doi.org/10.1109/GlobalSIP.2016.7905838>
- Lv et al. 2018** Lv, P., Wang, M., Xu, Y., Peng, Z., Sun, J., Su, S., Zhou, B. and Xu, M. (2018) "USAR: An interactive user-specific aesthetic ranking framework for images", in *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1328–1336. Available at: <https://doi.org/10.48550/arXiv.1805.01091>
- McLean 2015** McLean, T. (2015) "SpongeBob gets real", *Animation Magazine*, 29(3), pp. 10–12. Available at: <https://www.animationmagazine.net/2015/02/spongebob-gets-real/> (Accessed: May 27, 2026)
- Moen 2015** Moen, K. (2015) "Imagination and natural movement: The Bray Studios and the "invention" of animated film", *Film History: An International Journal*, 27(4), pp. 130–150. Available at: <https://doi.org/10.2979/filmhistory.27.4.130>
- Murray, Marchesotti, and Perronnin 2012** Murray, N., Marchesotti, L. and Perronnin, F. (2012) "AVA: A large-scale database for aesthetic visual analysis", in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp.

- Nielsen 2024** Nielsen (2024) “Time spent streaming surges to over 40% in June, the highest share of TV usage in the history of Nielsen’s The Gauge™”, *Nielsen*, 16 July. Available at: <https://www.nielsen.com/news-center/2024/time-spent-streaming-surges-to-over-40-percent-in-june-2024/>.
- Pavlidis 2022** Pavlidis, G. (2022) “AI trends in digital humanities research”, *Trends in Computer Science and Information Technology*, 7(2), pp. 26–34. Available at: <https://doi.org/10.17352/tcsit.000048>
- Schein et al. 2002** Schein, A.I., Popescul, A., Ungar, L.H. and Pennock, D.M. (2002) “Methods and metrics for cold-start recommendations”, in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 253–260. Available at: <https://doi.org/10.1145/564376.564421>
- Simonyan and Zisserman 2014** Simonyan, K. and Zisserman, A. (2014) *Very deep convolutional networks for large-scale image recognition*. [arXiv preprint] Available at: <https://doi.org/10.48550/arXiv.1409.1556>
- Stefanini et al. 2019** Stefanini, M., Cornia, M., Baraldi, L., Corsini, M. and Cucchiara, R. (2019) “Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain”, in *International Conference on Image Analysis and Processing*. Cham: Springer International Publishing, pp. 729–740. Available at: https://doi.org/10.1007/978-3-030-30645-8_66
- Talebi and Milanfar 2018** Talebi, H. and Milanfar, P. (2018) “NIMA: Neural image assessment”, *IEEE Transactions on Image Processing*, 27(8), pp. 3998–4011. Available at: <https://doi.org/10.1109/TIP.2018.2831899>
- Tan and Le 2019** Tan, M. and Le, Q. (2019) “EfficientNet: Rethinking model scaling for convolutional neural networks”, in *International Conference on Machine Learning*. PMLR, pp. 6105–6114.
- Washington et al. 2022** Washington, A.L., Rhue, L.A., Nakamura, L. and Stevens, R. (2022) “Uncoupling inequality: Reflections on the ethics of benchmarks for digital media”, in *Proceedings of the Annual Hawaii International Conference on System Sciences*, Vol. 2022, p. 2846.
- Widayanti et al. 2023** Widayanti, R., Chakim, M.H.R., Lukita, C., Rahardja, U. and Lutfiani, N. (2023) “Improving recommender systems using hybrid techniques of collaborative filtering and content-based filtering”, *Journal of Applied Data Sciences*, 4(3), pp. 289–302. Available at: <https://doi.org/10.47738/jads.v4i3.115>
- Yu et al. 2018** Yu, W., Zhang, H., He, X., Chen, X., Xiong, L. and Qin, Z. (2018) “Aesthetic-based clothing recommendation”, in *Proceedings of the 2018 World Wide Web Conference*, pp. 649–658. Available at: <https://doi.org/10.1145/3178876.3186146>
- Yu et al. 2021** Yu, W., He, X., Pei, J., Chen, X., Xiong, L., Liu, J. and Qin, Z. (2021) “Visually aware recommendation with aesthetic features”, *The VLDB Journal*, 30(4), pp. 495–513. Available at: <https://doi.org/10.48550/arXiv.1905.02009>
- Zdenek 2015** Zdenek, S. (2015) “Reading sounds: Closed-captioned media and popular culture”, in *Reading Sounds*. University of Chicago Press.
- Zhang et al. 2019** Zhang, X., Gao, X., Lu, W. and He, L. (2019) “A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction”, *IEEE Transactions on Multimedia*, 21(11), pp. 2815–2826. Available at: <https://doi.org/10.1109/TMM.2019.2911428>

