# AI for Image Compositional Guidance, a Knowledge-Based Approach

**Blake Sanie**
Georgia Institute of Technology
bsanie3@gatech.edu

## Abstract

Image composition is an essential component of the artistic process behind photography, graphic design, and other visual spaces of creation. To achieve maximal aesthetic quality, artists aim to carefully capture a composition that is compelling, balanced, familiar, and comfortable to the viewer. The abstract nature of this task, at first, seems resistant to the aid of computer systems backed by artificial intelligence. However, this work explores this unlikely intersection of art, visual aesthetics, and Knowledge-Based Artificial Intelligence (KBAI) in the design and implementation of a real-world compositional guidance tool. Specifically, this work proposes a concrete algorithm that models the infamous SOAR cognitive architecture, combining Working, Episodic, Semantic, and Procedural Memory for deliberation and action over absorbed knowledge. The sections following motivate the need for and curiosity behind KBAI in an artistic setting and explain the technical design of the underlying framework and its processes. Then, the work presents visually insightful results to support in-depth analyses and conclusions, altogether evaluating the chosen KBAI approach to compositional guidance and future work in the visual AI and computer vision domain.

## Contents

# Introduction and Motivation

## Image Composition

At its core, photography is the art of producing impressionable images. Deeper, photographers aim to collect compelling captures that foster a connection with its viewer; telling a story, characterizing a subject, or emphasizing perspective. Achieving this task requires photographers to leverage their creative visions, manipulating capture parameters at their will, often through improvisation. Aside from lighting and choice of subject matter, compelling images are grounded in the aesthetics of their composition.

Effective image composition can be formulated as discovering the *optimal framing* for a given scene. *Framing* refers to focal length (degree of zoom) and orientation (direction the camera faces), with subject matter placed within the resulting field of view. The pursuit of *optimal* framing is subjective by nature - each photographer develops their own compositional eye according to their distinguishing artistic style.

## Research Objective

This work develops a framework for applying computational methods to aid composition selection, both in the field (at capture time, through change in framing) or in post-production (through strategic cropping). Given the current composition, the proposed system will creatively assist the user by suggesting a directional frame shift (up, down, left, right) and focal length adjustment (zoom in, zoom out) if deemed likely to yield an aesthetic improvement. Naturally, this task may appear contradictory - computers are designed to produce objective insight through logical algorithms, though artistic vision is an abstract concept with no closed-form solution.

Creating a direct leap from algorithm design to optimal composition selection is a daunting, arguably inapproachable task. However, leading Human-Cognitive theory helps bridge this gap. Firstly, the basis for photographic visual appeal can be broken down into several observable components. Consequently, one can develop computer models to detect the prominence of such features. In all, the cognitive process of explaining composition aesthetic quality can be reversed engineered; inspect an input image, evaluate aesthetic-driving features, and draw an actionable conclusion to maximize such criteria for visual appeal.

# Related Topics and Foundation

The pursuit of the research objective begins with establishing a foundation of the following topics. Then, their perspectives and methods are combined to form the basis of the proposed AI-based approach for image compositional evaluation and guidance.

## Human Cognition and Aesthetic Perception

The field of Cognitive Psychology studies the mental processes behind understanding and applying knowledge. Furthermore, this domain represents otherwise complex and unbounded human reasoning behavior as a digestible logical framework.

By the conclusions of Aesthetics (2018), user perception for image aesthetics can be reduced into three observable qualities: Symmetry, averageness, subject conformity, and curvature. Symmetry encompasses visual balance across distinguishing image features, including captured objects, textures, shapes, etc. In the most aesthetic case, the average viewer attention region is center aligned at an appropriate scale to neither dominate or be dominated by negative space. Averageness refers to image features inducing little shock among the audience. Viewers find comfort in identifying familiar color, texture, and compositional patterns; perception of over-the-top effects or unimaginable perspectives results in viewer distress. Similarly, subject conformity involves image content itself. To promote aesthetic quality, image objects are to be represented according to their expected prototype, especially cognitively significant contents such as faces and body parts. Lastly, it is found that humans exhibit a preference for curved patterns and contours, as opposed to the same objects bounded by sharp edges.

Despite this proposed cognitive framework for modelling visual appeal, the authors still note that personal preference plays an equal role in their own aesthetic evaluation [1].

## Image Saliency

In order to observe the described image aesthetic features, a critical prerequisite task must be addressed: determining which regions of an image attract viewer attention, and to which degree. The Computer Vision industry refers to this process as saliency detection. Concretely, given an input image, output a saliency map, or expected attention heatmap over the input pixels. Saliency and viewer attention will be used anonymously going forward.

Though many advanced techniques have emerged in recent years, the most intuitive remains the GradCAM algorithm. Originally proposed in Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (revised, 2019), this process consists of identifying a target object class, and propagating model loss with respect to such class loss backward through the input input itself. As a result, a saliency map identical in dimension to the input is formed, reflecting the classification's sensitivity with respect to each pixel. Further, in images where multiple subjects are prominently detectable, GradCAM can be repeated with respect to each of these class outputs to estimate viewer attention region contributed to each respective class [5].

# Knowledge-Based AI Approach

## Evaluating Approaches in Machine Learning

Previous works in surround research areas leverage machine learning advances to directly infer image aesthetic quality from input images themselves. Below is a list of notable submissions:

- A multi-scene deep learning model for image aesthetic evaluation (2016)
- Adaptive Fractional Dilated Convolution Network for Image Aesthetics Assessment

- Analysis of Deep Features for Image Aesthetic Assessment (2021)
- Image Aesthetic Assessment Assisted by Attributes through Adversarial Learning (2019)

These are strictly data-driven methods, where data quality and volume bias model effectiveness. Additionally, the results of such works may highlight discovered image patterns correlated with highly aesthetic images, but these insights do not deeply reflect human thought processes - such features only arise by blindly optimizing towards the ground-truth labels [6] [2] [3] [4].

### Strengths & Relevance of KBAI

Employing Knowledge-Based Artificial Intelligence (KBAI) that directly mirrors human-cognitive reasoning and behavior is viable alternative. Ideally, data considerations do not limit model performance, thus embracing a data-free workflow has the ability to nurtures both simplicity and effectiveness. After all, the labelled datasets utilized in the mentioned studies are mined from photo community sites flooded with subjective taste. By contrast, a KBAI aesthetic evaluation framework can explainably and deterministically produce the expected (grand-majority) visual appeal by following a guided reasoning method - this is the closest possible link between objective and subjective evaluation. Finally, the previous studies did not dissect image aesthetics into more granular components. This work addresses the subproblem of evaluating and acting upon image composition, where a specific cognitive framework can be more rigorously defined.

### The SOAR Cognitive Architecture

One framework for modelling human application of knowledge is the SOAR Cognitive Architecture. This theory delineates role-specific cognitive units. These blocks are tightly coupled in interaction to yield explainable agent behavior.

According to SOAR, the deepest point of cognition is long-term memory, partitioned into three parallel units:

- **Procedural Memory**: A routine of reasoning-focused steps used as the basis for deliberating agent action.
- **Semantic Memory**: A knowledge storage model for embedding and retrieving generalizations, facts, and other prior context of the agent's world.
- **Episodic Memory**: A series of distinct knowledge snapshots, such as events or one-time observations.

Placed below long-term memory is its complementing cognitive layer: **Working Memory**. This SOAR unit is responsible for direct interaction with the agent's actionable world. Working memory accepts precepts, queries the three components of long-term memory, combines the gathered knowledge, and concludes a final action. The SOAR architecture simultaneously enables continuous learning. As new observations cause internal impasses, the process of chunking strengthens each component of long-term memory to both propose an immediate response, and protect against future impasses when the precepts are encountered in the future.

## Theoretical Framework & Approach

In short, the proposed approach to image composition evaluation is to computationally model a SOAR-driven framework. Luckily, industry-standard data structures, machine learning models, and computational methods lend themselves to each component of SOAR as it stands. The task at hand is to purposefully combine such computational units into a greater workflow, evaluating image composition by directly reflecting the studied thought process of the human mind.

### GradCam to Support Procedural Memory

Procedural Memory will is modelled as a programmatic sequence of conditions and operations (often called control flow units). The following reasoning steps are employed for deliberation:

1. Apply GradCam on an input image to obtain a saliency map with respect to its plausible classifications (plural in case the scene is dominated by multiple objects). This forms a global saliency map, which will grow as GradCam is later applied with respect to other class predictions. Leverage the academic-standard image classification model Resnet50 as the base model.
2. Repeat (1) for each result class beyond the top classification, from greatest to least confidence. Only combine the saliency map with the global max if a new viewer attention region is highlighted - this prevents the same object in the scene (ex. a cat with competing classifications of Cheetah and Leopard) from unnecessarily updating the global saliency map multiple times. With this condition, the multi-class GradCam loop terminates in reasonable time without loss of effectiveness.
3. If multiple dominant saliency sub-regions are found (typically for multiple present objects), mathematically estimate the center and spread of the saliency distribution. Propose a zoom and reposition operation to center the attention regions symmetrically, and achieve a comfortable degree of attention spread over the image frame.
4. If a single dominant saliency region is found, an impasse arises. Because the image focuses on a presumed single subject, the artist may deliberatily break exact symmetry to build a compelling composition. For instance, in the domain of portraits, viewer attention is focused on the facial region. However, if the subject's face is directly centered in the image, too much negative space is reserved above the head, with not enough of the subject's body captured below the head. In this scenario, the ideal composition will more likely place the head between the top and middle of the frame to account, strengthening the averageness quality promoting aesthetic perception. When these impasses occur, rely on the Episodic Memory module to resolve the impasse.

### KNN to Support Episodic Memory

Episodic Memory operates by ”soft-matching” a case similar to the presented query, and applying the prior action.

Computationally, a K-Nearest-Neighbor (KNN) unsupervised model fits the task. Concretely, given a representation of an image, search through the indexed catalogue of aesthetic images to find an attention-region match for the same primary classification. Then, return a zoom and reposition operation to mimic the stored case's saliency distribution. The KNN model is fit on a professional image dataset (Unsplash), thus the assumption that the fetched cases offer a compelling composition is reasonable. Lastly, KNN operates on indexing and comparing vector representations - this image representation, or descriptor, will consist of a heavily downsized and flattened image saliency map. This formulation allows for cases to generalize to each other, as retaining directly flattening a high-pixel dimension image will encourage overfitting.

### Computational Working Memory Model

The computational implementation of working memory is an agent that commands the Procedural (GradCam) and Episodic (KNN) sub-models. First, the agent attempts to resolve image compositional guidance through the Procedural module, falling back onto the Semantic module as necessary (impasse). The Working module's logic is trivially developed.

As noted, Semantic Memory is a cognitive module that houses theoretical information. Because our task is safely reduced to the assumptions that drive human perception of visual aesthetics, our computation SOAR model does not require access to greater image context or photography understanding beyond image contents and representations already observable by the Procedural and Episodic modules. Our approach will omit this component of the SOAR framework for simplicity.

## Hypotheses

### Approach Effectiveness

Before embarking on experimentation and framework (re)iteration, a study of the objective's feasibility is in order. Fundamentally, this approach is guided by knowledge-based, rule-based methods. As a result, the quality of the agent's suggestions for improving image composition are strictly bounded by the prerequisite assumptions used to guide compositional improvement. In this work, the abstract, complex, and ultimately subjective task is greatly reduced. If multiple attention regions hold, the agent reduces its scope to mathematically solving for predefined saliency symmetry and scale. If the saliency map forms a single dominant cluster, the agent simply defers its suggestion by recalling the nearest professional-quality image, and matching to its saliency parameters.

With this degree of complexity reduction, the knowledge-based agent will surely and deterministically succeed. Any dissatisfaction that arises from agents actions are attributed to low quality of underlying assumptions, not the framework of reasoning forming the agent's backbone.

### Anticipated Edge Cases

When more than one saliency region is prominent, compositional advice is guaranteed to be generated. If an impasse is raised instead, resolution is dependent on the existence of relevant cases in the KNN-backed image saliency database. Compositional guidance cannot be resolved in two cases:

1. The dominant classification result is rare enough such that mass image dataset used to populate the episodic image descriptor index does not contain any cases corresponding to that class.
2. The dominant classification result exists in the mass image dataset, but no class-corresponding image in the dataset contains a single dominant saliency region. Thus, the Episodic module does not store any relevant cases for this class.

The occurrence of these edges cases resulting in no actionable output can be minimzed by storing a greater volume of cases in the Episodic memory component. This requires finding, loading, and processing supplemental datasets to build a more thorough and complete case index. Theoretically, there is no upper limit to the number of images shown to and stored by Episodic memory, though this notion supports a hypothesis that the image class coverage will exhibit diminishing marginal returns with respect to computational resources (compute and memory) inherently utilized by the Episodic module.

## Preliminary Experiments & Findings

### Extending GradCam for Multi-Class Saliency

### Multiple Attention Region Example

As mentioned, the GradCam method yields a saliency map contributing to a target output class. This work aims to build upon single-class saliency by combining maps across a variable number of prominent output classes. As a key building block for future agent functionality, this process is implemented and its offerings (versus drawbacks) analyzed first.

The image shown in Figure 1, framing a dog and a cat, serves as a multi-class example. Given as input to the base Resnet50 model, the output classes ordered by descending probability are:

1. German Shepherd: 45%
2. Egyptian Cat: 18%
3. Tiger Cat: 3%
4. Tabby Cat: 2%
5. Eskimo Dog: 2%

And so on.

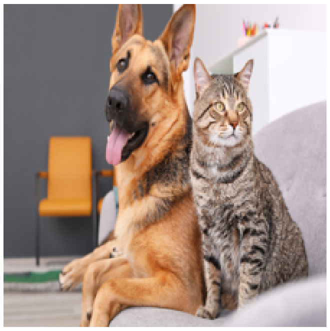


Figure 1: An example multi-class input image

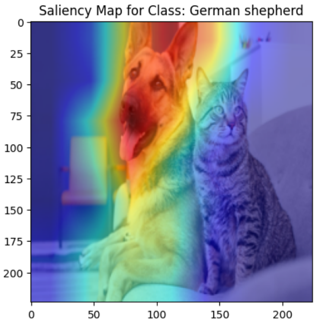


Figure 2: Saliency map for output class ”German Shepherd”

Figure 2 visualizes the saliency map for the most probable class: German Shepherd. The most impactful region of visual attention appears around the dog’s face and ears. Since this is the first output class inspected, this saliency map serves as the initial global map. The same process single-class GradCam process continues for subsequent class predictions.

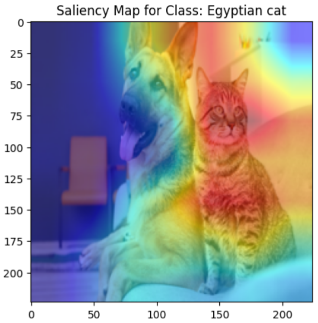


Figure 3: Saliency map for output class ”Egyptian Cat”

The saliency map for the next most confident class, shown in Figure 3, suggests a visual attention distribution across the cat subject. Interestingly, the model found the dog’s ear useful in forming this classification. This observation highlights one weakness of GradCam in estimating real-world saliency: machine learning model parameters are trained to strictly optimize their objective, often taking shortcuts if justified by the end-result. These shortcuts may be illogical or lack explanation altogether, thus resulting in artifacts when gradients are propagated back through the input image.

Assuming the saliency map for class ”Egyptian Cat” is combined with the global map (apply pixel-wise maximum), total collected units of saliency would grow 1.7x. This scalar surpasses the experimentally-set threshold of 1.4, thus the saliency map union is accepted and set as the new global map.

Applying GradCam over the next most-confidently predicted class, ”Tiger Cat”, yields a saliency map that would yielding a 1.0x saliency growth over the global map. Falling under the 1.4 threshold, the proposed addition of saliency is rejected, and no further output classes are inspected. This action demonstrates the ability of the proposed method to handle competing classes for the same object in the scene (the cat, in this case, since the previous prediction ”Egyptian Cat” was already considered).

Now having computed the final global multi-class saliency map shown in Figure 4, such map can be decomposed into visual attention regions. This is accomplished by forming an image-wide mask, applying a condition to the saliency gathered for each input pixel. Experimentally, it was found that a suitable condition for detecting separate regions is ignoring the 88% least salient pixels.

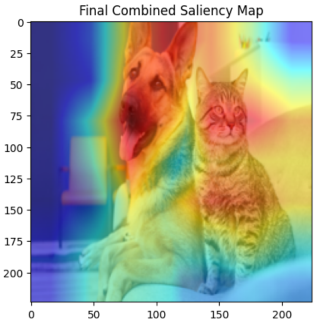


Figure 4: Global saliency map for input image

In other words, only the top 12% most attention-grabbing pixels continue to the next layer of processing. The resulting mask, visualized by Figure 5 contains at least one continuous segment(s), each of which corresponds to a unique visual attention region. Since such boolean segments are bounded (all surrounded by ignored pixels), the number of visual attention regions are deterministically countable with trivial methods.

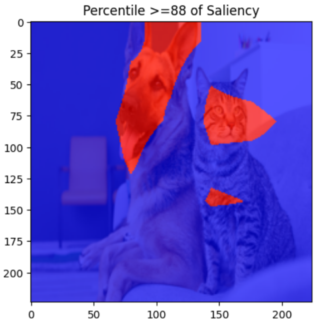


Figure 5: Boolean mask reflecting pixels above 88th percentile level of saliency

In this example, three significant attention peaks are found. Though expected to find more than one, a theoretical detection of two attention regions is more logical, given the two principle actors in the scene. This observation highlights another potential flaw to the proposed method in its elementary state: not all attention regions are equal. The filtration of pixels below the 88th percentile of saliency attempts to address this concern, but even once the mask is formed, regions themselves exhibit varying characteristics: size, shape, and mutual proximity. As suggested by the saliency peak mask generated over the input image, the attention region over the dog is noticeably more prominent than that over the cat’s mid-figure, by our human intuition. However, these regions are counted equally. Yes, this serves the purpose of separating single-attention-region cases from multi-region cases, but a gap in region expectation versus reality still remains.

Having detecting multiple attention regions, the knowledge-based agent will delegate the compositional guidance task to the Procedural module for further processing.

## Single Attention Region Example

The same processing workflow from above is applied to the image presented in Figure 6. Resnet50 predicts the following class distribution, ordered by descending probability:

1. Castle: 69%
2. Cliff: 6%
3. Palace: 3%

and so on.

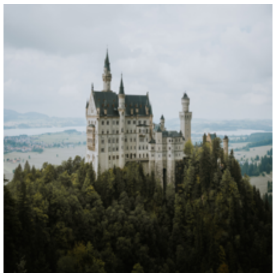


Figure 6: An example single-class input image

Note that in the castle example, the runner-up classification result, ”Cliff”, generated a saliency map very similar to that of ”Castle”, thus the proposed saliency map combination process terminated with only the dominant class prediction influencing the global map (Figure 7).

Here, a single peak visual attention region is found across the mask visualized in Figure 8. The knowledge-based agent will interpret this finding, falling onto the Episodic module for compositional guidance resolution.

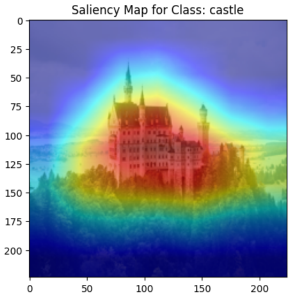


Figure 7: Saliency map for output class ”Castle”, also serving as the final global map

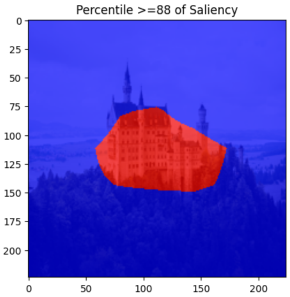


Figure 8: Saliency peak mask, utilizing 88th percentile

## Image Descriptor Engineering & KNN Considerations

The agent’s Episodic module depends on the ability to recall similar, previously seen images. This fundamental task requires storing and comparing a consistent vector representation of image saliency, called a descriptor. The size of such designed descriptor obeys a compromise between detail and generaliability. That is, larger sized descriptors effectively embody high-granularity features, but a match is only produced when the query descriptor happens to reflect the same degree of feature detail. By contrast, smaller sized descriptors smooth over more granular features, but the simplified nature of the representation allows query descriptors to consistent yield more reasonable, ballpark matches.

Ideally, a suitable middle ground is chosen between these extremes. Through this work’s initial tests, the most appropriate descriptor dimension is 36 - resulting from a 6x6 downsized saliency map, flattened into a single dimension, pictured in Figure 9. This choice of descriptor size allows for encoding of high-level visual information, such as greater attention region position and spread, without being dominated by small-scale, image-specific idiosyncrasies.

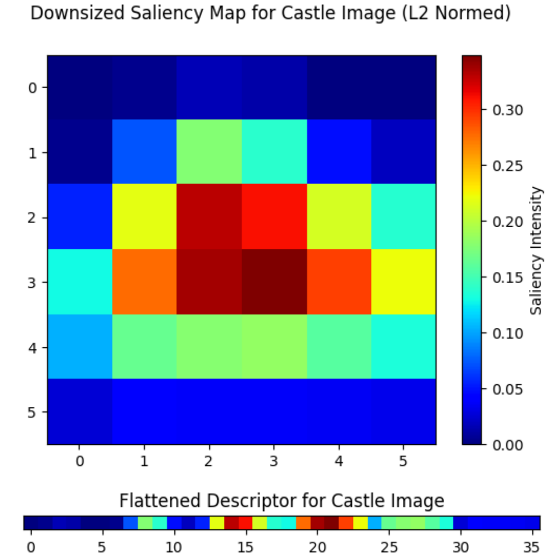


Figure 9: Downsized saliency map for castle image, L2 normed and flattened to form a descriptor vector

A well-designed saliency descriptor alone cannot guarantee effective retrieval of memorized high-aesthetic image cases; the volume and diversity of Episodic cases themselves must exhibit substantive coverage over the space of plausible queries. To analyze the spread of cases by primary prediction class, a preliminary data profiling study is run over 1000 images randomly downloaded from the Unsplash professional image database. Of the 1000 collected images, 724 exhibit a single visual attention region (following the previously outlined workflow). Descriptors are formed and pooled by top classified category.

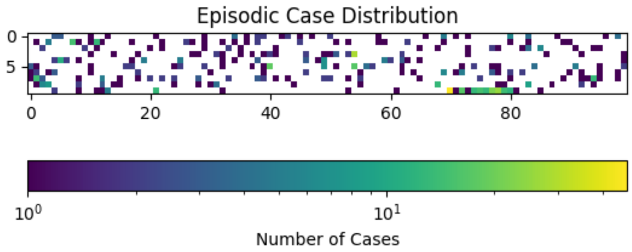


Figure 10: Downsized saliency map for castle image, L2 normed and flattened to form a descriptor vector

As suggested by Figure 10, the collected prediction class distribution exhibits higher frequency for some classes more than others. For instance, the most commonly observed dominant class (”Alp”) spans 47 case. On the other hand, 766 of Resnet50’s 1000 output classes are unpopulated.

These initial findings encourage a degree of skepticism regarding the agent’s ability to construct a robust, complete case index within its Episodic module. With one KNN built to support saliency search functionality for each output class, many queries will meet no resolution if output class is fully unrepresented.

Deeper, even if the class' corresponding KNN contains a few examples, low case volume will produce uninsightful matches, since a match of higher similarity is less likely. Optimistically, one may simultaneously expect image classes to be unevenly distributed; although many output classes are left unrepresented, this phenomenon may occur because those classes are less frequently observed themselves in real-world images. In this experiment, the output class ”seashore” was observed 27 times, while ”lab-coat”, ”letter-opener”, and ”jigsaw puzzle” were never encountered. In all, though the complete 1000-class distribution is empty in majority, only rarely captured, unanticipated subjects are left underrepresented; expected, commonly observed subjects tend to maintain adequate case volume for insightful match retrieval.

# Application of Methodology

## Experiment Setup

The proposed computational framework is implemented as a series of dependent subroutines. This reflects the modular structure of the SOAR cognitive model and its inner workings. As performed in the working memory module, a multi-class saliency map is generated for an input image, on top of which the number of saliency regions is inferred by Figure 5. If multiple classes are inferred, the subroutine encompassing procedural memory is invoked. This will directly return composition guidance instructions back to the caller. If a single class is invoked, the subroutine encompassing a lookup over the KNN image library is performed, also directly returning back to the caller.

For ease of experimentation, all subroutines are defined within distinct code cells of a Jupyter Notebook. This allows code segments to be executed ad-hoc, with complete integration of graphical output. This methodology is powerful, for instance, when simply regenerating graphics following a variable change, instead of re-executing the entire script. As a result, experiment efficiency remains high while computational cost and runtime are kept to a practical minimum.

This code structure not only enables, but encourages further experimentation and results reproduction. This work intends to stand as a foundational contribution to the Knowledge-Based AI field of research. That is, parameters may be tuned, new images may be evaluated, and new features may be integrated as future research requires.

## Executing Compositional Guidance Deliberation

Up to this point, image preprocessing stages (saliency map generation, region count, etc) are complete. However, a logical process is required to evaluate the underlying features and compute a compositional guidance result.

In the case of multiple saliency regions, a global attention region is modelled by a 2D Normal distribution overlaying the image. This is formed by collapsing the saliency map across the x and y axes, respectively, each used to model a 1D Normal distribution. From here, a 2D Normal Distribution is formed with the center coordinate formed by the respective x and y centers, and uniform covariance as the maximum of variances along the x or y axes. Note that covariance is designed to be uniform in order to form an attention distribution with equal spread along the x and y axes.

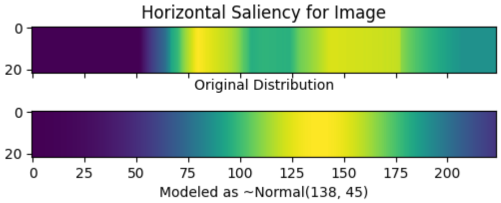


Figure 11: Saliency of Cat and Dog image (Figure 1) collapsed to the x direction, modelled as 1D Normal Distribution

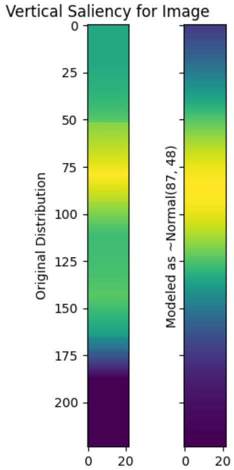


Figure 12: Saliency of Cat and Dog image (Figure 1) collapsed to the y direction, modelled as 1D Normal Distribution

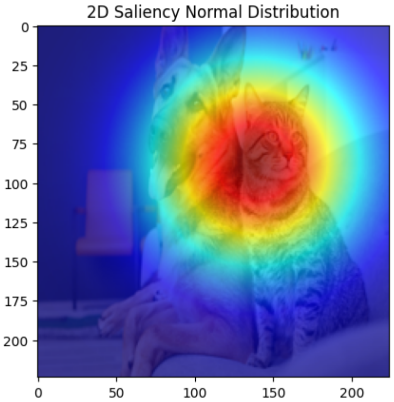


Figure 13: Saliency of Cat and Dog image (Figure 1) modelled as 2D Normal Distribution, with covariance $48^2$ because y variance is greater than x variance.

This allows cases where the true global saliency region is ”long and skinny” to avoid being ”cutoff” by the suggested composition due to the smaller directional variance skewing the Normally Distributed saliency region's size.

Figures 11, 12, and 13 illustrate this process visually.

From here, if a significant enough compositional adjustment is necessary to center the Normalized attention region (move left, right, up, or down) with appropriate scale (zoom in or out) to justify compositional guidance, such instructions are generated according to the following procedure:

1. Initialize horizontal action, vertical action, and zoom action to empty (no instruction)
2. If the modeled horizontal center deviates from the image center by tolerance% of image width, set horizontal action to ”left” or ”right”, as appropriate.
3. If the modeled vertical center deviates from the image center by tolerance% of image width, set vertical action to ”up” or ”down”, as appropriate.
4. If the modeled attention covariance surpasses 7000 (assuming image width 224 pixels), the attention region has significant spread; set zoom action to ”zoom out”. If the covariance falls below 3000, by contrast, suggest ”zoom in”. Note that 7000 and 3000 were determined experimentally, and may be treated as a parameter defined by subjective preference for degree of zoom.

In these experiments, tolerance% is set to 5%, meaning the composition is considered ”good enough” if within 5% of the optimal configuration. This tolerance can be tuned by in future work as well.

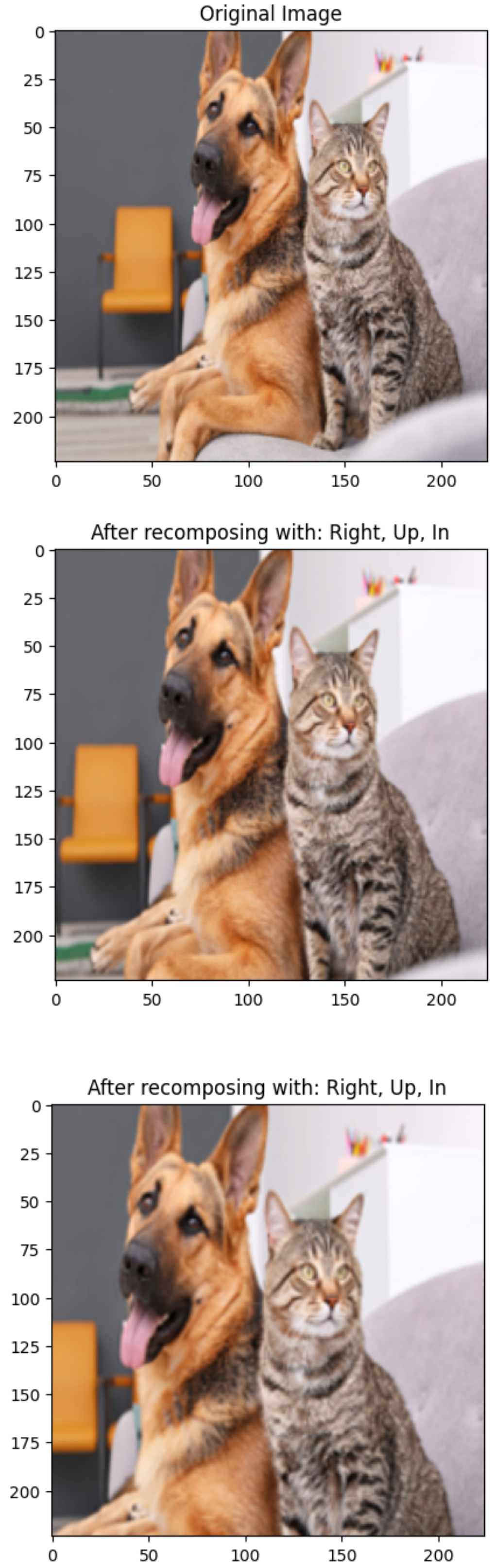
In the case of a single detected saliency region, obtain the closest episodic result by forming a saliency description and querying the KNN case library. Perform the same 2D Normal Distribution modelling over the attention region for both the input image and the closest case. Then, treating the attention distribution over the input image as the image center, execute the logical process outlined above.

Concretely, if the attention region of the closest case holds a different position and scale relative to that of the input region, the agent will generate and return compositional guidance such that the input image may more closely resemble the retrieved case.

## Presentation of Results

The below results illustrate the image given as input to the agent, and the resulting composition achieved by following the agent’s compositional guidance. This inference process may be repeated multiple times, continually improving image composition in the pursuit of maximizing aesthetics of the capture.

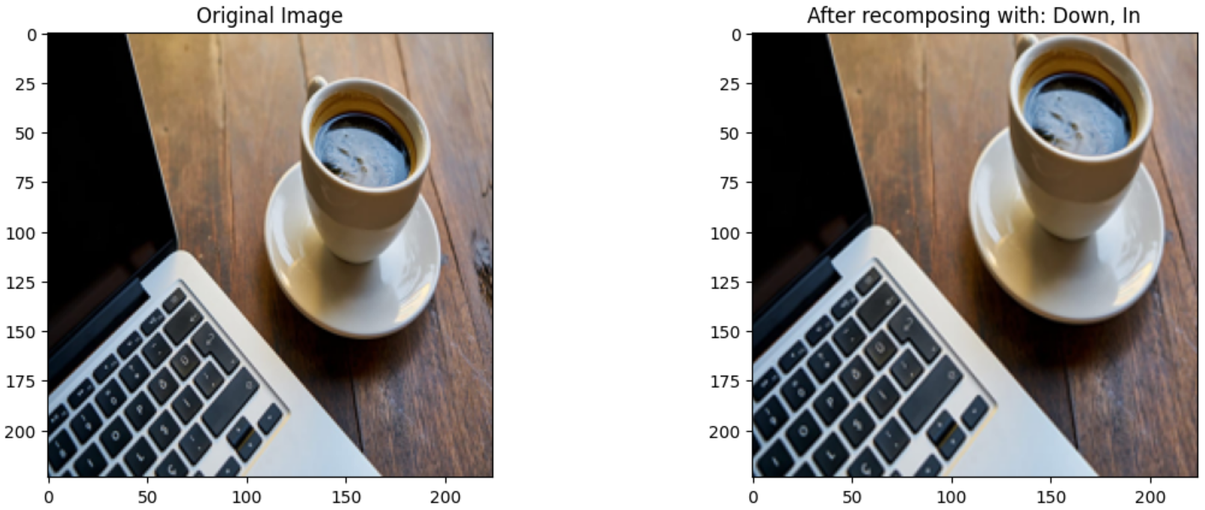
### Multiple Saliency Regions


Original Image

After recomposing with: Right, Up, In

After recomposing with: Right, Up, In

Figure 14: Compositional guidance computed and applied to Figure 1


Original Image

After recomposing with: Down, In

Figure 15: Compositional guidance computed and applied to an image of a notebook computer and coffee on a desk

### Single Saliency Region

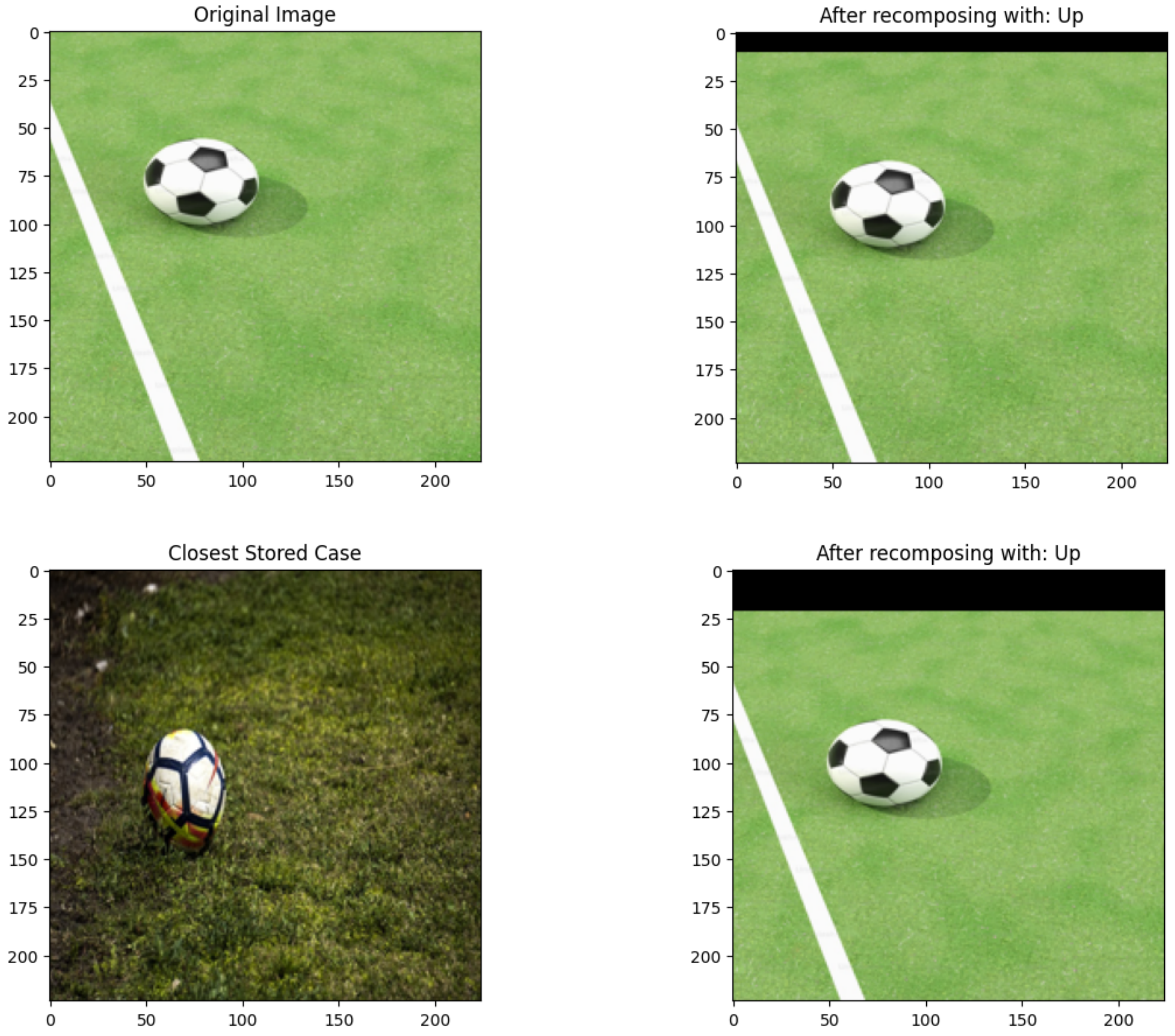
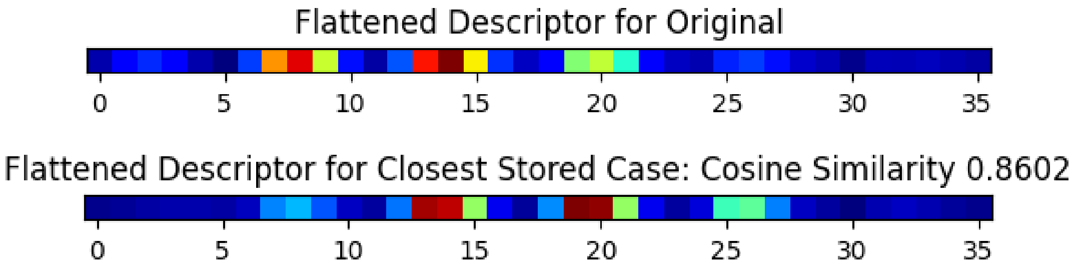


Figure 16: Compositional guidance computed and applied to an image of a soccer ball



# Evaluation of Results

## Drawing Conclusions

Because the underlying idea of this project is to model human intuition, it is expected that the agent's generated compositional guidance follows a reasonable and explainable deliberation process. In the first example (Figure 14), the agent expresses a desire to achieve higher aesthetic quality by repositioning up and to the right, zooming in to fill the scene with the eye-catching subjects. In the second example (Figure 15), the agent believes no horizontal shift is necessary, but a downward repositioning and inward zoom are beneficial. When a single attention region dominates the input image, as in the example encompassing Figure 16, the agent wishes to reposition the frame upward to match the closest known example of an aesthetic image. No zoom is suggested, as the attention distribution spread of the input image is already similar enough to that of the closest retrieved case.

On the surface, the model generated productive guidance instructions. Yes, the internally leveraged constants and thresholds can be tuned to fit any subjective eye, but improvements are made across the general rules of composition and aesthetic quality expressed earlier.

However, the AI does not always perform perfectly. For instance, by zooming in the on image behind (Figure 15), the top of the coffee mug is slightly clipped by the upper bound of the frame, thus creating tension in the image. This is a minute detail the model does not explicitly address, thus exhibiting that real-world edge cases are certain to arise and result in situation compositional flaws.

Through my experiments, I noticed a second shortcoming of this computational model - When compounding the agent's compositional guidance (that is, applying guidance to an image repetitively in search of the optimally aesthetic composition), the number of saliency regions may switch from multiple to singular, or vice versa. As a result, the cognitive model re-routes the deliberation knowledge to different modules (procedural and episodic), and the agent's compositional feedback may attempt to achieve a different compositional vision.

Lastly, the success of single-saliency-region queries depend on a sufficient space of known and preprocessed cases. With regards to Figure 16, the agent found a clearly similar case to relate to the query, but through experimentation, it is clear that this is not always the case. This notion reinforces the SOAR cognitive architecture's standing as a theoretic framework - implementing as a concrete technical model would require a high degree of quantitative knowledge (image datasets) to function with high quality in a grand majority of real-world cases.

## Comparison with Existing Literature

The literature review conducted in the previous discussion of related topics suggested that humans perceive high aesthetic quality when balance, symmetry, and normalness are visually present. This work validates this literature review by computation implementation, and evaluation of generated results. Beyond simply suggesting productive compositional changes, this project's environment provides the unique capability of simulating such advice (by applying the repositioning and zoom to each image) to directly evaluate the quality of guidance. In all cases, the simulated result following guidance execution reflects a greater degree of symmetry, interest, or relevance to a ”typical”, or ”normal”, known case.

Of course, more work is required to bridge the gap between the current state of human-cognitive theory and computational models. This project's approach is a simple, foundational, prototype. Many edge cases are not address, just as many implications still arise that could benefit from advancements in this method.

## Potential Implications

- **Inflated ease of artistic creation** - With an AI assistant to build effective photographic compositions, less emphasis is placed on the photographer to adopt and strengthen this skill. Later on, in a scenario where this technology is not available, the artist cannot maintain the same effectiveness in the creation of their works.
- **Lack of quality guarantee** - as concluded, though the model is effective in its simplified task, it may struggle against the unique, fine-grain perspectives of the real-world. As a result, the agent may suggest potentially unaesthetic compositional suggestion. This may arise due to incorrect or inaccurate subject recognition, lack of similar known cases, idiosyncrasies within the scene, etc. Included in these list of reasons is the bounded effectiveness of the proposed method itself. In all, always relying on this method for achieving high-quality composition is not the answer to always optimizing image aesthetics.
- **Discrimination of subjects** - due to the agent's dependence on other classical Machine Learning methods with their own implications, there stands a possibility that distinct subjects in the scene are discriminated by gender, age, skin color, etc. Deeper, with the agent potentially placing a greater degree of ”emphasis” on one subject compared to another, the resulting compositional guidance will aim to highlight or focus on that subject, leaving out the other.
- **Aesthetics as an objective measure** - Aesthetic quality is unarguably subjective by human nature. Any given image and its composition may be pleasing or displeasing to two different individuals. This work aims to provide a middle-ground innovation, a perspective that is most relatively agreeable out of all perceptions. Still, the framework proposed in this project relies on several fixed parameters and other human-behavioral assumptions. It should be clearly expressed that the compositional guidance generated by this agent is not objective, and each individual's personal opinion of image aesthetics remains completely valid.

# Conclusions

## Reflection on Results and Main Claims

Looking back to the early conception of this proposed method and the chosen scope of this work, one can argue that success is achieved for the set goal - providing compositional guidance to improve image composition aesthetics through a KBAI approach. Deeper, subtasks (multi-class saliency, fetching recorded cases) and objectives are handled on a situational basis, reflecting the hypothetical SOAR cognitive architecture in executable program form. Still, one must not forget that this success is only achieved partly because the selected task was overly simplified.

Though the main claim is supported by this report, the framework does not operate with high proficiency at a semantic level, that of human understanding and cognition. This is apparent by the pitfalls of the generated compositional guidance analyzed previously.

Additionally, I am not confident that this method applied to a simplified target task applies as appropriately to more refined, granular, complex compositional guidance tasks. Not only does this framework rely on key assumptions such as the accuracy of the GradCAM algorithm, its underlying image classifier, mutli-class saliency being normally distributed across the image plane etc. Also, as task complexity grows, a more intelligent model will require an even greater degree of experimentation and edge case evaluation and handling - a degree of experimentation that may require significant time for even entire teams of AI engineers.

## Future Work

If this project continued along a timeline greater than the length of the semester, I would place emphasis on and further explore the below considerations.

- **Single-Object Case Coverage**. The current state of this work processes a segment of the Unsplash image dataset to construct the basis for the episodic memory module. Beyond leveraging more memory to support a larger KNN index, more time can be spent to process the Unsplash dataset in its entirety, or even encode episodic cases over the massive ImageNet or Cifar datasets used to train large state-of-the-art computer vision models.
- **Saliency Encoding Scheme Engineering**. This work encodes image saliency into vector descriptor form by downsizing the saliency map, flattening into a single dimension, and normalizing by euclidean distance into a unit-magnitude vector. This is a simple encoding scheme used to intuitively build descriptor vectors, but that does not mean the vector descriptor output is the most information dense for the given vector size. I hypothesize that there exists a more efficient saliency descriptor representation that can be achieved by leveraging unsupervised methods such as dimensionality reduction by Principal Component Analysis (PCA) or Auto-Encoder Neural Networks as sub-modules to the episodic component of SOAR. Further, this reduction in descriptor size will allow more recorded cases to fit into the memory-backed KNN, allowing for image queries to draw from a larger episodic library.
- **Continuous Learning and Chunking Functionality**. One strength of the SOAR cognitive model is the ability to impose continuous learning. That is, as the compositional guidance agent yields incorrect or generally unfavorable outputs, the agent learns to reinforce other examples or logical pathways to reduce occurrence of the error or similar errors in the future. Currently, this implementation of SOAR is focused on performing inference from the devised procedural and episodic workflows. Deeper, chunking is the process of instituting a new rule to resolve procedural conflicts - when multiple asserts collide and contradict, or no assertions arise to begin with. Recall that the episodic case distribution is not complete across all classes. For the agent to be maximally robust, it needs a mechanism to handle when the single-object query image pictures an unrepresented, unstored class. In the future, design and integration of a continuous learning process will not only build a smarter agent, but allow it to adapt appropriately to increase guidance quality as agent use increases!
- **Integration of the Semantic Memory Module**. Recall that the currents state of this work approximates the function of the SOAR architecture by ignoring the role of Semantic Memory. This is because representing, storing, and drawing from semantics is a complex task for computational systems. In all, this is the core struggle and area of innovation in the field of Knowledge-Based AI. However, concrete knowledge representations do exist, such as semantic networks or frames, that allow semantic information to be stored and queried. Currently, the agent deliberates a compositional action based on its visual perception. However, if the agent can develop the ability to develop a semantic understanding of the scene's setting, scenario, relationship between complimentary or competing subjects, etc., more informed composition guidance will result. More generally, the current work observes the image's ”what” before performing inference, though an even more robust framework would build upon the ”why” and ”how” of collective scene contents themselves.

## Key Points and Takeaways

Image composition lies in an open-ended artistic domain, with computation and AI existing in a closed-form, calculated space. This work demonstrates the ability of Knowledge-Based AI in bridging the gap between the two seemingly disjoint disciplines. The theory-driven SOAR cognitive architecture, as shown, can successfully occupy the form of programmatic logic when applied to a specific concrete task such as composition guidance. Still, this agent prototype finds success in a simplified world model, where perception is limited to the given image itself and a preprocessed visual object knowledge base. Further improvements are encouraged by later development of SOAR mechanisms such as continuous learning and a knowledge base with greater semantic, not recorded case, emphasis. Ultimately, this work lays a firm groundwork of logical design and experimentation atop which future innovations between AI and the visual and artistic domain are possible, motivated, and encouraged.

# References

[1] Aenne A Brielmann and Denis G Pelli. "Aesthetics". In: *Current Biology* 28.16 (2018), R859–R863.

[2] Qiuyu Chen et al. "Adaptive fractional dilated convolution network for image aesthetics assessment". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 14114–14123.

[3] Hyeongnam Jang and Jong-Seok Lee. "Analysis of deep features for image aesthetic assessment". In: *IEEE Access* 9 (2021), pp. 29850–29861.

[4] Bowen Pan, Shangfei Wang, and Qisheng Jiang. "Image Aesthetic Assessment Assisted by Attributes through Adversarial Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 679–686. DOI: `10.1609/aaai.v33i01.3301679`. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/3845`.

[5] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[6] Weining Wang et al. "A multi-scene deep learning model for image aesthetic evaluation". In: *Signal Processing: Image Communication* 47 (2016), pp. 511–518. ISSN: 0923-5965. DOI: `https://doi.org/10.1016/j.image.2016.05.009`. URL: `https://www.sciencedirect.com/science/article/pii/S0923596516300662`.