Automatic Visual Citation Generation for Text-to-Image Generation

Ning Xu
Advanced R&D
Adeia Inc.
San Jose, USA
ningxu01@gmail.com

Serhad Doken
Advanced R&D
Adeia Inc.
San Jose, USA
serhad.doken@adeia.com

***Abstract*—As generative artificial intelligence (GenAI) systems increasingly influence the fields of art and design, they raise critical challenges regarding copyright and artistic attribution. This paper introduces a novel system for generating visual citations within AI-generated images, thereby addressing copyright concerns while fostering ethical use and sharing of digital art. The proposed system aims to efficiently identify elements and styles that are possibly derived from existing artworks within the outputs of text-to-image AI systems. This system enhances privacy and security as it operates on image features and embeddings, rather than accessing the images directly, thus avoiding the handling of copyrighted materials. The system automatically embeds metadata into the images, detailing the origins of the included artistic elements. Moreover, the system enhances compliance by enabling automatic or manual modifications of generation prompts, ensuring that generated images are free of copyright infringements. This functionality not only protects artists' intellectual property but also supports transparent acknowledgment in digital media creation. The proposed system has the potential to facilitate respectful interactions between AI technologies and creative content.**

Keywords—Generative AI, Text-to-Image, Visual Citation, Copyright Compliance, Digital Art

I. INTRODUCTION

Generative artificial intelligence (AI) technologies are revolutionizing numerous industries, with particularly profound impacts in the fields of art and design. Recent advancements in text-to-image generation, exemplified by systems like DALL-E [1], MidJourney [2], and Stable Diffusion [3], have demonstrated the potential to transform ideas into visual representations with unprecedented ease and flexibility. These technologies not only democratize artistic creation but also open new avenues for commercial and educational applications, enhancing creative processes and enabling novel forms of expression. As these AI models become more accessible and their usage more widespread, they increasingly influence how content is created and consumed, setting the stage for significant shifts in digital media landscapes.

Despite the remarkable capabilities of generative AI in producing visually compelling images, there exists a critical deficiency in how these technologies handle the intellectual property rights of artists. The core issue arises from the AI's ability to replicate and remix styles and elements from existing artworks without proper attribution or respect for copyright laws. This capability, while innovative, often leads to the unauthorized use of copyrighted materials, potentially infringing on the rights of original creators. Artists and copyright holders face significant challenges in controlling the use and distribution of their works, which are frequently used as training data for these AI systems without explicit consent or compensation. This situation not only poses legal risks but also ethical concerns, undermining the rights of artists and devaluing their contributions to the cultural and creative sectors.

Current methods to address these copyright concerns in digital artworks—such as watermarking and restrictive licensing of datasets—are inadequate. These approaches often fail to prevent infringement before it occurs or do not provide a mechanism for proper attribution once an artwork is generated. Watermarking can be obtrusive and easily removed, while restrictions on datasets limit the AI's learning potential and creativity, constraining the development of generative technologies. Furthermore, existing systems do not dynamically adapt to the evolving landscape of copyrights and artistic creations. There is a pressing need for an integrated solution that not only detects potential copyright infringements but also adapts the generative process to respect intellectual property proactively. Such a system would ensure compliance with copyright laws while supporting the sustainable and ethical use of generative AI in art, providing a balanced approach that protects creators' rights without stifling innovation.

The primary objective of this research is to introduce a system designed to integrate visual citations directly within AI-generated images. Our aim is to ensure that every digital artwork generated by AI technologies does so in compliance with copyright laws, thereby safeguarding the intellectual property rights of artists. This objective extends beyond mere compliance; it seeks to foster an environment of ethical use and acknowledgment within the digital creative industries. By automating the process of attributing artistic influences and copyrighted elements, the proposed system promises to transform how generative AI respects and interacts with existing artworks.

The proposed solution leverages an architecture that utilizes advanced machine learning techniques to analyze and extract

features and embeddings from images without the need for direct access to the images themselves. This approach circumvents the privacy and security issues associated with handling copyrighted materials directly. By employing a combination of neural network-driven embeddings and feature detection algorithms such as SIFT (Scale-Invariant Feature Transform) and RANSAC (Random Sample Consensus), our system can accurately identify and attribute the origins of artistic elements within newly generated images. Once these elements are identified, the system embeds a visual citation within the image metadata, detailing the original artist and the specific aspects of their work that influenced the generation process.

Moreover, the system is designed to be proactive in its compliance; it not only detects potential copyright infringements but also modifies the AI's generative process in real-time. This is achieved through dynamic prompt adjustment and image regeneration strategies that ensure the final output adheres to legal and ethical standards without compromising artistic integrity. This dual capability—to detect and adapt—sets our system apart from existing solutions, making it a significant advancement in the field of generative AI.

II. Related Work

A. Text-to-Image Technologies

The advent of text-to-image generation technologies such as DALL-E [1] , MidJourney [2] and Stable Diffusion [3] has catalyzed a revolution in digital art creation. By enabling intricate translation of textual prompts into visual content, these systems have blurred the boundaries between the written word and visual expression, facilitating new forms of creativity and challenging the conventional authorship in art.

B. CLIP (Contrastive Language–Image Pre-training)

CLIP exemplifies the synergistic potential of combining linguistic and visual elements [4]. By effectively mapping textual descriptions to corresponding images, CLIP has become a cornerstone for interpretive tasks, enabling AI to generate images that reflect complex narratives and abstract concepts conveyed in natural language.

C. Style Embedding

The concept of style embedding has further expanded the capabilities of generative AI, allowing for the replication and assimilation of distinct artistic styles within generated imagery [5]. This process encodes stylistic elements such as texture, brushwork, and color schemes, which can then be applied to novel content, fostering new artistic creations imbued with the essence of historical and contemporary art styles.

D. Vector-Based Embedding for Image Search

Moving beyond style alone, vector-based embeddings have revolutionized image search technologies [6]. By converting images into high-dimensional vectors, AI can perform nuanced searches that go beyond simple pattern recognition, allowing for sophisticated queries based on stylistic and compositional content. This vectorization process is essential for the quick retrieval of relevant images from extensive databases, crucial for establishing visual citations.

E. Bag of Visual Words

The 'bag of visual words' model translates the principles of text analysis into the visual domain [7]. This method breaks down images into feature descriptors that can be matched with a pre-defined visual vocabulary, enhancing the efficiency and accuracy of image search and comparison, which is especially beneficial for identifying and cataloging elements for visual citations.

F. Feature Detection and Matching Techniques (SIFT, RANSAC)

Algorithms like SIFT [8] and RANSAC [9] have become fundamental in identifying consistent features across varying conditions for images and robustly matching these features. The reliability of these techniques is important for determining the influence or derivation of elements within AI-generated images.

G. Non-Maximum Suppression (NMS)

In the final stage of visual citation identification, NMS plays a significant role in discerning distinct, non-overlapping elements [10]. By applying NMS, the system can effectively resolve instances where multiple similar features are detected, ensuring that each visual citation is unique and accurately localized within the generated image.

III. Approach

The overview of our approach to automating the generation of visual citations in using GenAI for text-to image generation is visually represented in the sequence diagram of Figure 1.

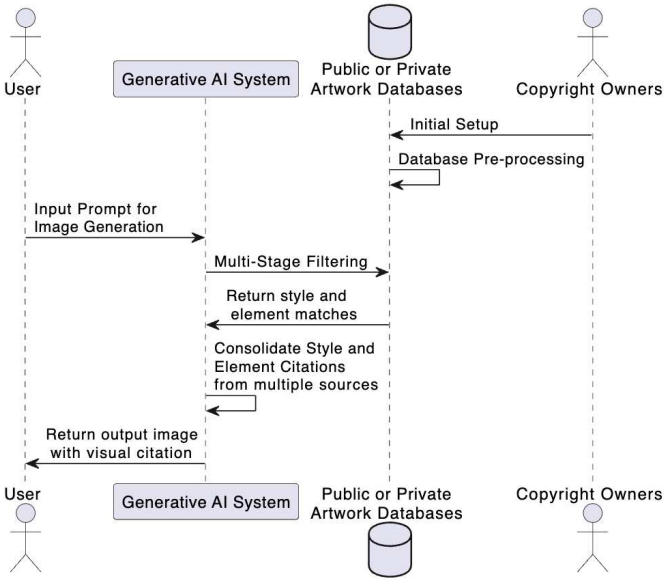

Fig. 1. Overview of the system diagram.

In our proposed system, the repositories of artworks are diverse

and range from openly accessible to restricted databases. These collections may be widely available to the public or secured behind access controls, such as subscription-based models. In addition, databases may include proprietary training datasets that are maintained—and potentially owned—by the entity operating the Generative AI (GenAI) system. Artworks in these databases may include metadata that includes the creator's name, the artwork's title, its distinctive style, and the terms under which it may be licensed, among other pertinent information.

A. Initial Setup and pre-processing

Copyright owners can upload their artwork to designated public or private databases, which may vary from openly accessible repositories to more restricted collections. Alongside the upload, owners articulate the licensing terms for each artwork. This sets clear guidelines on how the artwork can be used, specifying any restrictions or permissions that apply.

Copyright owners also categorize their artwork by style. This categorization is crucial as it aids the GenAI system in later identifying and matching similar styles in generated images.

The system then performs clustering of artworks based on the specified styles. These clusters may be grouped by specific artists or artistic movements, creating sets such as Monet-Impressionist or Author_A-Expressionist, which facilitates more nuanced matching.

For each artwork, content embeddings, SIFT features, and visual word histograms are computed and stored. This step ensures that each artwork can be effectively matched with generated images based on visual similarities.

The above steps are illustrated in Figure 2.

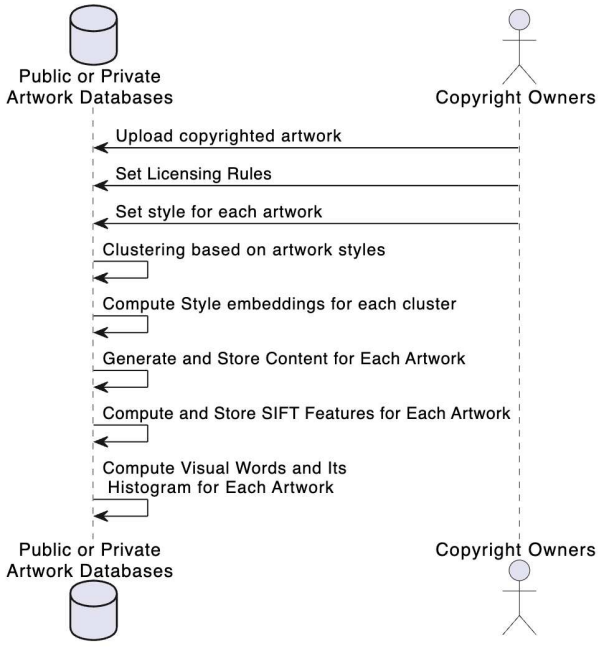

Fig. 2. Initial setup and pre-processing.

B. Multi-Stage Filtering and Matching Process

Figure 3 illustrates the sequence diagram of this multi-stage filtering and matching process.

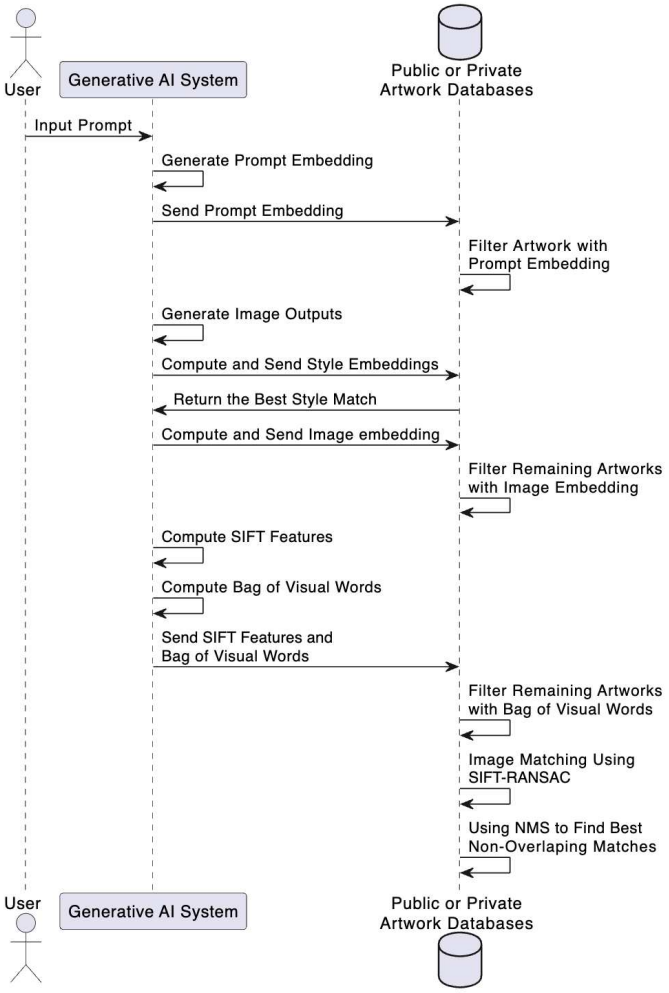

Fig. 3. Multi-stage filtering and matching

- User Interaction

Users interact with the GenAI system by inputting textual prompts, which the system will use to generate new artwork. The prompt acts as a seed for the creative process that follows.

- Initial filtering of databases based on CLIP

As the GenAI system receives a text prompt, it generates an embedding from this input using CLIP. This could be used for image generation, and at the same time, it can serve to filter the database for artworks that match the textual description. As the text-to-image takes time to generate output, this initial filtering of the databases can occur simultaneously and significantly reduce the candidate image numbers before the image is generated, thus increasing the efficiency of the system.

- Post-Generation Filtering Based on Image Embeddings

Once the image is generated, the GenAI system will first compute the style embeddings for the new image, using them to find the best matches in the databases which contain the embeddings of each style cluster.

At the same time, the GenAI system will also compute the image embeddings, which are used to filter the databases for artworks with similar content, using vector-based image search technology.

- Feature-based filtering and matching

While the databases are performing the post generation filtering, the GenAI system calculates visual features like SIFT and the bag of visual words for the generated image.

Using the bag of visual words, the databases filter the remaining artworks to those most visually similar to the generated image. This approach treats images as collections of local features or "visual words" extracted using algorithms like SIFT. The generated image is decomposed into its visual words, which are then compared against the visual vocabulary built from the database images. Images containing a high number of matching visual words are considered visually similar and retrieved for further matching.

The system then applies SIFT-RANSAC to conduct refined image matching, identifying specific locations and masks where features match, often represented as bounding boxes. SIFT (Scale-Invariant Feature Transform) detects distinctive local features like corners, blobs, and edges that are invariant to scaling, rotation, and other transformations. RANSAC (Random Sample Consensus) is used in conjunction with SIFT to robustly estimate the geometric transformation between the matching keypoints, filtering out outlier matches.

Matching keypoint locations are enclosed by bounding boxes, indicating corresponding regions of high visual similarity between the AI-generated and database images. This mapping enables localization of specific visual elements borrowed or derived from the database images within the new AI-generated composition. The bounding boxes can then be used for accurate visual attribution and citation of borrowed content.

- Find best matches with NMS

To find the best non-overlapping matches using non-maximum suppression (NMS), the following steps can be taken after obtaining the bounding boxes from the SIFT-RANSAC pipeline:

1. Sort Bounding Boxes: Sort all the bounding boxes detected across the database images in descending order based on a confidence score (e.g., number of inlier SIFT matches, transformation residual error).

2. Initialize Best Matches: Create an empty list to store the final set of best non-overlapping matches.

3. Iterate Through Sorted Boxes:

a. Take the highest confidence bounding box that is not already processed.

b. Add this bounding box to the Best Matches list.

c. Compare this box with the remaining unprocessed boxes:

If the Intersection over Union (IoU) with another box exceeds a threshold (e.g., 0.5), discard the lower confidence box as it significantly overlaps.

Else, keep the other box for further consideration.

4. Repeat step 3 until all boxes are processed.

The NMS procedure ensures that the final Best Matches list only contains high confidence bounding boxes that have minimal overlap with each other. This is important because:

a) It avoids reporting redundant visual matches from the same source region. b) It prioritizes matches with higher confidence/quality over lower ones when they overlap. c) It provides a concise set of distinct visual citations covering different borrowed elements.

The IoU threshold can be tuned based on how conservative the system should be in allowing overlap between matches. A higher threshold will result in fewer final matches but greater separation between them.

Additionally, NMS can be applied hierarchically by first obtaining non-overlapping matched regions, and then detecting non-overlapping keypoint matches within each region for precise visual attribution.

C. Consolidation and Generate Visual Citation

If the best style match across different databases exceeds a certain threshold, preset by the system, it will be cited as a visual citation for style similarity. This threshold can be determined empirically based on the desired level of confidence required to make a style attribution claim.

Non-maximum suppression (NMS) will be applied again to the match results aggregated from different databases. This ensures that the final set of cited style and element matches have minimal overlap, prioritizing the highest confidence, distinct matches. The NMS parameters like intersection-over-union threshold can be tuned based on how conservative the system should be in allowing overlap between visual citations.

The final citations will include:

- Style citation(s) indicating the artistic style(s) most strongly represented in the generated image, based on the top style database match(es).
- Element citations specifying localized regions within the image that exhibit strong visual similarity to database images. These will be denoted by bounding box coordinates around the matched regions.

Once the set of relevant visual citations is consolidated, the GenAI system will embed this citation data into the metadata of the generated image file. Typical metadata formats like Exif, XMP, IPTC can be leveraged to store the citation details.

The structure of the embedded citation data could be:

- Style Citation(s): List of matched styles with confidence scores. This section cites the overall style influences from artists or art movements on the AI-generated image. It includes the artist's name, the specific influence (e.g., technique, color palette), and a URL for

further reference. This type of citation acknowledges broader stylistic inspirations drawn from the works of noted artists.

- Element Citation(s): List of bounding boxes with coordinates, confidence, and source image ID. This category is more specific and includes citations for distinct elements within the image that closely resemble or are inspired by copyrighted artworks. It provides detailed information, including the artist, the title of the artwork, the similarity score, and crucially, a bounding box that precisely locates the influenced element within the image. The bounding box coordinates (x, y, width, height) define the area of the image where the similarity is noted, offering clear, visual acknowledgment of the specific inspiration.

This metadata embedding ensures that the critical visual attribution information stays attached to the image as it gets shared or transmitted.

The final image file, now seamlessly integrating the generated visual content alongside the corresponding citation metadata, is then returned to the user. Users can choose to extract and display the citation details as needed through metadata readers.

By consolidating visual citations and embedding them into the image metadata itself, this approach provides a robust, self-contained way to facilitate transparent artistic attribution for AI-generated imagery.

D. License Compliance and Image Regeneration

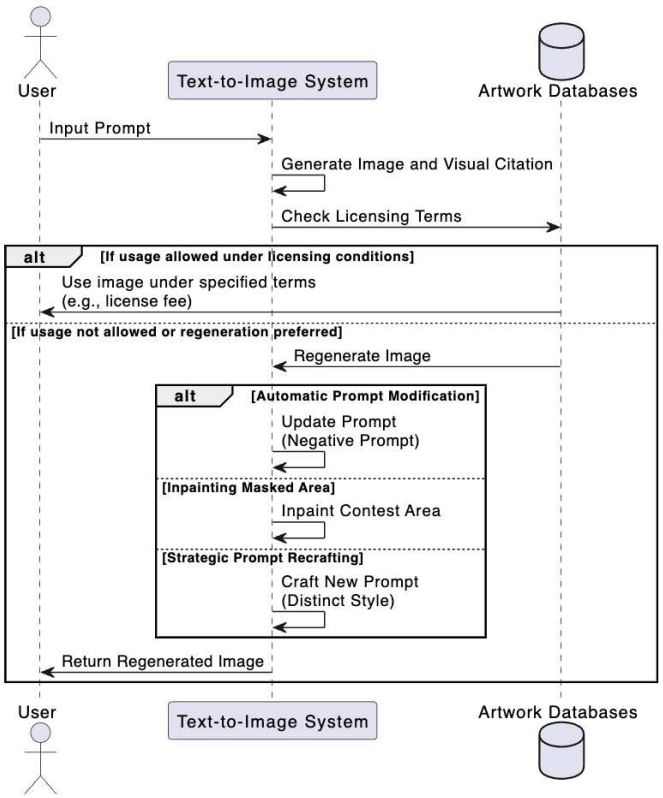

Fig. 4. License compliance and image regeneration.

Upon generating visual citations, the GenAI system can perform an additional application to handle instances where an AI-generated image exhibits significant element similarity to copyrighted artworks, based on the visual citations. This application can be integrated seamlessly with the text-to-image generation system, and the artists' predefined licensing conditions in the database, determining the permissibility and terms of use for each influenced element within the generated images. Figure 4 illustrates this application after visual citations are generated.

When a visual citation is identified, the system assesses it against the artist's licensing conditions stored within the database. If the conditions permit use under specific terms, such as paying a licensing fee, the image may be used accordingly.

If the usage is not permitted or desired by the user under the stipulated terms, the text-to-image generation system is tasked with regenerating the image. This can be achieved through various strategies to change the output image while retaining the creative intent of the original prompt.

The system can automatically update the prompt to exclude the cited visual element, utilizing a negative prompt that describes the element to be avoided [11]. This ensures that the regenerated image does not replicate the identified element. Alternatively, the user can manually modify the prompt.

For images where specific areas are cited, inpainting techniques [12] can be applied to regenerate only the contested sections of the image, seamlessly integrating new content to generate an output without this visual citation.

Strategic Prompt Recrafting: Leveraging the metadata associated with the matched artwork, including details about the author and style, the system can craft a new prompt. This prompt deliberately seeks inspiration from styles that are distinct from or inversely related to the identified artist [13], ensuring the new output diverges from the visually-cited style or elements.

By integrating licensing compliance checks and regeneration strategies, this application ensures that AI-generated images respect intellectual property rights and adhere to the stipulated terms of use defined by the original artists. It offers a comprehensive solution for responsible and ethical generation of visual content while protecting the rights of creators whose works have influenced the AI system.

IV. Experimental Results

To validate the proposed visual citation system, we conducted experiments on a curated dataset comprising AI-generated images and a database of artworks from various styles and genres. It is important to note that the results presented here are based on simulations and controlled experiments, as integrating and validating the system on a large-scale, real-world GenAI system with extensive databases is an ongoing effort.

Figure 5 illustrates an example of the embedded visual citation for an output image generated by our system. The metadata section clearly displays the identified visual citations, including the style match (Impressionist) and the localized element matches represented by bounding box coordinates and confidence scores.

In Figure 6, we present a sample output image from GenAI system on the left and an artwork from the database that has been identified as a source of visual influence on the right. The red bounding boxes are drawn based on the visual citation, highlighting the specific region within the database artwork that exhibits significant visual similarity to a corresponding area in the generated image.


```
{
  "VisualCitationMetadata": {
    "CreationDate": "2024-01-25",
    "GeneratedBy": "GenAI System v3.0",
    "Citations": [
      {
        "Type": "Style",
        "Details": [
          {
            "Artist": "Claude Monet",
            "Influence": "Impressionist technique and color palette",
            "ArtworkURL": "https://examplemuseum.org/monet-impressionism",
            "License": "Public Domain"
          }
        ]
      },
      {
        "Type": "Element",
        "Details": [
          {
            "Artist": "An Artist",
            "ArtworkTitle": "A Sunday Afternoon on the Island",
            "Year": "1984",
            "ArtworkURL": "https://examplegallery.org/sundayafternoon",
            "SimilarityScore": "90%",
            "BoundingBox": {
              "x": 100,
              "y": 150,
              "width": 200,
              "height": 100
            },
            "LicensingConditions": {
              "AllowedUse": "Non-commercial purposes only",
              "RequiresLicense": true,
              "LicenseFee": "Variable based on use case",
              "ContactInfo": "licensing@examplegallery.org"
            }
          }
        ]
      }
    ]
  }
}
```


Figure 5. An example of the embedded visual citation.

Figure 6. Left: GenAI output Image; Right: cited image from database. Red boxes are drawn based on the match.

V. Conclusion

The rise of powerful AI systems for generating visuals from text has unlocked new frontiers in digital creativity while raising concerns around artistic attribution and intellectual property rights. This work proposes a visual citation system that promotes transparency by identifying and attributing visual elements and styles within AI-generated images that may have been influenced by existing artworks. Through robust computer vision techniques, metadata embedding, and licensing compliance checks, the system enables responsible content generation while respecting artists' rights. Experimental results demonstrate the system's feasibility, though challenges remain in fully capturing nuanced artistic influences and establishing fair compensation models. As AI-driven art continues advancing, collaborative efforts across disciplines are crucial to develop ethical guidelines and technological frameworks that uphold artistic integrity while fostering innovative expression. Responsible deployment of such citation systems can navigate the complex landscape of AI creativity while preserving intellectual property principles.

References

[1] A. Radford et al., "DALL·E: Creating Images from Text," OpenAI Blog, Jan. 5, 2021. [Online]. Available: https://openai.com/blog/dall-e/

[2] MidJourney, "MidJourney: An Independent Research Lab," MidJourney. [Online]. Available: https://www.midjourney.com/

[3] Stability AI, "Stable Diffusion: A Latent Text-to-Image Diffusion Model," Stability AI, 2022. [Online]. Available: https://stability.ai/blog/stable-diffusion-public-release

[4] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," arXiv:2103.00020, 2021.

[5] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," arXiv:1812.04948, 2018.

[6] J. Johnson et al., "Image Retrieval using Scene Graphs," in CVPR, 2015, pp. 3668-3678.

[7] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in ICCV, 2003, pp. 1470-1477.

[8] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91-110, 2004.

[9] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," Commun. ACM, vol. 24, no. 6, pp. 381-395, 1981.

[10] A. Neubeck and L. Van Gool, "Efficient Non-Maximum Suppression," in ICPR, 2006, pp. 850-855.

[11] Guo, D., et al. "Generating Adversarial Examples with Adversarial Attacks." arXiv preprint arXiv:2103.06624 (2021).

[12] Nazeri, K., et al. "EdgeConnect: Structure Guided Image Inpainting using Edge Prediction." arXiv preprint arXiv:1901.00212 (2019).

[13] Tan, J., et al. "Text-guided Neural Artistic Style Transfer." arXiv preprint arXiv:2202.03057 (2022).