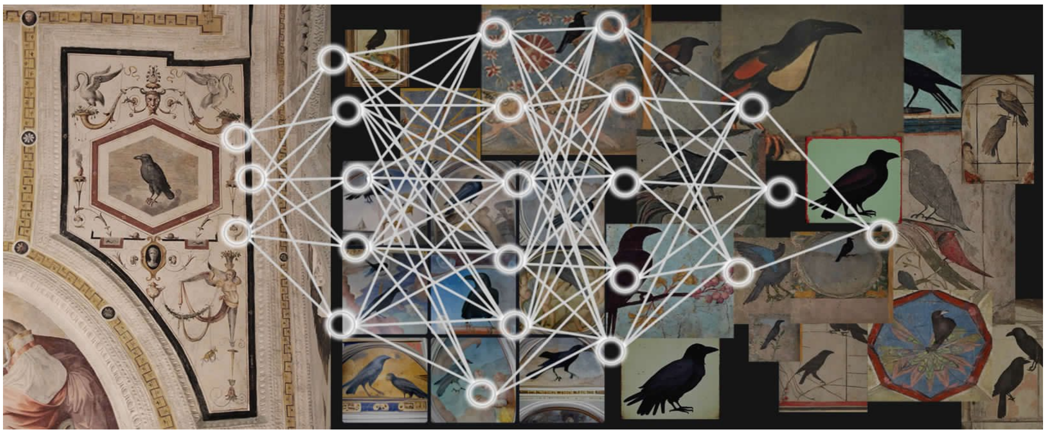
AI Text-To-Image for the Representation of Treaties Texts. The Case Study of *Le Vite* by Vasari

Assunta Pelliccio
Marco Saccucci
Virginia Miele

Abstract

A.I. Neural Networks (NN) constitute interesting support for the analysis of the text-image relationship, which, since ancient times, has stimulated essential reflections on the resulting ontological value. Among these, the Text-To-Image (TTI), designed to learn the transformation of a text into an image, is the most suitable to make an innovative contribution to this complex investigation. In the literature, the applications of the TTI are many and concern both the learning of the network and its training in constructing images starting from *ad hoc* descriptions. This paper instead investigates the training of the NN in transforming the descriptions of treatise texts into images. This operation is complex because it requires knowledge of the two different constructive systems of the 'text architecture', one governed by the grammar of the sentences and the other by the grammar of the text, which introduces a multitude of variables that are not easily decodable by neural networks. The contribution presents the results of the first phase of the research with the development of a procedure based on the interception of sentences and paradigmatic relationships to be decoded for the formation of the network. The experimentation is conducted on the treatise *Le Vite* by Giorgio Vasari, which contains an accurate description of the grotesques, created by the author in the church of Sant'Anna dei Lombardi in Naples for which graphic and geometric analyses have already been carried out.

Keywords
AI, Neural Networks, Text-To-Image, Representation, Treaties



Symbolic representation of the network connection between real images and AI. Graphic elaboration by the authors.

doi.org/10.3280/oa-1016-c380

Introduction

The oldest relationship between words and images dates back to the 2nd century when Hermogenes introduced ékphrasis, as descriptive speech which puts the object before the eyes. In particular, in modern thought, ékphrasis finds its most natural place in the verbal description of art, tracing "figurative and literary *tòpoi* (...) from the descriptions of the great masterpieces" [Albanese 2013, p. 2]. Over the centuries, analysing the delicate intertwining of text and images has aroused essential reflections on the resulting ontological value. All images, graphic, optical and perceptive, more objective, mental and verbal, and less objective, assume the same value according to Wittgenstein's theory of images [Wittgenstein 1922] and W.J.T. Mitchell [Mitchell 1995]. If the signs of language are "images of what they represent" [Wittgenstein 1922, p. 3], the image itself is nothing more than a graphic sign of the object it represents [Mitchell 1995]. Images, in the broad sense of text, image or idea, are just semblances of the real world expressed differently. The investigation into the relationship between 'image and text' finds, in the opinion of the authors, the most exhaustive outcome in the triple 'typographic convention' proposed by Mitchell himself [Mitchell 1995]: the 'image/text', with the slash, highlights a problematic void, a fracture in the representation of their relationship; 'imagetext', without glyphs, designates composite and synthetic works (or concepts), which combine image and text; 'image-text', with the insertion of the hyphen, underlines the relationship between what is visual and verbal. The latter relationship is investigated in this paper, which presents the first reflections on the topic with the awareness that the visual-verbal relationship is only a fractal of the more complex 'imagetext', and, therefore, also requires the use of different sensory channels (eye and ears), the definition of semiotic functions (iconic aspect and arbitrariness of the symbol), the identification of the cognitive modality (space-time) and the application of operative codes (analogue, digital, A.I.). The tool used is the artificial intelligence (A.I.) of neural networks, i.e., mathematical-IT calculation models, which, inspired by the biological functioning of the human brain, can build processes based on information interconnections. The science that deals with the definition and management of interconnections is called 'connectionism' and is based on the Parallel Distributed Processing (PDP) of information.

In summary, at the basis of the artificial neural network, there are algorithms which, in an 'adaptive' way, can connect external data (training) with internal design information (learning) of the network itself, modifying its structure (nodes and interconnections or arches) from time to time. The experimentation took place on the treatise *Le Vite* by Giorgio Vasari, which contains an interesting description of the grotesque iconographic apparatus he created in the church of Sant'Anna dei Lombardi in Naples. Among the numerous neural networks offered by the technological market, Text-To-Image (TTI), i.e., networks capable of transforming a text, formulated in natural language, into an image, are the most suitable for the research. The choice is due to the possibility of training the NN thanks to comparing texts and images. In fact, from the images, it is possible to acquire significant data thanks to the geometric and graphic analysis previously performed through a digital photogrammetric survey. [Miele et al. 2022].

A.I. neural networks. State of the art

The use of A.I. Neural Networks was born in the early 40s of the last century with the demonstration of the implementation of the algorithm underlying the Turing machine [McCulloch and W. Pitts 1943]. For several years, the term neural networks (NN) has included biological and artificial without distinction. The NNs, whose application involves many sectors of the soft and hard sciences, are intelligent systems capable of artificially reproducing the performance of an expert person in a specific domain of knowledge or field of activity and, therefore, capable of identifying the solution for any complex problem. The functionality of NNs is organised on three levels, involving thousands of nodes and tens of thousands of connections (arcs). For the various levels, the nodes and the arches have the task of receiving and processing the incoming signals and adapting them to the requests coming from the next level of the network. For this process to be efficient, it is necessary to 'train' the neural networks, that is, to guide them in their behaviour when faced with a solution to a problem of any kind. The main characteristics of neural networks are the complexity of the structure, for which it is almost impossible to trace the logical procedure that led to a specific solution because they 'deduce' rules and activities automatically through learning and not through the deduction of reasoning [Perlovsky 2001]. Among the artificial intelligence technologies, the Text-To-Image (TTI) can generate quality images starting from a descriptive text in 'natural language', which in computer science translates as the ability of algorithms to understand the text and the words spoken to humans nowadays. For this, they represent innovative support for the definition of a systematic analysis of the relationship between image and text. The applications of the TTI in the literature are many. Generally, they concern the quality and control of the generation of images based on the description of natural language [Li et al. 2019] or the definition of desired attributes while preserving content irrelevant to the text [Li et al. 2020]. In S. Reed [Reed et al. 2016], GAN adversarial networks capable of translating visual concepts of characters into pixels generating qualitatively valid images have been designed and developed. The transformation of a description of a historical text, such as a treatise, into an image inevitably requires a close relationship with linguistics [Dardano 2017] is a procedure so far to be studied. Among the open-source NNs, Stable Diffusion was the most suitable for the research. It is a Deep Learning model released under an Open-Source licence by the startup StabilityAI. The model is built on a GAN network and, therefore, is deeply generative thanks to the two opposing networks inside, which generate the so-called 'contradictory'. Trained on nearly 6 billion pairs of images and captions taken from the public LAION-5B dataset, sorted by language, resolution and watermark quality. The network generates new images using the text-image sampling script known as 'txt2img', a text prompt, and assorted optional parameters ranging from sampling to image size and randomness agent values (seeds). The system can also randomise the seed to explore different generated outputs and uses so-called 'negative hints', which eliminate or add elements necessary to improve the image's realistic quality.

Text-To-Image procedure for the recognition of the figurative and literary tòpoi of the treatises

Translating written text into an image is a complex operation. In the text, "two different constructive systems coexist the one that regulates the grammar of the sentence, based on substantially obligatory rules and the one that regulates the grammar of the text, based on choices made by the speaker taking into account criteria of balance and good training" [Palermo 2016, p. 217]. Punctuation also plays a fundamental role because it acts as a textual guide helping to identify the boundaries between units. In summary, the complexity of reading a text is mainly due to the vision of language as a modular system that interconnects the linguistic and extralinguistic reality in the productive and receptive phase [Palermo 2016]. Modulating the language of the text to communicate to artificial intelligence the most suitable description for the reconstruction of an image, therefore, requires the identification of a multitude of variables that are not easily decodable, above all because neural networks, as previously mentioned, do not deduce reasoning. The decoding of the variables that contribute to the most congruous definition of the text-image relationship is the goal of future research developments, for which the training of the network to the elements of the visual language is already envisaged, such as balance, configuration, shape, development, space, light, colour, movement, dynamics, expression, which help translate the 'architecture of the text'. In this first phase of activity, the research aims to focus on the lexical semantics to intercept the sentences and the paradigmatic relationships in a sentence to be decoded, in a simplified way, for a neural network. For this purpose, a procedure divided into 5 phases was defined (fig. 1) The first phase – 'text' – identifies the portion of text that contains the description of the image to be created through the A.I. This operation requires careful linguistic analysis, con-

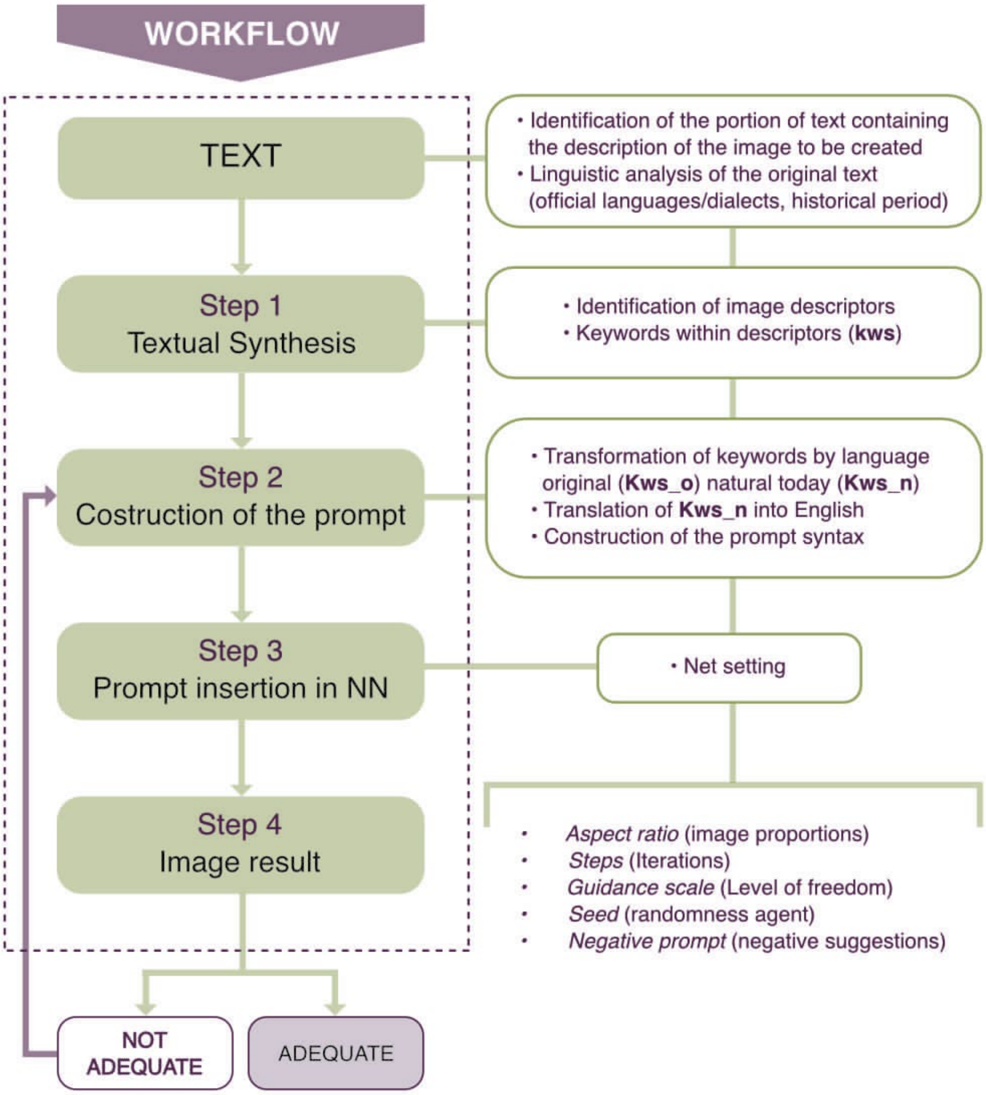

WORKFLOW
TEXT
Step 1
Textual Synthesis
Step 2
Costruction of the prompt
Step 3
Prompt insertion in NN
Step 4
Image result
NOT ADEQUATE
ADEQUATE
• Identification of the portion of text containing the description of the image to be created
• Linguistic analysis of the original text (official languages/dialects, historical period)
• Identification of image descriptors
• Keywords within descriptors (kws)
• Transformation of keywords by language original (Kws_o) natural today (Kws_n)
• Translation of Kws_n into English
• Construction of the prompt syntax
• Net setting
• Aspect ratio (image proportions)
• Steps (Iterations)
• Guidance scale (Level of freedom)
• Seed (randomness agent)
• Negative prompt (negative suggestions)

Fig. 1. Workflow of the procedure. The graph shows the interdisciplinarity of the image-text relationship at the basis of visual culture. The training operations of the network are iterable according to the result obtained. Elaboration by the authors.

textualised to the historical period of the text to understand if the semantic units used by the author can be assimilated to the 'natural language' previously defined. The second phase – 'textual synthesis' – isolates a finite number of words (nodes) that describe the image. They can be thought of as keywords that categorise image elements. The third phase – 'construction of the prompt' – modulates the language, which must be transformed from the original (the treatises) into natural language. This operation is essential to obtain a result as close as possible to the real one. Furthermore, the NNs have been trained in English, so the untranslatability of some words from one language to another plays an important role. The transformation from the mother tongue (e.g., vernacular) to the natural one and, therefore, to the English language generates various degrees of arbitrariness, which must be controlled during the formation of the network. Once the words have been decoded in this phase, the syntax (arcs) construction is necessary to identify the most suitable morphological signals. The process generally requires several attempts before getting the optimal solution. The fourth phase – 'prompt insertion in NN' – consists of processing the prompt obtained in the previous phases using the trained neural network. In this phase, the machine must set the operating parameters, such as:

- 'Aspect ratio' is the width to the height of an image, expressed as two numbers separated by colons, such as 16:9 or 1:1. For the x:y aspect ratio, the image is x units wide and y units high;
- 'Steps' are the number of steps required by A.I. during creation. The more passes, the better the overall image quality;

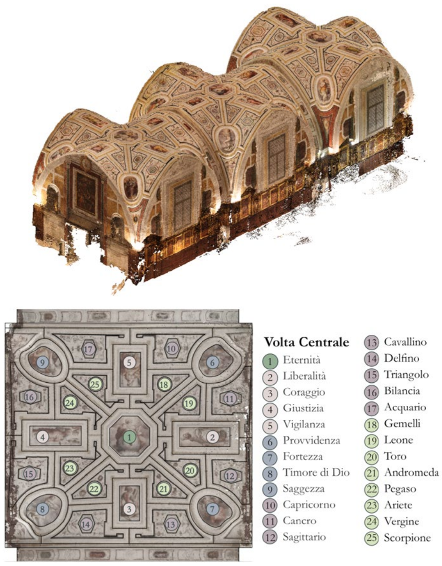
- 'Guidance scale' is the level of freedom (or precision) attributed to the A.I. in the training phase, starting at the prompt. Higher levels of values force the A.I. automatically follow the request more rigorously;
- 'Seed or agent of randomness' for the AI. The same seed and the same prompt will produce the same picture. Reusing the same seed with different suggestions can produce a consistent style;
- 'Negative prompt' is the description of elements not contained in the image. The A.I. that way does not use concepts/terms listed in the negative prompt.

The fifth stage – 'image result' – shows the result of the network processing. It is subject to the operator's judgement, who may deem the result of the treatment adequate or not. In the latter case, the process must be repeated from the third stage. The prompt must be reformulated by modifying the syntax and setting the parameters of the neural net.
The procedure was applied to the case study of Vasari's grotesques in the sacristy of Sant'Anna dei Lombardi in Naples, described by the author in his treatise *Le Vite*.

Case study. The images of the grotesques from Vasari's treatise, Text-To-Image procedure

In 1544 Giorgio Vasari from Arezzo, a Mannerist painter and historian, was commissioned to decorate the vault of the Sacristy of Sant'Anna dei Lombardi in Naples and did so by inserting grotesque decorations. The interesting aspect of research is the description of these decorations in Vasari's treatise, *Le Vite*, considered the first modern treatise in art history. An accurate analysis of the decorative apparatus can be found in [Miele et al. 2022] (fig. 2).
The possibility of comparing the description of the text on the grotesque apparatus with the images of the actual decorations, acquired with a digital photogrammetric survey, has allowed the application of the previously described procedure. The operations are carried out on the vault, which has anthropomorphic and phytomorphic decorations. Once the semantic units in the description of the treatise were identified, the neural networks were applied for the production of images from the text. The comparison between the original text, written in the sixteenth-century language, with the natural language on the first two images has generated only one level of arbitrariness in the original-natural-English language flow. In the third image, on the other hand, based on words such as *scarpelloni*, which are currently obsolete, the level of arbitrariness increases (fig. 3).
The first elaboration concerns the image of the 'horse with legs made of leaves'. Various attempts were necessary for lexical and syntactic choice, modifying the prompt several times (fig.4). The application pointed out some network limitations. Notably, the Stable Diffusion neural network trained with DB LAION exhibits failure rates of 25% on animal limb generation. In order to obtain a result more similar to reality, we finally proceeded with impainting, i.e., partial training of the network with specific reference images chosen by the operator (fig. 5).
For the second sentence of the treatise (man-legs-crane), the A.I. handles the text as shown in figure 6. In particular, many images are inadequate as the network censors various proposals for the non-explicit request for nudity. Furthermore, it is observed that the images are far from Vasari's representations without a forced network training and inpainting procedure.
Several attempts have been made for the third sentence of the treatise, three of which are shown in figure 7. The network has substantial limitations on managing words that express quantities (Infinities), but above all, in managing multiple degrees of arbitrariness, which distances the image from the desired results.
If the level of arbitrariness is 1, depending on the operator's interpretation, the image is very close to the real one. By increasing the levels of arbitrariness, the final image is very distant from the real one.

Fig. 2. Digital photogrammetric model of the sacristy and description of the iconographic apparatus of the central vault. Graphic elaboration by the authors.

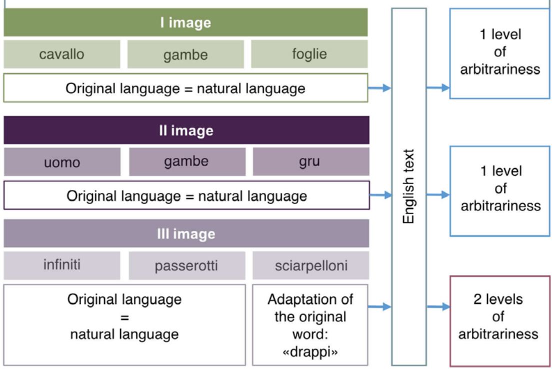


Volta Centrale

1	Eternità	13	Cavallino
2	Liberalità	14	Delfino
3	Coraggio	15	Triangolo
4	Giustizia	16	Bilancia
5	Vigilanza	17	Acquario
6	Provvidenza	18	Gemelli
7	Fortezza	19	Leone
8	Timore di Dio	20	Toro
9	Saggezza	21	Andromeda
10	Capricorno	22	Pegaso
11	Cancro	23	Ariete
12	Sagittario	24	Vergine
		25	Scorpione

Fig. 3. The procedure for choosing semantic units (kw_o), comparing the original language and the natural one (kw_n), and translating words into English. Each pass generates levels of arbitrariness that affect network training. Elaboration by the authors.

*"[...] Le grottesche sono una spezie di pittura licenziose e ridicole molto, fatte dagl'antichi per ornamenti di vani, dove in alcuni luoghi non stava bene altro che cose in aria; per il che facevano in quelle tutte sconciature di monstri per strattezza della natura e per gricciolo e ghiribizzo degli artefici, i quali fanno in quelle cose senza alcuna regola, apiccando a un sottilissimo filo un peso che non si può reggere, a un **cavallo** le **gambe** di **foglie**, a un **uomo** le **gambe** di **gru**, et **infiniti sciarpelloni** e **passerotti**; e chi più stranamente se gli immaginava, quello era tenuto più valente [...]". (Vasari, 1986, p. 73)*



I image

cavallo	gambe	foglie

Original language = natural language

II image

uomo	gambe	gru

Original language = natural language

III image

infiniti	passerotti	sciarpelloni

Original language = natural language

Adaptation of the original word: «drappi»

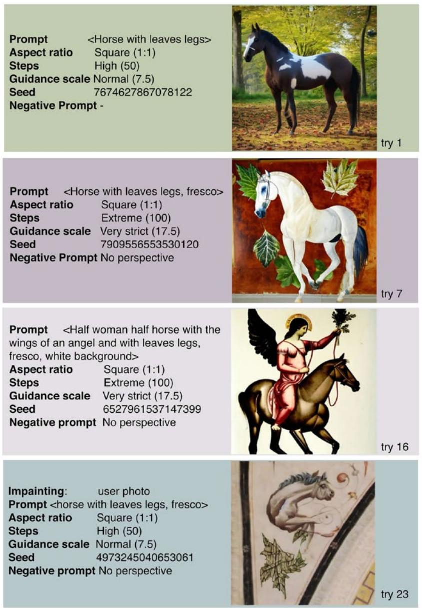
English text

1 level of arbitrariness

1 level of arbitrariness

2 levels of arbitrariness

Fig. 4. Processing attempts of the first image, 'on a horse the legs of leaves', with the AI Open-Source Stable Diffusion, modifying the network setting and the prompt. Elaboration by the authors.



Prompt <Horse with leaves legs>
Aspect ratio Square (1:1)
Steps High (50)
Guidance scale Normal (7.5)
Seed 7674627867078122
Negative Prompt -

try 1

Prompt <Horse with leaves legs, fresco>
Aspect ratio Square (1:1)
Steps Extreme (100)
Guidance scale Very strict (17.5)
Seed 7909556553530120
Negative Prompt No perspective

try 7

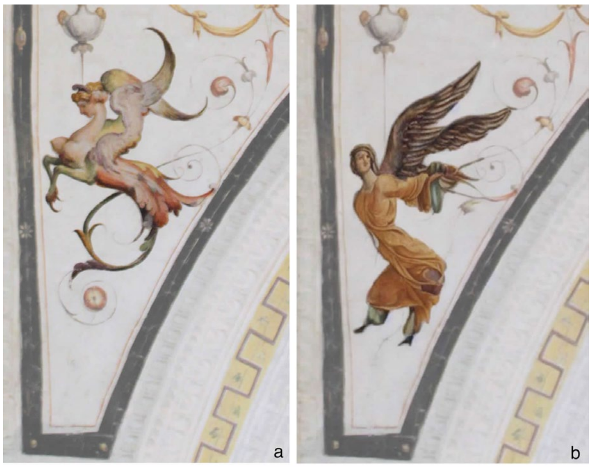
Prompt <Half woman half horse with the wings of an angel and with leaves legs, fresco, white background>
Aspect ratio Square (1:1)
Steps Extreme (100)
Guidance scale Very strict (17.5)
Seed 6527961537147399
Negative prompt No perspective

try 16

Impainting: user photo
Prompt <horse with leaves legs, fresco>
Aspect ratio Square (1:1)
Steps High (50)
Guidance scale Normal (7.5)
Seed 4973245040653061
Negative prompt No perspective

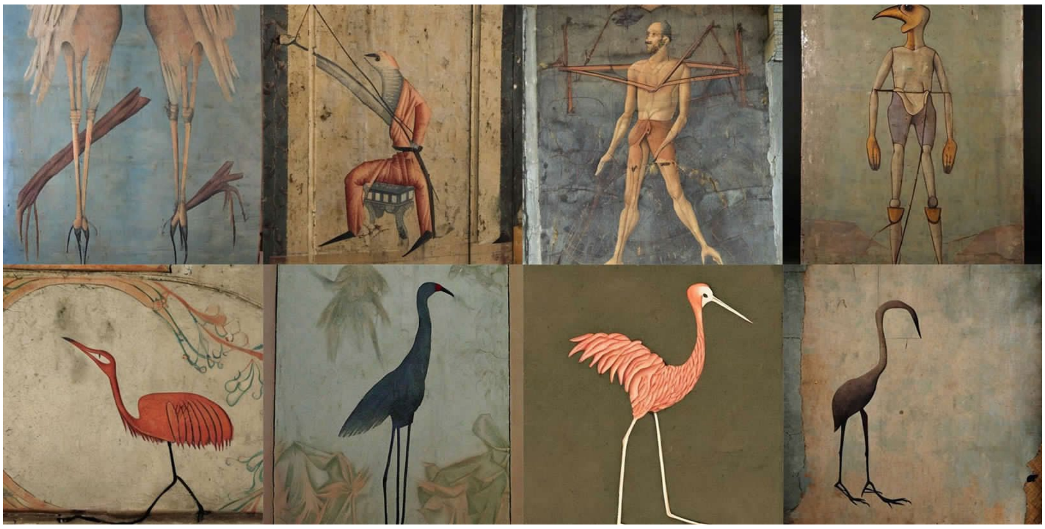
try 23

Fig. 5. Adequate result: a) Original photo entered by the operator in the network; b) 'Impainting' with AI processing. Elaboration by the authors.



Prompt <Woman/horse leaves legs with the wings of an angel, fresco>
Aspect ratio Tablet (2:3)
Steps High (50)
Guidance scale Normal (7.5)
Seed 5431238313039876
Negative prompt No perspective

Fig. 6. AI elaborations on the second chosen sentence of the treatise. Graphic elaboration by the authors.



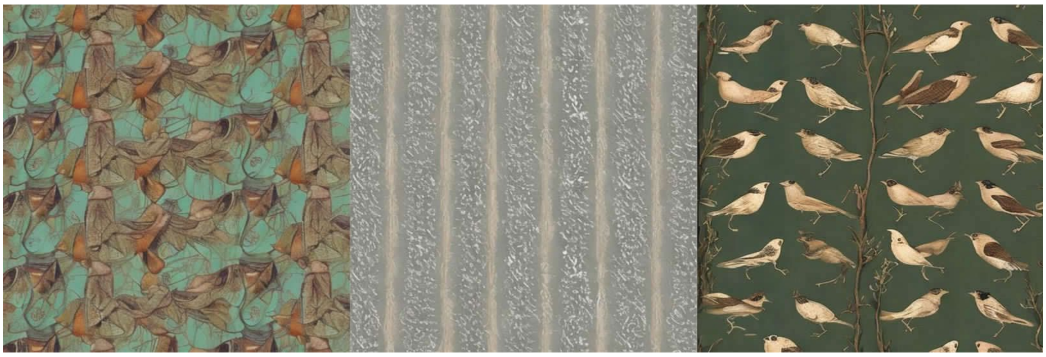


Fig. 7. AI elaborations on the third chosen sentence of the treatise, with multiple levels of arbitrariness. Graphic elaboration by the authors.

Conclusion

Since ancient times, the debate on the relationship between text and image has continuously developed exciting themes, which have taken on different connotations depending on the knowledge developed and the tools used. Currently, the study of the relationship converges in the discipline called visual culture, defined by Mitchell as 'interdisciplinary' or rather 'indiscipline', thus including numerous specificities, all of which are fundamental for understanding the connections between texts and images. Starting from this assumption, the paper presents the first results of a new analysis of the image-text relationship based on using the most recent AI-neural network technology. The goal is to investigate network training to create images derived from treatises, which require in-depth knowledge of 'text architecture'. The first results highlight the numerous criticalities linked to the historicity of language and the imprinting of the image construction. They also suggest that the process produces much more realistic results if the components of the visual language, which determine the compositional structure of the image, are also included in the formation of the network. Future research developments want to investigate this latter aspect.

References

Albanese A. (2013). Michele Cometa, La scrittura delle immagini. Letteratura e cultura visuale. In *Between*, Vol. III, No. 5.

Bellosi L., Rossi A. (Eds.). (1986). *Vasari G. Le vite de' più eccellenti architetti, pittori, et scultori italiani, da Cimabue insino a' tempi nostri. Nell'edizione per i tipi di Lorenzo Torrentino, Firenze 1550*. Turin: Einaudi.

Cometa M. (2012). *La scrittura delle immagini: letteratura e cultura visuale*. Milan: Raffaello Cortina.

Dardano M. (2017). *La prosa del Cinquecento: studi sulla sintassi e la testualità*. Pisa: Fabrizio Serra.

Li B., Qi X., Lukasiewicz T., Torr P. (2019). Controllable text-to-image generation. In *Advances in Neural Information Processing Systems*, No. 32, pp. 1-11.

Li B., Qi X., Lukasiewicz T., Torr P. H. (2020). Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7880-7889. IEEE

McCulloch W. S., Pitts W. (1943). A logical calculus of the ideas immanent in nervous activity. In *The bulletin of mathematical biophysics*, No. 5, pp. 115-133.

Miele V., Saccucci M., Pelliccio A. (2022). Le grottesche di Sant'Anna dei Lombardi a Napoli. Analisi geometrica dell'apparato decorativo nello spazio architettonico. In E. Cicalò, F. Savini, I. Trizio (Eds.). *Linguaggi Grafici. Decorazione*, pp. 390-413. Publica.

Mitchell W. T., Mitchell W.J.T. (1995). *Picture theory: Essays on verbal and visual representation*. Chicago: University of Chicago Press.

Palermo M. (2016). La dimensione testuale. In S. Lubello (Ed.). *Manuale di linguistica italiana*, pp. 222-241. Berlin, Boston: De Gruyter.

Perlovsky L. I. (2001). *Neural networks and intellect: Using model-based concepts*. London: Oxford University Press.

Reed S., Akata Z., Yan X., Logeswaran L., Schiele B., Lee H. (2016). Generative adversarial text to image synthesis. In *International conference on machine learning*, pp. 1060-1069. PMLR.

Wittgenstein L. (1922). *Tractatus Logico-Philosophicus*. London: Kegan, Trench, Trubner & Co.

Authors
Assunta Pelliccio, Università degli Studi di Cassino e del Lazio Meridionale, pelliccio@unicas.it
Marco Saccucci, Università degli Studi di Cassino e del Lazio Meridionale, m.saccucci@unicas.it
Virginia Miele, Silesian University of Technology (Poland), vmiele@polsl.pl

To cite this chapter: Pelliccio Assunta, Saccucci Marco, Miele Virginia (2023). AI Text-To-Image for the Representation of Treaties Texts. The Case Study of Le Vite by Vasari. In Cannella M., Garozzo A., Morena S. (Eds.). *Transizioni. Atti del 44° Convegno Internazionale dei Docenti delle Discipline della Rappresentazione/Transitions. Proceedings of the 44th International Conference of Representation Disciplines Teachers*. Milano: FrancoAngeli, pp. 1824-1831.

Isbn 9788835155119