STUDIUM EDUCATIONIS
XXV - 2 - December 2024 | ISSN 2035-844X
DOI: 10.7346/SE-022024-08

Available on-line at
https://ojs.pensamultimedia.it/index.php/studium

OPEN ACCESS

Ricevuto: 30/09/2024
Accettato: 15/11/2024
Pubblicato: 30/12/2024

FOCUS

# From Images to Narrations: The InVisible Project's AI-Powered Image Captioning for Transforming Architectural Learning and Supporting Blind Students

**Fotis P. Kalaganis, Kostas Georgiadis, Nikos Lazaridis, Spiros Nikolopoulos, Ioannis Kompatsiaris**
Centre for Research and Technology Hellas, Information Technologies Institute, MKlab, Thermi-Thessaloniki, 57001, Greece

# Dalle Immagini alle Narrazioni: Il Progetto Invisibile e la Descrizione delle Immagini con AI per Trasformare l'Apprendimento Architettonico e Supportare Studenti Non Vedenti

## Abstract

*Architecture involves the art and science of designing buildings and structures, blending aesthetic and technical expertise. It involves the understanding of both design principles and historical/cultural contexts of architectural styles. However, this poses challenges for visually impaired individuals, as much of the information is perceived visually.*

*To this end, the InVisible project has realized an image captioning system integrated into architecture curricula. Such an approach allows us to provide contextualized narrations of architectural images. Tailored to the field of architecture, these captions enrich educational resources like image databases and interactive learning tools, benefiting both students and educators by making content more accessible. Ultimately, InVisible's approach enhances learning for all and promotes inclusivity in architectural education.*

**Keywords**
**Computer vision, image captioning, blind, architecture, accessibility**

L'architettura combina arte e scienza nella progettazione di edifici e strutture, fondendo competenze estetiche e tecniche. Richiede la comprensione dei principi di design e dei contesti storici e culturali degli stili architettonici. Tuttavia, ciò rappresenta una sfida per le persone non vedenti, poiché gran parte delle informazioni è percepita visivamente.

A tal fine, il progetto InVisible ha sviluppato un sistema di didascalie automatiche integrato nei corsi di architettura. Questo approccio fornisce narrazioni contestualizzate per immagini architettoniche. Adattate al campo, le didascalie arricchiscono risorse educative come database di immagini e strumenti di apprendimento interattivi, rendendo i contenuti più accessibili per studenti e insegnanti. L'approccio di InVisible migliora l'apprendimento per tutti e promuove l'inclusività nell'educazione architettonica.

**Parole chiave**
**Visione artificiale, didascalia delle immagini, cieco, architettura, accessibilità**

## 1. Introduction

Architecture, as both an art and a science, requires a delicate balance between aesthetics and technical expertise. It involves the design and construction of physical structures, with a focus on functionality, visual appeal, and cultural significance. Architectural education plays a pivotal role in equipping students with the skills to navigate these complexities. Students must learn to create designs that are both practical and visually compelling, interpret intricate blueprints, and appreciate the historical and cultural contexts underlying various architectural styles. However, a significant challenge arises when this visually dependent field intersects with accessibility needs, particularly for visually impaired or blind individuals, as much of the essential information is predominantly visual.

One of the key barriers in architectural education for visually impaired individuals is the reliance on visual elements, such as drawings, diagrams, and blueprints, to convey architectural concepts. These individuals often face difficulties in accessing critical information, which in turn limits their ability to engage fully with the material. Despite advancements in inclusive education, there remains a notable gap in how architectural content is presented to visually impaired learners.

In response to this challenge, image captioning systems have emerged as a transformative tool, revolutionizing how visual content is made accessible. The domain has evolved from basic object recognition models (Georgiadis, 2021) to sophisticated algorithms capable of generating detailed and contextually relevant descriptions of complex scenes. Early approaches focused primarily on identifying individual objects within images, using standard object detection techniques. However, as the field matured (Lazaridis, 2024), the need for systems capable of understanding and describing entire scenes became more apparent. One of the early breakthroughs in this domain was the Show and Tell model by Vinyals et al. (Vinyals, 2015), which introduced the concept of using a Convolutional Neural Network (CNN) for image feature extraction and a Recurrent Neural Network, specifically Long Short-Term Memory, for generating textual descriptions of images. This model demonstrated that deep neural networks could generate coherent captions by learning from large datasets like MSCOCO (Lin, 2014). The Show, Attend and Tell model (Xu, 2015) further improved this approach via attention mechanisms, enabling the system to focus on specific regions of an image while generating corresponding descriptive text. Another significant milestone was the development of the Places365 dataset (Zhou, 2017). This dataset, which contains 1.8 million images across 365 different scene categories, has been instrumental in advancing scene understanding for various tasks. Places365 has been widely used to train CNNs that are highly effective at predicting scene categories, environmental conditions (e.g., indoor/outdoor), and scene attributes (e.g., lighting conditions), which are essential for generating detailed, context-aware descriptions of architectural spaces. Recent advances in Vision-Language models have significantly improved the quality and contextual relevance of generated captions. Models like CLIP (Radford, 2021) have introduced a new paradigm by training visual and textual encoders together, allowing these systems to align images and descriptions without requiring explicit labels for every training sample. This multimodal approach enables better generalization and contextual understanding, making these models highly effective for tasks such as image captioning in diverse domains. Natural language generation models have also played a critical role in enhancing the quality of image captioning. T5 (Raffel, 2020) and similar models have proven highly effective in generating coherent, contextually appropriate text when combined with image features extracted by CNNs. These models leverage transfer learning and are pre-trained on vast amounts of textual data, enabling them to generate descriptions that are not only accurate but also fluent and natural.

Despite the advancements in image captioning technologies, several challenges remain. One persistent issue is the contextual accuracy of the generated captions. While current systems can describe basic scenes and objects with reasonable accuracy, they often struggle with more complex scenarios where subtle distinctions or abstract concepts are involved. Architecture is one such typical example where more sophisticated models capable of understanding domain-specific knowledge (e.g. differentiating between architectural styles or structures) are required. These tasks demand advanced models capable of interpreting domain-specific knowledge, ensuring a deeper understanding of the architectural context, which is critical for accurately reflecting the nuanced details that characterize different architectural elements.

In this direction and aiming to address the aforementioned challenge within the architectural context, this paper presents an advanced Image Captioning service, developed as part of the InVisible Project (https://www.invisible-eplus.com). This service was designed to generate comprehensive textual descriptions of architectural interest (e.g. architectural elements, buildings, etc.) from images, specifically tailored to provide accessible descriptions for visually impaired users. The service integrates advanced machine learning models, Natural Language Processing techniques, and appropriately curated datasets to deliver rich, context-aware descriptions of architectural nature. The generated content covers various architectural aspects and is provided in accessible formats, such as detailed text descriptions or audio guides that can eventually lead to a more comprehensive understanding of architectural concepts and improved access to architectural education for the visually impaired individuals. Finally, the paper presents the insights gained throughout the development process, which was conducted in close collaboration with visually impaired individuals, instructors, and other key stakeholders. This collaborative approach provided valuable feedback and highlighted critical considerations in ensuring the system's effectiveness, usability, and inclusivity. The lessons learned from this process emphasize the importance of user-centered design and interdisciplinary input in creating a solution that not only meets accessibility needs but also enhances educational outcomes in architectural studies and beyond.

## 2. Methodology

### 2.1 *Datasets*

#### 2.1.1 *Places 365 Dataset*

The Places365 dataset (Zhou, 2017) is a large-scale benchmark designed for scene recognition in computer vision. Created as part of the MIT Places project, it includes a diverse collection of images representing various environments, both indoors and outdoors. The dataset consists of approximately 1.8 million training images, along with 36,500 validation images and 328,500 test images. It spans 365 different scene categories, capturing a wide range of settings, from natural landscapes like desert, lawn, and forests to man-made locations such as castles, churches (both indoor and outdoor), and catacombs.

Each image in the dataset is annotated with a label that describes the overall scene rather than focusing on specific objects. Places365 is widely used in machine learning tasks such as scene recognition, image classification, and transfer learning. Its comprehensive labeling of scenes allows researchers and developers to train deep learning models to understand and classify complex environments. Places365 has become an essential resource for advancing research in high-level visual recognition and scene classification.

#### 2.1.2 *Architectural Heritage Elements Dataset*

The Architectural Heritage Elements Dataset (Jose, 2017) is a specialized image dataset designed to support the recognition and classification of architectural elements from historical and cultural heritage sites. It focuses on the detailed features of architectural structures such as facades, columns, arches, windows, and decorative elements commonly found in heritage buildings and monuments.

This dataset contains annotated images of various architectural elements, providing both visual and semantic information about these components. It is designed to assist in tasks like automatic detection, classification, and preservation of architectural heritage, supporting efforts in areas such as heritage conservation, restoration, and digital archiving.

The dataset is particularly useful for training machine learning models that need to recognize architectural styles and elements, contributing to the fields of computer vision, cultural heritage preservation, and architectural analysis.

#### 2.1.3 *Sun Attribute Dataset*

The SUN Attribute dataset (Patterson, 2014) is a large-scale image dataset designed for scene understanding, focusing on the attributes that describe various environments. Unlike typical scene recognition datasets that classify entire scenes, SUN Attribute annotates specific characteristics or attributes within the scenes. These attributes include visual properties like "sunny," "open area," "natural lighting," or "wooden," providing a more detailed understanding of the scene's composition.

The dataset contains over 14,000 images across 717 different scene categories. For each image, multiple attributes (102 in total) are annotated, capturing both physical and semantic features of the scene. These attributes provide rich descriptive information, which can be combined to recognize and differentiate between different types of environments.

The SUN Attribute dataset is widely used for tasks such as scene understanding, object detection, and image retrieval, and it helps improve models that need to understand the underlying properties of an environment rather than just classifying it into a broad category.

### 2.2 *Scene and Attribute Recognition*

Our methodological approach is based on employing the most suitable deep learning model for the task at hand (i.e., scene and attribute recognition). Two separate models were employed, namely ResNet-Scene and ResNet-Attr, pretrained on Places365 (for scene recognition) and SUN Attribute (for attribute recognition) datasets respectively. As presented in table below, among the wide variety of available models for these datasets, the Resnet based-model is the one achieving the highest top-5 accuracy (Xiao, 2018; Zhou 2016).

| | Validation Set of Places 365 | | Test Set of Places365 | |
|---|---|---|---|---|
| | Top-1 accuracy | Top-5 accuracy | Top-1 accuracy | Top-5 accuracy |
| AlexNet | 53.17% | 82.89% | 53.31% | 82.75% |
| GoogLeNet | 53.63% | 83.88% | 53.59% | 84.01% |
| VGG | **55.24%** | 84.91% | **55.19%** | 85.01% |
| ResNet | 54.74% | **85.08%** | 54.65% | **85.07%** |

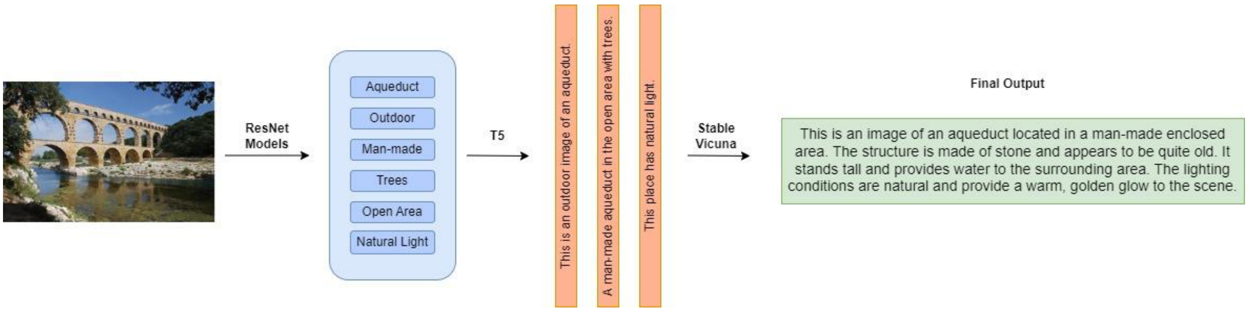
While for attribute recognition the pretrained model was proved to be sufficient, for scene recognition it was of paramount importance to also support fine-grained architectural elements of the AHE dataset. Therefore, we exploited the ResNet-Places365 approach which was pretrained on Places365 and further fine-tuned the model on the AHE dataset as well. However, this approach posed some important issues that should be addressed. Our post-training analysis revealed a major misclassification of architectural elements attributed to the imbalanced nature of datasets and the overlapping of categories. Moreover, dataset biases – such as the underrepresentation of certain architectural styles or regions – also posed challenges for training models that can generalize well across diverse contexts.

To this end, inspired by Yan et al. (Yan, 2015), we explored methods for improving fine-grained classification in architectural datasets by using hierarchical models that account for the structural relationships between different architectural elements. Moreover, we also dealt with dataset biases (i.e., the underrepresentation of certain architectural styles or regions) by assigning different class weights to each category during the training phase, so that each category's contribution was even. Ultimately, this approach led to a custom classification layer that could accurately recognize and describe architectural components in detail. Finally, to deal with the overlapping categories between the two datasets we employed a heuristic scheme that assigned architectural features as additional descriptors when detected in prominent scene categories leading to more detailed image tagging.

### 2.3 *Scene Captioning*

In order to be able to automatically provide detailed descriptions to images of architectural interest, we followed a two-step approach. The first step consists of recognizing the scenery (concepts and attributes as described in Section 2.2) and then transforming these keywords into meaningful, human-like, phrases with contextual information. Therefore, we took advantage of the advances in the field of Natural Language Processing and in particular the developments of Large Language Models (LLMs).

In more detail, upon receiving an image, the system first identifies the scene category (e.g., aqueduct, cathedral, etc.) and the environmental attributes (e.g., outdoor, natural light, etc.) using the pre-trained models. This information forms the base for generating a structured, hierarchical, description. In parallel, the system detects any prominent architectural elements (e.g., stained glass, flying buttress) using the fine-tuned AHE-based model that are treated as additional descriptive keywords. Finally, a base language model (namely T5) is employed to create a structured description of the image which consequently is fed into the Stable Vicuna language model (https://huggingface.co/CarperAI/stable-vicuna-13b-delta) (an open source LLM capable of exhibiting quality similar to that of OpenAI ChatGPT for such a task), which generates a detailed text description of the scene and architectural features. The model is prompted with contextual information such as lighting, environment type, and detected attributes to ensure a rich and accurate narrative. The aforementioned pipeline is schematically depicted in Figure 1.



This is an outdoor image of an aqueduct.

A man-made aqueduct in the open area with trees.

This place has natural light.

This is an image of an aqueduct located in a man-made enclosed area. The structure is made of stone and appears to be quite old. It stands tall and provides water to the surrounding area. The lighting conditions are natural and provide a warm, golden glow to the scene.

**Figure 1. Conceptual diagram of the overall methodological approach**

## 3. Implementation

The following section provides an in-depth explanation of the implementation process for the image captioning service. It outlines the key components and the developed models used to develop the system, as well as the strategies employed to ensure accessibility and inclusivity.

### 3.1 *Inclusive Service Design: Best Practices for Accessibility*

Creating a truly accessible digital platform requires careful attention to a range of design elements that ensure all users can fully engage with the content and functionality. Therefore, to ensure that the online platform will be accessible for visually impaired individuals, the following general guidelines [19] were incorporated in the service's design process:

- **Use of appropriate color contrast:** Ensure that the text and background colors offer sufficient contrast to make content easily readable for all users, especially those with low vision. Utilize a color contrast checker to validate that the design meets the recommended contrast ratios for accessibility, helping to prevent eye strain and improve overall usability.
- **Use of clear and concise language:** Communicate using straightforward and unambiguous language on the platform. Avoid the use of jargon, slang, or complex terms that could confuse users. Prioritizing clarity and brevity ensure that the information is accessible to a wider audience.
- **Make navigation simple:** Design the platform to support intuitive navigation, allowing users to easily move through content using a keyboard. Ensure that links, buttons, and other interactive elements are clearly labeled and appropriately spaced, enabling users to quickly identify their purpose without unnecessary confusion or complexity.
- **Use accessible form elements:** Ensure that form fields are properly labeled and easy to navigate, with clear instructions on what information is required. This helps all users, including those using assistive technologies like screen readers, to complete forms efficiently and without errors.

- **Provide audio options:** Offer audio alternatives for content, such as spoken descriptions or narration, to accommodate visually impaired or blind users. This ensures they can fully access the platform's information and features without relying on visual cues.
- **Provide text resizing options:** Allow users the flexibility to adjust text size to a comfortable level. Offering this option helps those with visual impairments or reading difficulties, ensuring that the content remains accessible and legible across a variety of devices and display settings.
- **Test the platform with screen readers:** Conduct thorough testing with popular screen readers such as NVDA and JAWS to confirm that all elements of the platform are accessible and functional. This helps identify potential barriers and ensures a seamless experience for users who rely on assistive technologies.
- **Follow accessibility guidelines:** Adhere to established standards such as the Web Content Accessibility Guidelines (WCAG) to ensure the platform complies with best practices for accessibility. This guarantees that the platform meets the needs of all users, regardless of their physical or cognitive abilities, promoting inclusivity and equal access to information.

### 3.2 *Technical Guidelines for Enhancing Web Accessibility*

Ensuring accessibility is paramount to delivering an inclusive user experience for all individuals, particularly those who are visually impaired or blind. This section outlines essential technical strategies for enhancing accessibility within the web service, focusing on best practices that facilitate seamless interaction and navigation. In detail, the following technical specifications were embodied in the development phase of the web service:

- **Use HTML5:** Implement HTML5 to structure the platform, as it provides semantic tags that help visually impaired and blind users better navigate the content. These tags improve the clarity of the page structure when accessed through assistive technologies, making it easier to understand the layout and purpose of each section.
- **Use ARIA:** Integrate Accessible Rich Internet Applications (ARIA) attributes to enhance the accessibility of user interface components. ARIA roles, states, and properties provide assistive technologies with additional context about buttons, forms, or dynamic content, ensuring users can interact with the platform's features more effectively.
- **Use CSS for layout:** Rely on Cascading Style Sheets (CSS) to control the visual layout and formatting of the platform, rather than using tables. This approach allows for a more flexible and responsive design while maintaining accessibility. Tables should be reserved for tabular data, as they can confuse screen readers when used for layout purposes.
- **Use appropriate font sizes:** Choose font sizes that are easily readable and ensure that text is legible without strain. Provide users with the option to adjust font sizes via their browser's settings or by including a built-in text resizing feature on the platform, accommodating individual visual needs.
- **Use alt tags for images:** Include descriptive alt tags for all images, conveying the context and purpose of the image to visually impaired and blind users. Alt tags should offer clear, concise descriptions that help users understand the visual content, particularly when the images contain important information.
- **Use keyboard navigation:** Ensure that all platform features and functionality are accessible using only a keyboard. This is essential for users with motor impairments or those relying on assistive technologies, enabling them to navigate the platform without needing a mouse or other pointing device.
- **Use skip links:** Implement skip links at the top of the page to allow users, especially those using screen readers, to bypass repetitive navigation menus and go directly to the main content. This enhances efficiency for visually impaired and blind users, who might otherwise need to listen to the same elements repeatedly.

– **Test with assistive technologies:** Regularly test the platform with a range of assistive technologies, including screen readers (such as NVDA or JAWS), screen magnifiers, and braille displays. Testing ensures that the platform is truly accessible and identifies potential barriers that need to be addressed to improve usability for all users.

### 3.3 *Co-Creation in Design*

The initial design of the image captioning service was heavily informed by discussion groups held with visually impaired individuals and instructors, where participants shared their challenges in accessing visual content, particularly in architectural education. This process provided valuable insights into the specific needs of the target users, such as the preference for detailed, structured descriptions and the importance of providing voice-based outputs for easier consumption.

Based on this feedback, the development team prioritized the creation of a web-based platform that was both visually accessible and easy to navigate. The interface was designed to, among others, be minimalistic, with clear instructions, large text, and screen reader compatibility to ensure that visually impaired users could interact with the service independently. Special attention was paid to creating an intuitive workflow where users could upload images and receive both text and audio descriptions without requiring complex interactions.

### 3.4 *Development of the Image Captioning Service*

The image captioning system was developed as a web-based service to ensure broad accessibility, allowing users to access the tool from any device. This approach was specifically chosen to facilitate flexibility and adaptability across various learning environments, ensuring that users can seamlessly integrate the service into their educational or professional workflows. The underlying models, as detailed in Section 2.2, were meticulously tailored to meet the identified requirements, with a particular focus on maintaining a user-centered approach that prioritizes the needs of visually impaired individuals and other key stakeholders.

The service enables users to upload architectural images, which are then processed through a sophisticated methodological framework (outlined in Section 2.3) to generate comprehensive, contextually accurate descriptions of the architectural scenes and elements depicted. The system's ability to deliver rich, detailed descriptions ensures that users, regardless of their visual abilities, can engage meaningfully with the architectural content. Moreover, the service also integrates a voice output capability, which was identified as a key component by the participants of the discussion groups, providing this way auditory descriptions of images. In detail, once an image is processed and its description is generated, users have the option to listen to the description read aloud via a simple button click.

Finally, throughout the creation of the service, multiple rounds of testing were conducted to ensure that the platform was functional, accessible, and aligned with user expectations. Based on the test results, several refinements and addition of features were made, with each iteration aiming to create a more personalized and user-friendly experience.

### 3.5 *Human Feedback Loop for Continuous Improvement*

To ensure that the captions generated by the system were useful, understandable, and met the needs of visually impaired users, a human feedback loop mechanism was incorporated in the web platform. After the initial deployment of the service, users were encouraged to provide feedback on the quality, clarity, and relevance of the descriptions they received and also suggest changes by providing alternative and more accurate textual descriptions. The received feedback can be used to iteratively improve the system by refining employed models, leading to more accurate and detailed descriptions.

## 4. Conclusion and Lessons Learnt

In this paper, we presented the InVisible’s deep-learning-based system for captioning images in the field of architectural history. Such systems hold significant promise for applications in education and heritage preservation. They can assist in cataloging architectural features or generating descriptions for educational tools, making architectural knowledge more accessible. However, achieving the accuracy and depth of description needed for these applications remains a challenge, emphasizing the need for ongoing refinement of these models.

Developing a deep-learning-based system for captioning images in architectural history has revealed several key challenges, including the need for domain-specific knowledge, high-quality and diverse training data, and the ability to balance high-level descriptions with fine-grained details. These systems struggle with capturing the historical and cultural context of architectural features and must often rely on transfer learning and interdisciplinary collaboration with experts to ensure accuracy. Ambiguity in architectural styles and the need for explainable AI further complicate the process, while combining visual and textual data can enhance contextual understanding.

In detail, our effort towards the creation of the Invisible’s image captioning system has underscored the importance of integrating domain-specific knowledge into the models. Architecture, particularly its historical study, is deeply tied to specialized terminology, styles, and contextual understanding. A system trained with generic datasets would struggle to differentiate between specific architectural elements or periods, such as distinguishing Gothic from Baroque styles or identifying unique features like Corinthian columns. To address this, it is crucial to train models on datasets that contained a rich vocabulary and detailed visual examples from a wide range of architectural elements and eras.

Moreover, the diversity and quality of training data were also key factors in the success of our system. In architecture, the richness of styles, materials, and techniques across different time periods must be reflected in the training images. Without a diverse dataset, models may generate oversimplified captions that miss the complexity of architectural elements. Additionally, finding the right balance between high-level descriptions, such as identifying a building’s style, and more detailed captions that focus on specific architectural elements can be a challenge. Deep learning models often struggle with providing such detailed descriptions unless they have been explicitly trained on attributes like “vaulted ceilings” or “flying buttresses.”

Through the development of the presented system, we were able to uncover the importance of contextual understanding. Architectural history goes beyond visual recognition to include the cultural and historical significance of a structure, which is difficult for AI models to grasp without external data. A building’s importance in history may not always be visible in its facade, and models need to be equipped with both visual and textual information to fully capture these nuances. This is where techniques like transfer learning, which involve fine-tuning pre-trained models on domain-specific datasets, become useful. However, careful calibration is needed to avoid losing architectural specificity when relying on general datasets for training.

Interdisciplinary collaboration also proved to be essential when developing AI for architectural history. Historians and architects play a vital role in annotating the training data and ensuring the models accurately capture the necessary details and nuances. This collaboration ensures that the system is not only able to recognize architectural elements but also to interpret them in a meaningful way. At the same time, explainability in AI is crucial for fostering trust, particularly in fields like history and heritage conservation, where academic rigor is required. Understanding how a deep learning system arrived at a particular conclusion or description can help researchers (or users) assess its reliability and accuracy.

In addition, it was within our scope to also tackle another significant challenge concerning the handling of ambiguity. Many architectural styles evolve over time or merge elements from different periods, and deep learning systems need to be flexible enough to accommodate such complexity. This could mean providing multiple interpretations when a structure has mixed features or offering a level of uncertainty when the visual data alone is inconclusive. Similarly, combining visual data with textual information from historical documents or databases can significantly improve a system’s performance, providing it with richer context and enabling it to generate more informed captions.

Finally, given that InVisible's system was tailored to help blind students with architectural history, close collaboration with blind associations proved to be essential in its developments. Blind associations can provide critical insights into the specific needs, preferences, and challenges faced by visually impaired students, ensuring that the system is designed with accessibility in mind (Georgiadis, 2019; Kalaganis, 2021). Engaging with these organizations, the development team gained valuable insights on how the system can best convey architectural information in a useful and meaningful way for these individuals. Moreover, blind associations, also assisted the development process, by outlining the most effective methods for communicating complex visual concepts, such as architectural styles or features, to blind students. This included determining the appropriate level of detail for descriptions or tailoring the language to be more accessible and intuitive. In addition, they provided essential guidelines with respect to practical considerations towards a seamless user experience. Furthermore, working with these associations fosters a participatory design approach, ensuring that the system is not only functional but also empowering for blind students. By aligning the development process with the actual needs of the end users, the system can better promote independent learning and enhance educational opportunities for blind students in the field of architectural history.

## Acknowledgements

The research presented in this manuscript is part of the InVisiBLe (Inclusive and Innovative learning tool for Visually Impaired and Blind people) project, Co-funded by the Erasmus + programme of the European Union under Call: 2021 Round 1 KA2, (Project Code: 2021-1-IT02-KA220-HED-000031139, 02/2022 – 07/2024)

## References

Georgiadis K., Kalaganis F., Migkotzidis P., Chatzilari E., Nikolopoulos S., Kompatsiaris I. (2019). A computer vision system supporting blind people-the supermarket case. In *Computer Vision Systems: 12th International Conference, ICVS 2019*, Thessaloniki, Greece, September 23–25, 2019, Proceedings 12 (pp. 305-315). Springer International Publishing.

Georgiadis K., Kordopatis-Zilos G., Kalaganis F., Migkotzidis P., Chatzilari E., Panakidou V., ... Kompatsiaris I. (2021, June). Products-6k: a large-scale groceries product recognition dataset. *In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference* (pp. 1-7).

Jose L. (2017). *Architectural Heritage Elements image Dataset*, available online.

Kalaganis F. P., Migkotzidis P., Georgiadis K., Chatzilari E., Nikolopoulos S., Kompatsiaris I. (2021, July). Lending an artificial eye: beyond evaluation of CV-based assistive systems for visually impaired people. In *International Conference on Human-Computer Interaction* (pp. 385-399). Cham: Springer International Publishing.

Lazaridis N., Georgiadis K., Kalaganis F., Kordopatis-Zilos G., Papadopoulos S., Nikolopoulos S., Kompatsiaris I. (2024). The Visual Saliency Transformer Goes Temporal: TempVST for Video Saliency Prediction. *IEEE Access*.

Lin T. Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., ... Zitnick C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.

Patterson G., Xu C., Su H., Hays J. (2014). The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108, 59-81.

Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., ... Sutskever I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., ... Liu P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1-67.

Vinyals O., Toshev A., Bengio S., Erhan D. (2015). Show and tell: A neural image caption generator. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156-3164).

Xiao T., Liu Y., Zhou B., Jiang Y., Sun J. (2018). Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 418-434).

Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhudinov R., Zemel R. &amp; Bengio Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the 32nd International Conference on Machine Learning, in Proceedings of Machine Learning Research,* 37, 2048-2057.

Yan Z., Zhang H., Piramuthu R., Jagadeesh V., DeCoste D., Di W., Yu Y. (2015). HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 2740-2748).

Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).

Zhou B., Lapedriza A., Khosla A., Oliva A., Torralba A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6), 1452-1464.