Generation of Annotated Data Using Pre-Trained Diffusion Models

Alleviating the data pain in the context of architectural façade segmentation

Alejandro Rueda[1], Yao Sun[2], Ivan Bratoev[3] and Frank Petzold[4]
[1,3,4]*Technical University of Munich.*
[2]*Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling 82234, Germany.*
[1]*alejandro.rueda@tum.de, ORCID: 0009-0001-3855-0601*
[2]*yao.sun@dlr.de, ORCID: 0000-0003-2757-1527*
[3]*bratoev@tum.de, ORCID: 0000-0002-7908-917X*
[4]*petzold@tum.de, ORCID: 0000-0001-8974-0926*

Abstract. The success of modern machine learning methods such as deep learning has broadened its influence on almost all research fields, including architecture and the built world. A common requirement for these data-centric technologies is the availability of data for training. Dataset creation and curation is a time-consuming and error-prone process. Depending on the complexity of annotations needed, this can become a significant bottleneck. Particularly in tasks such as in semantic segmentation, where labelling a single urban image may require several minutes. Segmented data is essential for analysing and developing methodologies for objects in the built world. Yet existing façade segmentation datasets often suffer from inconsistencies, limited size and varying annotation standards. This work proposes an approach to bridge the data gap, using generative models. A pretrained stable diffusion model is finetuned on a custom dataset created with the purpose of providing a wide range of architectural styles and image types. In a subsequent stage, the pretrained model is used to train an ensemble of pixel classifiers to predict class labels for generated images. A data generation pipeline is proposed and prototypically tested, focusing on the generation of segmented data for facade labels.

Keywords. Generative AI, Façade Segmentation, Synthetic Data, Stable Diffusion, Urban Informatics

1. Introduction

High quality annotated datasets are still a scarcity in the context of architecture, hindering the research progress in this field. The issue is aggravated for data that has costly manual labour demands, such as is the case for segmentation tasks. The study of architectural façade structure requires images of labelled façade elements making it a textbook example for this challenge. Existing dataset for façade segmentation contain

ARCHITECTURAL INFORMATICS, Proceedings of the 30th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2025, Volume 1, 59-68.

multiple issues starting with their size. Currently the largest ones, IRFs (Wei et al,. 2024) and LabelMeFacade (Fröhlich et al., 2010), contain only 1057 and 945 datapoints respectively. Others such as ECP (Teboul et al, 2011), eTRIMS (Korc and Förstner 2009), CMP (Tyleček and Šára 2013), Graz50 (Riemenschneider et al., 2012) and ADE20K (Zhou et al., 2017) rarely surpass the 100 datapoints mark.

Moreover, there are also some challenges regarding the quality of the contents of those datasets. Figure 1 shows two example cases depicting wrong and missing labels. Another common problematic is the lack of shared standards. There are discrepancies not only on the way that facades are represented, but also how distortions in images are handled. Each dataset defines its own set of segmentation classes and image resolutions, making it difficult to combine and use them together. Figure 2 depicts some examples of relevant existing façade segmentation datasets.

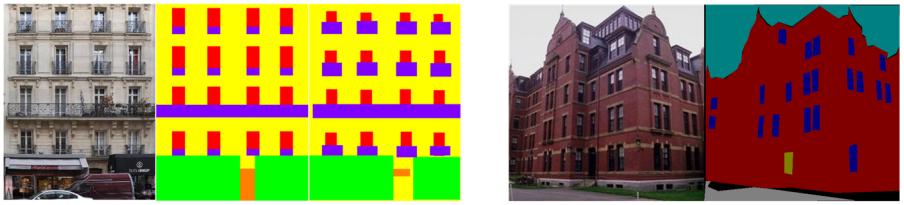

Figure 1. Errors in existing datasets. Left: Mislabelling of a car as a door and wall in the original and updated ECP dataset. Right: Missing window annotations in a LabelMeFacade datapoint.

This shortage of data arises directly from the error-prone and time-intensive nature of dataset creation and curation processes. Annotating a single urban image may require up to 60 minutes (Wu et al., 2023). Furthermore, the rewards of investing time publishing a high-quality dataset are overshadowed by those acquirable researching other parts of the field, such as deep learning model architectures and their applications (Sambasivan et al., 2021, Wei et al., 2024)

To tackle the data challenge in this context, pipelines have been proposed to generate annotated data using procedural generation and deep generative models (Kikuchi et al., 2023, Wu et al. 2023). Whilst the results of such endeavours are promising, demonstrating efficient generation of acceptable annotated images, they are still far from becoming a tangible solution to the data issue. Beyond the evident quality issues with the generated images, there are additional, less noticeable shortcomings. These include concerns about the diversity of the output and the overall adaptability of the method. Similar to the missing architectural style variation in the existing annotated datasets (Wei et al., 2024), synthetic pipelines also suffer from a lack in diversity. Moreover, adapting methods that rely on procedural modelling requires a significant amount of effort and expertise (Martinovic and Van Gool 2013).

These factors discourage a wide adoption of these approaches. This paper introduces a novel approach that leverages state-of-the-art deep generative models, such as Stable Diffusion (Ronneberger et al., 2015), to create high-quality annotated façade segmentation datasets. Building on the work of (Baranchuk et al., 2022), our method directly addresses critical limitations of existing pipelines, including data

diversity and adaptability while streamlining the annotation process. By addressing these issues, this method mitigates challenges associated with traditional and synthetic dataset generation while providing a scalable framework for architectural research.

Figure 2. Examples from existing datasets. From left to right: IRFs, LabelMeFacade, CMP, ADE20K, ECP, eTRIMS & Graz50. The datasets vary significantly in terms of annotation granularity, image type, and architectural style.

2. Objective

The main aim of this work is to provide an adaptable method for generating high-quality 2D annotated segmentation data using open source, state of the art generative models. In contrast to previous works (e.g. Kikuchi et al. 2023), the goal is to propose a methodology that doesn't rely on any procedural modelling and has a very small data annotation requirement (less than 5 datapoints per class).

To achieve this, we aim to tackle multiple sub goals that represent unavoidable hurdles needed to be addressed. Starting with the selection of the raw data used in the finetuning process of the generative model. To avoid the common absence of diversity in façade datasets the selected raw data must represent a wide variety of architectural styles from different cultures. Additionally, the raw data should also include multiple types of photos to enhance the flexibility and adaptability of the proposed method. By mixing different perspectives, distortions and artistic quality in the raw data selection, multiple target domains are covered. To allow a wide range of weather conditions, clutter and settings, a high-quality generative text to image model (Stable Diffusion v1.4) is used via finetuning.

Such models require the training data to be structured in image and text tuples, both describing the same content. Currently, there are no published façade datasets that meet the requirements described for the raw data selection. As a result, this works also aims to process the raw data and generate captions for each datapoint. These captions also represent the building blocks of any prompt used to guide the generation of the labels.

Lastly, as a validation of the proposed methodology, this work aims to generate a dataset based on requirements adaptable to other settings.

3. Methods

Figure 3. Examples of images from the different selected domains and their generated captions.

The proposed methodology derives from the works started by the DatasetGAN (Zhang et al., 2021) , BigDatasetGAN (Li et al., 2022) and its adaptation to Diffusion models (Baranchuk et al., 2022). All these works follow the same workflow. First a pretrained generative model is selected. Next, a small dataset representative of the targeted segmentation scheme is created. Finally, a segmentation branch is trained on top of the frozen generative model. This is achieved by using multiple feature maps extracted from the inner blocks of the underlying generative model and the created segmentation dataset. Our proposed approach expands on these works by greatly improving the quality of the output and providing means to control the generated output via prompts. Instead of using a baseline denoising diffusion probabilistic model (DDPM), we utilize a stable diffusion model trained on the large-scale dataset LAION-5B and fine-tune it on the target domain.

3.1. STABLE DIFFUSION FINE-TUNING

As input for the fine tune process a curated dataset containing around 100k data points (images and text pairs) was created. Following a bootstrapping methodology, a clean subset of the Places dataset (Zhou et al. 2017), a collection of google street view images (Kang et al., 2018) and openly available exterior architecture photographs (Paulat 2022) were used. The Places dataset consists of photos commonly found on the internet captured by non-experts. Such images display different perspectives of the façades and contain a large diversity of clutter. Street view images depict façades from a distinct and consistent vantage point. Such images provide extensive global coverage and serve as a valuable resource for urban analysis, architectural studies, and dataset creation in machine learning applications. Architectural photography depicts façades in a very artistic, clean and well thought manner, providing a clear view of the structures and features of built objects. Integrating these three existing datasets offers a diverse array of architectural styles and image types, making the base generative model highly flexible.

As part of the curation preprocess, images unrelated to façades were removed and the image resolution of the dataset determined. To provide an acceptable quality for the generation process, a resolution of 512 by 512 pixels was selected. Images stemming from the Places dataset needed an extra preprocessing step, since their native resolution was only half of the target, i.e. 256 by 256 pixels. To achieve better results than those attainable via interpolation, a super-resolution model, GigaGAN (Kang et al., 2023), was utilized to preprocess all Places images.

To produce the text part i.e. the captions of the created image dataset, an open-source visual language model CogVLM (Wang et al., 2023), was utilized. The captioning process itself was guided by prompting the CogVLM model to follow

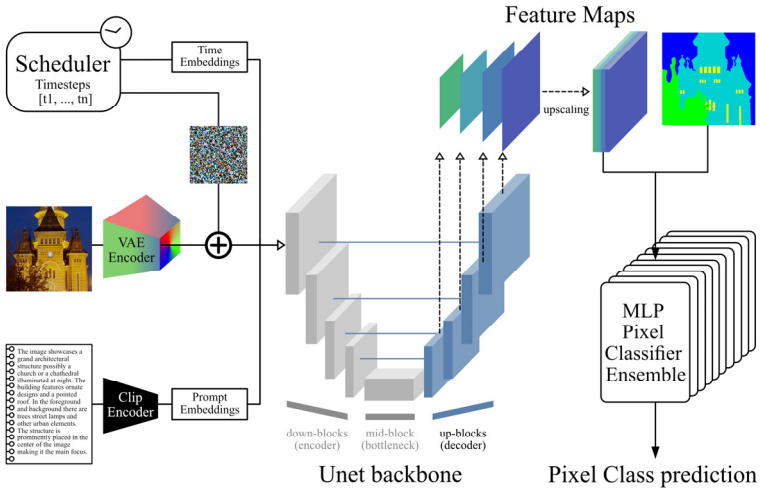

Figure 4. The segmentation training pipeline. Embeddings for a chosen set of noise steps are created as usual in the stable diffusion model. Activations from the upsampling blocks are then extracted during the denoising step to train the segmentation head. The diffusion model and the segmentation head generate both the images and annotations.

typical best practices for alternative text descriptions of images in architecture (Washington University in St. Louis [WashU], n.d.). To improve the performance of the captioning model, image meta-data available was utilized. The image datasets merged contained information about the building type. The resulting textual data, also known as prompts, paired with their corresponding images, constitute the dataset used to fine-tune the stable diffusion model. Figure 3 depicts example images of each domain and their corresponding generated caption.

3.2. IMAGE ANNOTATIONS GENERATION PIPELINE

To segment the generated data, an ensemble of multi-layer perceptrons (MLPs) is trained as a lightweight segmentation head attached to the image generation backbone. The pixel segmentation head is trained not only on a ground truth segmentation mask, but also on different features stemming from the interior of the backbone. As observed by (Baranchuk et al., 2022) the representations learned in the autoencoder part of these models, can be used to attain segmentation masks at different granularities depending on the features selected at different noise steps during the diffusion process.

The current implementation of stable diffusion has multiple blocks to choose from. To perform the denoising process in latent space, an autoencoder is used to project input images into latent space and to decode the results back to image space. Between the encoding and decoding parts of this autoencoder, the generative backbone model, a U-Net, is located. The upsampling blocks from this backbone contain both the encoding and decoding signals thanks to the skip connections. Based on this fact, these blocks are the targets of the feature extraction.

Subsequently, the segmented image generation is achieved by manually or automatically prompting the finetuned model using a set of desired facade categories and characteristics. The automatic process follows the same guidelines used in the initial dataset captioning i.e., by following guidelines of architecture photograph alternative text creation. By reusing the same captioning scheme better results are attained. Figure 4 presents the process for training the segmentation head and for generating annotated images.

3.3. EXPERIMENT

To validate the proposed methodology, we conducted an experiment with the objective of generating annotated data replicating the characteristics of the façade segmentation dataset TMBud (Orhei et al., 2021). This dataset was selected as the target due to the high quality of its labels and its limited size. There are a total of 820 unlabelled photos and 300 annotated images, using the following classes to segment their content: background, building, door, window, sky, vegetation, ground and noise.

To apply the proposed generation pipeline, the segmentation head must be trained on the target data. It's important to note that, in contrast to the previous works, our methodology allows for better control of the generation process via prompting. However, it also imposes a new requirement for image captions during the segmentation branch training.

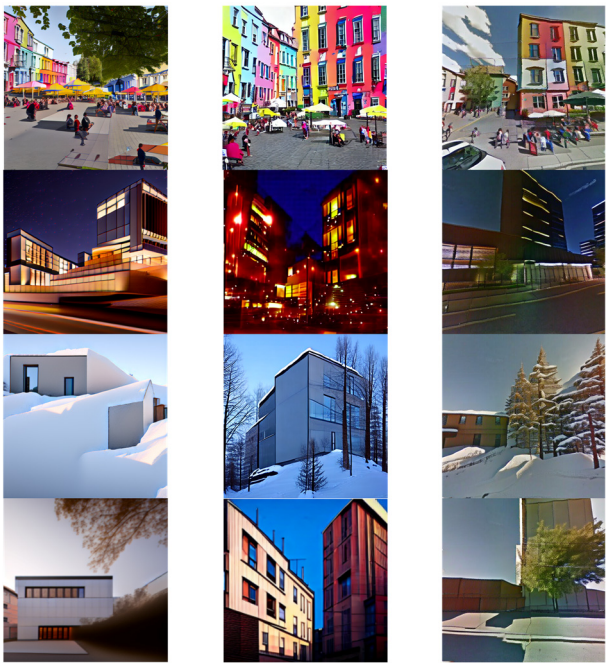

Figure 5. Images generated using the fine-tuned model. Each column represents a different domain: architectural photography, normal photos and street view images. To demonstrate the flexibility of the model, the same prompt was used with minor changes related to daytime and weather.

4. Results

The proposed methodology is capable of generating a wide range of annotated facade images, mirroring the different domains used to fine-tune the generative model. This demonstrates the usability of the created dataset for fine-tuning text-to-image models such as Stable Diffusion and directly addresses parts of the existing dataset issues in the field of architecture. Figure 5 shows multiple images generated using different image types and settings.

The quality of the generated image data surpasses that of the previous work; however as illustrated in Figure 5, the accuracy of the generated annotations is inferior to than that of the TMBuD dataset. Due to the generative nature of the data, exact evaluation against a ground truth is not feasible. Visual inspection highlights discrepancies in the segmentation results. Despite this, this pipeline provides an initial step towards tackling the data issue by providing high fidelity and varied images of architectural façades. In its current state, generated images annotations yield results comparable to existing datasets, primarily because the actual datasets are very limited in size. The generated annotations require further validation, but this was not pursued due to their evident imprecision. Nevertheless, the proposed pipeline can be adapted to any segmentation scheme as demonstrated by the experiment without incurring

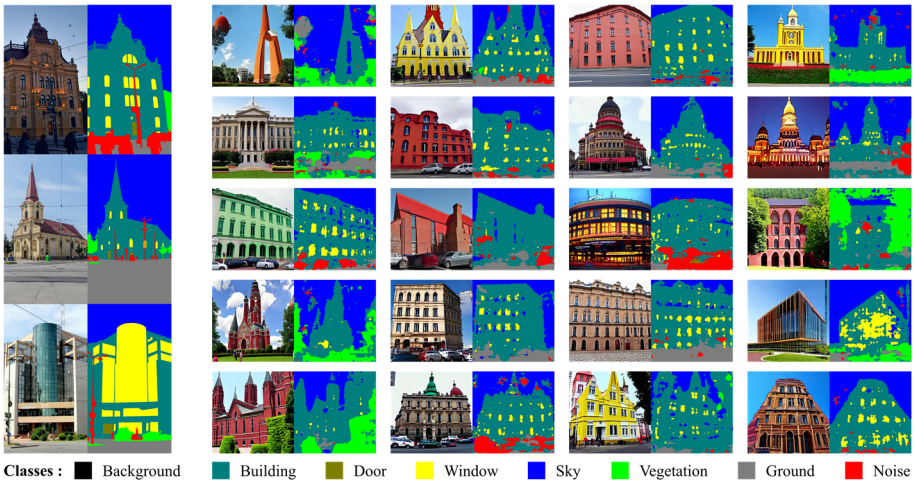

Figure 6. Randomly picked results from the generation process. The first column depicts some examples from the target dataset, TMBuD.

significant costs to the user. As in previous works, only a small amount of annotated data and textual descriptions are required to generate any façade segmentation data. Figure 6 presents randomly selected generated images and annotations.

5. Conclusion

The methodology proposed adapts to various annotation requirements with minimal overhead and can be improved by utilizing different segmentation heads. It emphasizes the usability of openly available diffusion models, such as stable diffusion, which serve as robust tools for generating high quality annotated data. Using models trained on vast datasets, significantly enhances the model's expressiveness and representation space compared to models trained solely on a specific domain. In the context of annotated facade image generation, it allows the model to efficiently handle clutter and the diverse range of objects present in such images.

Beyond the proposed methodology, both the fine-tuned generative model and the data used in the process serve as valuable tools for other downstream tasks in the architectural research field. For instance, the features learned in the backbone provide a robust representation of façade characteristics, which can be further utilized in various applications, including architectural style classification, building age prediction, and structural damage detection, among others. This highlights the importance of creating foundation models for architectural research such as DINO (Caron et al. 2021); models which would offer valuable learned representation of architectural features, such as those present in façades, serving as a backbone to multiple downstream tasks.

There are multiple possible reasons behind the poor annotation results. Starting with the selection of the blocks for the feature extraction and the selection of the noise steps utilized during the denoising process. The current selection was guided by the findings of previous work; however, due to differences in the underlying models, this

choice may be suboptimal. Additionally, the segmentation head could be redesigned to better utilize the input signals. A potential enhancement could involve employing methods aimed at achieving interpretability of language models by extracting learned concepts via sparse autoencoders (Cunningham et al., 2023, Templeton et al., 2024).

Although the data generated using this approach is not perfect, it represents a significant step towards alleviating the data scarcity in the context of facade image segmentation. The generation of annotations within the proposed pipeline necessitates further refinement and more robust validation.

Acknowledgements

The research presented is supported by the TUM George Nemetschek Institute Artificial Intelligence for the Built World.

References

Baranchuk D., Voynov A., Rubachev I., Khrulkov V., & Babenko A. (2022). Label-Efficient Semantic Segmentation with Diffusion Models. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600.

Fröhlich B., Rodner E., & Denzler J. (2010). A Fast Approach for Pixelwise Labeling of Facade Images. In Proceedings of the International Conference on Pattern Recognition (ICPR 2010).

Kang J., Körner M., Wang Y., Taubenböck H, & Zhu X. X. (2018). Building instance classification using street view images. ISPRS Journal of Photogrammetry and Remote Sensing, 145, 44-59. https://doi.org/10.1016/j.isprsjprs.2018.02.006

Kang, M., Zhu, J. Y., Zhang, R., Park, J., Shechtman, E., Paris, S., & Park, T. (2023). Scaling up gans for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10124-10134).

Korc, F., & Förstner, W. (2009). eTRIMS Image Database for Interpreting Images of Man-Made Scenes (Report No. TR-IGG-P-2009-01). Dept. of Photogrammetry, University of Bonn.

Li D., Ling H., Kim S. W., Kreis K., Fidler S., & Torralba A. (2022). BigDatasetGAN: Synthesizing ImageNet with Pixel-wise Annotations. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022 (pp. 21298–21308). IEEE.

Martinovic A, & Van Gool L. (2013). Bayesian Grammar Learning for Inverse Procedural Modeling. In 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013 (pp. 201–208). IEEE Computer Society. https://doi.org/10.1109/CVPR.2013.33

OpenAI. (2024). ChatGPT Model: GPT-4 | Temp: 0.7 [Large Language Model]. https://openai.com/chatgpt/

Orhei, C., Vert, S., Mocofan, M., & Vasiu, R. (2021). TMBuD: a dataset for urban scene building detection. In International Conference on Information and Software Technologies (pp. 251–262).

Paulat, T.(2022). Modern Architecture (100k Images). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/tompaulat/modernarchitecture

Riemenschneider, H., Krispel, U., Thaller, W., Donoser, M., Havemann, S., Fellner, D., & Bischof, H. (2012). Irregular lattices for complex shape grammar facade parsing. In 2012

IEEE Conference on Computer Vision and Pattern Recognition CVPR (pp. 1640-1647). https://doi.org/10.1109/CVPR.2012.6247857

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery. https://doi.org/10.1145/3411764.344551

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P., & Paragios, N. (2011, June). Shape grammar parsing via reinforcement learning. In CVPR 2011 (pp. 2273-2280). IEEE.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., ... & Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread.

Tyleček, R., & Šára, R. (2013). Spatial Pattern Templates for Recognition of Objects with Regular Structure. In Proc. of German Conference on Pattern Recognition (GCPR), pp. 364-374, 2013.

Wang W., Lv Q., Yu W., Hong W., Qi J., Wang Y., Ji J., Yang Z., Zhao L., Song X., Xu J., Xu B., Li J., Dong Y, Ding M., & Tang J.. (2023). CogVLM: Visual Expert for Pretrained Language Models. arXiv preprint. https://doi.org/10.48550/arXiv.2311.03079

Washington University in St. Louis (n.d.). *Alt-Text for Architecture Images.* Sam Fox School of Design & Visual Arts, Washington University in St. Louis. https://insidesamfox.wustl.edu/about/communications/digital-accessibility/describing-architecture-images/

Wei, J., Hu, Y., Zhang, S., & Liu, S. (2024). Irregular Facades: A Dataset for Semantic Segmentation of the Free Facade of Modern Buildings. Buildings, 14(9), 2602. https://doi.org/10.3390/buildings14092602

Wu, W., Zhao, Y., Shou, M. Z., Zhou, H., & Shen, C. (2023). Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1206-1217).

Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., & Fidler, S. (2021). DatasetGAN: Efficient Labeled Data Factory With Minimal Human Effort. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021 (pp. 10145–10155). Computer Vision Foundation / IEEE.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 633-641).

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6), 1452-1464. https://doi.org/10.1109/tpami.2017.2723009

ChatGPT (OpenAI, 2024) was used to improve flow and correct errors in spelling and grammar.