ENCODING TRANSITIONS

Self-Supervised Learning of Architectural Thresholds with Masked Spatiotemporal Transformers

Sihui Lin[1]
[1] *University of California, Los Angeles.*
[1] *sihui1in@ucla.edu, 0009-0008-0187-5091*

Abstract. Architectural thresholds are often treated as fixed components in digital workflows, despite being perceptual events that unfold through movement. This paper addresses the problem of how to computationally model these dynamics. The aim is to develop a method that encodes and classifies threshold moments directly from sequential visibility data, asking whether spatiotemporal learning can capture the structural logic of perceptual change. The study contributes a framework that integrates 3D isovist sampling, transition detection, and masked spatiotemporal transformers to treat thresholds as learnable patterns. Using depth-encoded panoramas derived from spherical isovists, a TimeSformer encoder is pretrained on synthetic typologies and applied to diverse architectural case studies. Results show coherent latent organisation and successful transfer to real environments, revealing how thresholds vary across interior and garden contexts. The study concludes that thresholds possess consistent geometric signatures that can be learned computationally, while noting limitations related to path dependence, synthetic bias, and cultural variability.

Keywords. Deep learning, Isovist analysis, Representation learning, Spatial experience, Spatial Transition, Spatiotemporal transformer

1. Introduction

Architectural discourse has long been captivated by the in-between—those charged instants where spatial experience shifts, and visibility, proximity, and orientation subtly reconfigure the world before the moving body. Evans (1978) demonstrates how doors and passages choreograph the unfolding of built form, while Bachelard (1969) frames the threshold as an intensified zone mediating inside and outside. Norberg-Schulz (1971) likewise underscores transitions as elemental to how places structure meaning and dwelling. Across these accounts, thresholds accumulate cultural, symbolic, and ritual significance, yet contemporary digital workflows rarely register this experiential richness. Building Information Modeling (BIM) environments reduce thresholds to functional tags, most commonly "doors," which abstracts away the momentary reorganisation of perception that gives transitions their architectural resonance. As a result, present-day design tools provide limited means to analyse or compare thresholds as dynamic, time-based events.

Recent computational design research has increasingly focused on developing methods to model spatial experience in a more systematic and analysable form. Still, these approaches typically assess continuous spatial fields or isolated metrics rather than the discrete perceptual transitions that punctuate movement. This research addresses that gap by proposing a methodology for encoding thresholds as moments of perceptual change captured through 3D isovists, transition detection, and masked spatiotemporal transformers. A TimeSformer (Time–Space Transformer) encoder is first pretrained to classify transitions into synthetic typologies defined by systematic variations in height and width. The pretrained model is then applied to architectural case studies in order to examine how these learned representations generalise across cultural and spatial contexts. Together, these components establish a computational foundation for treating thresholds as dynamic and evolving events that can be analysed, compared, and categorised with a level of precision that existing architectural tools do not offer.

2. Background

Computational approaches to architectural analysis increasingly draw on visibility-based methods to connect spatial configuration with perceptual experience, and isovist modelling has been central to this development. Benedikt (1979) defines an isovist as the set of all points visible from a specific vantage point in an environment, a formulation that transforms subjective, momentary perception into a quantifiable geometric field. Subsequent research expanded this foundation through new descriptors and analytical frameworks. Turner et al. (2001) introduced visibility graphs and isovist fields, demonstrating how local visibility structure aggregates into higher-level spatial organisation. Franz and Wiener (2005) further established behavioral relevance by showing that specific isovist measures reliably correlate with wayfinding performance and patterns of spatial decision-making. Krukar et al. (2020) extended these ideas into embodied 3D isovists that incorporate eye height and directional sensitivity, thereby improving alignment between modeled visibility and pedestrian experience. More recent work by Peponis et al. (2025) proposed refined descriptors such as depth and distributedness and defined architectural thresholds through the systematic analysis of occluding edges. Together, these studies show that isovist-based methods function not only as geometric abstractions but as computational instruments capable of identifying perceptual discontinuities, spatial partitions, and potential transitional zones.

Machine learning has opened new possibilities for deriving perceptual structure directly from spatial data, offering analytical capacities that extend beyond handcrafted geometric descriptors. Peng et al. (2017) showed that depth-encoded isovist panoramas can be classified by convolutional neural networks (CNN), demonstrating that spatial types can emerge from learned visibility patterns rather than predefined metrics. Sedlmeier and Feld (2018) similarly used classic 2D isovist measures reduced through Principal Component Analysis (PCA) to distinguish indoor spatial types such as passageways, while Miao et al. (2024) expanded this line of inquiry by training a neural architecture to learn “spatial memory” from agent-driven trajectories, revealing how perceptual transitions accumulate into temporally structured representations. Additional work has annotated movement paths through sequential changes in isovist structure (Feld et al., 2018) and identified perceptually salient decision points by examining 3D visibility discontinuities and visual transitions within architectural environments (Bhatia et al. 2012; Krukar et al., 2020). Together, these studies show that learned or data-driven approaches can uncover perceptual tendencies embedded in visibility patterns that static geometric analyses often overlook. Building on this trajectory, this study advances isovist-based research from isolated viewpoints to sequential, movement-dependent perception. The focus shifts from global categorisation of space to the localised transitions that constitute threshold experience, positioning spatial change itself as a computationally learnable event.

3. Methodology

This section outlines the methodological framework for identifying and modelling threshold transitions along movement paths. The pipeline integrates 3D isovist sampling, threshold detection, and hybrid representation learning to capture the structure of spatial change with continuity, establishing a coherent basis for quantifying thresholds as dynamic events embedded in the evolving spatial field.

3.1. 3D ISOVIST SAMPLING

In this study, thresholds are defined as statistically significant change points in spatial perception along path sequences (Feld et al., 2018). While spatial transitions are experienced through multiple sensory channels, including light, texture, and material, this work focuses on geometric configuration as a stable, quantifiable basis for computational modeling that remains invariant to those perceptual variables.

Following the movement-based spatial perception framework of Miao et al. (2024), agent trajectories are generated using PedSim (Wang, 2019), a Grasshopper plugin for simulating goal-oriented movement with collision avoidance. Each path is designed to intersect potential threshold regions to ensure consistent sampling of transitions. Along each trajectory, 3D isovists are computed at 0.5 m intervals. From an eye height of 1.6 m, 2048 rays are cast uniformly over a sphere using Fibonacci sampling (Cao et al., 2022), intersecting with mesh geometries up to a 10 m distance. The intersection distances are recorded as ordered vectors, preserving the directional continuity of spatial structure for subsequent encoding.

3.2. DETECTING THRESHOLD EVENTS

Drawing on established computer vision methods that use image-based representations to assess spatial qualities, a panoramic depth encoding method (Peng et al., 2017; Miao et al., 2024) is adapted to project spherical isovists into equirectangular greyscale images. Ray distances are mapped onto a 64×32 grid where pixel intensity reflects distance to surrounding geometry, with logarithmic scaling and percentile-based normalisation applied to enhance local contrast. While human visual perception exhibits anisotropic spatial attention (Krukar et al., 2021), with different sensitivity to horizontal and vertical space, the depth panorama encoding treats all spatial directions uniformly. This preserves complete geometric information for computational analysis rather than imposing a priori assumptions about perceptual salience.

Sedlmeier and Feld (2018) successfully classify the passageway spatial type by extracting classic isovist metrics and applying principal component analysis (PCA). Yet this dimensional reduction neglects spatial correlations fundamental to 3D threshold perception. To preserve such structure, the current approach measures frame-to-frame dissimilarity between consecutive depth panoramas using two complementary image metrics. (1) Frame-wise L2 distance (Euclidean norm) measures the magnitude of visibility change between consecutive isovists, capturing how much the visible environment has changed. (2) Structural Dissimilarity Index Measure (DSSIM, or 1-SSIM), a perceptually-motivated metric originally developed for image quality assessment (Wang et al., 2004), evaluates whether the spatial configuration has been fundamentally reorganised. While these two metrics generally exhibit similar trends and identify comparable transition points, they are each sensitive to geometric magnitude and perceptual structure respectively, thus providing a more comprehensive picture of spatial transitions (Figure 1). Both signals are then smoothed, normalised, and subjected to prominence-based peak detection. Threshold candidates are identified as local maxima in either metric, prioritizing frames where numerical magnitude or structural pattern indicate significant spatial transition. Each validated threshold is then centered within a 7-frame window (±3 steps) to capture the full perceptual transition sequence for classification.

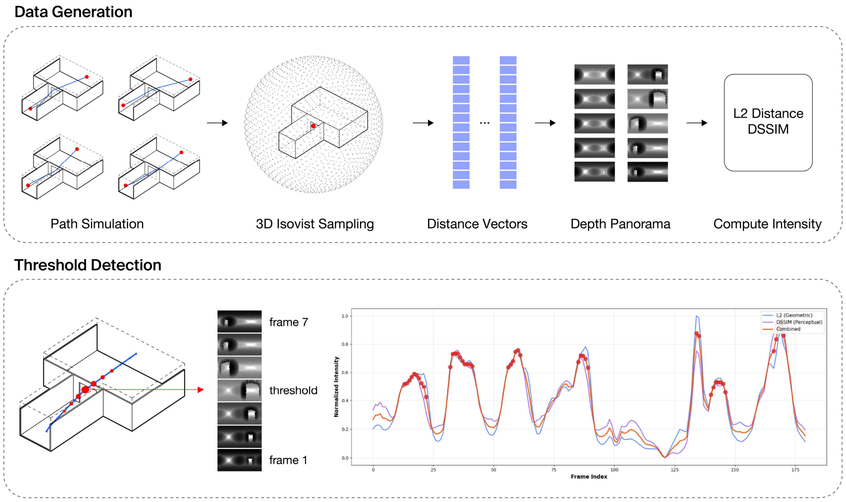

Figure 1. Data Generation and Threshold Detection Pipeline

3.3. LEARNING THRESHOLD REPRESENTATIONS

These 7-frame sequences are treated as short video clips and serve as the input to a TimeSformer encoder (Bertasius et al., 2021). Each frame is divided into non-overlapping patches that are linearly embedded and augmented with positional encodings, allowing the model to interpret each patch as a spatiotemporal token. Within each transformer block, divided space–time attention alternates between temporal and spatial operations, capturing both the evolution of visibility across movement and the organisation of spatial boundaries within each frame. This structure enables the encoder to learn how perceptual continuity emerges from sequences of local spatial change, allowing thresholds to be represented as distributed transitions instead of discrete geometric events.

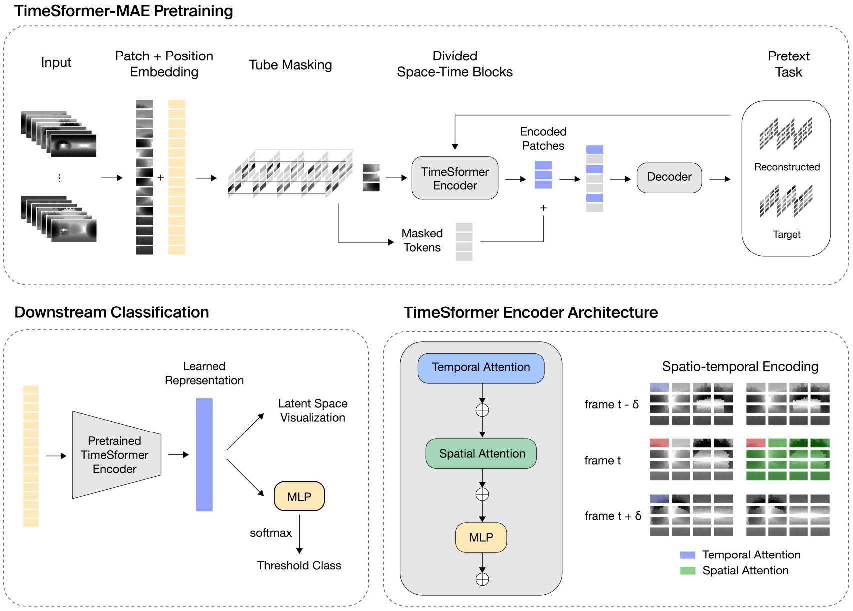

Figure 2. Pretraining and Downstream Task Pipeline

The encoder is pretrained using a masked autoencoding framework adapted from VideoMAE (Tong et al., 2022). The approach adopts a self-supervised learning strategy in which the model reconstructs masked regions of a video sequence from the remaining visible patches. During pretraining, a high ratio (~90%) of spatiotemporal tokens is randomly hidden using tube masking, and the encoder learns to infer the missing spatial information through contextual reasoning across space and time. A lightweight decoder subsequently reconstructs the masked portions, and training minimises the mean-squared error between the predicted and original depth panoramas. This training process encourages the model to internalise structural regularities of spatial change—how depth gradients, enclosure, and visibility evolve through movement. The framework is particularly suited to isovist-based depth data, where common visual augmentations such as cropping or rotation would disrupt geometric integrity. By reconstructing from partial information, the encoder learns the intrinsic logic of spatial configuration, forming a representation that captures both local geometry and global perceptual transitions. After pretraining, the decoder is removed, and the pretrained TimeSformer encoder serves as the backbone for downstream threshold classification (Figure 2).

4. Experiments

4.1. SYNTHETIC TYPOLOGIES

Experiment 1 examines whether a self-supervised encoder can learn the underlying structure of spatial transitions defined by systematic changes in height and width along movement paths. To construct a controlled setting, eight threshold types are generated, each representing a distinct configuration of dimensional change. Five geometric variations are generated for each type, and 100 paths are simulated per variation, yielding 4,000 modelled thresholds. Applying transition detection across these paths yields 10,514 seven-frame sequences, which serve as the basis for representation learning. The experiment evaluates whether self-supervised training produces a latent space that reflects how different modes of geometric change relate to one another and whether these relationships emerge as coherent structural groupings.

To analyse this learned space, embeddings from the pretrained encoder are clustered using k-means (Lloyd, 1982) and visualised with t-SNE (van der Maaten & Hinton, 2008) (Figure 3). The resulting clusters align with ground-truth typologies with 95.4% accuracy, echoing prior work showing that self-supervised learning can capture latent organisation in spatial and movement-based data (Johanes & Huang, 2021; Miao et al., 2024). Most ambiguities occur between t2 and t7 and between t4 and t6, pairs that share similar width-change profiles. Their proximity suggests that the learned representation identifies recurring geometric tendencies and arranges them according to shared structural characteristics.

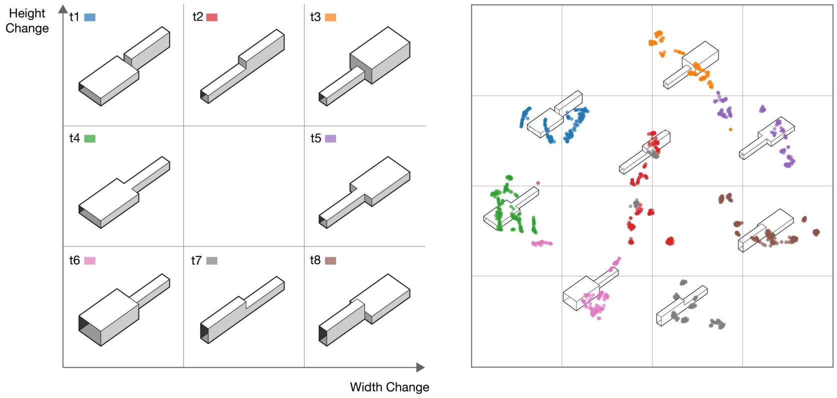

Figure 3. Left: eight synthetic typologies; right: t-SNE visualisation showing K-means cluster assignments performed on the latent representations

In the downstream stage, a lightweight Multi-Layer Perceptron (MLP) classifier is attached to the pretrained encoder and fine-tuned on the synthetic dataset. This step allows the system to map learned spatiotemporal features to human-interpretable typological categories. The resulting classifier achieves 99.75%, demonstrating that the pretrained representation is both highly discriminative and well suited for downstream threshold detection in real-world environments.

4.2. CASE STUDIES IN REAL-WORLD ARCHITECTURE

Experiment 2 evaluates how well the pretrained encoder transfers to architectural environments where thresholds are embedded in complex spatial arrangements and cultural histories. For each building, 100 movement paths are generated, producing 3,777 threshold windows. These paths sample multiple modes of traversal so that a range of threshold events can be captured. The aim is to evaluate whether a model trained on controlled geometric variations can categorise more intricate spatial transitions across different architectural contexts.

Four case studies are selected for their distinct threshold structures. Villa Müller (Prague, 1930) by Adolf Loos is selected for its spatial organisation under the Raumplan, where vertical offsets, partial-height partitions, and layered openings structure how interior thresholds guide and redirect the viewer's gaze (Colomina, 1990). House Without Qualities (Cologne, 1995) by O. M. Ungers represents a minimalist interior where transitions occur through precise dimensional adjustments. Sir John Soane's Museum (London, 1813) offers a highly articulated nineteenth century domestic environment where thresholds accumulate through partial views, stair shifts, and unexpected openings (Evans, 1978). Zhuozheng Yuan (Suzhou, Ming Dynasty), or the Garden of the Humble Administrator, introduces a classical Chinese garden whose framed views and colonnade-like passages differ fundamentally from Western interior thresholds. Together, these cases provide a diverse set of spatial conditions through which the model's capacity for transfer can be assessed.

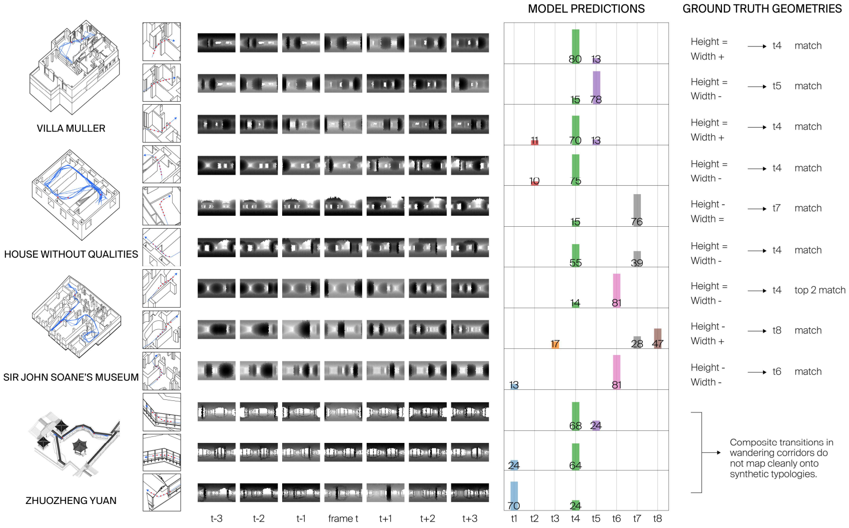

Figure 4. Sampled paths and corresponding typology prediction probabilities for case studies

Across the first three case studies, the encoder performs well and reveals spatial tendencies that align with the architectural logic of each environment. In Villa Müller, thresholds from the 100 sampled paths are predominantly classified as t5 (58.8%) and t1 (40.8%), which aligns with the width-based constrictions and releases characteristic of the Raumplan sequence. In House Without Qualities, predictions concentrate around t4 (77%) and t7 (9%), a pattern also seen in the selected thresholds in Figure 4, indicating that the building's controlled dimensional adjustments are consistently recognised. In Sir John Soane's Museum, the distribution is broader, with t6 (28.0%), t4 (24.3%), and t8 (19.4%) appearing most frequently. This variety reflects the more complex layering of rooms, levels, and partial enclosures, and the model's classifications track these overlapping geometric cues.

The results for Zhuozheng Yuan reveal the limits of the synthetic typology system. Many passages in the garden consist of repeated columns, partial occlusions, and open-frame transitions that do not correspond cleanly to controlled height or width changes. As a result, more than 95% of thresholds are predicted as t4, and only a small number of instances, illustrated in Figure 4, show any variation in the distribution. This outcome suggests that while the encoder transfers well to interiors with explicit dimensional modulation, it struggles to classify threshold conditions generated through landscape framing and episodic scene construction.

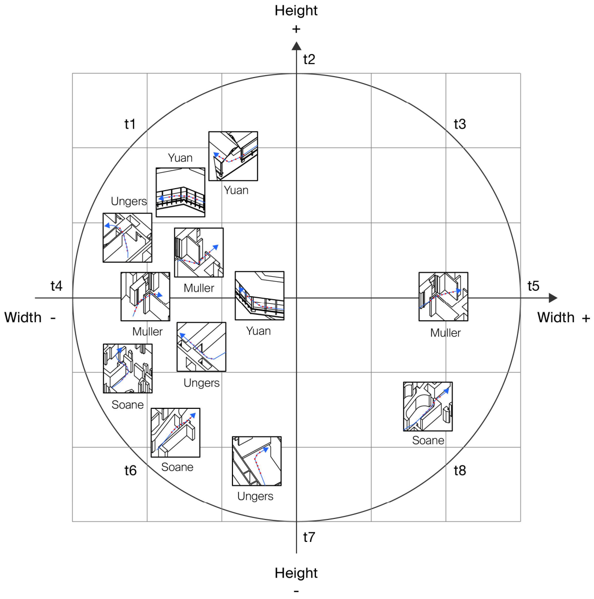

Figure 5. Classification map of case-study thresholds across eight synthetic typologies

To synthesise these observations, a classification map is constructed that positions twelve representative thresholds from the case studies within the eight synthetic typologies (Figure 5). This map visualises how real thresholds interpolate among the synthetic types and where hybrid or ambiguous cases emerge. Interpolation patterns reveal both alignment and deviation between modelled typologies and architectural conditions, providing insight into the structure of the learned representation and the degree to which synthetic geometries approximate threshold behaviour in diverse spatial settings.

5. Conclusion

This paper has introduced a computational framework that treats architectural thresholds as dynamic, movement-dependent events learned directly from sequences of 3D isovists. By combining depth panorama encoding, transition detection, and masked spatiotemporal transformers, the method demonstrates that consistent geometric signatures of thresholds can be captured in a latent space, yielding coherent clustering of synthetic typologies and transferable patterns across diverse case studies. The classification maps further show how real thresholds interpolate between canonical geometric types, revealing both alignment and friction between designed architectures and an abstracted typology space. By applying threshold classification across multiple buildings, the framework also offers a quantitative way to describe how different threshold types populate an environment, providing architects with a more precise tool for understanding and shaping transitional space.

However, several limitations refine the scope of these claims. Threshold classification remains path-relative, making results sensitive to the particular trajectories sampled. Although extensive 360° indoor datasets exist, they typically consist of static panoramas rather than continuous movement, limiting opportunities to validate sequential perception in real-world scans. The synthetic training data introduces modelling assumptions and geometric simplifications that shape what the encoder internalises as a "typical" threshold. Moreover, the emphasis on interior, dimension-based transitions narrows applicability to environments where thresholds arise through more diffuse or open-ended spatial cues. Finally, threshold experience is culturally and perceptually variable in ways that geometric visibility cannot fully represent. Future work will need to expand path sampling, incorporate real sequential datasets, and integrate human studies to evaluate how these learned representations align with lived, culturally situated experiences of transition.

Acknowledgement

ChatGPT (OpenAI, 2024) was used to improve the structure and flow of the manuscript; Claude (Anthropic, 2025) was used for code cleanup.

References

Anthropic. (2025). Claude 4.5 Sonnet [Large language model]. https://www.anthropic.com/claude/sonnet

Bachelard, G. (1969). *The poetics of space* (M. Jolas, Trans.). Beacon Press. (Original work published 1958)

Benedikt, M. (1979). To take hold of space: isovists and isovist fields. *Environment and Planning B: Planning and Design*, 6(1), 47–65. https://doi.org/10.1068/b060047

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In *Proceeding of the 38th International Conference on Machine Learning (ICML)*, 139, 813–824. https://arxiv.org/abs/2102.05095

Bhatia, S., Chalup, S. K., & Ostwald, M. J. (2012). Analyzing architectural space: identifying salient regions by computing 3D isovists. In *Proceedings of the 45th Annual Conference of the Architectural Science Association (ANZAScA)*, 53–62.

Cao, Y., Zheng, H., Liu, S. (2022). Measurement of Spatial Openness of Indoor Space Using 3D Isovists Methods and Fibonacci Lattices. In Computer-Aided Architectural Design.

Design Imperatives: The Future is Now. CAAD Futures 2021. Communications in Computer and Information Science, vol 1465. Springer, Singapore. https://doi.org/10.1007/978-981-19-1280-1_26

Colomina, B. (1990). Intimacy and spectacle: the interior of Adolf Loos. *AA Files*, 20, 5–15. http://www.jstor.org/stable/29543700

Evans, R. (1978). Figures, doors and passages. *Architectural Design*, 48(4), 267–278.

Johanes, M. & Huang, J. (2021). Deep learning isovist: unsupervised spatial encoding in architecture. In *Proceedings of the 41st Annual Conference of the Association of Computer Aided Design in Architecture (ACADIA)*, 134–141.

Krukar, J., Manivannan, C., Bhatt, M., & Schultz, C. (2020). Embodied 3D isovists: a method to model the visual perception of space. *Environment and Planning B: Urban Analytics and City Science*, 48(8), 2307–2325. https://doi.org/10.1177/2399808320974533

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), pp. 129–137. https://doi.org/10.1109/TIT.1982.1056489

Miao, S., Peng, W., Tsai, D., & Nagakura, T. (2024). Deep spatial memory: quantifying architectural spatial experiences through agent-driven simulations and deep Learning. In *Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA),* 1, 109–118. https://doi.org/10.52842/conf.caadria.2024.1.109

Norberg-Schulz, C. (1971). *Existence, space and architecture*. Praeger.

OpenAI. (2024). GPT-4o [Large Language Model]. https://openai.com/chatgpt/

Peng, W., Zhang, F., & Nagakura, T. (2017). Machine's perception of space: employing 3D isovist methods and a convolutional neural network in architectural space classification. In *Proceedings of the 37th Annual Conference of the Association of Computer Aided Design in Architecture (ACADIA)*, 474–481.

Sedlmeier, A., & Feld, S. (2018). Learning indoor space perception. *Journal of Location Based Services*, 12(3–4), 179–214. https://doi.org/10.1080/17489725.2018.1539255

Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 10078–10093. https://doi.org/10.48550/arXiv.2203.12602

Turner, A., Doxa, M., O'Sullivan, D., & Penn, A. (2001). From isovists to visibility graphs: A methodology for the analysis of architectural space. *Environment and Planning B: Planning and Design*, 28(1), 103–121. https://doi.org/10.1068/b2684

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.

Wang, P. (2017). PedSim: Pedestrian simulation in Grasshopper and Rhino.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. In *IEEE Transactions on Image Processing*, 13(4), 600–612. https://doi.org/10.1109/TIP.2003.819861