ShadeBench: A Benchmark Dataset and Method for Shade Simulation in Sustainable Society

Longchao Da[†], **Xiangrui Liu**[†], **Mithun Shivakoti**,
Thirulogasankar Pranav Kutralingam, Yezhou Yang, Hua Wei [*]
Computer Science
Arizona State University
longchao@asu.edu

Abstract

Heatwaves pose a significant threat to public health, especially as global warming intensifies. However, current routing systems (e.g., online maps) fail to incorporate shade information due to the difficulty of estimating shades directly from noisy satellite imagery and the limited availability of training data for generative models. In this paper, we address these challenges through two main contributions. First, we build an extensive dataset covering diverse longitude-latitude regions, varying levels of building density, and different urban layouts. Leveraging Blender-based 3D simulations alongside building outlines, we capture building shadows under various solar zenith angles throughout the year and at different times of day. These simulated shadows are aligned with satellite images in terms of the areas, providing a rich resource for learning shade patterns. Second, we propose the DeepShade, a diffusion-based model designed to learn and synthesize shade variations over time. It emphasizes the nuance of edge features by jointly considering RGB with the Canny edge layer, and incorporates contrastive learning to capture the temporal change rules of shade. Then, by conditioning on textual descriptions of known conditions (e.g., time of day, solar angles), our framework provides improved performance in generating shade images. We demonstrate the utility of our approach by using our shade predictions to calculate shade ratios for real-world route planning in Tempe, Arizona. We hope this work could provide a reference for urban planning in extreme heat weather and reveal its potential practical applications in the environment.

1 Introduction

Extreme weather is causing an increasing number of deaths worldwide, with heatwaves being a major contributing factor. According to a report [1], the frequency and intensity of extreme heat events have surged over the past two decades, with more than 178,700 deaths occurring annually (average from 2000 to 2019) as a direct result of high temperatures [2]. Research from the World Health Organization highlights that extreme heat is now one of the leading causes of weather-related deaths [3], disproportionately affecting vulnerable populations such as the elderly and those working or staying in outdoor areas with limited access to cooling infrastructure. This trend underscores the urgent need for adaptive measures, including heat-resilient urban design and shade-aware route-planning methods [4], to mitigate the public health impact of rising temperatures globally.

Since shade acts as a natural shelter that reduces direct exposure to solar radiation, understanding how shade changes in real-time is crucial for preparing outdoor activities and aiding urban planning

[*]Corresponding Author. [†] Authors contributed equally.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

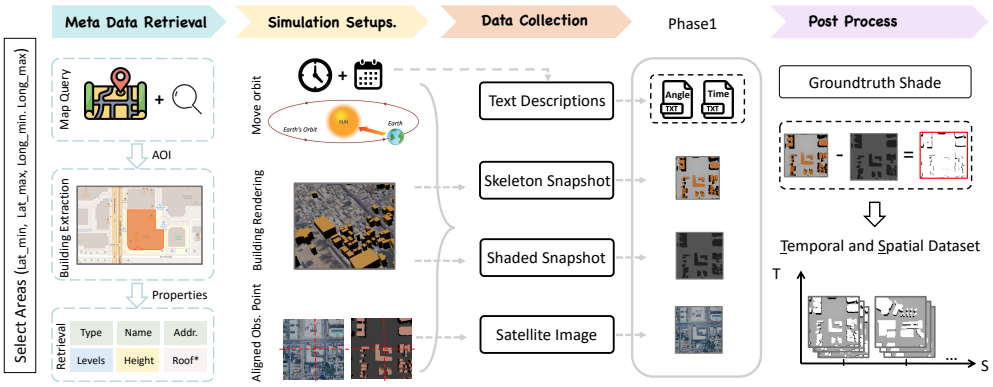

Figure 1: The overview of the pipeline, note that 'Addr.' is an abbreviation of Address, and 'Roof*' includes: roof shape and roof height, the building shape is approximated by the bird's-eye view of the roof shape.

in establishing artificial shelters in areas lacking natural shade. As introduced in [5], by identifying shaded areas and integrating this information into route-planning systems, individuals can make more comfortable travel choices, reducing their risk of heat-related illnesses. However, the existing research faces significant limitations: **First**, shade analysis using urban simulations is only on static maps, which lack good generalizability. **Second**, most of the methods are localized, relying on resource-intensive LiDAR data; they also lack scalability across different or more extensive areas. **Third**, they are unable to capture the real-time shade dynamics, limiting their utility in intelligent routing. These limitations impede the development of accurate shade modeling and timely planning, affecting the practical impact of existing studies.

To tackle the above challenges, a more adaptive and scalable approach to dynamically model shade variations is required. Given the success of Generative AI [6] (particularly diffusion models), in capturing spatial-temporal patterns in urban scene synthesis [7] and environment simulation [8], they present a promising direction for overcoming these challenges by generating shades for satellite images. However, to effectively leverage this method, a well-structured dataset that accurately aligns ground-truth shade-variance with its satellite-image level geographic information is necessary. Such a dataset is left blank in the current research domain.

In this paper, we **first** developed a rigorous and systematic pipeline to construct a comprehensive dataset for training shade-generation models. The dataset is designed to encompass three critical dimensions while adhering to a unified standard. The three dimensions include: (1) Geographical Diversity, covering a wide range of continents with varying latitudinal and longitudinal distributions; (2) Urban Layout Variability, capturing diverse urban configurations such as dense high-rise building places and sparse flat areas; and (3) Traffic Rule Variation, accounting for differences in driving orientations, including left-hand and right-hand traffic systems. The unified standard is that the cities must suffer from **significant heat events** as defined by the World Meteorological Organization (WMO), which guarantees the dataset focuses on regions with high temperatures and their associated impacts on shade dynamics are most pronounced. We **secondly**, designed a novel contrastive learning-based diffusion model approach, with a fine-granularity edge conditional module, that learns the shade variations based on the skeleton representations of corresponding satellite images. The model effectively maps shade dynamics to text prompts containing temporal and geographic information, allowing it to generate realistic shade predictions for arbitrary times of day or specific solar angles. This solution is unique for its ability to train on a limited satellite image dataset while generalizing to unseen buildings, provided an available satellite image.

In conclusion, this work contributes to advancing shade-generation research by addressing critical challenges in data preparation, model design, and real-world application. **First,** we developed a comprehensive, globally representative dataset, meticulously crafted to align satellite imagery with the dataset. **Second,** we introduced a text-conditioned image generation model leveraging edge conditioning and contrastive learning, enabling accurate and generalizable shade predictions. **Third,** we conducted extensive experiments across diverse cities, showcasing the model's robustness in handling varying landscapes, urban layouts, and geospatial features. **Finally,** we demonstrated a

practical application of the model in shaded route planning, where it generates shade maps for different times of the day based on satellite images and textual prompts. These contributions collectively provide a robust framework for urban planning and heat mitigation strategies.

2 The ShadeBench Dataset

The archived geographical information might be out-of-date, which hinders the feasibility of simulating shade changes, but it is possible to use accessible 'satellite image + generalizable models' to infer the shade areas.

We construct a comprehensive, generalization-oriented dataset that aligns satellite imagery with OSM [9] building data across Geographical Diversity, Urban Layout Variability, and Traffic Rule Variation. The pipeline as in Figure 1, proceeds as follows in a single pass: `Metadata Retrieval` gathers OSM-based building attributes (e.g., property type, address, height, levels, and polygonal geometry) as the 3D skeleton input; `Simulation Setups` then uses Blender [10] to simulate sun-driven illumination via a controller that follows solar declination or angle at arbitrary dates and times, rendering the Area-of-interest (AOI) scaled to match prevalent maps (Google map tile level 13); `Data Collection` produces four aligned modalities: *shaded snapshots* x^{shade} (with sunlight; cast shadows and illuminated surfaces), *skeleton snapshots* x^{sk} (sun off; pure structure), *satellite images* x^{sat} (real scenes co-registered with simulation), and *text descriptions* T capturing temporal/solar conditions; finally, `Post Process` extracts a clean ground-truth shade mask by subtracting structure and thresholding noise from the shaded render, enabling precise supervision. The following example showcases structured text prompts for different temporal conditions: We summarize the text-conditioning and ground-truth extraction below:

$$T = f(\theta_{sun}, t_{day}) \quad (1)$$

$$x^{gt} = x^{shade} - x^{sk} - \mathbb{I}(x^{shade} \leq \alpha) \quad (2)$$

Example Text Prompts:
Prompt 1: Solar declination: -20.7°
Prompt 2: Angle: 45°
Prompt 3: Right now, it is 6:00 PM in a day.

This process improves alignment between simulation and real-world imagery and enables controllable shade generation via images, textual prompts, or their combination. We provide a benchmark split of 70% training and 30% testing (a validation can be split as needed).

3 A Text Conditioned Shaded Image Generation Model

Given a shade dataset $D = \{x^{shade}, x^{sk}, x^{sat}, T\}$, where x^{shade} is a sunlit shaded snapshot, x^{sk} is a skeleton snapshot, x^{sat} is an aligned satellite image, and T encodes solar angle θ_{sun} and timestamp t_{day}. Our goal is to synthesize accurate shade images conditioned on a base map and text prompts. We adopt the diffusion-based ControlNet [11] as the backbone, leveraging diffusion models' strengths in image synthesis and editing [12]. As illustrated in Figure 2, our method builds upon the ControlNet with (i) edge-information incorporation and (ii) a contrastive-learning module.

Reflection on ControlNet in Our Task ControlNet extends diffusion models by injecting visual conditions into generation and builds on pretrained Stable Diffusion [13] through an additional control pathway for precise guidance. For shade synthesis, we condition on structure and description: the base map x^{sk} gives scene skeleton, and the text prompt T encodes solar angle θ_{sun} and timestamp t_{day}. These inputs form the conditioning c that drives generation: $x^I = G(T, x^{sk})$. At decoder layer i, ControlNet integrates the conditions as: $\mathbf{h}^{i+1} = D^i(\mathbf{h}^i + \mathbf{h}^i_{\text{cond}})$, where $\mathbf{h}^i$ is the current feature map and $\mathbf{h}^i_{\text{cond}}$ is derived from c. Vanilla ControlNet produces shade images yet struggles with subtle shade signals and time-dependent changes, which motivates the enhancements below.

3.1 Edge Enhanced Conditional Generation

We introduce an edge-enhanced condition to capture fine shade boundaries. The base map $x^{sk} \in \mathbb{R}^{H \times W \times 3}$ is a three channel mask of building footprints. Canny detection yields a single channel

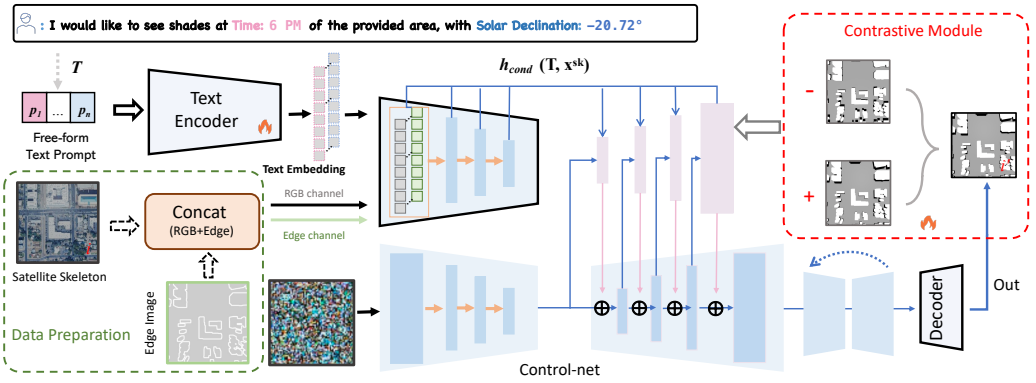

Figure 2: The structure of the proposed DeepShade method, based on the basic control net, we design two parts, **first** one on the left hand is the data preparation module, which concatenates the skeleton image with the Canny edge features, it helps the model focus on the shade edges and capture the overall skeleton of the building structures. The **second** on the right part is the contrastive part, the constructed contrastive buffer takes positive and negative pairs, during the training period, it optimizes the model by comparing the generated image i' with two contrastive pairs, and effectively learns the temporal difference reflected on the edge features.

edge map $x^{edge} \in \mathbb{R}^{H\times W\times 1}$. We concatenate them to form a four-channel condition: $x^{cond} = \left[x_R^{sk}, x_G^{sk}, x_B^{sk}, x^{edge}\right] \in \mathbb{R}^{H\times W\times 4}$ This tensor and the text prompt T feed the ControlNet U Net:

$$x^I = G\left(T,\ x^{cond}\right) \tag{3}$$

which encourages straight and well-aligned shadow boundaries.

3.2 Contrastive Based Shade Generation

We adopt a contrastive learning paradigm to promote temporal consistency in the image generation.

Contrastive Buffer Pairs Creation Let $D = \{x_1, x_2, \ldots, x_n\}$ with timestamp t_i and location l_i. Pairs follow:

$$\text{Label}_{ij} = \begin{cases} 1, & \text{if } l_i = l_j \text{ and } abs(|t_i - t_j|) = h, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Positive pairs share location with a timestamp gap of h hours. Negative pairs differ in location or have a larger gap. For each x_i we sample up to k_+ positives and up to k_- negatives[2]. The training set is $P = \{(x_i, x_j, \text{Label}_{ij}), x_i^{edge}, x_i^{sk}, T\}$. Unless stated, x_i and x_j are skeleton images.

Contrastive Learning for Shade Generation For each pair (x_i, x_j) we obtain embeddings h_i and h_j from a pretrained U-Net head and compute $S_{uv} = \frac{h_u \cdot h_v}{\|h_u\| \cdot \|h_v\|}$, $u, v \in \{1, \ldots, N\}$; we then optimize temporal consistency with InfoNCE $\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N}\exp(S_{ij}/\tau)}$; and use the total objective $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ControlNet}} + \lambda_1 \mathcal{L}_{\text{contrastive}}$ with $\lambda_1 = 0.1$. This joint loss improves temporal coherence and yields shade patterns that better match real world evolution over a day.

4 Experimental Study

In this section, we conduct experiments to verify the effectiveness of our proposed method. We have designed three sets of experiments, the **first** is to verify the model's performance in dense building cities, such as Beijing, Phoenix downtown, and São Paulo, The **second** is to show the performance in the relatively sparse environment, such as Tempe. The **third**, we perform an ablation study to understand the contribution of different components in our methods.

[2]We set $k_+ = 5$ for efficiency.

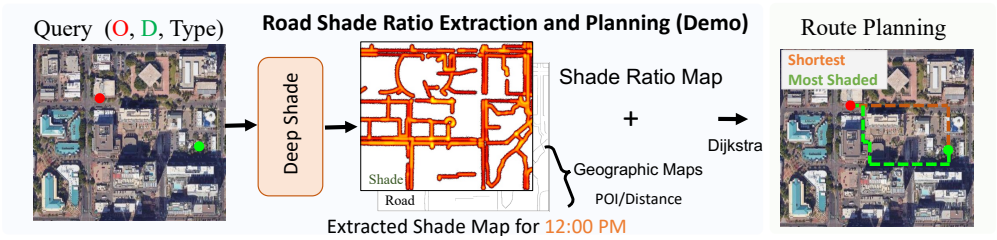

Figure 3: The demo of using our trained DeepShade in Tempe city for shade ratio extraction based on the generated shade maps, and we can effectively leverage this information to conduct planning that balances the distance and shade exposure at the Noon of a day (the preference score of shade is set as 50%, so distance and shade take half of the weight during planning).

Baselines on Scenario1	Beijing (CHN)		Phoenix (USA)		São Paulo (BRA)		Madrid (ESP)		Cairo (EGY)		Mumbai (IND)	
	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓
Diffusion Model	0.610	0.518	0.411	0.446	0.475	0.440	0.388	0.417	0.357	0.437	0.352	0.399
ControlNet	0.941	0.225	0.941	0.265	0.951	0.291	0.936	0.277	0.944	0.265	0.915	0.254
Edge Control	0.934	0.225	0.934	0.254	0.954	0.284	0.946	0.243	0.942	0.267	0.929	0.273
DeepShade	**0.945**	**0.194**	**0.946**	**0.164**	**0.959**	**0.210**	**0.948**	**0.239**	**0.954**	**0.257**	**0.931**	**0.250**
Baselines on Scenario2	Xi'An (CHN)		Tempe (USA)		Brasilia (BRA)		Seville (ESP)		Aswan (EGY)		Jaipur (IND)	
	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓
Diffusion Model	0.425	0.468	0.381	0.347	0.402	0.400	0.342	0.434	0.356	0.456	0.339	0.357
ControlNet	0.930	0.236	0.969	0.330	0.973	0.297	0.934	0.281	0.967	0.352	0.948	0.297
Edge Control	0.930	0.237	0.968	0.335	0.967	0.299	0.936	0.276	0.966	0.379	0.947	**0.273**
DeepShade	**0.932**	**0.233**	**0.969**	**0.329**	**0.979**	**0.241**	**0.936**	**0.242**	**0.973**	**0.283**	**0.956**	0.275

Table 1: **Test results from two scenario types (dense* and sparse* building cities)**: DeepShade consistently outperforms other baselines on both experimental scenarios, demonstrating superior language understanding and segmentation accuracy across in-domain and out-of-domain datasets, even when applied with random transformations. ↑ means the larger, the better, while ↓ vice versa.

5 Demonstration

Besides the quantitative analysis in experiments, it is important to demonstrate the real-world impact of the work. Thus, we design a proof-of-concept demo as shown in Figure. 3 using a subarea of Arizona State University. In this demo, we tackle the problem of integrating the shade ratio as a factor when making the routing suggestions. The input is a skeleton image containing the building outline, extracted from a satellite image of the interested area. Then, based on the time for planning, we describe as the text prompt T together with the image that is processed to x^{cond} as in Eq. 3, the shape map will be generated as output, it will be used for shade ratio calculation by overlaying the shade map with the road using longitude and latitude ranges. Given the shade ratio, the planning is made by jointly considering the user's preference (weight) on shade and distance by a variant of the Dijkstra algorithm. The green shows a more shaded plan, while the red means the shortest path. This demo reveals the potential of a real-world application; given that shade-involved planning is crucial for areas that suffer from extreme heat waves, this demonstration shows a way that could possibly help decrease heat stroke cases and improve the health of outdoor people.

6 Conclusion

We simulate realistic urban shade patterns by introducing both a novel dataset and a generative framework. We developed a diverse dataset of building layouts with aligned satellite imagery and timestamped shade snapshots, paired with text-condition encoding solar zenith angle and time of day. We then propose DeepShade, a text-conditioned, edge-enhanced diffusion model built on ControlNet that fuses building skeletons and Canny edges and incorporates a contrastive learning module to enforce temporal consistency. We have conducted extensive experiments across multiple cities (**Appendix A**), showing that DeepShade could function reasonably well in in-domain and out-of-domain tests. An ablation study further validates the importance of edge conditioning and contrastive loss, and lastly, we show a proof-of-concept application as **Figure. 3**, in shaded route planning task, which demonstrates the work's practical utility for heat-aware urban navigation.

References

[1] H. Ritchie, "How many people die from extreme temperatures, and how this could change in the future: Part two," *Our World in Data*, 2024, https://ourworldindata.org/part-two-how-many-people-die-from-extreme-temperatures-and-how-could-this-change-in-the-future.

[2] Q. Z. Shanshan Li, "World's largest study of global climate related mortality links 5 million deaths a year to abnormal temperatures," 2021.

[3] NIEHS, "Heat and health impacts of climate change," 2022, accessed: February 1, 2025. [Online]. Available: https://www.niehs.nih.gov/research/programs/climatechange/health_impacts/heat

[4] K. Ma, "Parasol navigation: Optimizing walking routes to keep you in the sun or shade," *Parasol Navigation: Optimizing walking routes to keep you in the sun or shade*, 2018.

[5] L. Da, R. Chhibba, R. Jaiswal, A. Middel, and H. Wei, "Shaded route planning using active segmentation and identification of satellite images," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 5205–5209.

[6] L. Da, M. Gao, H. Mei, and H. Wei, "Prompt to transfer: Sim-to-real transfer for traffic signal control with prompt learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 82–90.

[7] J. Tang, Y. Nie, L. Markhasin, A. Dai, J. Thies, and M. Nießner, "Diffuscene: Denoising diffusion models for generative indoor scene synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 20 507–20 518.

[8] L. Da, J. Turnau, T. P. Kutralingam, A. Velasquez, P. Shakarian, and H. Wei, "A survey of sim-to-real methods in rl: Progress, prospects and challenges with foundation models," *arXiv preprint arXiv:2502.13187*, 2025.

[9] F. Ramm, J. Topf, S. Chilton *et al.*, "Open street map," https://www.openstreetmap.org/, 2025, accessed: 2025-07-22.

[10] R. Hess, *Blender foundations: The essential guide to learning blender 2.5*. Routledge, 2013.

[11] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[12] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[13] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[14] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (ssim) revisited: A data-driven approach," *Expert Systems with Applications*, vol. 189, p. 116087, 2022.

[15] H. L. Tan, Z. Li, Y. H. Tan, S. Rahardja, and C. Yeo, "A perceptually relevant mse-based image quality metric," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4447–4459, 2013.

[16] S. Ghazanfari, S. Garg, P. Krishnamurthy, F. Khorrami, and A. Araujo, "R-lpips: An adversarially robust perceptual similarity metric," *arXiv preprint arXiv:2307.15157*, 2023.

[17] Y.-J. Cho, "Weighted intersection over union (wiou): a new evaluation metric for image segmentation," *arXiv preprint arXiv:2107.09858*, 2021.

[18] K. Deja, A. Kuzina, T. Trzcinski, and J. Tomczak, "On analyzing generative and denoising capabilities of diffusion-based deep generative models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 218–26 229, 2022.

A Experimental Study Details

A.1 Metrics

To evaluate the quality of generated shadow shades, we employ five commonly used metrics: Structural Similarity Index Measure (SSIM), Mean Squared Error (MSE), mean Intersection over Union (mIoU), Boundary Intersection over Union (B-IoU), and Learned Perceptual Image Patch Similarity (LPIPS) as in [14], [15] [16]. A more detailed explanation of the latter two metrics is as follows.

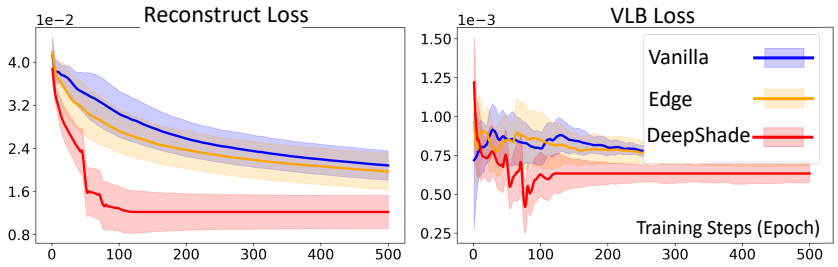

Figure 4: The training loss curves show that DeepShade shows obvious improvement regarding the convergence speed, in comparison to two baselines: edge-conditional generation and vanilla control net. This is attributed to the integration of edge features and the contrastive framework to improve training efficiency.

Boundary Intersection over Union (B-IoU) ↑ Inspired by the work [17], we proposed a new metric named: B-IoU, which measures the alignment of boundaries between the generated and ground truth shadow masks. If M_{pred} and M_{gt} are the binary shadow masks, and K is a 3×3 structuring element. We extract their boundaries via morphological dilation and erosion:

$$\partial M \;=\; \text{dilate}(M, K) \;-\; \text{erode}(M, K) \tag{5}$$

Let's denote that ∂M_{pred} and ∂M_{gt} are the predicted and ground-truth boundaries, B-IoU is computed as the Intersection over Union of these boundary sets:

$$\text{B}-\text{IoU} \;=\; \frac{\left|\partial M_{\text{pred}} \cap \partial M_{\text{gt}}\right|}{\left|\partial M_{\text{pred}} \cup \partial M_{\text{gt}}\right|} \tag{6}$$

where $|\cdot|$ denotes the number of boundary pixels. The range of B-IoU is $[0, 1]$, where higher values indicate better boundary alignment. Detailed implementation can be found in the evaluation code.

Learned Perceptual Image Patch Similarity (LPIPS) ↓ LPIPS quantifies perceptual similarity by comparing feature embeddings of the generated and ground truth images extracted from a pretrained network (e.g., AlexNet). It is computed as:

$$\text{LPIPS}(x, y) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} ||\phi_l(x)_{h,w} - \phi_l(y)_{h,w}||_2^2 \tag{7}$$

where ϕ_l denotes the feature map at layer l, and H_l, W_l are its height and width. LPIPS ranges from $[0, \infty)$, where lower values indicate greater perceptual similarity.

A.2 Result Analysis

A. The Efficient Convergence Speed. As shown in Figure. 4, the plotted curves are the reflection of the mean and standard deviation of the 5-rounds result for each method across two commonly adopted losses (reconstruction and VLB loss [18]). Our model DeepShade demonstrates a much more efficient convergence performance in comparison to baseline methods: edge-condition diffusion and vanilla controller. The training of diffusion models is essentially difficult; however, our proposed method provides a more rational way in the shade prediction setting.

Model	SSIM↑	mIoU↑	B-IoU↑	MSE↓	LPIPS↓
Backbone Model (direct)	$0.4252_{\pm 0.01}$	$0.0358_{\pm 0.00}$	$0.0213_{\pm 0.00}$	$41.2666_{\pm 1.65}$	$0.7967_{\pm 0.00}$
Vanilla Control Net	$0.9690_{\pm 0.04}$	$0.2736_{\pm 0.13}$	$0.0812_{\pm 0.05}$	$18.3388_{\pm 3.37}$	$0.3304_{\pm 0.03}$
Edge Condition	$0.9684_{\pm 0.01}$	$0.2898_{\pm 0.04}$	$0.1040_{\pm 0.01}$	$18.6686_{\pm 0.70}$	$0.3358_{\pm 0.01}$
Ours (DeepShade)	$\mathbf{0.9692}_{\pm 0.04}$	$\mathbf{0.2903}_{\pm 0.20}$	$\mathbf{0.1240}_{\pm 0.07}$	$\mathbf{18.1721}_{\pm 4.09}$	$\mathbf{0.3024}_{\pm 0.29}$

Table 2: Comparison of different models across various metrics: SSIM, MSE, mIoU, LPIPS, and B-IoU. In this experiment, each of the trained models is fed with a bird's-eye view satellite skeleton image and a text prompt describing the time and solar angle, and our method consistently performs better than these baseline methods.

B. Accurate Generation Across 12 Cities in the World. In this part, we focus on the Table. 1. The upper part is the dense areas in the world selected dataset for testing, and the lower part is the sparse areas. We can observe that our model outperforms most of the baseline methods in the various metrics (with a total of 50 epochs of training), regardless of the dense or relatively sparse scenario land covers. It demonstrates the vivid simulation of the conditioned area using the ControlNet-based method by training on one dataset, with good transferability, it better indicates that this work has great potential for real-world applications.

C. Ablation Study. We also included an ablation study in the paper, as in Table 2, it is the training conducted in the Tempe dataset, and the test is the other 30% from the original dataset given by the default split. This result reveals the importance of each component in our method. We can see that the edge condition (our method without contrastive learning) suffers the most performance drop in comparison to the DeepShape full model structure, and if we also remove the edge conditions, the performance further decreases, the backbone model is the stable diffusion model (all of the above models are trained 5 times with mean and std reported).

B Dataset Creation

In order to develop generalizable models, we construct a formal pipeline that creates a comprehensive dataset covering three dimensions: Geographical Diversity, Urban Layout Variability, and Traffic Rule Variation. This dataset aligns satellite images with Open Street Map (OSM) [9] building information. The pipeline involves four major steps as shown in Figure 1, which illustrates the overall framework from input to the collected outcome.

First, `Metadata Retrieval` obtains all necessary metadata used for the 3-D simulation. To align with the research community and utilize the large open-source data, we adopt OSM data as the metadata for shade simulation. Moreover, based on the longitude and latitude information, we can extract the building's geographical information, including "property type, address, height, levels, shape, and geometry set of locations", this typically provides information such as a list of points that form a polygon of building's 3D skeleton.

Second, `Simulation Setups`. With the necessary data collected, we performed the simulation to capture the shade changes for our dataset creation. In addition, we use Blender [10] for large-scale city-wide shade simulation. Based on scientific rules, we set up a controller as the sun's movement engine, which can generate the sun's trajectory following the sun's solar declination or angle on any day of the year, and at any time of the day. Then, we rendered the area of interest (AOI) within the simulation platform and adjusted the scale to ensure it aligns with the prevalent maps. Here, we align images with Google map tile level 13.

Third, `Data Collection`. We collected four types of data: (1) skeleton snapshots, (2) shaded snapshots, (3) satellite images, and (4) text descriptions. These diverse data representations enable learning models to understand shade variations and their dependencies on environmental factors.

C Limitation and Future Work

The paper proposes a very first dataset that aligns shades to the satellite view of a location; however, there might be missing building issues in ground truth shades since the OSM data is not always updated as the time the satellite image is taken. Future development to tackle this problem is

important; meanwhile, the real-time shade planning would require the pre-generated shade ratios, and they should be regularly indexed to the real-world road maps to enable large-scale support for the planning systems. A more accurate and robust framework is worth exploring.

D Reproducibility

The dataset `https://huggingface.co/datasets/DARL-ASU/DeepShade` and codebase `https://github.com/LongchaoDa/DeepShade_repo.git` will be released.