Leveraging Zero-Shot Learning on Street-View Imagery for Built Environment Variable Analysis

Siyuan Yao[1][0000−0002−4093−193X], Siavash Ghorbany[1][0000−0002−9588−0527], Matthew Sisk[1][0000−0002−4141−9655], Ming Hu[1][0000−0003−2583−1161], and Chaoli Wang[1][0000−0002−0859−3619]

University of Notre Dame, Notre Dame, IN 46556, USA
{syao2,sghorban,msisk1,mhu1,chaoli.wang}@nd.edu

Abstract. We present a novel approach to analyzing built environment variables (BEVs) using deep learning and Google Street View (GSV) images. By identifying and classifying BEVs, we aim to assist architecture professionals in understanding the relationship between heat-related health risks and BEVs. Traditional methods require extensive finetuning with human-labeled datasets, which is inefficient for analyzing diverse BEVs. Our approach integrates open-set object detection models with vision-language models to accurately identify buildings and classify wall materials without additional finetuning on our own human-labeled datasets. This versatile model can efficiently handle mixed materials, offering a cost-effective and scalable solution for analyzing the comprehensive built environment. The results will support architecture professionals in developing effective mitigation strategies for vulnerable populations living in less resilient housing, addressing public health risks associated with climate change.

Keywords: Machine learning · Sustainable cities · Built environment variables · Google street view · Heat-related health risks.

1 Introduction

The increasing frequency and intensity of climate change-related events pose significant risks to public health, especially for vulnerable populations residing in older, less resilient housing. By leveraging *Google Street View* (GSV) images and advanced computer vision techniques, we aim to automate the identification and classification of *Built Environment Variables* (BEVs), including wall materials, roof types, external shades, window-to-wall ratios, and housing conditions, which are critical indicators of a building's sustainability. The automated collection of BEV data using our method will enable architecture professionals to analyze the relationship between these BEVs and heat-related health risks, facilitating the development of effective mitigation strategies.

However, existing methods often rely on finetuning pretrained models with thousands of customized, human-labeled images for each object detection or classification task [28]. This approach is not cost-effective when analyzing multiple aspects of buildings or incorporating new variables, as each new task requires a specially labeled dataset. Additionally, current models typically use one-hot encoding for categorization, which lacks flexibility and cannot accurately detect overlapping categories (e.g., a wall containing wood and brick).

To address these issues, our approach integrates *Open-Set Object Detection* (OSOD) models with *Vision-Language Models* (VLMs), significantly enhancing accuracy and eliminating the need for additional finetuning with customized human-labeled images. This versatile model can analyze various BEVs, including mixed materials, without requiring multiple specialized models. It offers a cost-effective, scalable, and flexible solution for built environment analysis, enabling more comprehensive and nuanced assessments of building sustainability and associated health risks.

2 Related Work

Recent advances have significantly expanded the use of computer vision techniques in architectural analysis and urban studies. Starzyńska-Grześ et al. [28] conducted a comprehensive review, highlighting five key applications of computer vision in architecture. In these studies, while such analyses could traditionally be performed by human experts, computer vision models help accelerate and scale the process effectively. Comprising nearly half of the studies reviewed, street-view imagery has emerged as an ideal source for automated visual analysis. We will particularly focus on works involving street-view imagery closely related to our approach.

One application of computer vision in architecture is the classification of building styles and typologies [16, 10, 13, 8, 1]. These approaches automate the labor-intensive process of architectural surveys and historic preservation efforts by accurately identifying and classifying building styles. Another application is detecting and classifying building details [6, 34, 35, 5, 4]. These capabilities with street-view images are essential for various analyses, including urban modeling and maintenance planning, as they allow for detailed architectural features to be extracted efficiently and accurately. The qualitative analysis of urban environments using computer vision is also gaining traction. Studies have employed such algorithms to assess the quality of streetscapes, evaluate environmental aesthetics, and correlate visual data with socio-economic indicators [23, 33, 3, 15, 29, 9, 32, 22, 31]. Furthermore, computer vision has been applied to assessing property values [12, 17]. By automating the evaluation of building conditions and property values, researchers can provide more accurate and efficient assessments, benefiting stakeholders across the real estate and urban planning sectors.

Integrating computer vision in analyzing buildings and built environments offers promising opportunities for automating and enhancing various architectural and urban studies. Unlike the previous methods, by leveraging OSOD models and LLMs, we can assist researchers in gaining deeper insights into the built environment, leading to more informed decision-making and improved urban sustainability without requiring large-scale specialized datasets.

3 Our Approach

Our primary objective is to detect buildings in GSV images and classify their wall material types as representative BEVs. This aids architecture professionals in analyzing the relationship between BEVs and heat-related health risks. Traditional methods for similar architectural analysis tasks often rely on training with human-labeled datasets, which can be costly and inefficient when dealing with various BEVs, such as wall material, roof material, window types, and housing conditions. We aim to accomplish this in a zero-shot manner, eliminating the need for specially created human-labeled datasets.

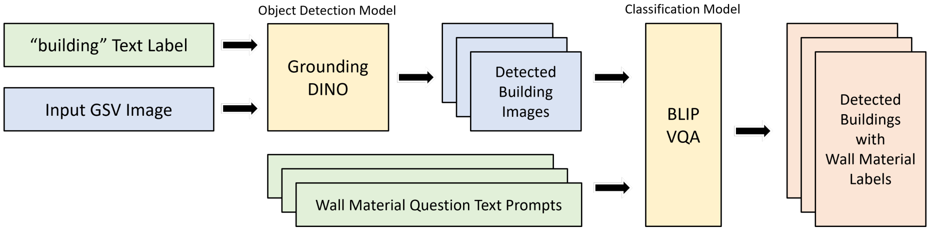

Fig. 1. Overview of our zero-shot approach for detecting wall material types of individual buildings in GSV images.

We employ a two-step approach for detecting wall material types of individual buildings in GSV images, as illustrated in Figure 1. First, we feed "building" text input and GSV image as input to Grounding DINO [19], a state-of-the-art OSOD model, to accurately identify buildings in GSV. Next, we leverage *Bootstrapping Language-Image Pre-training* (BLIP) [14] with its *Vision Question Answering* (VQA) feature to classify the wall materials. This is achieved by posing a series of questions related to wall materials and using the responses to determine the wall material type. This method enables us to determine wall material types efficiently without relying on extensive human-labeled datasets. We justify and explain the method in the following.

3.1 Open-Set Object Detection

Traditional *Closed-Set Object Detection* (CSOD) models, such as Faster R-CNN [26] and YOLO [25], can only detect objects belonging to the categories in their pretrained datasets without finetuning on customized datasets. For instance, the commonly used Faster R-CNN pretrained model is trained on the ImageNet 1K V1 dataset, which comprises 1000 different categories, with only a few related to buildings. The lack of semantic knowledge connecting these categories makes it difficult to generalize a single category for all buildings using Faster R-CNN without finetuning. Similarly, YOLOv1, the first version of YOLO, is pretrained on ImageNet with 1000 classes and cannot be used directly without finetuning.

YOLOv8, the most recent version, provides a pretrained model with 80 classes but does not include buildings. Thus, the primary rationale for using OSOD instead of CSOD lies in the constraints of the pretraining datasets, which OSOD does not have, and the flexibility OSOD provides through semantic classes in object detection.

We chose Grounding DINO as a state-of-the-art OSOD model. We found that it effectively detects various buildings, including houses, townhouses, apartment buildings, stores, gas stations, schools, churches, etc. Its robustness and versatility make it ideal for our detection task. More importantly, it can be used straight out of the box. In Section 4.3, we compare Grounding DINO with other methods to show its exceptional zero-shot capability for handling this task.

3.2 Vision-Language Models for Zero-shot Image Classification

Image classification tasks typically use CNN-based models such as VGG [27] and ResNet [7]. However, similar to the constraints faced by CSOD models, the classification categories are limited to those in the pretrained dataset if we want to use them out of the box. Since most of these models are pretrained on the ImageNet dataset, they cannot be directly applied to wall material classification tasks without finetuning on our own human-labeled dataset. Therefore, we must utilize models that learn image representations using natural language supervision to achieve zero-shot classification.

Integrating vision and language modalities, vision-language models are commonly used for zero-shot image classification. One of the most well-known models is *Contrastive Language-Image Pre-training* (CLIP) [24], which achieves performance comparable to ResNet on the ImageNet dataset while being significantly more robust across other natural image distributions. However, when analyzing wall materials as a BEV, it is essential to account for cases where walls contain wooden and brick parts. In such scenarios, even CLIP may not be sufficiently accurate in classifying the wall material types.

As capable of answering open-ended questions based on images, VQA models are a type of vision-language model that can retrieve images with specific characteristics. Such a capability is essential for identifying wall materials and mixed types. Given the primary question prompt about the material of a wall, VQA models respond by selecting the word associated with the highest likelihood score, even if the score is relatively low. This approach is sufficient for a basic classification, where the goal is to categorize the wall into a single material type. In our experiment, for example, approximately 98% of wooden walls were correctly identified by the answer from the BLIP model to the primary question, and 93% of brick walls were accurately identified.

However, to confirm if the wall is a mixture of multiple materials, in addition to the wall material, we also ask for the presence of wooden or brick parts as a supplement. We then compile the responses to categorize the wall material. For example, if VQA models indicate the presence of wooden and brick parts on a wall initially identified as wooden, we classify the wall material as a combination of wood and brick. This decision-making process is detailed in Table 3.

For the VQA models, we selected several feasible options. First, we chose ViLT [11], a unified vision-language transformer trained on comprehensive datasets, including Google Conceptual Captions (GCC), Stony Brook University Captions (SBU), Visual Genome (VG), and COCO Captions (COCO), and finetuned on the VQAv2 dataset. Second, we selected BLIP [14], which employs a novel vision-language pretraining framework that learns from noisy image-text pairs from web data. It is pretrained on datasets such as COCO, VG, CC, SBU, and an additional 115 million images with noisy texts from the web dataset LAION, and it is also finetuned on VQAv2. In addition to these specialized VQA models, we tested Large Language-and-Vision Assistant (LLaVA) [18], which combines a pretrained CLIP ViT-L/14 visual encoder with Vicuna, a variant of the large language model LLaMA.

Based on our findings from the comparison in Section 4.4, we determined that BLIP is the most effective model for our wall material classification task.

Table 1. Decision-making process for determining wall materials based on responses to questions by VQA models.

What Material?	Has Wooden Part?	Has Brick Part?	Class
Wood	Yes/No	No	Wood
Brick	No	Yes/No	Brick
Wood	No	Yes	Wood&Brick
Brick	Yes	No	Wood&Brick
Wood/Brick	Yes	Yes	Wood&Brick
Neither (e.g., Concrete)	Yes/No	Yes/No	Other

4 Experiments

4.1 Dataset

For building detection, we collected 672,526 GSV images from a county in the United States. Using the National Structure Inventory, we obtained the locations of 105,885 buildings in the county. We matched these images to the buildings and found that 97,130 buildings had GSV images within 100 meters. To validate the effectiveness of our method and compare it with other approaches, as shown in Figure 2, we randomly sampled 2,476 images and carefully annotated them into four categories: wood, brick, wood and brick, and other. The annotation process included drawing bounding boxes around the buildings in the images and labeling each building with its corresponding wall material. We used 2,000 images for training and the remaining 476 for testing. Note that our method and other zero-shot baselines did not use the training images.

For classification evaluation, we utilized 2,476 sampled GSV images to generate 6,130 cropped building images within the bounding boxes. We allocated

80% of these images for training and 20% for testing. Given the random sampling of GSV images, the number of images per category is not equivalent: there are 4,537 images for wood, 700 for brick, 678 for wood and brick, and 215 for other. The proportion of images in each category within the dataset essentially reflects the actual distribution found in the county.

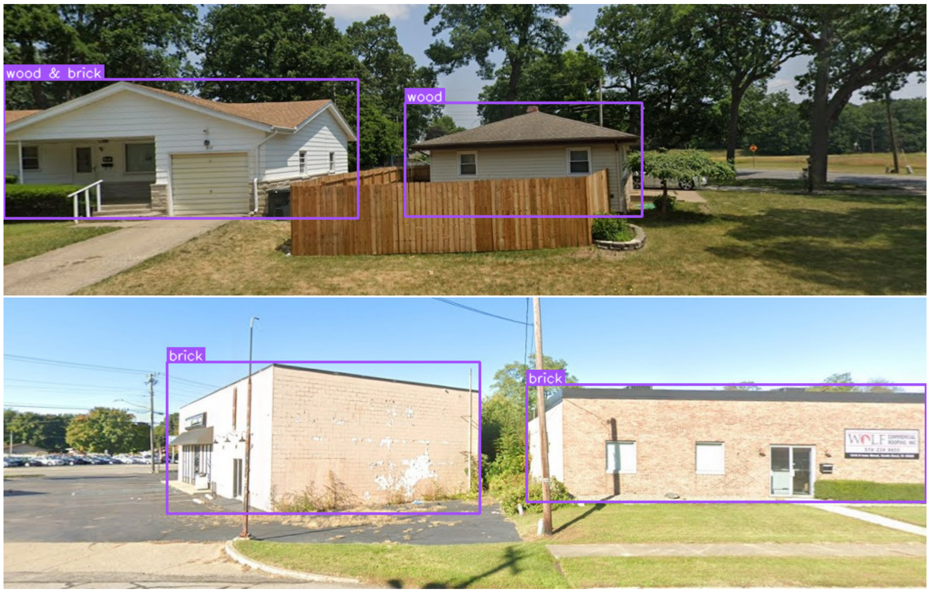

Fig. 2. Sample annotated GSV images.

4.2 Metrics

For building detection, we employ *mean Average Precision* (mAP) to rigorously evaluate the accuracy and effectiveness of each detection method. This metric provides a comprehensive measure of precision and recall across different *Intersection-over-Union* (IoU) thresholds, ensuring a robust assessment of object detection performance. mAP is computed as

$$\text{mAP} = \frac{1}{|C|} \sum_{c \in C} AP_c, \tag{1}$$

where C represents all classes, and c is one of the classes. The *Average Precision* (AP) for each class is given by

$$\text{AP} = \sum_{i} (R_i - R_{i-1}) \times P_i, \tag{2}$$

where P_i and R_i denote the precision and recall at threashold index i.

For building classification, we utilize accuracy to evaluate each classification method. The accuracy is computed as

$$\mathrm{ACC} = \frac{1}{N}\sum_{i}^{N} 1(y_i = \hat{y}_i), \tag{3}$$

where N represents the total number of predictions, y_i is the ith prediction and $\hat{y}_i$ is the corresponding ground truth.

The time for finetuning and inference is recorded in minutes, with all experiments running on a single NVIDIA A40 GPU with 48 GB of VRAM.

4.3 Single Category Building Object Detection

We compared Grounding DINO against Faster R-CNN (which utilizes a ResNet-50-FPN backbone) and YOLOv8, focusing solely on building detection. As shown in Table 2, despite Faster R-CNN and YOLOv8 being finetuned on our dataset of 2,000 human-labeled images, Grounding DINO demonstrates an mAP that is on par with these models. While Grounding DINO operates slower than its counterparts, the results underscore its efficacy as a robust choice for zero-shot building detection.

Table 2. Comparison of object detection models on buildings. We report mAP, Training Time (TT), and Inference Time (IT) in minutes. Best values are highlighted in bold.

Model	mAP	TT	IT
Faster R-CNN (ResNet-50-FPN)	0.80	24.33	0.85
YOLOv8	**0.92**	20.46	**0.05**
Grounding-DINO-Swin-T	0.90	—	1.56

4.4 Wall Material Classification

Considering the complexity of our task to classify wall materials, we evaluated various models, incorporating both zero-shot techniques and methods that require finetuning on our human-labeled datasets. For the models requiring finetuning, we explored the ResNet family, utilizing ResNet-18, ResNet-50, and ResNet-152 to represent the small, base, and large configurations, respectively. We also include EfficientNet [30] and ConvNeXt [21], which are designed to offer competitive accuracy and enhanced efficiency compared to Vision Transformers (ViTs) [2]. From the EfficientNet family, we selected EfficientNetV2-S and EfficientNetV2-L as the base and large models. We also tested the newer ConvNeXt series, employing the base model ConvNeXt-S and the larger ConvNeXt-L. ViTs were also included, with ViT-B/32 as the base model and ViT-L/16 as the larger option. Finally, we incorporated Swin-B [20] as the Swin Transformer series base model. This extensive comparison provides a thorough benchmark of the performance of leading classification methods when finetuned and evaluated on our datasets.

For the zero-shot methods, we examine models including CLIP, ViLT, BLIP, and LLaVA, as discussed in Section 4.4. Specifically for CLIP, we evaluate the base model ViT-B/32 and the larger model ViT-L/14, where the numbers 32 and 14 refer to the input patch size. For ViLT and BLIP, we select the models finetuned on the VQAv2 dataset to optimize VQA performance. Additionally, we include LLaVA, specifically version 1.6 and its largest 34b variant. In this comparison, we evaluate models of varying complexities to find the most suitable model for our task.

The results detailed in Table 3 reveal that while all finetuned models surpass 80% overall accuracy, only BLIP achieves greater than 80% accuracy, with ViLT closely approaching this mark. This superior performance can likely be attributed to both models being finetuned on the VQAv2 dataset, which is known for enhancing VQA capabilities. Interestingly, due to the lack of finetuning on VQA datasets, despite its considerable model size, LLaVA fails to predict the material of building walls accurately. Our analysis indicates that the multiple-question decision-making process described in Section 3.2 was particularly effective with BLIP. This process was also applied to ViLT and LLaVA but with less success. BLIP also demonstrates superior performance overall due to its higher and well-balanced accuracy rates in distinguishing between mixed wood and brick materials and the "other" category (non-wood or brick).

In terms of supervised models, the largest configuration of ViT performs optimally, albeit with significant training time and slower operational speed compared to other models. A more efficient alternative with comparable performance is the larger version of EfficientNet. The dataset shows a category bias towards buildings with wooden walls, particularly in the selected county, which aids in achieving high wood material accuracy (around 95%) and, consequently, an overall accuracy easily above 80%. However, the results might vary in areas with different class distributions, yet BLIP consistently maintains over 70% accuracy across all classes. This consistency suggests that BLIP is well-suited for robust inference in extended experiments on a larger scale. Such robustness makes it a reliable tool for architecture professionals to analyze the relationships between heat-related risks and BEVs, using wall material as a key BEV indicator.

4.5 Building Object Detection and Wall Material Classification

We evaluated our object detection and classification strategy, which integrates Grounding DINO with BLIP, against traditional models like Faster R-CNN and YOLO. As shown in Table 4, despite Faster R-CNN and YOLO being finetuned on our datasets covering four wall material classes, our method demonstrates superior performance, achieving a mAP of 0.82 with all categories scoring above 0.7 AP. This surpasses the performance of supervised models, underscoring the

Table 3. Comparison of classification models on buildings across four wall material categories. We report overall accuracy (%), category-specific accuracy (%), Training Time (TT), and Inference Time (IT) in minutes. Best values are highlighted in bold.

Model	All	Wood	Brick	Wood&Brick	Other	TT	IT
ResNet-18	85.17	95.81	68.57	42.65	48.84	8.55	**0.12**
ResNet-50	86.55	95.04	72.14	52.94	60.47	15.37	0.19
ResNet-152	86.31	94.93	69.29	56.62	53.49	20.2	0.34
EfficientNetV2-S	85.98	94.16	73.57	55.15	51.16	32.95	0.43
EfficientNetV2-L	87.53	94.82	**79.29**	58.09	53.49	122.62	0.81
ConvNeXt-S	85.66	94.49	75.71	50.74	41.86	113.48	0.28
ConvNeXt-L	86.55	**96.26**	65.00	55.88	48.84	187.47	0.71
ViT-B/32	84.76	95.59	65.00	43.38	51.16	16.78	0.18
ViT-L/16	88.02	95.48	75.00	60.29	60.47	363.25	1.39
Swin-B	86.88	95.93	72.14	53.68	48.84	44.68	0.47
CLIP-ViT-B/32	62.71	71.15	35.00	25.00	67.44	—	0.32
CLIP-ViT-L/14	68.38	81.83	12.86	38.97	58.14	—	0.49
ViLT-VQAv2	79.71	94.82	49.29	17.65	55.81	—	2.06
BLIP-VQAv2	**89.49**	94.49	77.86	**70.59**	**81.40**	—	12.70
LLaVA-1.6-34b	61.53	66.08	60.71	43.38	25.59	—	50.57

robustness and effectiveness of our zero-shot approach in handling complex detection tasks without specific training on the dataset.

These findings indicate that our method, combining Grounding DINO and BLIP, provides a compelling zero-shot solution, effectively managing classification demands with higher precision and adaptability than traditional, fully-trained models. This highlights the potential for zero-shot learning methods in practical applications, achieving high accuracy without extensive retraining on specialized datasets.

Table 4. Comparison of object detection models on buildings across four wall material categories. We report overall mAP, category-specific AP, Training Time (TT), and Inference Time (IT) in minutes. Best values are highlighted in bold.

Model	mAP	Wood	Brick	Wood&Brick	Other	TT	IT
Faster R-CNN (ResNet-50-FPN)	0.61	0.77	0.33	0.21	0.25	67.95	0.87
YOLOv8	0.74	**0.93**	**0.79**	0.64	0.59	29.28	**0.05**
Grounding-DINO-Swin-T + BLIP-VQAv2	**0.82**	0.86	0.72	**0.71**	**0.75**	—	6.21

5 Conclusions and Future Work

We have presented a novel zero-shot method combining an OSOD model, Grounding DINO, and a VLM, BLIP, to classify wall materials on buildings detected in GSV images. We compared our method to other models and justified our choices based on detailed evaluations. Our approach exhibits great robustness and reliability compared to methods finetuned on our own human-labeled dataset, demonstrating its capability to identify wall materials as a BEV for further analysis by architecture professionals.

Our method has a few limitations. It is slower than supervised methods, so scaling up the dataset, such as for the entire county could take days. To address this, we could use our method to create a machine-labeled dataset that can finetune more efficient models such as YOLO, ViT, or EfficientNet. VQA-based decision-making poses another concern, as its accuracy heavily depends on how questions are framed. Poorly designed questions can significantly affect the classification results. Thus, we should establish clear guidelines for crafting questions that elicit the most accurate responses from various VQA methods, especially when applying this technique to other BEVs. Furthermore, our current approach can identify combinations of wood and brick on walls but cannot determine the exact proportions of each material. Achieving this level of detail requires image segmentation, a more complex task. We plan to develop a zero-shot segmentation technique to estimate the proportions of wall materials better.

Acknowledgments. This research was supported in part by the U.S. National Science Foundation through grants IIS-1955395, IIS-2101696, OAC-2104158, TI-2317971, and IIS-2401144, the University of Notre Dame's Lucy Family Institute for Data & Society Health Equity Data Lab, and Notre Dame-IBM Technology Ethics Lab.

References

1. Alhasoun, F., González, M.: Urban street contexts classification using convolutional neural networks and streets imagery. In: Proceedings of IEEE International Conference on Machine Learning and Applications. pp. 1198–1204 (2019)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., J., U., Houlsby, N.: An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations (2021)
3. Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E.L., Li, F.F.: Using deep learning and Google street view to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences **114**(50), 13108–13113 (2017)
4. Ghorbany, S., Hu, M., Sisk, M., Yao, S., Wang, C.: Passive over active: How low-cost strategies influence urban energy equity. Sustainable Cities and Society **114**, 105723 (2024)
5. Ghorbany, S., Hu, M., Yao, S., Wang, C., Nguyen, Q., Yue, X., Alirezaei, M., Tasdizen, T., Sisk, M.: Examining the role of passive design indicators in energy burden reduction: Insights from a machine learning and deep learning approach. Building and Environment **250**, 111126 (2024)
6. Gong, F.Y., Zeng, Z.C., Zhang, F., Li, X., Ng, E., Norford, L.K.: Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. Building and Environment **134**, 155–167 (2018)

7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
8. Hu, C.B., Zhang, F., Gong, F.Y., Ratti, C., Li, X.: Classification and mapping of urban canyon geometry using Google street view images and deep multitask learning. Building and Environment **167**, 106424 (2020)
9. Ilic, L., Sawada, M., Zarzelli, A.: Deep mapping gentrification in a large Canadian city using deep learning and Google street view. PLoS ONE **14**(3), e0212814 (2019)
10. Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X.: Building instance classification using street view images. ISPRS Journal of Photogrammetry and Remote Sensing **145**(A), 44–59 (2018)
11. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: Proceedings of the IEEE International Conference on Machine Learning. pp. 5583–5594 (2021)
12. Law, S., Paige, B., Russell, C.: Take a look around: Using street view and satellite images to estimate house prices. ACM Transactions on Intelligent Systems and Technology **10**(5), 54:1–54:19 (2019)
13. Law, S., Seresinhe, C.I., Shen, Y., Gutierrez-Roig, M.: Street-Frontage-Net: Urban image classification using deep convolutional neural networks. International Journal of Geographical Information Science **34**(4), 681–707 (2020)
14. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of International Conference on Machine Learning. pp. 12888–12900 (2022)
15. Li, L., Tompkin, J., Michalatos, P., Pfister, H.: Hierarchical visual feature analysis for city street view datasets. In: Proceedings of IEEE VIS Workshop on Visual Analytics for Deep Learning (2017)
16. Li, X., Zhang, C., Li, W.: Building block level urban land-use information retrieval based on Google street view images. GIScience & Remote Sensing **54**(6), 819–835 (2017)
17. Lindenthal, T., Johnson, E.B.: Machine learning, architectural styles and property values. The Journal of Real Estate Finance and Economics (2021)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems (2023)
19. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10012–10022 (2021)
21. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
22. Min, W., Mei, S., Liu, L., Wang, Y., Jiang, S.: Multi-task deep relative attribute learning for visual urban perception. IEEE Transactions on Image Processing **29**, 657–669 (2020)
23. Naik, N., Philipoom, J., Raskar, R., Hidalgo, C.: Streetscore – predicting the perceived safety of one million streetscapes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 793–799 (2014)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable

visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning. pp. 8748–8763 (2021)

25. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of Advances in Neural Information Processing Systems. pp. 91–99 (2015)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations (2015)
28. Starzyńska-Grześ, M.B., Roussel, R., Jacoby, S., Asadipour, A.: Computer vision-based analysis of buildings and built environments: A systematic review of current approaches. ACM Computing Survey **55**(13s), 284:1–284:25 (2023)
29. Suel, E., Polak, J.W., Bennett, J.E., Ezzati, M.: Measuring social, environmental and health inequalities using deep learning and street imagery. Scientific Reports **9**, 6229:1–6229:10 (2019)
30. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of International Conference on Machine Learning. pp. 6105–6114 (2019)
31. Wang, M., Vermeulen, F.: Life between buildings from a street view image: What do big data analytics reveal about neighbourhood organisational vitality? Urban Studies **58**(15), 3118–3139 (2021)
32. Ye, Y., Zeng, W., Shen, Q., Zhang, X., Lu, Y.: The visual quality of streets: A human-centered continuous measurement based on machine learning algorithms and street view images. Environment and Planning B: Urban Analytics and City Science **46**(8), 1439–1457 (2019)
33. Yin, L., Wang, Z.: Measuring visual enclosure for street walkability: Using machine learning algorithms and Google street view imagery. Applied Geography **76**, 147–153 (2016)
34. Zeng, Z., Wu, M., Zeng, W., Fu, C.W.: Deep recognition of vanishing-point-constrained building planes in urban street views. IEEE Transactions on Image Processing **29**, 5912–5923 (2020)
35. Zhong, T., Ye, C., Wang, Z., Tang, G., Zhang, W., Ye, Y.: City-scale mapping of urban façade color using street-view imagery. Remote Sensing **13**(8), 1591:1–1591:17 (2021)