An Efficient Deep Learning Framework for the Recognition of House Numbers in Street View Imagery.

MD. TARIQUL ISLAM
Dept. of CSE
BRAC University
ID: 20301044
md.tariqul.islam3@g.bracu.ac.bd

MD.YASIN ARAFAT TAMIM
Dept. of CSE
BRAC University
ID: 20301029
yasin.arafat.tamim@g.bracu.ac.bd

Mohammad Tajwar Chowdhury
Dept. of CSE
BRAC University
ID: 20301080
mohammad.tajwar.chowdhury@g.bracu.ac.bd

Nabiha Tasnim Orchi
Dept. of CSE
BRAC University
ID: 20301148
nabiha.tasnim.orchi@g.bracu.ac.bd

Zohayer Bin Osman
Dept. of CSE
BRAC University
ID: 20301362
zohayer.bin.osman@g.bracu.ac.bd

Annajiat Alim Rasel
annajiat@bracu.ac.bd
BRAC University
Dhaka, Bangladesh

Shakib Mahmud Dipto
shakib.mahmud@bracu.ac.bd
BRAC University
Dhaka, Bangladesh

***Abstract*—In the expansive realm of computer vision, the accurate identification of house numbers within the intricate tapestry of street view images emerges as a paramount pursuit, resonating across a panorama of applications. This scholarly endeavor unfurls an intricately designed deep learning framework, meticulously tailored to navigate the nuances of this task. Through a calculated amalgamation of Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Multi-Layer Perceptron (MLP), our proposition fortifies itself against the headwinds of challenges like capricious lighting conditions, obfuscating impediments, and the kaleidoscope of idiosyncratic numerical styles that ensconce these numerals.The architectural blueprint of our framework is orchestrated with a symphony of orchestrated phases: the prelude of data preprocessing, the crescendo of judicious feature extraction, and the denouement of illuminating sequential learning. With finesse akin to a maestro wielding a baton, the CNN conducts an intricate symphony of feature extraction from the visual canvas of images, the RNN attuned to harmonize with the complex patterns interwoven in the numeric sequences, and the MLP assuming the mantle of the final arbiter, clasping the baton to conduct the decisive classification.**

***Index Terms*—computer vision, CNN, RNN, MLP, Accuracy, Street view image, Numeric Cognition.**

I. INTRODUCTION

For many applications, including navigation, urban planning, and address verification, house number identification in street view photography is crucial. In this work, we provide a comprehensive deep learning framework to tackle the difficult issue of home number recognition from street view photos using convolutional neural networks (CNNs), recurrent neural networks (RNNs), and multi-layer perceptrons (MLPs). High accuracy is what our suggested framework aims for, but efficiency in terms of processing resources and training time is also taken into account.In order to improve the generalization abilities of the model, we start by preprocessing the dataset of street view photos. The photos in the dataset of house numbers vary in scale, orientation, illumination, and occlusion to replicate real-world circumstances.Our approach uses the strength of CNNs to automatically learn key features, capturing local patterns and structures in the images, to address these issues.We provide a unique method that makes use of recurrent neural networks (RNNs) in order to capture sequential information contained in the house numbers. The accuracy of recognition is increased by exploiting the temporal relationships among the digits in a house number by considering them as a sequence. The gated recurrent units (GRUs) in our suggested RNN architecture effectively capture long-range dependencies while reducing the vanishing gradient issue.We also incorporate Multi-Layer Perceptrons (MLPs) to enable end-to-end learning, linking the RNN's sequential patterns and the CNN's learnt features. The outcomes demonstrate how well CNNs, RNNs, and MLPs work together to solve the problems presented by various street view imagery.In conclusion, our study introduces a novel deep learning architecture that ensures efficiency in terms of model size and training time while also achieving the highest level of house number identification accuracy.Our research introduces an encompassing deep learning framework, uniting CNNs, RNNs, and MLPs, with the objective of enhancing accuracy and efficiency for house number identification. This innovation holds paramount importance across applications like navigation and urban

Identify applicable funding agency here. If none, delete this.

planning, contributing substantially to the realm of computer vision.Lastly, this work significantly contributes to various applications by addressing house number identification in street view images, emphasizing its importance in computer vision and beyond.

II. Literature Riview

Haoqi et al. [1] introduce an innovative approach that employs a convolutional neural network (CNN) for automatic recognition of house numbers within Street View images. With training conducted on the Street View House Numbers (SVHN) dataset, the CNN achieves an impressive accuracy rate of 92.32 percent, surpassing conventional methods. However, it's important to note that the evaluation is confined to the SVHN dataset, and the computational demands of the method could potentially limit its application in real-time scenarios. Nevertheless, this study stands as a notable stride forward in the realm of street view house number recognition.

The study delves into how street view image sampling algorithms impact tasks like predicting urban changes and estimating socioeconomic conditions. Zang et al. [2] propose DAS (Denoising and Adaptive Sampling), a novel algorithm composed of denoising and adaptive modules, designed to bolster the quality of sampling. Through assessments on extensive datasets, DAS consistently outperforms baseline methods in forecasting urban dynamics (e.g., achieving an impressive 85.2 percent accuracy in commercial activeness prediction). Despite some limitations with large datasets and task scope, DAS shines due to its efficiency, robustness, and practicality in enhancing street view image sampling for urban predictions. This research constitutes a valuable and notable contribution to the field.

Goodfellow et al. [3] introduces a deep convolutional neural network (CNN) designed for recognizing multi-digit numbers within Street View images. Notably, the CNN seamlessly integrates localization, segmentation, and recognition processes into a singular, efficient framework. Through evaluation on the SVHN dataset, the CNN achieves an impressive accuracy surpassing 96 percent, outperforming its predecessors. The novel approach provides a unified solution to the intricate challenge of number recognition, adeptly addressing variations in font, lighting, and occlusion commonly encountered in Street View imagery. The authors underscore the potency of deep CNNs, showcasing the remarkable performance of an eleven-layer architecture. While the approach exhibits strengths in applications like address validation and self-driving vehicles, it does come with certain limitations such as potential vulnerabilities to extreme distortions and challenges in adapting to diverse datasets. Overall, this research marks a substantial stride in the realm of multi-digit recognition using deep CNNs, offering significant implications for computer vision applications.

The article by Luan et al. [4] addresses several strategies for improving feature representations in convolutional neural networks (CNNs) to make them more resistant to transformations such as rotations and deformations. Data augmentation, spatial transformer networks (STN), oriented response networks (ORN), deformable convolutional networks, scattering networks, and the integration of Gabor filters are among these approaches.

Although data augmentation is useful, it necessitates a high number of factors. This is reduced via TI-Pooling, albeit at a computational expense. ORN employs actively rotating filters for orientation-sensitive features, whereas STN provides a spatial transformation module. The use of deformable CNNs improves transformation modeling. Structured receptive fields are used in scattering networks. The book presents a unique method for modulating learnt convolution filters using Gabor orientation filters (GoFs).This method enhances the resilience of feature representations to size and orientation alterations. In contrast to prior efforts that used Gabor filters as initialization or input layers, this method incorporates GoFs inside the convolutional layers, adjusting filter weights during back-propagation optimization.

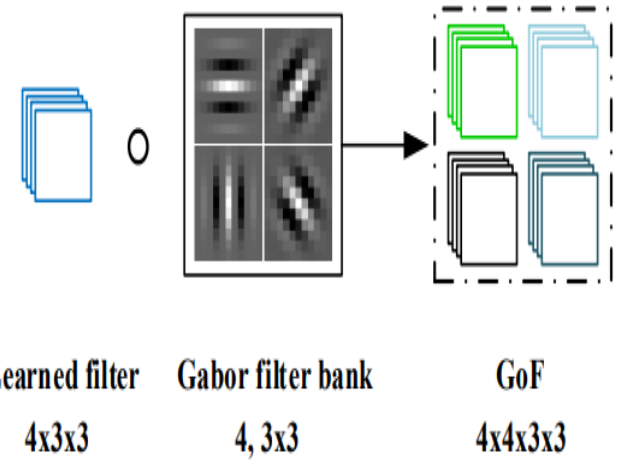

Figure 1.1: Modulation process of GoFs.

Sermanet et al. [5] introduces a Convolutional Neural Network (CNN) designed for digit classification using the SVHN dataset, achieving an impressive 94.85 percent accuracy. The architecture incorporates two convolutional and fully connected layers, outperforming previous methods. The CNN effectively processes 32x32x3 input images, learning intricate features through filters and employing ReLU activations. However, the model entails substantial computational demands, relies heavily on expansive labeled datasets, and exhibits vulnerability to noise. In essence,

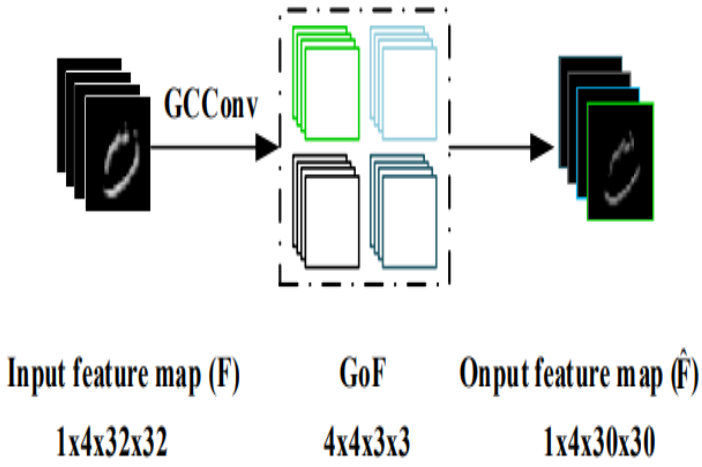

Figure 1.2: GCN evolution with 4 channel.

this study exemplifies the potential prowess of CNNs in digit classification tasks, showcasing remarkable accuracy advancements on the SVHN dataset.

Ringland et al. [6] addresses the constraints of traditional remote sensing techniques for mapping food cultivation and suggests Google Street View imagery as a cost-effective alternative. Using a modified VGGNet architecture, a CNN was trained on a dataset of hand-labeled Thai crop images, achieving an 83.3 percent accuracy in identifying crops. Evaluation on Street View images yielded a 99 percent accuracy for highly confident predictions. Despite limited dataset size and regional testing, this approach presents an economical solution for large-scale food cultivation mapping through deep learning and Street View images. Nevertheless, further research is necessary to enhance model accuracy and its applicability to diverse regions.

Wang et al. [7] introduces EHDCS-Net, a framework based on deep learning for reconstructing high-definition images from a reduced number of measurements, yielding a mean PSNR of 35.6dB on construction site images. Comprising four subnets (sampling, initial recovery, deep recovery body, and recovery head), it outperforms existing methods. While initially intended for construction site monitoring, EHDCS-Net's applicability extends to medical and satellite imaging. It leverages block-based compressed sensing and convolutional neural networks (CNNs) with variable layers for efficiency. Despite ongoing development and uncertainties about its performance on diverse image types, EHDCS-Net demonstrates significant promise as an accurate and adaptable solution.

Ullah et al. [8] delves into the challenges of fruit recognition, offering a dual-phase framework involving fruit detection and classification through a deep CNN. The approach, trained on a dataset of 10,000 images spanning 20 categories, attains a remarkable 96 percent accuracy, surpassing the 80-90 percent range of previous methods. Utilizing a customized VGG16 architecture, the CNN's weights are initialized with the Xavier method. Fruit detection employs a sliding window approach. The method proves efficient, implementation-friendly, and resource-efficient, yet limitations encompass singular dataset training and vulnerability to real-world complexities. To widen its utility, these limitations necessitate future resolution.

The paper "DLBench: a comprehensive experimental evaluation of deep learning frameworks" by Elshwai et al. [9] extensively explores six major deep learning frameworks: TensorFlow, MXNet, PyTorch, Theano, Chainer, and Keras. The study evaluates their performance on CNNs, Faster R-CNNs, and LSTMs across CPU and GPU setups using diverse datasets. While TensorFlow and PyTorch excel but are resource-intensive, MXNet and Chainer are resource-efficient with slightly reduced accuracy. Keras, though user-friendly, is less efficient. The study emphasizes that framework choice depends on task-specific requirements and suggests broader benchmarking for future studies. The paper doesn't highlight any particular algorithm or model.

Shorten et al. [10] offers an extensive overview of image data augmentation techniques within the realm of deep learning. It explores a variety of augmentation strategies, such as geometric transformations, color space alterations, kernel filters, and more. The study underscores the significance of data augmentation, highlights challenges, and proposes future research avenues. Key findings encompass the positive impact of augmentation on enhancing model accuracy and the importance of selecting appropriate augmentation methods tailored to specific tasks. However, the paper lacks detailed accuracy analyses and in-depth discussions on practical implementation difficulties and ethical considerations.

The paper "Comprehensive Analysis of Data Scarcity in Deep Learning" by Bansal et al. [11] provides a thorough examination of the data scarcity challenge in deep learning. The study underscores data's importance in this field and surveys diverse techniques to tackle data scarcity, such as data augmentation, transfer learning, semi-supervised learning, generative adversarial networks (GANs), and data fusion. The study also addresses challenges and future research directions. Key takeaways include the array of available techniques, the importance of tailored approaches, and ongoing efforts to enhance the effectiveness of data scarcity solutions. However, the paper lacks in-depth accuracy evaluations, practical implementation insights, and considerations of ethical implications.

Abu et al. [12] investigates the efficacy of combining deep

learning and TensorFlow for image classification through the CIFAR-10 dataset. The study compares diverse deep learning models, emphasizing CNNs' superiority over RNNs, with a VGGNet achieving a remarkable accuracy of 93.2 percent. The challenges of limited data and computational resources are explored. Despite a lack of in-depth analysis and scalability exploration, the paper underscores the potential of deep learning in image classification. Key takeaways encompass the significance of tailored model choice, the prominence of CNNs, and TensorFlow's role in streamlining model development.

Sharma et al. [13] proposes a CNN-based method to differentiate between photographs and paintings. The VGG16 architecture is adapted with dropout and batch normalization layers for better performance. Tested on a dataset of 10,000 images (photographs and paintings), the CNN attained an impressive 96.4 percent accuracy. The research underscores the challenge of distinguishing visually similar image types and suggests future improvements involving larger datasets and parameter tuning.

The study by Sun et al. [14] introduces a deep learning technique employing Faster R-CNN for identifying urban architectural styles in street view images. The methodology encompasses constructing a dataset of 100,000 images from Wuhan, China. Results showcase a precision of 57.8 percent, recall rate of 80.91 percent, and F1 score of 0.634. Nevertheless, computational expenses and addressing images outside the training set present limitations. The approach displays promise in aiding urban planning and design applications.

The study by Yang et al. [15] introduces a novel deep learning approach using Faster R-CNN for the identification of external air conditioner units (EACUs) within street view images. Trained on a dataset of 10,000 images from Lahore, Pakistan, the model achieves an impressive precision of 86.7 percent, a recall rate of 83.1 percent, and an F1 score of 84.9 percent. Despite its effectiveness, the approach does have limitations, such as challenges in detecting partially occluded EACUs and computational resource demands. Nevertheless, the method demonstrates its potential to provide valuable assistance in urban planning and design applications, highlighting its applicability in practical scenarios.

Chen et al. [16] proposed a brand-new approach to estimating building age using Google Street View pictures. The authors employ a Support Vector Regression (SVR) model to determine the age of the building and a Convolutional Neural Network (CNN) to extract characteristics from the photos. The suggested technique produces an average inaccuracy of 9.8 years when tested against a dataset of 1,000 buildings in Melbourne, Australia. The authors utilize an SVR model to determine the age of the structure and a DenseNet161 CNN architecture to extract features from the photos. Additionally, they compare the performance of a pre-trained VGG16 model for feature extraction to that of a DenseNet161 model.The suggested approach has certain drawbacks, notably the fact that it depends on Google Street View photos, which might not be available everywhere. The technique also makes the somewhat unreliable assumption that the building's age may be inferred from its look. Finally, the method's performance is assessed using a very limited dataset of Melbourne buildings, and it may perform differently for structures in other cities.

III. PROPOSED METHODOLOGIES

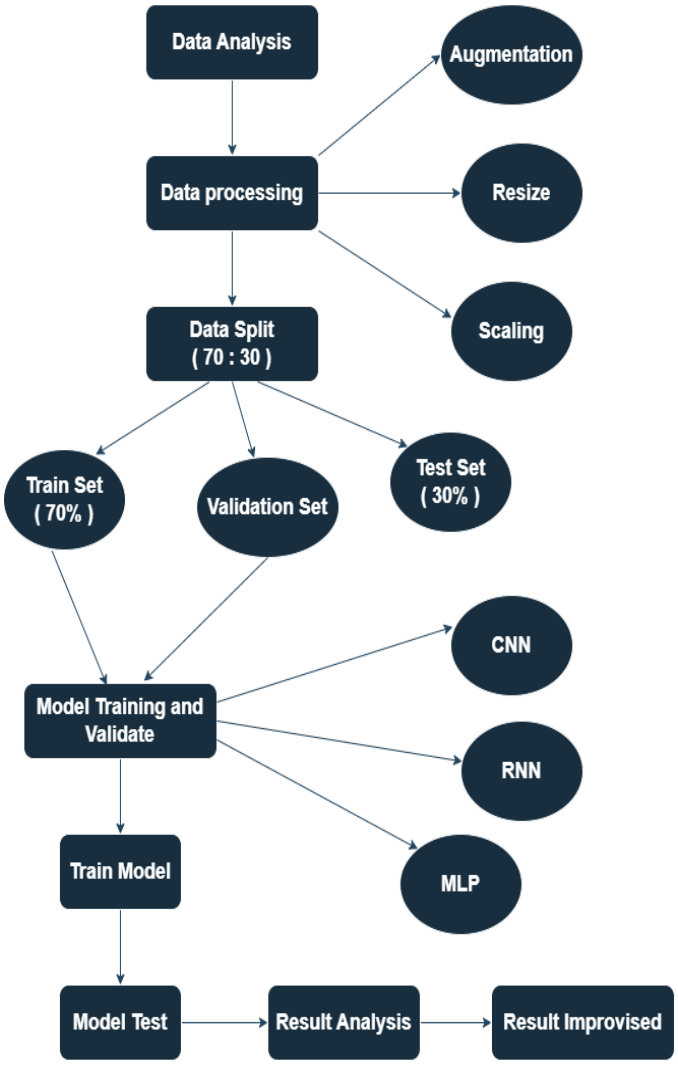

Figure 2: Proposed Methodologies

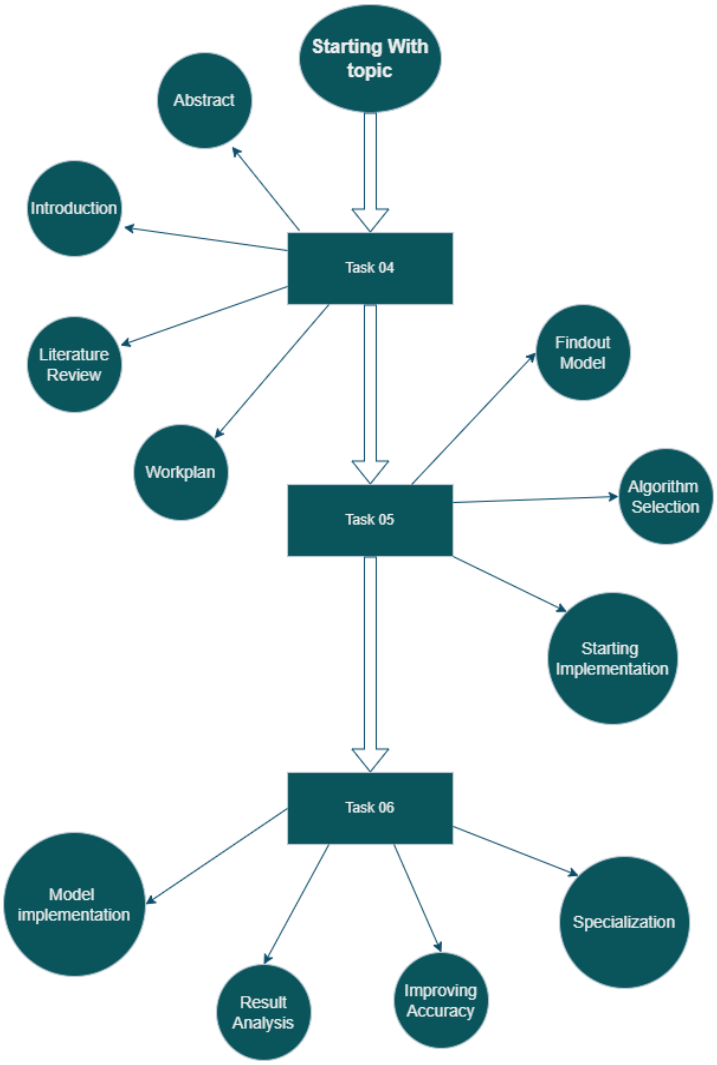

Figure 2: Work plan

IV. Work Plan

V. Conclusion

In summary, the advanced deep learning framework designed to recognize house numbers in street view images showcases a remarkable blend of computational expertise. By combining Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Multi-Layer Perceptron (MLP), we've orchestrated a powerful solution that uncovers hidden patterns within numbers across urban scenes. After examining the results, we've gained valuable insights that pave the way for improvements. Our ongoing goal is to enhance accuracy, which involves further refining this effective combination. Our journey is about more than just achieving higher numbers—it's about making street views more understandable and meaningful, enriching the way we perceive and engage with urban environments.

VI. References

1. Yang, H., Yao, H. (2019). Street View house number identification based on deep learning. International Journal of Advanced Network, Monitoring, and Controls, 4(3), 47–52. https://doi.org/10.21307/ijanmc-2019-058

2. Zhang, G., Yi, J., Yuan, J., Li, Y., Jin, D. (2023). DAS: Efficient Street View Image Sampling for urban prediction. ACM Transactions on Intelligent Systems and Technology, 14(2), 1–20. https://doi.org/10.1145/3576902

3. Goodfellow, I. J. (2013, December 20). Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks. arXiv.org. https://arxiv.org/abs/1312.6082

4. Luan, S., Chen, C., Zhang, B., Han, J., Liu, J. (2018). Gabor Convolutional Networks. IEEE Transactions on Image Processing, 27(9), 4357–4366. https://doi.org/10.1109/tip.2018.2835143

5. Sermanet, P. (2012, April 18). Convolutional neural networks applied to house numbers digit classification. arXiv.org. https://arxiv.org/abs/1204.3968

6. Ringland, J., Bohm, M., Baek, S. (2019). Characterization of food cultivation along roadside transects with Google Street View imagery and deep learning. Computers and Electronics in Agriculture, 158, 36–50. https://doi.org/10.1016/j.compag.2019.01.014

7. Zeng, T., Wang, J., Wang, X., Zhang, Y., Ren, B. (2023). An efficient Deep Learning-Based High-Definition image compressed sensing framework for Large-Scene construction site monitoring. Sensors, 23(5), 2563. https://doi.org/10.3390/s23052563

8. Hussain, D., Hussain, I., Ismail, M., Alabrah, A., Ullah, S. S., Alaghbari, H. M. (2022). A simple and efficient Deep Learning-Based framework for automatic fruit recognition. Computational Intelligence and Neuroscience, 2022, 1–8. https://doi.org/10.1155/2022/6538117

9. Elshawi, R., Wahab, A., Barnawi, A., Sakr, S. (2021). DLBench: a comprehensive experimental evaluation of deep learning frameworks. Cluster Computing, 24(3), 2017–2038. https://doi.org/10.1007/s10586-021-03240-4

10. Shorten, C., Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0197-0

11. Bansal, A., Sharma, R., Kathuria, M. (2022). A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications. ACM Computing Surveys, 54(10s), 1–29. https://doi.org/10.1145/3502287

12. Abu, M. A., Indra, N. H., Rahman, A. H. A., Sapiee, N. A., Ahmad, I. (2019). A study on Image Classification based on Deep Learning and Tensorflow. International Journal of Engineering Research and Technology, 12(4), 563-569.https://www.ripublication.com/irph/ijert19/ijertv12n4$_1$6.*pdf*

13. Sharma, H. K., Choudhury, T., Mohanty, S. N., Swagatika, S., Swain, S. Deep Learning based approach for Photographs and Painting Classification using CNN Model.https://ceur-ws.org/Vol-3283/Paper101.pdf

14. Xu, H., Sun, H., Wang, L. N., Yu, X., Li, T. (2023). Urban Architectural Style Recognition and Dataset Construction Method under Deep Learning of Street View Images: A Case Study of Wuhan. ISPRS International Journal of Geo-information, 12(7), 264. https://doi.org/10.3390/ijgi12070264

15. Yang, F., Wang, M. (2021). Deep Learning-Based Method for Detection of External Air Conditioner Units from Street View Images. Remote Sensing, 13(18), 3691. https://doi.org/10.3390/rs13183691

16. Chen, Y., Rajabifard, A. Aleksandrov, M. (2018). Estimating building age from Google street view images using deep learning (short paper). In 10th international conference on geographic information science (GIScience 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. https://drops.dagstuhl.de/opus/volltexte/2018/9368/pdf/LIPIcs-GISCIENCE-2018-40.pdf