FINE-GRAINED CLASSIFICATION OF URBAN BUILDINGS USING LOW-RESOLUTION OVERHEAD IMAGES

Zhiyi He[1], Wei Yao*[1], Jie Shao[1], Puzuo Wang[1]
[1]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University
Email: zhiyihe@polyu.edu.hk; Email: wei.hn.yao@polyu.edu.hk;
Email: jie.shao@polyu.edu.hk; Email: puzuo.wang@connect.polyu.hk

KEY WORDS: Super-resolution, Image classification, Denoising diffusion probabilistic model, Building types, Urban remote sensing

ABSTRACT: Fine classification of urban buildings into detailed (sub-)functional types based on satellite images is a crucial research area with significant implications for urban planning, infrastructure development, and population distribution analysis. However, this task faces huge challenges due to the low resolution and overhead view limitation of high altitude satellite images and urban buildings across multiple fine-grained categories exhibit significant variations in the number and distribution, leading to class imbalance problem. To address these issues, we propose a new approach to fine-grained classification of urban buildings from low-resolution overhead imagery in two steps. In the first phase, we introduce a Denoising Diffusion Probabilistic Model (DDPM) based super-resolution method to enhance the quality of low-resolution satellite images. This approach leverages the transfer of features across different domains to improve the visual quality and level of detail in the images, enabling more accurate building classification in the next phase. In the second phase, we develop a new fine-grained building classification network with two improvements: Category Information Balancing Module (CIBM) and Contrast Supervision (CS) technique. The CIBM module addresses the issue of class imbalance by adaptively adjusting the weights of different building categories, ensuring a more balanced class representation in the training process. Additionally, the CS technique provides contrastive learning-based supervision sources, enabling the network to capture more discriminative features for improving classification accuracy. We evaluate our method on a newly constructed urban building fine-grained classification dataset for Hong Kong city with 11 building fine categories and achieve promising results, our method's mean Top-1 accuracy reached 60.45%, superior to existing state-of-the-art classification method. Extensive ablation experiments are conducted to validate the effectiveness of our method, which demonstrates that CIBM and CS are able to improve Top-1 accuracy by 2.6% and 3.5% against the baseline method, respectively. We are also glad to find out a positive coupling effect between these two modules, which improved the network performance by 7.8% once combined to work on the data at the same time. And they can also be easily inserted to other classification networks to achieve similar enhancements.

1. INTRODUCTION

Buildings are crucial urban structures that play a key role in human habitation and socio-economic activities. The rapid advancement of imaging technology for remote sensing has made it possible to detect large areas of buildings from remotely sensed images cost-effectively. Automatic identification of precise boundaries and fine-grained categories of buildings can aid emergency management, 3D reconstruction, mapping and urban planning[1]. However, much of the current research in building classification relies on either street view imagery or sub-metre satellite imagery, producing only coarse building categorisations. Therefore, acquiring cost-effective imagery is crucial for this research, and improving the accuracy of building classification is an outstanding issue to be resolved. This paper is dedicated to the study of a methodology for the fine-grained classification of buildings in high-density urban areas only using easily accessible low-resolution satellite imagery, to lower the threshold of access to useful data and reduce the economic and time costs.

2. RELATED WORKS

2.1 Building Classification with Street View Images

Dominik Laupheimer et al. categorized terrestrial images of building facades into five broad categories. They used Convolutional Neural Networks (CNNs) to classify the street view images. However, the error rate of 36% misclassified images highlights the necessity for further improvement[2]. Jian Kang et al. obtained Google Street View images from the USA and Canada to perform architectural semantic classification using CNN, instead of directly using the satellite imagery[3]. Recognizing buildings from Street View images and encoding them for image classification, as proposed by Kun Zhao et al, is also a ground-breaking but useful approach[4]. These methods offer useful guidance for the functional classification of urban buildings using streetscape images. However, they do have some limitations. Firstly, street view images rely on high-resolution images, which can be costly to obtain and may result in omissions when buildings are obstructed. So, not all buildings can be classified using this approach. Secondly, these methods currently have a limited number of functional building categories and cannot achieve a fine-grained classification of building functions.

2.2 Building Classification with Satellite or Aerial Images

C. Xiao et al. proposed the utilization of oblique-view images to categorize building functions. They classified the building functions into four distinct categories during their experimentation. Subsequently, the final test demonstrated a classification accuracy of 60%[5]. Xingliang Huang et al. conducted a study on building detection and classification using high resolution satellite images with a resolution of 0.5-0.8m. The focus was on the object-level interpretation of individual buildings, enabling a 5-category vocabulary classification of buildings. However, the data required extensive use of semantic information, which is known to be labour-intensive[6-7].

2.3 Building Classification with Satellite or Aerial Images

Vannucci et al. proposed a radial basis-based under-sampling technique: removing commonly occurring samples in the training set and adaptively determining the optimal imbalance rate for various datasets. This technique resulted in improved model classification performance and enhanced model generalisation abilities[8]. Hasib et al. proposed the Hybrid Sampling with Deep Learning Method (HSDLM). The dataset is pre-processed via label coding, with noise being removed through the under-sampling algorithm. They also use the SMOTE over-sampling technique to balance the data and implements three parallel types of LSTM to improve accuracy[9]. These works aim to address category imbalances through methods such as up-sampling and down-sampling. However, none of these approaches take into account the issue of sample similarity within categories.

3. METHODOLOGY

In this section, we outline the data processing methods employed in our study. Initially, we present a general overview of the data flow associated with our proposed scheme. This is subsequently followed by a detailed description of two enhancement modules designed to address the challenges posed by low resolution imagery and imbalanced category information in satellite images.

3.1 Overview of proposed method

In this section we will describe how our approach works to solve the problems mentioned above. As shown in the Figure 1, our processing flow is divided into two phases. The first phase is a super-resolution network of low-resolution satellite images using transfer learning-based DDPM, in this part, we fine-tuned a super resolution model with Hong Kong overview satellite image pairs, with which target low resolution images are transferred to high resolution images.

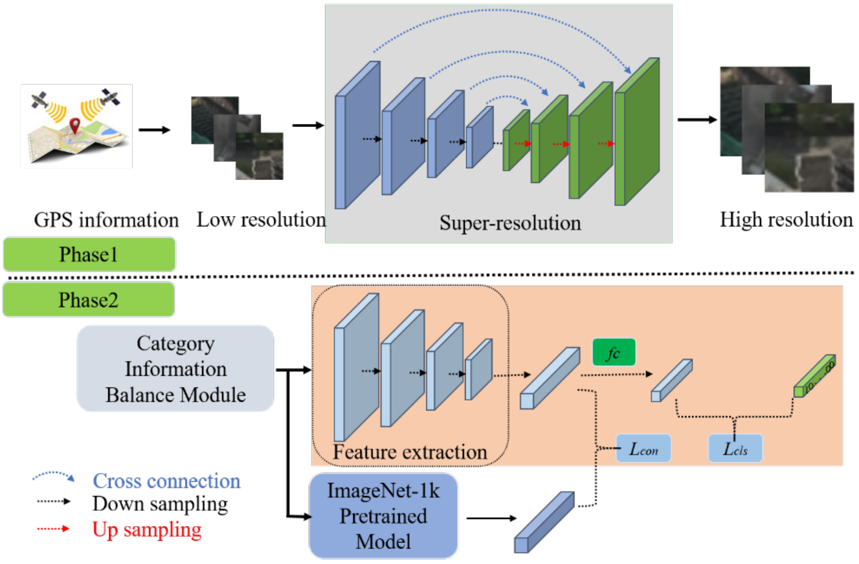

Figure 1. Overview of our proposed building category image classification network based on low resolution satellite.

Another phase is a novel network for urban building category fine-grained classification, considering efficiency and light weight, we adapt the ShufffleNetV2 backbone and proposed category information balanced module to alleviate category information imbalances and improve model robustness, and a distillation learning strategy is proposed to learn the features learned by the original ShufffleNetV2 model on the ImageNet-1k dataset, which improved the performance and convergence speed of the model. More details willed be discussed later in this paper.

3.2 Denoising Diffusion Probabilistic Model

We are given a dataset of input-output image pairs, denoted $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$, which represent samples drawn from an unknown conditional distribution $p(y|x)$. This is a one-to-many mapping in which many target images may be consistent with a single source image. We are interested in learning a parametric approximation to $p(y|x)$ through a stochastic iterative refinement process that maps a source image ***x*** to a target image $y \in R^d$.

The conditional DDPM model generates a target image y_0 in *T* refinement steps. Starting with a pure noise image $y_T \sim \mathcal{N}(0, I)$, the model iteratively refines the image through successive iterations $(y_{T-1}, y_{T-2}, \ldots, y_0)$ according to learned conditional transition distributions $p_\theta(y_{t-1}|y_t, x)$ such that $y_0 \sim p(y|x)$ [10-11].

The distributions of intermediate images in the inference chain are defined in terms of a forward diffusion process that gradually adds Gaussian noise to the signal via a fixed Markov chain, denoted $q(y_t|y_{t-1})$. While in principle this diffusion process can also condition on the source image ***x***, $q(y_t|y_{t-1}, x)$, in super-resolution all of the information of ***x*** is already contained in y_0, making the diffusion process unconditional on is a reasonable choice.

The goal of our model is to reverse the Gaussian diffusion process by iteratively recovering signal from noise through a reverse Markov chain conditioned on ***x***. In principle, each forward process step can be conditioned on ***x*** too. We learn the reverse chain using a neural denoising model f_θ that takes as input a source image and a noisy target image and estimates the noise.

3.3 Category Information Balanced Module

A common way to consider the unbalance problem is sampling, our proposed CIBM takes into account the cosine similarity information between samples of each category by adding features extraction and similarity calculation.

We assume that the unbalanced dataset has a total of *n* individual categories, each with the number of samples x_1,x_2,... ,x_n, then in the training sampling phase each category is assigned weights W_1, W_2,... ,W_n, so as to ensure that the number of samples sampled from different categories is balanced throughout the process, and the weights are calculated as follows:

$$p_i = \frac{x_i}{\sum_{i=1}^{n} x_n}, w_i = \frac{p_i^{-1}}{\sum_{i=1}^{n} p_i^{-1}}. \tag{1}$$

where p_i = the number of samples in category *i* as a proportion of the overall number of samples
w_i = the normalized weight

Our design of CIBM, taking into account information from different samples within each category as well. We will feed samples from each category into the Decode of the existing pre-trained model for category information extraction, which can be described by the following expression:

$$f(i,j) = D(I(i,j)). \tag{2}$$

where *D*(.) = the decode category feature vector extraction operation
I(*i*, *j*) = the *j*th sample in the *i*th category
f(*i*, *j*) = the features extracted from the corresponding image

We calculate the Euclidean distance between any two sample category features within each category, and finally each category generates a Euclidean distance matrix, which is calculated as shown below:

$$\text{dis}(i,j,k) = \sqrt{\sum_{x=1}^{d} \left(f(i,j)_x - f(i,k)_x \right)^2} \tag{3}$$

where *dis*(*i*, *j*, *k*) = the value of the *j*th row *k* columns in the distance matrix of the *i*th category
d = the length of the category feature vector

So the category distance weights S_i and the final sampling weights can be obtained as follows:

$$S_i = \sum_{j=1}^{x_i} \sum_{k=1}^{x_i} \mathrm{dis}(i, j, k), (\text{when } j \neq k),$$
$$W_i = \frac{S_i \cdot p_i^{-1}}{\sum_{i=1}^{n} \left(S_i \cdot p_i^{-1} \right)}. \tag{4}$$

4. EXPERIMENT

In this section, we introduce the methods for acquiring and allocating data. Next, we present the classification results of the building and compare them with existing methods. Finally, we demonstrate the plug-and-play nature and effectiveness of our approach through detailed ablation experiments.

4.1 Dataset

Our low-resolution satellite imagery was obtained from Google Map, with a spatial resolution of 4 metres and an image size of 32*32 for the top-view image of the building. This is very widely available and free. The boundary information of our building images was obtained from the Hong Kong Government's public GeoData Store. It covers the area shown on the left in Figure 2. We classify the buildings into 11 categories according to urban building functions, using the abbreviation 'CO' for Commercial & Office, 'Edu' for Education, 'HP' for High Private, 'Indus' for Industrial, 'LP' for Low Private, 'Medic' for Medical, 'Mix' for Mixed-Use Building, 'PH' for Public House, 'PS' for Public Service, 'Recre' for Recreation and 'Reli' for Religion. There are a total of 1000 samples for each category, including 800 training samples and 200 test samples.

4.2 Dataset Results and Discussion

In order to verify the validity of our method, we have conducted experiments on our data using MVit, EfficientFormer, EfficientNet and ShuffleNetV2, these are current state-of-art methods in image classification and compared them in detail with our proposed method. The classification results Figure 3, for the sake of fairness, the results are generated on the super-resolved images of the DDPM method.

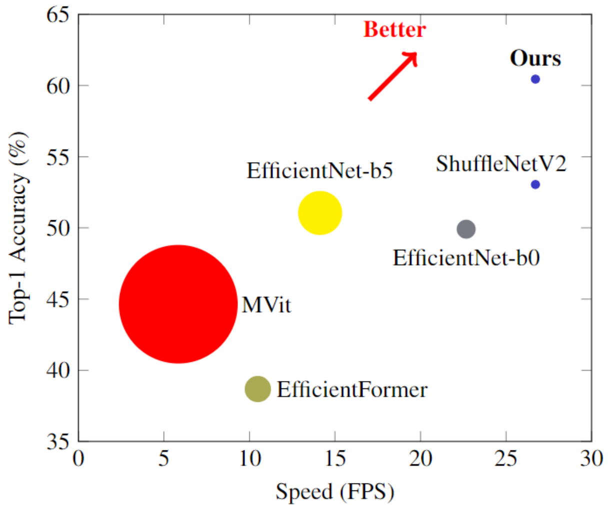

Figure 2. Parameters and performance comparison between our model and existing models.

In this draft, The size of the circle represents the size of the model, the smaller it is the fewer the model parameters. The closer the model is to the top right corner represents higher accuracy and faster inference. As the red arrow shows, it can be perceived as better performance. MVit and EfficientFormer, based on the transformer approach, do not perform as well as EfficientNet and ShuffleNetV2, based on the CNN approach, in terms of Top-1 Acc results on this dataset, which is consistent with the characteristics of these two network families of the CNN family. Transformer-structured networks tend to perform well in scenarios with high data quality and large data volumes, such as ImageNet and COCO, and their effectiveness decreases when the data volume is small, as is the case for our dataset. It should also be noted that CNN-based networks have a smaller Model Size, which means that they have fewer parameters to learn, and can be trained to produce good parametric models with less data, while also being lightweight. The Model Size of our model is only 11.1Mb which is one 54th of MVit, one 12th of EfficientFormer, and one third of EfficientNet's lightest b0. Our model size is the same as the baseline (ShuffleNetV2), but with a large improvement in performance, with Top-1 Acc improving by about 14.8%.

We conducted tests on the application in the three open test areas as demonstrated in dataset, these areas presented high building density, various building types, and random distribution, thus posed as challenging. Our method's outcomes are illustrated in Figure 3, the left-hand side (a) depicts a satellite image, while the right-hand side (b) displays a mask of the results of the map's individual building. As shown in the sub-figure (b), the classification of each house category utilizing only low-resolution satellite images, without any omissions. In comparison, the alternative method necessitates high-resolution street view images and still experiences omissions[12]. Our method is considerably more efficient and cost-effective compared to the methodology that necessitates sub-meter satellite imagery[6-7].

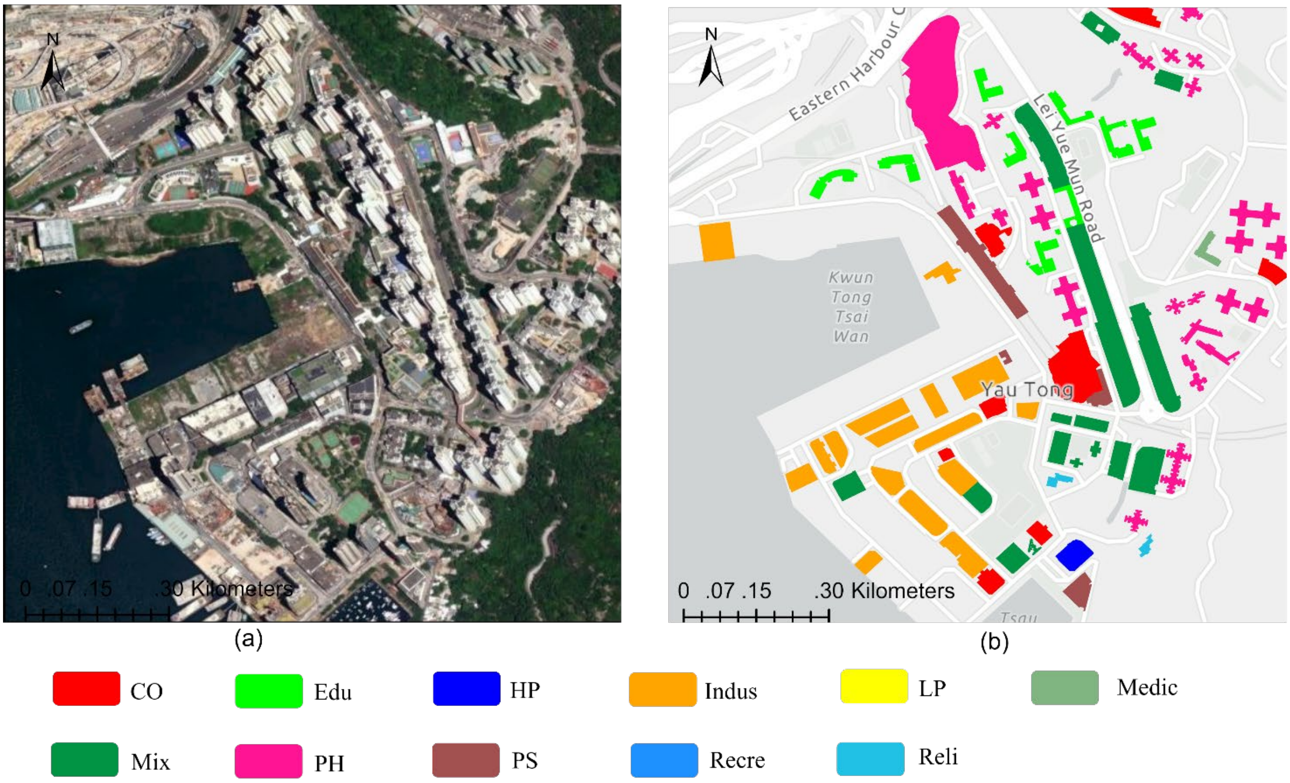

Figure 3. Classification results for high building-density areas.

4.3 Ablation Study

To illustrate the effectiveness of our proposed contrast supervised net and CIBM for building classification, we compared the results after adding these methods. As shown in Figure 4 and Table 1.

In order to verify the effectiveness and plug-and-play ubiquity of our proposed CIBM and CS, we conducted ablation experiments on different SOTA base networks, noting that all experiments in the table were conducted on the results of Phase1 processing to ensure data consistency. In the experiments, we add CIBM and CS to the base network one by one and test their performance. Specifically, the data in the table can be roughly summarized to show that the effect of CIBM and CS on the performance of different networks is not identical, and that there is a preference for different metrics, but also some common features, for example, the addition of a CS has a better effect on the network than the addition of a single CIBM. This is probably due to the fact that the sampling of samples by the CIBM during the training phase changes the relationship between the distance of each category of features and the distance of the categories in ImageNet-1K, making it possible for the CIBM to be more effective than the CS. This is beneficial information for comparison supervision, which in turn back-propagates to adjust the model to better fit the current training dataset. This is a very interesting phenomenon, as we did not expect them to be coupled with a 1+1>2 result when we designed the network.

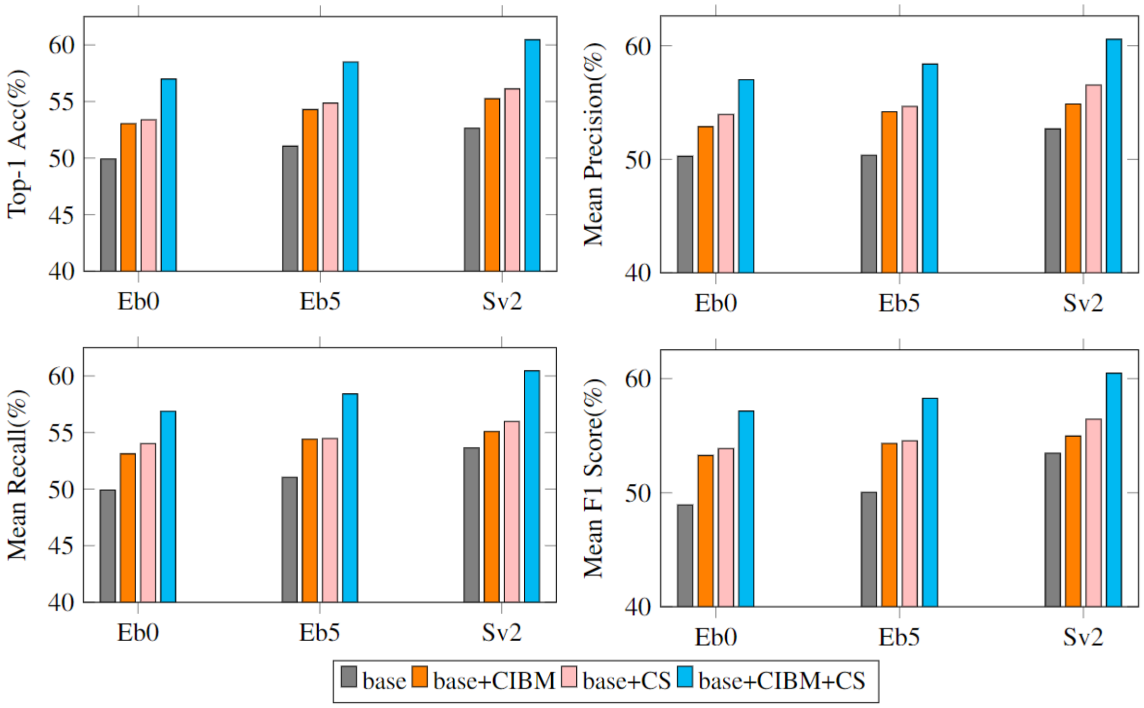

Figure 4. Comparison of results from different methods.

Table 1. Comparison of results from different methods, ↑ means higher is better.

Metric	Top-1 Acc ↑	Δ	Top-5 Acc ↑	Δ	Mean Precision ↑	Δ	Mean Recall ↑	Δ	Mean F1 Score ↑	Δ
Eb0	49.91	0	90.68	0	50.26	0	49.91	0	48.92	0
Eb0+CIBM	53.04	+6.3%	91.09	+0.4%	52.87	+5.2%	53.12	+6.4%	53.26	+8.8%
Eb0+CS	53.38	+6.9%	90.87	+0.2%	53.95	+7.3%	54.02	+8.2%	53.86	+10.1%
Eb0+CIBM+CS	**56.98**	+14.1%	**91.45**	+0.8%	**57.01**	+13.4%	**56.87**	+13.9%	**57.15**	+16.8%
Eb5	51.05	0	90.91	0	50.36	0	51.05	0	50.03	0
Eb5+CIBM	54.28	+6.2%	91.47	+0.6%	54.19	+7.6%	54.40	+6.5%	54.31	+8.5%
Eb5+CS	54.86	+7.4%	91.00	+0.01%	54.65	+8.5%	54.46	+6.6%	54.55	+9.0%
Eb5+CIBM+CS	**58.48**	+14.5%	**92.75**	+2.0%	**58.39**	+15.9%	**58.41**	+14.4%	**58.27**	+16.4%
Sv2	52.64	0	90.95	0	52.68	0	53.64	0	53.46	0
Sv2+CIBM	55.24	+4.9%	91.12	+0.2%	54.87	+4.2%	55.08	+2.7%	54.96	+2.8%
Sv2+CS	56.11	+6.6%	92.08	+1.2%	56.54	+7.3%	55.98	+4.3%	56.44	+5.6%
Sv2+CIBM+CS	**60.45**	+14.8%	**93.50**	+2.8%	**60.57**	+14.9%	**60.45**	+12.6%	**60.47**	+13.1%

5. CONCLUSIONS

The fine classification of buildings based on remote sensing images is a popular research topic, as the results are useful in giving a good idea of the economic, industrial and even population distribution within a city or region. This is essential for urban planning, road construction, etc. However, there are two main challenges: (1) the low resolution of overhead views from high altitude remote sensing, and (2) the wide variation in the number of different types of buildings, making it difficult to avoid class imbalance in the acquired training data. To address these two problems, we design a method for fine-grained classification of two phases of low-resolution building overhead views. In the first phase, we design a model migration-based DDPM method to enhance the low-resolution images, and in the second phase, we design a category information balanced module (CIBM) and contrast supervision (CS) to improve the performance of the fine-grained building classification network. We achieved excellent results on a Google Map-based intercepted dataset, while we conducted full ablation experiments to verify the effectiveness of our improvements.

Our research contributes to the field of urban analysis by providing a practical solution for fine classification of buildings in challenging scenarios. The proposed method can serve as a valuable tool for urban planners, aiding in the understanding of economic, industrial, and population distribution within cities and regions, ultimately facilitating informed decision-making processes in urban development and infrastructure planning.

6. ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foun-dation of China (Grant No.42171361) and by a grant from the Research Grants Council of the Hong Kong Special Admin-istrative Region, China (Project No. PolyU 15215023). This work was also funded by the research project (Project Num-ber: 2021.A6.184.21D) of the Public Policy Research Fund-ing Scheme of The Government of the Hong Kong Special Administrative Region. Thanks to KaYee Ng from The Hong Kong Polytechnic University for her assistance and advice during the acquisition of the dataset and the illustration process.

7. REFERENCE

[1] D. Frantz et al., 2021. National-scale mapping of building height using Sentinel-1 and Sentinel-2 time series. *Remote Sensing of Environment*. 252, pp. 112-128.

[2] Laupheimer. Neural Networks for the Classification of Building Use from Street-View Imagery. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 177-184.

[3] Kang, J., et al, 2018. Building instance classification using street view images. *ISPRS journal of photogrammetry and remote sensing*, 145, pp. 44-59.

[4] Zhao, K., et al, 2021. Bounding boxes are all we need: street view image classification via context encoding of detected buildings. *IEEE Transactions on Geoscience and Remote Sensing*, 60, pp. 1-17.

[5] Xiao, C., et al. Efficient building category classification using façade information from oblique aerial images. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 1309–1313.

[6] Huang X. et al., Urban Building Classification (UBC)-A Dataset for Individual Building Detection and Classification from Satellite Imagery. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1412-1420.

[7] Huang X. et al., 2023. Urban Building Classification (UBC) V2 - A Benchmark for Global Building Detection and Fine-grained Classification from Satellite Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, doi: 10.1109/TGRS.2023.3311093.

[8] Vannucci et al., 2016. Smart Under-Sampling for the Detection of Rare Patterns in Unbalanced Datasets. In: Intelligent Decision Technologies 2016. IDT 2016. *Smart Innovation, Systems and Technologies*. 56, pp. 395-404.

[9] Hasib, K. M. et al, 2021. HSDLM: A Hybrid Sampling With Deep Learning Method for Imbalanced Data Classification. *International Journal of Cloud Applications and Computing (IJCAC)*, 11(4), pp. 1-13.

[10] Jonathan, H. et al, 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, pp. 6840-6851.

[11] Saharia, C. et al, 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, pp. 4713-4726.

[12] Zhang, Y., et al, 2023. Knowledge and topology: A two layer spatially dependent graph neural networks to identify urban functions with time-series street view image. *ISPRS Journal of Photogrammetry and Remote Sensing* ,198, pp. 153-168.