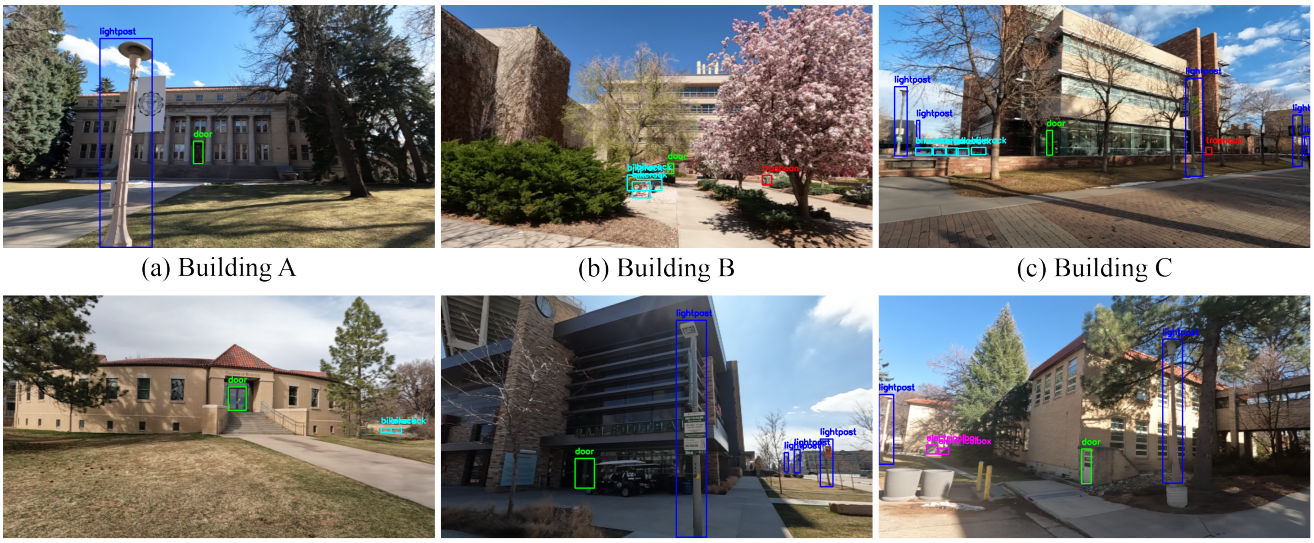
CVPR #11

[UNI]101: An Educational Dataset for Introductory Computer Vision

Anonymous CVPR submission

Paper ID 11



(a) Building A (b) Building B (c) Building C

(d) Building D (e) Building E (f) Building F

Figure 1. An introduction to *[UNI]101*. Six images of buildings containing labels for building classification, as well as labels for objects such as bike racks, light posts, trash cans, electrical boxes, and doors.

Abstract

Education in Machine Learning, particularly in Computer Vision, currently lacks modern, relatable datasets specifically designed to support a structured learning progression. Existing datasets such as MNIST or ImageNet are cornerstones in teaching Computer Vision, but they address isolated tasks and may not resonate with students' daily experiences. Addressing this gap, we introduce [UNI]101, a comprehensive video dataset consisting of 277,056 frames of university buildings, captured at 24 FPS in 1080p resolution. [UNI]101 is designed to scaffold an entire semester of introductory Computer Vision: students begin with image classification across 48 campus buildings, advance to object detection over five commonly seen campus object classes (bike racks, trash cans, doors, electrical boxes, and light posts), and progress to 3D scene reconstruction via Gaussian Splatting, all using a single, familiar dataset. The dataset includes 750 unique labeled objects, benchmark models for each task (six for classification, five for detection), and educational Jupyter Notebooks structured as lesson plans. By grounding an entire curriculum in scenes students encounter daily, [UNI]101 bridges the gap between foundational Computer Vision education and modern research topics, providing a progressive, hands-on learning experience within a unified framework.

1. Introduction

Machine Learning (ML) continues to evolve at an increased pace, attracting an unprecedented number of learners eager to explore its complexities [1]. Among its diverse subfields, Natural Language Processing (NLP) and Computer Vision (CV) stand out; however, the educational resources available, particularly datasets, often lag behind in terms of relevance and accessibility. Traditional datasets, while foundational, can often fail to resonate with students and may present steep learning curves [8, 9, 22]. The need for datasets that not only illustrate but also demystify the underlying principles of machine learning while being relatable to

Table 1. Comparison of [UNI]101 with other Building Datasets. This table highlights differences in the number of labeled frames, unique buildings, and objects, and indicates the presence of bounding box labels for object detection. Notably, [UNI]101 ranks second highest in the number of labeled frames and leads in the number of unique labeled objects.

Dataset	Video Data	Frames	Buildings	Objects	Bounding Boxes	3D Ready
Obeso et al. [27]	✗	284	–	–	✗	✗
Shalunts et al. [35]	✗	400	–	–	✗	✗
Shao et al. [36]	✗	1,005	201	–	✗	✗
Taoufiq et al. [40]	✗	1,033	–	–	✗	✗
Xu et al. [44]	✗	$\sim$ 5,000	–	–	✗	✗
Taoufiq et al. [40]	✗	6,297	–	–	✗	✗
Cordts et al. [6]	✓	25,000	–	–	✗	✗
Zheng et al. [46]	✗	$\sim$ 21.4 million	5,312	–	✗	✗
Philben et al. [28]	✗	5,062	11	11	✓	✗
Philben et al. [29]	✗	6,300	11	11	✓	✗
Barz and Denzler [5]	✗	9,485	9,346	566	✓	✗
[UNI]101(Ours)	✓	**277,056**	**48**	**750**	✓	✓

students is more pressing than ever [34]. Enhancements in dataset design could include more intuitive data representations, increased contextual relevance, and a stronger alignment with the current trends and applications in the industry. Recognizing these needs can lead to the development of resources that are not only more instructive but also more inspiring to a new generation of Computer Vision enthusiasts.

Improving the quality and applicability of educational datasets is not merely an academic exercise; it has implications for the entire field of Machine Learning. By equipping learners with better tools and more relatable experiences, we can accelerate the development of skilled practitioners who are prepared to tackle modern challenges. Moreover, datasets that better reflect the complexities of real-world data can enhance the robustness and applicability of Machine Learning models. This transition towards more effective educational resources is crucial for fostering innovation and ensuring that the next wave of Machine Learning advancements is grounded in a deep and practical understanding of the technology.

[UNI]101: In this work, we introduce [UNI]101, the first publicly available dataset tailored specifically for educational use in Computer Vision, designed to resonate with university students globally. Unlike existing educational resources that address isolated tasks, [UNI]101 is structured to support a full-semester curriculum spanning three core Computer Vision tasks of increasing complexity. Students begin with image classification of 48 campus buildings, advance to object detection of five common campus object classes, and progress to 3D scene reconstruction using Gaussian Splatting all within the same familiar dataset. This graduated structure enables learners to build progressively on prior knowledge while working with scenes they encounter daily. The dataset comprises video footage of university buildings captured at 24 FPS and 1080p resolution, annotated with both classification and object detection labels. We additionally provide starter Jupyter Notebooks for each task, structured as lesson plans, to facilitate initial learning and experimentation, which has been shown to be beneficial in a classroom environment [16].

Benchmarking & Maintaining: To demonstrate the feasibility and effectiveness of [UNI]101, we provide benchmarks for all three tasks within the dataset: building classification, object detection, and 3D Gaussian Splatting reconstruction. These benchmarks serve as baselines for educators and researchers to assess and build upon. Additionally, we propose expanding the dataset's scope through collaboration with other universities, enriching it with diverse architectural styles and increasing its variability and robustness.

Contributions: In summary, we make five main contributions: First, we construct, to the best of our knowledge, the first Computer Vision dataset specifically tailored to support a semester-long educational progression for university students. Second, we provide comprehensive labels for both image classification and object detection, allowing learners to utilize a single dataset for mastering multiple core tasks. Third, we include pre-processed images from video footage, enabling beginners to focus initially on simpler image-based tasks before advancing to video data processing. Fourth, we demonstrate that the same dataset naturally extends to advanced topics such as 3D Gaussian Splatting, providing a bridge from introductory to modern research-level techniques. Fifth, we supplement our dataset with documented Jupyter Notebooks that serve as structured lesson plans, facilitating immediate, practical engagement.

2. Related Works

For this work, we focus on two separate fields of related work: building datasets and educational datasets.

2.1. Building Datasets

Building recognition research has leveraged datasets ranging from photo-sharing platforms [46] to historic landmarks [2–4] and architectural style classification [5, 27, 35, 40, 44]. Table 1 compares [UNI]101 with these works. Most prior building datasets are image-only collections sourced from the web [5, 13, 27–29, 35, 36, 40, 44–46]; [UNI]101 is instead manually captured video, combining classification, object detection, and 3D reconstruction in a single educational resource. Notably, Barz et al. [5] also provide class and bounding box annotations, but their boxes capture architectural elements linked to the class label, whereas ours define a separate object detection task. Popular datasets such as Oxford 5k [28] and Paris [29] target content-based image retrieval rather than building recognition, while Taoufiq et al. [40] and others [13, 27, 35, 44] classify building *types* rather than identifying specific buildings.

Unlike prior works that prioritize state-of-the-art models, [UNI]101 is tailored for entry-level learners, akin to MNIST [21] and Caltech101 [12], which remain widely used despite high benchmark accuracies. While Cityscapes [6], Open Images V4 [20], and ImageNet [7] offer object detection or image classification, [UNI]101 uniquely combines both tasks with fine-grained building recognition and extends to 3D reconstruction.

2.2. Educational Datasets

While a variety of datasets exist for the purposes of Machine Learning education, many of these were not originally created with education as a primary objective. For example, the popular MNIST dataset [21] for handwritten digit recognition and the CIFAR datasets [19] for object recognition are frequently used in educational settings, but their initial focus was on advancing Machine Learning research. Similarly, the COCO dataset [22] for object detection and segmentation, while immensely valuable for learning, was primarily designed for large-scale challenges and competitions.

In contrast, there have been some deliberate efforts to curate datasets specifically for educational purposes. The UCI Machine Learning Repository [11] hosts a variety of datasets suitable for teaching and learning, covering a wide range of Machine Learning tasks. However, most of these datasets focus on tabular data and do not adequately address the unique challenges and opportunities presented by Computer Vision tasks.

To the best of the authors' knowledge, this work is unique in its combination of building recognition and distinct object detection into an introductory educational dataset.

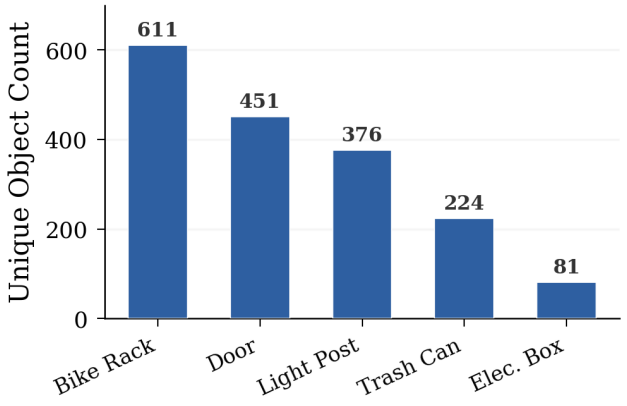


Figure 2. Unique objects across the [UNI]101 dataset. The distribution highlights a common campus trend: a high frequency of bike racks and a relatively low occurrence of electrical boxes.

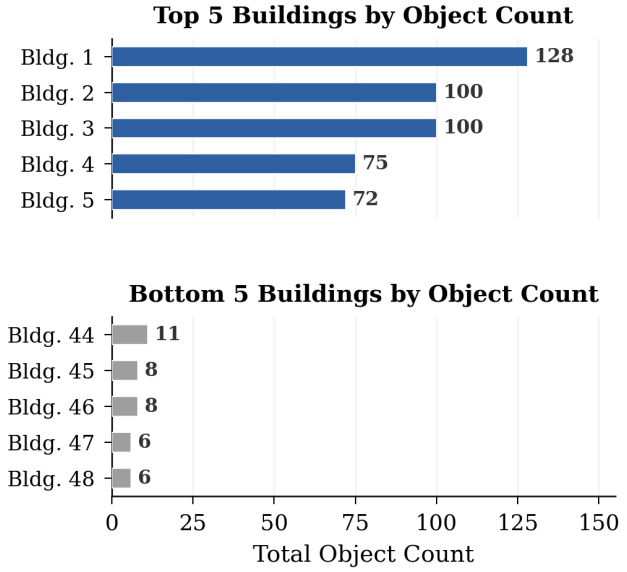


Figure 3. Top and bottom five building object counts. Larger campus areas such as stadiums contain up to 128 unique objects, while smaller buildings include as few as 6.

3. [UNI]101 Construction

Dataset Collection. Data was collected using a GoPro Hero 10 camera, set to record at 24 FPS with a resolution of 1080p. The data collection procedure involved systematically walking around each building's perimeter and capturing all of its faces using the GoPro cameras. In total, 48 university buildings were recorded, with researchers fully encircling each building when possible, resulting in a total of 252,837 labeled frames. During these recordings, care was taken to ensure that the entire building was within the frame, along with surrounding elements such as the ground, to facilitate the labeling of objects like bike racks and trash

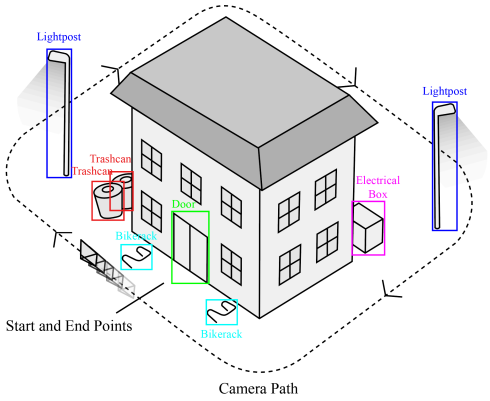


Figure 4. This figure illustrates the initial setup and approach used by the researcher during the data collection process for the [UNI]101 dataset. Starting at the main entrance of the building and positioned as far back as possible, the researcher began recording to capture both the building and surrounding objects in the frame. The recording continued as the researcher walked around the building, ensuring comprehensive coverage of the architecture and any visible objects such as bike racks, benches, and trash cans. The circuit was completed just short of the starting point to maintain continuity and minimize overlap in the visual data captured.

cans in addition to the buildings themselves. This comprehensive approach to filming was designed to capture a wide array of angles and perspectives, enhancing the dataset's utility for both object detection and building classification tasks. Figure 4 visualizes the data collection process.

Labels. For building classification, we provide JSON files containing the labels along with corresponding data splits. To ensure robust model training and evaluation, each building video was labeled at every 20th frame, with some final frames intentionally excluded to avoid overlap between training and testing datasets, thereby preventing wraparound issues. For object detection, labels are formatted according to the YOLO [30] standard. This involves labeling every frame of each video in a separate text file, where each line represents an object with five attributes: the class identifier, the center x and center y coordinates of the bounding box, and the dimensions (height and width) of the bounding box. This detailed labeling facilitates precise object localization and is crucial for training effective detection models.

Interrater Reliability. To validate annotation accuracy, we conducted an interrater reliability assessment using the Kappa coefficient [14, 26]. Five annotators independently labeled the same video (Forestry) without specific prior instructions, achieving a Kappa score of 0.80, demonstrating strong consistency across annotators.

Object Classes. One common issue with many educational datasets is the overwhelming number of object classes, which can complicate the learning process. To address this, we carefully selected only five classes that are often present on campus: bike racks, light posts, doors, trash cans, and electrical boxes. These classes were chosen for their familiarity to university students, ensuring that the objects are easily identifiable and relatable. This focused approach not only simplifies the learning experience but also makes it easier for students to understand how models function and to interpret their outputs. By reducing complexity, we aim to simplify the principles of object detection and enhance the practicality of experiments conducted with this dataset. The overall distribution of unique objects (counting only each occurrence of an object once per video) can be seen as right-tailed in Figure 2. As expected on a college campus, there are a large number of bike racks present (509 in total), while electrical boxes were the least present (68 in total). Additionally, per building, we saw averages of: bike racks (9.98), doors (7.71), light posts (6.27), trash cans (5.30), and electrical boxes (1.42), depicted in Figure 3.

Data Annotation. The annotation process was conducted using the Computer Vision Annotation Tool (CVAT). Five object categories relevant to campus infrastructure were annotated: light posts, bike racks, trash cans, electrical boxes, and doors (specifically of buildings). Annotations were performed by annotators who manually drew bounding boxes around each instance of the target objects. To ensure accuracy and consistency, the bounding boxes were adjusted every 20 frames. Each bounding box is represented using two points that define a square enclosing the object of interest.

Dataset Split (Building Classification). As each video contains a building that has been fully encircled, we needed to be careful about how we chose to split the data. To minimize scene overlap at 24 FPS, we selected every 80th frame from each video, resulting in a total of 7,440 frames. After pruning the frames, the average number of frames per video is 155, with a minimum of 31 frames and a maximum of 395 frames per video. From the selected frames, we distributed 60%, 20%, and 20% of the shuffled frames per video to the training, validation, and testing sets, respectively.

Dataset Split (Object Detection). Unlike image classification, not all buildings needed to be present in each split for object detection. As we had 56 videos (48 buildings; some were unable to be completed in a single video due to fencing, overlapping the same scenes, etc.), we decided to withhold 5 videos for validation and 5 videos for testing, while the remaining 38 videos went into the training set. To ensure consistency and quality, we decided to choose 10 videos for validation and testing that were a sizeable amount of frames ($\sim$4500) and contained a good distribution of objects. For example, for the Forestry building in the testing set, it contained 4728 frames, 10 light posts, 4 doors, 2 trash cans, 4 bike racks, and 4 electrical boxes. Similar to Section 3, we selected every 20th frame from each video and

(a) Lightpost (b) Doors (c) Trashcans

(d) Bikeracks (e) Electrical Box (f) Multiclass Objects

Figure 5. This figure presents an up-close view of each object type included in the dataset, along with an image featuring multiple objects. A common theme observed throughout the dataset is the grouping of bike racks, as exemplified in image (f).

(a) View 1 (b) View 2 (c) View 3

Figure 6. Three separate views illustrate the diversity of objects found in a single video, in this case, the Computer Science Building. View 1 displays four distinct objects, View 2 features two, and View 3 highlights a grouping of multiple objects, specifically bike racks. This demonstrates the wide variety of objects that a single building can present within the dataset.

then assigned the pruned frames to corresponding sets by the building name. The training, validation, and testing sets contain 7,334, 808, and 643 frames respectively.

[UNI]101 as an Educational Resource. [UNI]101 is designed to be an accessible and valuable resource for introductory Machine Learning courses. The dataset's manageable size allows for efficient training and experimentation on standard hardware, making it ideal for educational settings. The inclusion of diverse scenes and lighting conditions introduces students to the challenges of real-world object detection, preparing them for more complex tasks.

To further enhance its educational value, we provide a suite of Jupyter notebooks tailored for beginners. These notebooks offer step-by-step guidance, from dataset loading and visualization to model training and evaluation. Extensive comments and explanations within the notebooks aim to aid students in understanding the underlying concepts and techniques. Additionally, pre-trained models will be made available, enabling students to explore object detection without requiring extensive computational resources.

Ethical Considerations. The collection of a university building dataset raises concerns regarding the potential recording of students and their residential quarters. Mindful of privacy issues, we consciously avoided capturing footage that included dormitory buildings, thereby respecting the privacy of students residing in these areas. Additionally, while we made efforts to minimize recording students without their consent, it was sometimes unavoidable due to the public nature of the environments. To account for this, the faces of all individuals shown in the dataset have been blurred to ensure anonymity and to stop their likeness from being used for training purposes with the dataset.

Table 2. Baseline results for Building Classification and Object Detection tasks on the [UNI]101 dataset. Building classification performance is evaluated using accuracy and F1 score metrics, while object detection performance is assessed using $AP_{.50}$, $AP_{.75}$, and mean Average Precision (mAP).

Building Classification Baselines			**Object Detection Baselines**			
Model	Accuracy (%)	F1 Score	Model	$AP_{.50}$	$AP_{.75}$	mAP
VGG19	89.84	0.77	Faster R-CNN	0.62	0.26	0.44
EfficientNet_V2_M	93.96	0.86	FCOS	0.42	0.14	0.28
ResNet50	96.70	0.92	RetinaNet	**0.69**	**0.28**	**0.49**
ResNext50_32x4d	**98.63**	**0.95**	SSD	0.53	0.18	0.35
Vit_L_16	98.35	0.94	SSDlite	0.38	0.05	0.22
Swin_V2_S	96.43	0.90	–	–	–	–

Table 3. Per-scene 3D Gaussian Splatting results using Splatfacto on six [UNI]101 buildings. Metrics: PSNR (dB), SSIM, LPIPS, and rendering FPS. All models were trained on posed frames extracted via COLMAP from the original walkthrough videos.

Scene	PSNR↑	SSIM↑	LPIPS↓	FPS↑
Building 1	26.20	0.832	0.141	87.2
Building 2	29.64	0.908	0.123	94.6
Building 3	20.61	0.741	0.256	1.0
Building 4	27.41	0.846	0.167	94.5
Building 5	27.25	0.834	0.218	102.4
Building 6	28.57	0.898	0.116	96.9
Average	**26.61**	**0.843**	**0.170**	**79.4**

4. Baseline Experiments

We conduct simple baseline experiments for both image (building) classification and object detection, to demonstrate the feasibility of the dataset.

4.1. Building Classification

For our building classification baseline models, we chose to use six separate easily accessible models that are freely downloadable utilizing PyTorch [15]. We selected the following models: ResNet50 [18], Inception_V3 [38], ResNext50 [42], VGG-19 [37], ViT [10], SwinS [25], and EfficientNet [39]. During the fine-tuning process, we used a batch size of 8, a learning rate of 0.0001, and an input size of 224 x 224. Each image was normalized using the mean and standard deviation of ImageNet [7]. The fine-tuning proceeded for a maximum of 100 epochs, with early stopping applied after 10 epochs without improvement. All baselines were fine-tuned by cross-entropy loss. To evaluate the classification baselines, we proposed to use accuracy and F1 score (macro). The left table on Table 2 shows that ResNext50 outperformed other baselines on both Accuracy and F1 score.

4.2. Object Detection

The PyTorch documentation offers five object detection models, we decided to use all of them. These models included Faster R-CNN [31], FCOS [41], RetinaNet [23], SSD [24], and SSDlite [32]. We used the pre-trained weights for `fasterrcnn_resnet50_fpn`, `fcos_resnet50_fpn`, `retinanet_resnet50_fpn`, `ssd300_vgg16`, and `mobilenet_v3_large`, respectively. For fine-tuning the object detection baselines, we applied a batch size of 8, a learning rate of 0.0001, and an input size of 256 x 256. Both SSD and SSDlite used 512 x 512 input data. As in the classification task 4.1, each image was normalized by mean and standard deviation values of ImageNet [8] and fine-tuned with early stopping and 100 maximum epochs. To evaluate the baselines, we employed metrics; $AP_{.50}$, $AP_{.75}$, and mAP. The right half of Table 2 shows that RetinaNet outperformed other baselines on all metrics. More detailed results for each object are placed in Section A

4.3. In-the-Wild 3D Reconstruction

To demonstrate that [UNI]101 supports a full curriculum progression, we apply Splatfacto [43], a Gaussian Splatting [17] implementation, directly to the walkthrough videos. Posed frames are extracted via COLMAP [33]; no additional data collection is required beyond what the dataset already provides.

Table 3 reports per-scene results across six buildings. Five of six scenes achieve PSNR above 26 dB and SSIM above 0.83, with real-time rendering speeds exceeding 87 FPS. Building 3 is a notable outlier (20.61 dB, 1.0 FPS), which we attribute to heavy foliage occlusion and limited viewpoint coverage itself a useful teaching example of how capture conditions affect reconstruction quality. Figure 7 shows qualitative results: ground truth frames, novel views via camera path interpolation, and rendered depth maps that reveal meaningful geometric structure.

Educational value. This extension enables a natural pro-

Figure 7. Qualitative results of 3D Gaussian Splatting applied to three campus buildings from the [UNI]101 dataset. Each row shows a different building, with columns displaying (left to right) a ground truth input frame, a novel view synthesized via camera path interpolation, and the corresponding rendered depth map. These results demonstrate that the same video data used for classification and detection naturally extends to 3D reconstruction tasks, supporting a progressive educational curriculum.

gression from 2D classification to object detection to 3D reconstruction, mirroring the arc of a semester-long course. Because students already know the buildings from earlier tasks, the reconstructions offer a tangible connection to the underlying geometry, they can compare rendered views against scenes they walk through daily. The dense 24 FPS video provides ideal input for structure-from-motion, meaning instructors can introduce multi-view geometry and neural rendering without curating a separate dataset. Variance across scenes (e.g., Building 3 vs. Building 2) also provides natural discussion points around data quality, occlusion handling, and failure-mode analysis.

5. Discussion & Limitations

Semester-Long Curriculum A distinguishing feature of [UNI]101 is its capacity to serve as a unifying thread across an entire introductory Computer Vision course. In a typical semester, students can follow a structured progression: (1) weeks 14 focus on image classification, where students learn data loading, augmentation, transfer learning, and evaluation using the 48-building classification task; (2) weeks 59 introduce object detection, building on the same imagery but now requiring students to understand bounding boxes, anchor-based and anchor-free architectures, and metrics such as mAP; and (3) weeks 1014 advance to 3D reconstruction, where students leverage the raw video to explore structure-from-motion, neural rendering, and Gaussian Splatting. This progression is pedagogically motivated: each stage reuses the same familiar scenes while introducing fundamentally new concepts, reducing the cognitive overhead of switching datasets and allowing students to focus on the techniques themselves. The familiarity of campus buildings structures students walk past daily further lowers the barrier to engagement, as students can immediately connect model outputs to their lived experience.

Additional Uses. While [UNI]101 is specifically designed for academic settings, its utility extends far beyond the classroom. With object labels for items commonly found around university campuses and urban environments, this dataset can be instrumental in training models for industry applications or urban research. Potential areas of impact include smart city initiatives, where models trained with [UNI]101 could improve public safety, urban planning, and automated maintenance tracking. Additionally, it could be used in research for enhancing navigational and positioning aids, as well as developing advanced surveillance systems. The dataset's relevance to everyday objects and scenarios

also makes it a valuable tool for startups and technology companies focusing on the development of real-world Artificial Intelligence applications.

Open-sourced. To foster widespread educational use, we will publicly release all data and associated code with this project. This openness not only allows university students to learn using this dataset but also enables anyone around the world to easily download and begin their journey into computer vision.

Limitations. The primary limitation of the [UNI]101 dataset lies in its current scope. The dataset is confined to a specific geographical location, the [University] campus, and includes only a limited set of object categories. All videos were captured in broad daylight, limiting applicability to other lighting conditions. For the Gaussian Splatting task, the walkthrough-style capture provides good frontal coverage but limited overhead or aerial viewpoints, which may affect reconstruction quality for rooftops and occluded areas. Adding diverse weather conditions, seasons, and times of day would make the dataset more robust and provide additional challenges for learners at all levels.

6. Future Work & Maintenance

In future iterations of the [UNI]101, we plan to expand the dataset to extend its diversity and utility. This expansion will involve collecting additional videos across the campus, capturing a wider range of environmental conditions such as varying lighting (e.g., nighttime, overcast), weather (e.g., rain, snow), and seasonal changes (e.g., foliage).

Additional Labels. [UNI]101 serves as a foundational tool for introductory learning in object recognition, focusing on simpler categories such as bike racks. To challenge medium or advanced learners and increase the dataset's educational utility, future versions will expand the object categories to include more complex items like skateboard racks, windows, and various types of signage. This enhancement will introduce greater variability and subtlety into the object recognition tasks, catering to a broader spectrum of skill levels from beginners to intermediate learners. Additionally, we plan to enrich the dataset with image classification labels that indicate the age of buildings. This feature will enable students to develop models capable of predicting a building's age, thereby broadening their learning experience and application skills in practical scenarios.

Collaborations. To further enrich the [UNI]101 dataset, we envision collaborations with other universities, fostering a richer representation of diverse campus environments. By incorporating data from different institutions, we can expand the range of architectural styles, building types, and campus objects, thereby enhancing the dataset's generalizability and applicability to a broader educational context. Additionally this will allow for more type of image classification, such as architecture style or campus name.

Model Deployment & Community Engagement. We plan to develop tutorials for model deployment in real-world scenarios and organize competitions where students develop and deploy custom classification models. The dataset will continue to serve as a resource for the Computer Vision Club at [University], with undergraduate research assistants contributing to its maintenance and expansion.

7. Conclusions

In this paper, we introduced [UNI]101, a novel dataset comprised of video footage of 48 university buildings, broken down into 277,056 labeled images featuring a total of 750 unique objects across 5 distinct classes. Tailored specifically for introductory courses in Computer Vision, [UNI]101 facilitates hands-on learning through fundamental tasks such as image classification and object detection. The selection of objects and structures is particularly designed to resonate with university students, thereby enhancing the educational experience.

To support effective utilization of this dataset, we provide comprehensive tutorials via Jupyter Notebooks, complete with six baseline models for image classification and five for object detection. These resources are designed to aid students in not only understanding the basics of model implementation but also in appreciating the practical challenges of Computer Vision applications. Furthermore, we demonstrated that the same walkthrough videos naturally extend to in-the-wild 3D Gaussian Splatting, enabling a progressive curriculum from classification to detection to 3D reconstruction.

Our goal is for [UNI]101 to serve not only as a foundational tool for learners worldwide but also to inspire innovations in educational methodologies within the field of Computer Vision. By facilitating accessible and engaging learning experiences, we hope to encourage more students to pursue advanced studies and careers in this dynamic field. Looking forward, we plan to expand [UNI]101 by incorporating more diverse environments and enhancing the dataset with additional object classes and annotations to cover a broader range of conditions and scenarios. This expansion will aim to increase the datasets robustness and applicability, preparing learners to tackle real-world challenges more effectively.

References

[1] Karan Aggarwal, Maad M Mijwil, Abdel-Hameed Al-Mistarehi, Safwan Alomari, Murat Gök, Anas M Zein Alaabdin, Safaa H Abdulrhman, et al. Has the future started? the current growth of artificial intelligence, machine learning, and deep learning. *Iraqi Journal for Computer Science and Mathematics*, 3(1):115–123, 2022. 1

[2] Donna Agius. *Automatic recognition of historical buildings in Valletta using smartphone technology*. bache-

lorThesis, University of Malta, 2016. Accepted: 2016-08-31T09:37:51Z. 3

[3] Stefano Alletto, Davide Abati, Giuseppe Serra, and Rita Cucchiara. Wearable vision for retrieving architectural details in augmented tourist experiences. In *2015 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)*, pages 134–139, 2015.

[4] Stefano Alletto, Davide Abati, Giuseppe Serra, and Rita Cucchiara. Exploring Architectural Details Through a Wearable Egocentric Vision Device. *Sensors*, 16(2):237, 2016. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. 3

[5] Björn Barz and Joachim Denzler. Wikichurches: A fine-grained dataset of architectural styles with real-world challenges. *CoRR*, abs/2108.06959, 2021. 2, 3

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 6

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 6

[9] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 6

[11] Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. 3

[12] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 3

[13] Abhinav Goel, Mayank Juneja, and C. V. Jawahar. Are buildings only instances? exploration in architectural style categories. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, New York, NY, USA, 2012. Association for Computing Machinery. 3

[14] Kilem L Gwet. Intrarater reliability. *Wiley encyclopedia of clinical trials*, 4, 2008. 4

[15] Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. Pytorch. *Programming with TensorFlow: Solution for Edge Computing Applications*, pages 87–104, 2021. 6

[16] Jeremiah W Johnson. Benefits and pitfalls of jupyter notebooks in the classroom. In *Proceedings of the 21st annual conference on information technology education*, pages 32–37, 2020. 2

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM Transactions on Graphics (ToG)*, 2023. 6

[18] Brett Koonce and Brett Koonce. Resnet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72, 2021. 6

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 3

[20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 3

[21] Yann LeCun, L eon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *PROC. OF THE IEEE*, page 1, 1998. 3

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 3

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. 6

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. *SSD: Single Shot MultiBox Detector*, page 2137. Springer International Publishing, 2016. 6

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 6

[26] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012. 4

[27] Abraham Montoya Obeso, Jenny Benois-Pineau, Alejandro Álvaro Ramirez Acosta, and Mireya Saraí García Vázquez. Architectural style classification of mexican historical buildings using deep convolutional neural networks and sparse features. *Journal of Electronic Imaging*, 26(1):011016–011016, 2017. 2, 3

[28] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2, 3

[29] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2, 3

[30] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 4

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 6

[32] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019. 6

[33] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 6

[34] Corey Seemiller, Meghan Grace, Paula Dal Bo Campagnolo, Isa Mara Da Rosa Alves, and Gustavo Severo De Borba. What makes learning enjoyable? perspectives of todays college students in the us and brazil. *Journal of Pedagogical Research*, 5(1):1–17, 2020. 2

[35] Gayane Shalunts, Yll Haxhimusa, and Robert Sablatnig. Architectural style classification of building facade windows. In *Advances in Visual Computing*, pages 280–289, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. 2, 3

[36] H. Shao, T. Svoboda, and L. Van Gool. ZuBuD — Zürich buildings database for image based recognition. Technical Report 260, Computer Vision Laboratory, Swiss Federal Institute of Technology, 2003. 2, 3

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 6

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. 6

[39] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[40] Salma Taoufiq, Balázs Nagy, and Csaba Benedek. Hierarchynet: Hierarchical cnn-based urban building classification. *Remote Sensing*, 12(22):3794, 2020. 2, 3

[41] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection, 2019. 6

[42] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017. 6

[43] Congrong Xu, Justin Kerr, and Angjoo Kanazawa. Splatfacto-w: A nerfstudio implementation of gaussian splatting for unconstrained photo collections, 2024. 6

[44] Zhe Xu, Dacheng Tao, Ya Zhang, Junjie Wu, and Ah Chung Tsoi. Architectural style classification using multinomial latent logistic regression. In *Computer Vision – ECCV 2014*, pages 600–615, Cham, 2014. Springer International Publishing. 2, 3

[45] Wei Zhang and Jana Koecká. Hierarchical building recognition. *Image and Vision Computing*, 25(5):704–716, 2007.

[46] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. Tour the world: Building a web-scale landmark recognition engine. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1085–1092, 2009. 2, 3

A. Additional Results

We further break down the performance for each object class. Table 4 shows results for two IoU thresholds ($AP_{.50}$ and $AP_{.75}$).

	$AP_{.50}$					$AP_{.75}$				
Method	Lightpost	Door	Trash Can	Bike Rack	Electric Box	Lightpost	Door	Trash Can	Bike Rack	Electric Box
Faster R-CNN	0.38	0.73	0.88	0.72	**0.63**	**0.14**	**0.40**	0.30	0.21	0.27
FCOS	0.25	0.51	0.45	0.46	0.48	0.07	0.22	0.12	0.04	0.20
RetinaNet	**0.46**	**0.77**	**0.92**	**0.97**	0.55	0.13	0.39	**0.35**	**0.35**	**0.29**
SSD	0.22	0.67	0.68	0.72	0.50	0.07	0.24	0.18	0.22	0.25
SSDlite	0.22	0.52	0.41	0.46	0.21	0.04	0.06	0.05	0.03	0.12

Table 4. Object detection performance across two IoU thresholds.