VEHICLE CLASSIFICATION IN URBAN REGIONS OF THE GLOBAL SOUTH FROM AERIAL IMAGERY

M. Mühlhaus*, F. Kurz, A. R. Guridi Tartas, R. Bahmanyar, S. M. Azimi, J. Hellekes

Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
- manuel.muehlhaus@dlr.de

KEY WORDS: Aerial Images, Dataset Annotation, Deep Neural Networks, Global South, Object Detection, Vehicle Classification

ABSTRACT:

Land transport is a major contributor to the human-caused climate change; knowing the total number and composition of the vehicle fleet is key for estimating its emissions. Especially for countries of the Global South, emission inventories are associated with high uncertainties because fleet data are often unknown or outdated – classifying vehicles on remote sensing has the potential to change this. We present the XWHEEL dataset based on annotated vehicles in aerial images with six classes depending on the number of wheels, size and motorization. The dataset consists of 73 annotated aerial images of the city of Dar es Salaam (Tanzania) with 15,973 vehicles. To analyze the performance of the dataset, a convolutional neural network, ReDet, and a transformer-based neural network, DINO_{OBB}, are trained with different configurations and validated on the validation and test split, but also on aerial images from other regions. The transformer-based DINO architecture has been adapted to the remote sensing domain and modified to predict Oriented Bounding Boxes. Results show a good performance on the test split from Dar es Salaam, when the two-wheeled classes are merged and the non-motorized three-wheeled vehicles are excluded due to their rare occurrence. The best performing algorithm configurations with four classes were then tested on aerial images of Kathmandu (Nepal) and Kampala (Uganda). The performance drops for cycles and three-wheeled vehicles, as their appearance varies between countries. A main finding is that we can reliably detect the different vehicle classes in Dar es Salaam. When algorithms trained on XWHEEL are generalized to other regions of the Global South, performance decreases for the more difficult classes (bicycles and tricycles). To obtain results that are comparable across the board, we therefore recommend expanding the dataset with additional annotations from other regions of the Global South.

1. INTRODUCTION

Accurately classifying vehicles in remote sensing images, particularly in urban areas, has become increasingly crucial. Such a classification can provide valuable insights into traffic analysis and management, urban planning and development, environmental monitoring, and search and rescue operations. One application case is the estimation of land transport emissions in the Global South, where uncertainties in the absolute number of vehicles, the composition of the vehicle fleet and rapid settlement growth contribute to uncertainties in both the value and the spatial distribution of the calculated emissions.

To address these challenges, the combination of precise vehicle classification, mileage, and emission factors can improve the accuracy of total emission calculations compared to existing inventories. In addition, the use of remote sensing imagery facilitates the mapping of vehicles over large areas, allowing for more accurate and up-to-date modeling of land transport emissions on the country level. Currently, the majority of datasets available for vehicle detection algorithms are based on scenes from the Global North. As a result, the development of these algorithms has primarily been tailored to the conditions of this world region, potentially limiting their effectiveness in the Global South, where there are significant differences in the total number and types of vehicles. Consequently, relying solely on existing datasets to detect vehicles in Global South scenes may result in reduced accuracy, and any analysis based on the output of such algorithms would introduce significant uncertainty. Therefore, it is crucial to develop datasets that accurately reflect the unique conditions in order to improve the effectiveness of these algorithms in regions of the Global South. Figure 1 provides examples of some vehicles commonly found in the Global South but rarely seen in the Global North.

To address this issue, our paper introduces the XWHEEL dataset, which is a new aerial imagery dataset designed to improve vehicle detection in the Global South, focusing on Dar es Salaam (Tanzania). By creating this dataset, we aim to provide the community with a valuable tool for developing more accurate and effective vehicle detection systems that are tailored to the unique characteristics of this world region. In general, some areas of the Global South are covered with aerial imagery from public databases like OpenAerialMap. The challenge is that sensor quality and mosaicking are more heterogeneous and tend to be poorer than for the Global North. In particular when working with not-projected aerial images, the existing GNSS/inertial information must be further improved during a preprocessing step.

To create an effective dataset, it is crucial to begin by appropriately standardizing and categorizing vehicle types based on their appearance and relevance to different analytical applications. In addition, the process of annotating vehicles in remote sensing data involves several uncertainties that must be addressed by incorporating attributes. This task is particularly challenging in countries of the Global South, where vehicles exhibit diverse construction principles, making it difficult to assign them to specific types. Moreover, literature suggests that there can be overlaps when classifying vehicles into different

*Corresponding author

(a) combination bus – s4w (b) jeepney – l4w (c) microbus – s4w (d) tuk tuk – m3w (e) tricycle – m3w

(f) cycle rickshaw – n3w (g) motorcycle – m2w (h) man-powered rickshaw – n3w (i) motorela – m3w (j) cycle rickshaw – n3w

(k) tuk tuk – m3w (l) tuk tuk – m3w (m) small truck – s4w (n) motorcycle – m2w (o) small bus – l4w

(p) cycle rickshaw – n3w, diff (q) cycle rickshaw – n3w (r) cycle rickshaw – n3w, diff (s) tricycle – m3w (t) cycle rickshaw – n3w, diff

Figure 1. Examples of vehicle classes in the Global South (a–j) compared with their appearance in aerial images (k–t)

types (Chevre et al., 2022, Facchin, 2019). To solve this problem, we propose a new classification of vehicle types based on the number of wheels, such as cycles (2w), three-wheeled vehicles (3w), and four-wheeled vehicles (4w). Table 1 displays the categorization used to create the XWHEEL dataset with a few examples for each category. Figure 1 provides examples of how vehicle types are assigned to the proposed classes and how these vehicles look like in aerial images with a Ground Sampling Distance (GSD) of 5 cm, which refers to the the distance between the centers of two adjacent pixels measured on the ground.

Vehicle detection in aerial imagery poses unique challenges compared to ground imagery because vehicles are smaller, more numerous, and located in arbitrary orientations throughout the image. As a result, identifying and locating vehicles in aerial imagery requires advanced techniques and algorithms to overcome these challenges. To improve the accuracy of aerial object detection, Oriented Bounding Boxes (OBB) are often used instead of Horizontal Bounding Boxes (HBB), which are typically used to detect objects in ground images (Han et al., 2021).

Convolutional Neural Networks (CNN) have demonstrated impressive performance in object detection in recent years. However, since they are primarily designed for ground imagery, they often do not account for varying orientations. To address this problem, (Han et al., 2021) introduced a rotation-equivariant detector (ReDet) that incorporates both rotation equivariance and rotation invariance. ReDet's rotation-invariant network allows accurate prediction of OBBs without the need for additional network parameters or extensive orientation augmentation data. Transformer-based neural networks have emerged as the leading object detection methods in many ground imagery benchmark datasets, such as the COCO dataset (Lin et al., 2014). In particular, DINO (Zhang et al., 2022), a neural network based on Detection with Transformers (DETR) architecture, is often used as a detector in many of the top scoring methods such as InternImage (Wang et al., 2023) and FocalNet (Yang et al., 2022). Recent studies have explored the use of transformer-based architectures for object detection in aerial imagery, demonstrating promising results. For instance, (Wang et al., 2022) have applied plain vision transformers and ViTAE transformers to the DOTA dataset (Ding et al., 2021), a large remote sensing dataset for detecting multiple object classes including vehicles. Moreover, a recent study (Dai et al., 2022) proposes a method for adapting DETR (Carion et al., 2020) to predict OBBs.

We summarize our contributions as follows:

1. We publish the XWHEEL dataset to the community and demonstrate the potential and challenges of the dataset. The dataset is available here: https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-12760/22294_read-84375/

2. We train and test two vehicle detection methods: ReDet, a CNN-based approach, and DINO_{OBB}, a DETR-based method that we adapt from (detrex contributors, 2022). We add angle as a query/output to DINO for predicting OBBs and adjust the dimensions in the transformer and loss functions accordingly.

3. We demonstrate that we can reliably detect three-wheeled vehicles and four-wheeled vehicles based on different experiments. For cycles, the detections are less reliable as they are difficult to distinguish from pedestrians with shadows or other small objects.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-371-2023 | © Author(s) 2023. CC BY 4.0 License.

Cycles (2w)	
Motor. Cycles (m2w)	Non-motor. Cycles (n2w)
Motorcycle Motorcycle taxi	Bicycle
3-wheeled vehicles (3w)	
Motor. 3-wheeled (m3w)	Non-motor. 3-wheeled (n3w)
Auto rickshaws Motorcycle plus trailer Motorcycle rickshaw, front/back Motorela Tricycle Tuk-tuk	Cycle rickshaw front Cycle rickshaw back Cycle rickshaw side Man-powered rickshaw
4-wheeled vehicles (4w)	
Small 4-wheeled (s4w)	Large 4-wheeled (l4w)
Micro bus Shared cab Van Car Trucks	Combination bus Jeepney Lorry Large, medium, mini bus

Table 1. XWHEEL's vehicle classes with examples

2. XWHEEL DATASET

Our XWHEEL dataset consists of 73 annotated aerial images acquired over Dar es Salaam (Tanzania). These images are selected from a collection of $11,822$ aerial images[1] with an average size of 5472×3456 pixel covering $45\ \mathrm{km}^2$. These images were captured by drones in 2017 at two different flight altitudes: $210\ \mathrm{m}$ above ground for a smaller area near the city center, and $350\ \mathrm{m}$ for the rest of the city. This resulted in an average ground sampling distance (GSD) of $8\ \mathrm{cm}$ and $5\ \mathrm{cm}$ for the higher and lower flight altitudes, respectively.

In this dataset, we initially classify vehicles into three types based on the number of wheels: cycles (2w), three-wheeled vehicles (3w), and four-wheeled vehicles (4w). Additionally, we take into account other attributes such as size (large or small) and motorization (motorized or non-motorized). We also consider attributes that reflect the appearance of the vehicles, such as 'occluded' (occl), 'difficult' (diff), and 'uncertain' (un). If more than $10\,\%$ of a vehicle's total or estimated total surface is obscured by objects such as trees, buildings, or bridges, it shall be annotated with the attribute 'occluded'. Additionally, if the class or motorized status of a vehicle is difficult to discern or cannot be identified at all, it shall be annotated with the attribute 'difficult' or 'uncertain', respectively. The XWHEEL dataset's categorization, along with sample instances for each category, is presented in Table 1. As shown in the table, the category of cycles is further divided into motorized (m2w) and non-motorized (n2w) categories. Cycles are a common mode of transportation in this region; however, differentiating between bicycles and motorcycles in aerial images can be challenging. In addition, the category of three-wheeled vehicles can also be divided into motorized (m3w) and non-motorized (n3w) categories. These vehicles are usually smaller than four-wheeled vehicles and display a significant degree of variability in terms of their shapes and colors across different regions. Non-motorized three-wheeled vehicles tend to be smaller than the motorized ones. As with cycles, distinguishing between motorized and non-motorized three-wheeled vehicles can also be a challenge in aerial imagery. Additionally, the category of four-wheeled vehicles is further subdivided into small (s4w) and large (l4w) subcategories. Vehicles with more than four wheels are also classified as large and assigned to the four-wheeled category. Moreover, we classify vehicles that are 1.5 times larger than standard cars as large. However, this assumption can lead to a grey area between the small and large categories, which may make it challenging to accurately classify some vehicles in certain cases. The proposed classes for vehicle types and their assignment are illustrated in the examples shown in Figure 1.

[1] Worldbank, senseFly S.O.D.A.

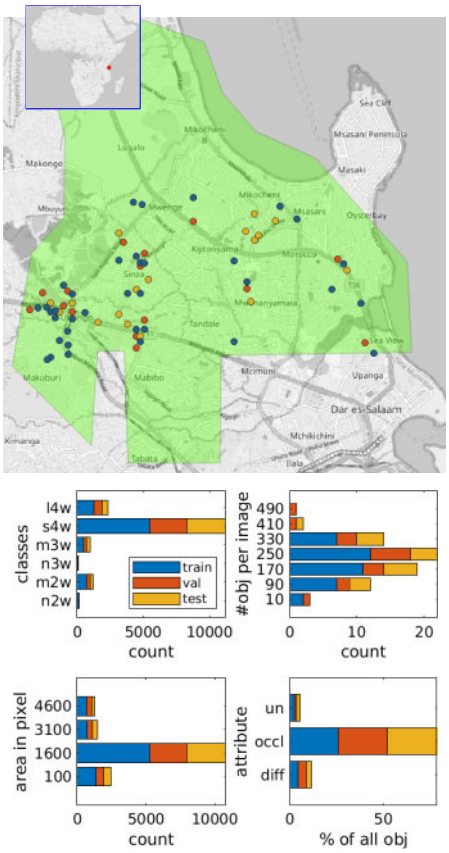


Figure 2. Spatial distribution and numerical statistics of XWHEEL dataset

To obtain annotated images for our dataset, we select 73 images with a GSD of $8\ \mathrm{cm}$ from various locations that cover different traffic situations. We annotate all vehicles in the selected images with OBB, and assigned each one its corresponding class and attributes. In total, we annotated $15,973$ vehicles. In order to ensure accurate annotation of data, we develop an annotation policy that specifies the defining characteristics of each class. Next, a team of experts performs a multi-level quality check to ensure that the annotations meet high standards. To illustrate the quality of our annotated data, we include several examples in Figure 4 that showcase our training dataset. In order to evaluate the performance of our model, we split the 73 images into three disjoint sets: the training set, which consists of 39 images, the validation set, which consists of 17 images, and the test set, which also consists of 17 images. This split results in 8,034 training vehicles, 3,906 validation vehicles, and 4,033 test vehicles. In addition, we prepare georeferencing information (interior and exterior parameters) for each image, which allows us to project the annotated vehicles onto world coordinates using an elevation model.

Figure 2 displays the spatial distribution of the selected images across the three sets, along with the distribution of the images with respect to the number of instances per class, the total number of instances, the annotated area per instance, and the at-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-371-2023 | © Author(s) 2023. CC BY 4.0 License.

tributes. Note that the images from the three sets do not overlap even though the location of images appear to be quite close in the figure. Based on the statistics, there is a significant class imbalance in the vehicle classes. The majority of all annotations are for small four-wheeled vehicles (s4w), and only a small number of non-motorized two- and three-wheeled vehicles are annotated. Additionally, the figure reveals that most images contain approximately 250 instances, with an average instance size of around 1600 pixels. Furthermore, less than $10\,\%$ of all instances have the attributes 'difficult' or 'uncertain' set to true. Based on a visual survey, it appears that the most common type of three-wheeled vehicle in Dar es Salaam is the tuk-tuk, which can be identified by its black roof. However, it is difficult to distinguish other types of three-wheeled vehicles, especially non-motorized ones. Although cycles outnumber three-wheeled vehicles, only motorized cycles are easy to distinguish.

As a means to assess the transferability of the algorithms trained on the Dar es Salaam images to other regions, we use an orthorectified aerial image mosaic of Kampala, Uganda[2]. This image mosaic covers a representative area of $0.013\,\mathrm{km}^2$ near the city center, with a GSD of $5\,\mathrm{cm}$. In addition, we use an image mosaic of Kathmandu, Nepal[3], taken in 2014, with a GSD of $10\,\mathrm{cm}$ and covering an area of $0.013\,\mathrm{km}^2$. In order to quantitatively evaluate the performance of vehicle detection, we partially annotate the vehicles in the image of Kampala. However, we left the vehicles in the remaining images unannotated due to the significant cost and human effort involved. Thus, our evaluation of the vehicle detection performance in those images is qualitative. Our observations show that in Kampala and Kathmandu the m3w class is less represented and less homogeneous compared to Dar es Salaam, where almost all m3w are characterized by a black roof.

3. EXPERIMENTS AND RESULTS

Prior to training, we perform offline augmentation on the dataset. The augmentation process involves resizing the original images by factors of 0.5, 1.0, and 1.5. After resizing, the images are split into patches of size 1024×1024 with a stride of 824. Since there are no large objects in the images, there is no need for significant overlap between patches.Any objects that are truncated by tiling are preserved if more than $50\,\%$ of the object remains within the tile. When testing scaled images, we apply Non Maximum Suppression (NMS) to suppress overlapping results of the same class with an Intersection over Union (IoU) greater than 10%. Results with lower confidence levels are suppressed. To prepare the model for a variety of scenarios, we apply standard online data augmentation techniques during training. These include cropping, flipping, and rotation augmentation.

3.1 Object detection methods

We used a state-of-the-art CNN-based architecture for the automatic detection of vehicles from aerial images, called ReDet (Ding et al., 2021). The main advantage of ReDet are rotation-invariant features, which reduces the model size and data needed to train the model compared to other CNNs.

As an alternative neural network architecture, we utilize the state-of-the-art transformer model DINO (Zhang et al., 2022). This model is initially optimized for detecting HBBs on the COCO dataset, but we aim to adapt it to the remote sensing domain.

Specifically, we modify the output dimension of DINO to predict OBBs by adding a rotation parameter to the existing four HBB parameters, resulting in five parameters. To ensure that our 5D OBB is compatible with the DINO model, we change the dimensions of the queries throughout the transformer. In addition, we adjust the Hungarian matching process by changing the generalized IoU loss to a distance IoU loss (Zheng et al., 2019). This change improves the accuracy of object detection by refining the matching process. To further improve the stability of the matching process, DINO incorporates a denoising loss by adding noise to the ground truth inputs during training. Since we are changing the dimensions of the output to 5D OBBs, we must also change the dimensions of the inputs to match the new format.

3.2 Evaluation metrics

In order to evaluate our experiments, we use mean Average Precision (mAP), which is calculated by computing Precision (P) and recall (R):

$$P = \frac{TP}{TP + FP}, \tag{1}$$

$$R = \frac{TP}{TP + FN}, \tag{2}$$

where (TP) are the True Positives, (FN) the False Negatives and (FP) the False Positives. These metrics depend on the confidence threshold required to count as a detection and can be plotted against each other for every confidence threshold, the so-called Precision-Recall curve. The area under this curve for each class is then calculated as AP (see Figure 3):

$$AP = \int_0^1 P(R)dR \tag{3}$$

The mAP score is then obtained by taking the mean value of all APs across all classes:

$$mAP = \frac{1}{N} \sum_n^N AP_n \tag{4}$$

3.3 Experiments

In order to analyze the potential of the XWHEEL dataset for training ReDet and $\mathrm{DINO}_{\mathrm{OBB}}$, we conduct several experiments denoted as A to L in Table 2. Due to the relatively small size of the XWHEEL dataset, we train all models for 36 epochs. We trained the object detection algorithms with different numbers of classes and evaluated the results on the validation and test sets of the XWHEEL dataset as well as the aerial images from Kampala and Kathmandu. In our experiments, we fuse the non-motorized cycle class with the motorized cycle class (n2w + m2w → 2w) as the number of annotated non-motorized cycles is very low. This allowed us to increase the number of samples in the 2w class, which could help improve the model's performance.

Experiments A and B in Table 2 represent the results of ReDet in detecting five vehicle classes on the validation and test sets.

[2]OpenAerialMap, CC-BY 4.0
[3]German Aerospace Center, MACS camera system

						AP (%)					
exp.	model	backbone	trained on	evaluated on	classes	mean	2w	n3w	m3w	s4w	l4w
A	ReDet	ReResNet50	train	XWHEEL val	5	67.0	55.4	5.9	85.2	95.9	92.4
B	ReDet	ReResNet50	train&val	XWHEEL test	5	65.8	55.6	2.2	85.3	96.0	89.3
C	ReDet	ReResNet50	train	XWHEEL val	4	82.2	55.4	–	85.2	95.9	92.4
D	ReDet	ReResNet50	train	XWHEEL test	4	80.3	53.9	–	82.9	96.1	88.2
E	ReDet	ReResNet50	train&val	XWHEEL test	4	80.9	54.1	–	83.9	96.4	89.1
F	ReDet*	ReResNet50	train&val	XWHEEL test	4	67.2	37.2	–	63.2	91.4	71.1
G	DINO	Swin_tiny_224_4_scale	train	XWHEEL test	4	78.3	49.7	–	83.3	94.2	85.9
H	DINO*	Swin_tiny_224_4_scale	train	XWHEEL test	4	79.4	49.8	–	85.5	95.3	87.1
I	DINO*	Swin_base_384_4_scale	train	XWHEEL test	4	80.2	50.3	–	87.7	95.4	87.4
J	DINO*	Swin_large_384_4_scale	train	XWHEEL test	4	80.9	51.0	–	89.2	95.4	87.9
K	DINO*	Swin_large_384_4_scale	train&val	XWHEEL test	4	81.6	52.9	–	89.4	96.1	88.2
L	DINO*	Swin_large_384_4_scale frozen(3/4)	train&val	XWHEEL test	4	82.2	53.2	–	90.6	96.2	88.6
M	ReDet	ReResNet50	train&val	Kampala	4	50.3	12.3	–	4.0	91.5	93.6
N	DINO*	Swin_large_384_4_scale	train&val	Kampala	4	57.8	20.1	–	20.0	92.1	99.0
O	DINO*	Swin_large_384_4_scale frozen(3/4)	train&val	Kampala	4	51.9	17.2	–	0.0	91.8	98.8

Table 2. Mean average precision of ReDet and DINO_{OBB} for different number of classes evaluated on validation or test split with different pre-trainings and backbones

model	backbone	# param.	speed	mAP
ReDet	ReResNet50	31.6 M	1.78 MP/s	77.6
DINO_{OBB}	Swin_tiny_224_4_scale	48.1 M	9.31 MP/s	75.8
DINO_{OBB}	Swin_base_384_4_scale	108.1 M	5.41 MP/s	77.4
DINO_{OBB}	Swin_large_384_4_scale	217.9 M	3.79 MP/s	78.5

Table 3. Comparison of model parameters, inference speed on TITAN RTX (24GB) and mAP (%) of EAGLE pre-trains (12 epochs from scratch on EAGLE) for the different models

As can be seen, the model performance for the n3w class is very poor. This is due to the small number of samples available for this class and the high variability in its appearance. Thus, we have decided to exclude n3w from the experiments C onward. By excluding this class, we simplify the experiments and focus on the classes with more samples and less variability in appearance. In experiment F, we pretrain ReDet for 12 epochs on the EAGLE dataset (Azimi et al., 2021), a vehicle dataset with many instances of two vehicle classes with high heterogeneity, and then fine-tune it on the XWHEEL dataset. In Table 2, for all models marked with a star (*), we follow the same pretraining procedure.

In experiments G to L, we train and evaluate DINO_{OBB} with different Swin backbones on the XWHEEL dataset. In experiments O to M, we evaluate the best ReDet and DINO_{OBB} models (experiments E, K, L) on an aerial imagery from Kampala.Furthermore, we apply the best ReDet and DINO_{OBB} models (E, L) to an aerial imagery from Kathmandu and since we do not have annotations for this image, we perform a qualitative evaluation. Table 3 shows some details about the different model configurations we use in our experiments.

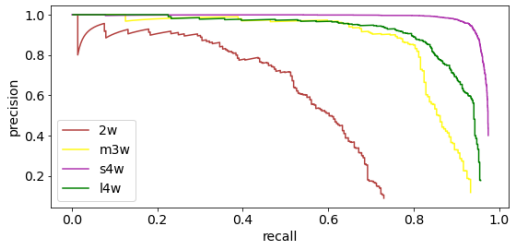
3.4 Results and discussion

According to Table 2, in experiments A and B, the AP scores for m3w, s4w, and l4w are notably high. However, the AP score for 2w is slightly lower, and the AP score for n3w is considerably lower. The reason for the poor results of n3w is the significantly low number of samples available in Dar es Salaam, which makes it very difficult to accurately identify and classify them. Therefore, we exclude them from our experiments and only consider the four other classes: 2w, m3w, s4w, l4w. Comparing experiments A to E, the test split seems to be more challenging for the algorithm than the validation split. Even if we add the validation part to the training in experiment E, we still get lower results compared to just evaluating on the validation split as in C. Based on the results, using the pre-trained model in experiment F significantly degrades performance. At this point, it is not clear why this is the case, or whether modifying the hyperparameters will solve the problem. However, since ReDet contains significantly fewer parameters (see Table 3) and is rotation invariant, it is unlikely that fine-tuning the model will yield substantial improvements over training it from scratch.

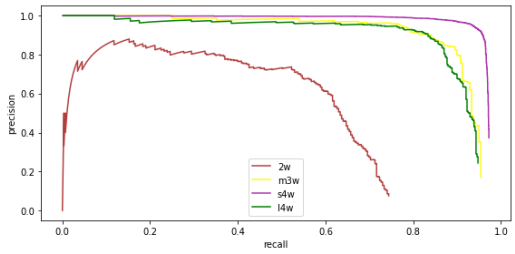
In contrast to ReDet, our DINO_{OBB} models exhibit improved performance when pre-trained backbones are used, even with smaller backbones. However, it is essential to utilize pre-trained backbones when employing larger backbones to achieve optimal results. For example, using the pre-trained backbone in experiment H results in a 1.1% improvement in mAP. As a result, we use pre-trained backbones for all subsequent DINO_{OBB} experiments. For experiments H through K, the backbone parameters are not frozen during fine tuning. However, in experiment H, 3 of the 4 backbone levels are frozen to preserve the lower-level features learned by EAGLE, and only the higher-level features are further trained. Given the improvement observed with the latter configuration, we plan to experiment with freezing different stages in the future, particularly as the initial test looks promising. Figure 3 shows the class-wise precision-recall curves for experiment E, the best result with ReDet and experiment L, the best result with DINO_{OBB}.

Additionally, the results demonstrate that DINO_{OBB} performs exceptionally well on m3w. Since 2w and m3w have similar frequencies in the dataset, it is unlikely that the different ability of DINO_{OBB} and ReDet to handle class imbalances is the primary reason. We currently assume that 3w objects are typically better recognized by context, as they often appear in clusters within images and are comparatively easier to identify in contrast to s4w objects. The qualitative results of different experiments are demonstrated in Figure 4. The ground truth is shown in the first column, while the second and third columns display the detections from ReDet and DINO_{OBB}, respectively. The confidence scores assigned to the detections from the network are indicated by the numbers on the bounding boxes in the second and third columns. For the qualitative results, we apply NMS across classes if there is an IoU overlap of more than 50%, while for the mAP calculation, NMS is only applied class-wise. Additionally, only detections with a confidence score higher than 30% are plotted.

In Figure 4, the first two rows show the qualitative results of the experiment E and L respectively. According to these results, it appears that DINO_{OBB} typically detects fewer false



(a) ReDet



(b) DINO$_{\mathrm{OBB}}$

Figure 3. Precision-Recall curves for experiment E and L.

positives and assigns lower confidence scores to detections than ReDet. This observation leads to a higher AP for the m3w class. The inference result of DINO$_{\mathrm{OBB}}$ on an image section of the XWHEEL test set is presented in the form of a confusion plot in Figure 5. The classes of s4w and l4w are accurately identified with mostly high confidence thresholds. The 3-wheelers in the upper middle are detected, but with lower confidence scores. While there are few false positives, some of them could potentially be true positives (such as the occluded 0.32 or 0.79 score detections); however, annotating them with high confidence is not feasible.

We perform experiments M to O to evaluate the transferability of the best ReDet and DINO$_{\mathrm{OBB}}$ models to a very small aerial image of Kampala. The image contains 178 annotated vehicles, and we do not train the models on this new image. The goal is to test the generalization ability of the models on different datasets. The Kampala image presents a challenge due to its diverse vehicle distribution and a GSD of $5\,\mathrm{cm}$, which differs from the GSD of the XWHEEL training data, which is mostly at $8\,\mathrm{cm}$. We observe that changing the resolution of the input image causes some small objects to be detected as cars, especially for the ReDet model (see Figure 4, the third row). Although DINO$_{\mathrm{OBB}}$ detects fewer false positives than ReDet, it also misses most of the 2ws and 3ws. We see that the mAP scores for s4v and l4w are extremely high, as they are quite easy to distinguish in the test image compared to some of the XWHEEL test images. Furthermore, the number of annotations in our dataset is insufficient to draw statistically significant conclusions beyond initial observations. For example, the AP of l4w appears high, but this is likely due to a recall of about 1 for most precision values, even with some false positives present in the results. Therefore, care should be taken when drawing firm conclusions based solely on these preliminary results. We also observe that there are very few three-wheeled vehicles in Kampala, and the few that are present have a different appearance than those in the Dar es Salaam training set. In the case of the 2w class, we find that the change in resolution contributes significantly to the degradation in performance, with many 2w vehicles being misclassified as small s4w vehicles or missed altogether.

The qualitative results for Kathmandu are presented in the fourth row of Figure 4. The performance of the model drops significantly in Kathmandu compared to Kampala, likely due to the worse GSD of the aerial images, which is $10\,\mathrm{cm}$ and lower than that of the training dataset. Consequently, many m3w and 2w vehicles are either not classified or classified with low confidence scores due to the low spatial resolution, poor lighting conditions, and mosaicing artifacts. Accurate annotation of these images would also be challenging due to the aforementioned difficulties.

4. CONCLUSIONS

Our paper presents the XWHEEL dataset, comprising of 73 high-resolution aerial images acquired from Dar es Salaam, Tanzania. The dataset contains 15,973 annotated vehicles, which are classified into six distinct categories based on their size, motorization, and number of wheels. In our study, we trained the ReDet and DINO$_{\mathrm{OBB}}$ detection algorithms on the XWHEEL dataset using different configurations. The results show a high detection performance on the two- and four-wheeled vehicles as well as the motorized three-wheeled vehicles. However, the non-motorized three-wheeled vehicles cannot be accurately detected due to insufficient training samples, despite our augmentation attempts. As a result, we have temporarily excluded this class from the analysis until additional annotated images become available. Nevertheless, the class remains part of the dataset and may pose a challenge for future studies.

When we test the ReDet and DINO$_{\mathrm{OBB}}$ detection algorithms with aerial imagery from other cities, we observe a significant drop in the performance of the two- and three-wheeled classes. This highlights the need to expand the dataset by including annotations from other cities. To achieve this, we conducted a thorough literature review on the fleet composition of countries in the Global South, as documented in (Salazar, 2014, Salazar, 2015). Based on our research, we anticipated a high occurrence of three-wheeled vehicles. However, during the annotation process, we discovered that the number of annotated three-wheeled vehicles was lower than expected. As a result, the imbalance of annotations between classes is one of the main reasons for the decreased performance of the two- and three-wheeled vehicle detection in our experiments. This suggests a possible change in the composition of the fleet in recent years. However, our study represents a crucial first step towards achieving accurate vehicle classification in the Global South. We have demonstrated that high-resolution aerial imagery can be used to classify vehicles by the number of wheels, which can have significant implications for modeling land transport emissions. Remote sensing technologies offer the possibility of mapping large areas accurately and in a timely manner, making it possible to model the impacts of different vehicle types on air quality and public health.

Our next step is to expand the XWHEEL dataset to include aerial imagery from other countries in the Global South. We plan to iteratively expand the dataset to cover different regions in the Global South, creating a representative dataset for the entire region. With the algorithm trained on this extended dataset, we can process vehicle detection on a city-wide basis for different countries in the Global South, thereby creating an inventory of land transport emissions. In particular, we plan to explore the potential of high-resolution satellite data with a GSD of $30\,\mathrm{cm}$ for this task. Additionally, we aim to improve the algorithms used for detection by experimenting with different

Annotations ReDet DINO_{OBB}

Figure 4. Results of ReDet and DINO_{OBB} on Dar es Salaam (first two rows), Kampala (third row), and Kathmandu (fourth row) for l4w, s4w, m3w, 2w classes. Additionally, the confidence scores of the inference are plotted next to each detection.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-371-2023 | © Author(s) 2023. CC BY 4.0 License.

Figure 5. Confusion plot of DINO_{OBB} on a test image crop from XWHEEL. The confidence threshold is set to 0.3 and the IoU threshold, overlap with the ground truth for a detection to count as a TP, is set to 50%. The different entries are color coded: TP, FP, FN, False Class. Additionally, the confidence scores of the inference are plotted next to each detection.

loss functions, freezing different stages, and further adapting DINO_{OBB} to the remote sensing domain. These efforts will help us achieve greater accuracy and coverage in our vehicle detection and emissions modeling efforts.

REFERENCES

Azimi, S., Bahmanyar, R., Henry, C., Kurz, F., 2021. EAGLE: Large-scale vehicle detection dataset in real-world scenarios using aerial imagery. *ICPR 2020*, 6920–6927.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (eds), *Computer Vision – ECCV 2020*, Springer International Publishing, Cham, 213–229.

Chevre, A., Williams, S., Plassat, G., Nyamador, E. S., Krambeck, H., Klopp, J. M., et al., 2022. Digital transport for africa. an open resource center for mapping public transport across africa. https://digitaltransport4africa.org/.

Dai, L., Liu, H., Tang, H., Wu, Z., Song, P., 2022. AO2-DETR: Arbitrary-Oriented Object Detection Transformer. *ArXiv*, abs/2205.12785.

detrex contributors, 2022. detrex: An research platform for transformer-based object detection algorithms. https://github.com/IDEA-Research/detrex.

Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2021. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1.

Facchin, A., 2019. Mapping and representing informal transport: the state of the art. https://shorturl.at/bmF89.

Han, J., Ding, J., Xue, N., Xia, G.-S., 2021. Redet: A rotation-equivariant detector for aerial object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2786–2795.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (eds), *Computer Vision – ECCV 2014*, Springer International Publishing, Cham, 740–755.

Salazar, P., 2014. Paratransit: A key element in a dual system. https://wedocs.unep.org/20.500.11822/16825.

Salazar, P., 2015. The challenge of finding a role for paratransit services in the global south. https://wedocs.unep.org/20.500.11822/16825.

Wang, D., Zhang, Q., Xu, Y., Zhang, J., Du, B., Tao, D., Zhang, L., 2022. Advancing Plain Vision Transformer Towards Remote Sensing Foundation Model. *IEEE Transactions on Geoscience and Remote Sensing*, PP, 1-1.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., Wang, X., Qiao, Y., 2023. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions.

Yang, J., Li, C., Gao, J., 2022. Focal Modulation Networks. *ArXiv*, abs/2203.11926.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J.-J., shuan Ni, L. M., yeung Shum, H., 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *ArXiv*, abs/2203.03605.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D., 2019. Distance-iou loss: Faster and better learning for bounding box regression. *AAAI Conference on Artificial Intelligence*.