Impact of Calibration Matrices on 3D Monocular Object Detection: Filtering, Dataset Combination and Integration of Synthetic Data

Alexandre Evain
ESIGELEC
Technopole du Madrillet
Avenue Galilee
France (FRA), 76800, Saint-Etienne-du-Rouvray
alexandre.evain@groupe-esigelec.org

Redouane Khemmar
ESIGELEC
Technopole du Madrillet
Avenue Galilee
France (FRA), 76800, Saint-Etienne-du-Rouvray
redouane.khemmar@esigelec.fr

Mathieu Orzalesi
SEGULA Technologies
19 rue d'Arras
France (FRA), 92000, Nanterre
mathieu.orzalesi@segula.fr

Sofiane Ahmedali
Universite d'Evry Val d'Essonne
2 Rue du Facteur Cheval
France (FRA), 91000, Evry-Courcouronnes
sofiane.ahmedali@univ-evry.fr

ABSTRACT

In traditional 2D object detection, augmenting datasets typically enhances model precision. However, 3D estimations from a 2D image are dependent on the camera's focal length, meaning that differences in focal length may undermine distance estimation, object dimension estimation, and subsequent 3D position estimation. In this article, we attempt to evaluate the impact of different calibration matrices on 3D monocular object detection. Firstly, we assess the impact of different calibration matrices within the same dataset by comparing the performance of filtered, non-filtered, and normalized datasets using the NuScenes dataset as a base. Our results show that filtering the dataset to only keep images sharing the same focal lengths results in increased depth and dimension estimations but at the expense of the other metrics. Then, we investigate the impact of dataset combination on 3D monocular object detection, focusing on the integration of datasets with varying focal lengths and matrices. Leveraging the NuScenes dataset, this time augmented with additional synthetic data from GTA, we evaluate the efficacy of dataset combination in improving model performance across a range of metrics. Contrary to our initial expectations, incorporating additional datasets does not consistently result in 2D performance improvements depending on their visual appearance, but also does not always result in decreased 3D performance either, despite their different focal lengths providing the model with contradictory 3D visual information, as long as the data contained is accurately labeled, showing that dataset combination has the potential to improve 3D monocular object detection.

Keywords

3D Monocular Object Detection, Dataset combination, Computer Vision, Camera Calibration Matrix, Focal Length, Dataset Filtering, Dataset Normalisation

1 INTRODUCTION

1.1 Dataset Combination

In machine-learning object detection, data availability is often the primary bottleneck in achieving optimal model performance, especially for real-life applications where diverse scenarios must be accurately captured. While numerous datasets (such as [1, 2, 3, 4]) exist for training machine learning models, each comes with its own set of limitations, necessitating the strategic combination of datasets to address these constraints effectively. Firstly, existing datasets exhibit variations in terms of the covered conditions and object classes. As an example, certain datasets might focus solely on daytime, clear weather conditions and do not have scenarios such as nighttime or adverse weather conditions. Other datasets are limited by the object class they cover, resulting in models able to detect cars but not buses as another example ([1] covers 3 classes while [2] cover 9). In addition, domains are also covered inequally by the existing datasets: While certain domains such as road situations might boast an abundance of datasets,

others like railroad scenarios suffer from scarcity. Furthermore, variations in the quality of images and annotations across datasets pose another challenge, as some datasets have their own specific image ratios or limited image quality.

Given these limitations, dataset combination emerges as a viable strategy to address the shortcomings inherent in individual datasets: Firstly, by aggregating multiple datasets, we can mitigate the incompleteness of coverage by incorporating diverse conditions, object classes, and domains into the training data, thereby enhancing the model's ability to generalize across a broader spectrum of scenarios. Then, dataset combination can increase the model's robustness. By amalgamating datasets with different scenarios, we can increase the representation of real-world scenarios.

Another interesting aspect of dataset combination is that it allows using synthetic datasets. These datasets, like [5], though providing perfect annotations, are still visually distinct from real-world images, potentially compromising their utility in practical applications. This limitation can be alleviated by combining synthetic datasets alongside real-world datasets. This way, we not only enrich the latter with additional scenarios but also imbue the synthetic datasets with greater realism. This amalgamation helps bridge the gap between synthetic and real-world data, enhancing the model's adaptability to real-life scenarios.

1.2 Focal Length and Contradictory Visual Information

In traditional 2D object detection tasks, the primary objective revolves around accurately identifying objects within an image. However, transitioning to 3D monocular object detection introduces additional challenges, such as estimating objects' distances, dimensions, sizes, and orientations.

While augmenting the dataset might bolster the 2D aspect of detection, the same approach may not yield commensurate improvements for the 3D predictions. This discrepancy is caused by the relationship between the camera's focal length, its field of view, the scene geometry, and the resulting image. Unlike in 2D detection, where object appearance suffices, the 3D estimations are fed potentially contradictory information:

- The size of an object within an image is dependant not only on its true dimensions but also on its distance from the camera and the camera's focal length, as expressed in the Equation (1):

$$d = \frac{f \cdot \mathrm{H}}{h} \tag{1}$$

  With h the heights of the object in pixels, H the actual width of the object, and f the camera's focal length.

  As a result, two images portraying objects of apparently identical dimensions might convey disparate distance estimations if captured using cameras with different focal lengths.

- In addition, the camera's field of view also affects the orientation estimations as well as the positions of the objects within the image. This effect is very noticeable when using wide-angle cameras, leading to side distortion, and method such as [6, 7] solve this problem by either making a FOV-independant detection model or by using sensor fusion.

- Finally, inaccuracies in distance or dimension estimation can reverberate through subsequent stages of 3D position estimation.

Some existing detection methods like [8, 9] take the focal length into account in their detection models. However, we aim to investigate the impact of different calibration matrices and dataset combinations on conventional 3D object detection models that do not explicitly consider focal length, thus highlighting the importance of these factors in enhancing detection performance. We aim to evaluate this impact through two primary avenues: Firstly, we investigate the ramifications of incorporating images captured using diverse camera setups within a single dataset, each equipped with its unique calibration matrix. Then, we assess the consequences of amalgamating datasets sourced from disparate sources, each characterized by distinct calibration matrices. We aim to determine the relationship between dataset composition and the efficacy of 3D monocular object detection and determine whether dataset composition can be overcome on the dataset level without directly modifying the object detection models.

In summary, our work has the following contributions:

- We examine how different calibration matrices within datasets affect 3D detection, and we assess

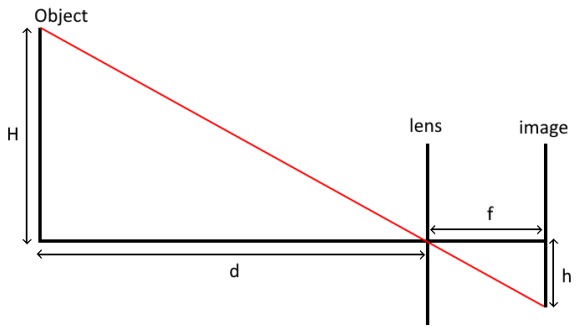

Figure 1: Relation between the camera's focal length f, the object's height in the image h, the real object's height H and the object's distance to the camera d, as explained in the Equation (1)

the efficacy of filtering methods in addressing contradictory information from diverse calibration matrices.

- We analyze combining datasets for 3D monocular object detection, understanding how different focal lengths affect model performance and 3D estimations.
- We analyze the effects of the visual normalization of the camera calibration matrices through distortion, particularly in the context of synthetic datasets, on prediction accuracy and model performance.

2 RELATED WORK

2.1 Monocular 3D Object Detection taking focal length into account

Monocular 3D object detection research has explored various methodologies to address inherent challenges in autonomous driving perception tasks. All the methods in this section focus on taking into account the focal length in the model itself for more accurate 3D predictions; we aim to differ from their approach by instead acting on the datasets themselves to see the effects of different focal lengths, of dataset filtering/normalization/combination on 3D object detection without modifying the object detection method itself. Our focus is to see what can be done to improve the detection while only changing the dataset.

The most recent method is MonoGDG[6], which proposes a geometry-guided domain generalization framework, addressing gaps at both camera and feature levels by incorporating geometry-based image reprojection and feature disentanglement techniques. The paper addresses most of the limitations of dataset combination by taking Focal Length, FOV Distortion, FOV Range, Camera Orientation, and Image Appearance into account in its architecture.

Another method, ODD-M3D[10] proposes object-wise dense depth estimation, improving depth estimation accuracy by randomly sampling points from the bounding box area of each object, and then using these using pre-generated sampled points for their depth estimation method, instead of relying on a single center point.

MonoEdge[8] proposes utilizing local appearance cues, particularly the edges of 3D bounding boxes, to estimate depth and global yaw angle directly from object appearance in images, enabling object depth and yaw angle derivation without requiring absolute size or position information and bypassing the need for explicit camera intrinsic parameters as well.

MonoUNI[11] introduces a unified optimization target, normalized depth, which addresses discrepancies between vehicle and infrastructure-side detection due to variations in pitch angle and focal length.

Advancements in depth estimation techniques, such as those explored by Deep Optics[12], integrate optics and image processing to improve depth estimation performance, with implications for 3D object detection tasks.

Flexibility and adaptability are crucial considerations, with approaches like Objects Are Different[13] offering frameworks that explicitly account for truncated objects and adapt multiple approaches for object depth estimation.

Additionally, incorporating motion cues for depth estimation and object detection presents promising avenues. Monocular 3D Object Detection with Depth from Motion[14] explores synergies between camera ego-motion and monocular understanding to improve accuracy and robustness in object detection tasks.

2.2 Datasets

The field of 3D object detection in computer vision has experienced notable advancements, driven by the availability of diverse datasets catering to various aspects of the task. These datasets have played a crucial role in benchmarking algorithms and propelling progress in the domain. However, each dataset typically focuses on specific scenarios, sensor modalities, or annotation techniques, leaving certain aspects of 3D object detection unexplored. In this section, we review related work spanning the development of diverse datasets for 3D object detection.

The KITTI dataset family [1, 15, 16, 17] has been instrumental in driving progress in various computer vision tasks relevant to autonomous driving. It introduced fundamental benchmarks for stereo, optical flow, visual odometry, and 3D object detection. KITTI-360 further extends this by focusing on suburban driving scenarios, while Virtual KITTI leverages computer graphics to propose an efficient real-to-virtual world cloning method.

Road datasets such as nuScenes [2] and Rope3D [18] provide diverse and challenging data for advancing roadside perception and autonomous driving technologies. These datasets offer extensive annotations and analysis for object detection and tracking, capturing diverse scenes and environmental conditions.

Efforts like A*3D and H3D [19, 20] aim to provide challenging real-world datasets with diverse scenes, varying weather conditions, and dense annotations, pushing the boundaries of autonomous driving research into more challenging environments.

Enhancing 2D datasets for 3D object detection, Cityscapes 3D [21] and PASCAL3D+ [22], augment existing datasets with 3D annotations, providing richer annotations and increasing variability for studying 3D detection and pose estimation.

Leveraging computer games for dataset creation has emerged as a cost-effective alternative to manual

data collection. Approaches like "Ground Truth from Computer Games" and "Free Supervision From Video Games" [5] demonstrate the feasibility of extracting pixel-level semantic labels and ground truth annotations from video games in real time, providing visually realistic images for training models on large-scale datasets.

3 METHOD

3.1 Dataset Configurations

Table 1: NuScenes[2] Matrix 1. The first three columns are the camera's rotation matrix (3,3), and the last column (1,3) is its translation matrix.

1252	0.0	826	0.0
0.0	1252	469	0.0
0.0	0.0	1.0	0.0
0.0	0.0	0.0	1.0

To understand the impact of having different camera calibration matrices present within the same dataset on 3D monocular object detection, we decided to use the NuScenes dataset[2] trimmed down to focus on three distinct classes (cars, pedestrians, and cyclists), because NuScenes images have three different camera calibration matrices:

- Matrix 1, with occurrences totaling 16,443 instances, constituting 56.71% of the dataset.
- Matrix 2, featuring 12,082 occurrences, representing 41.67% of the dataset.
- Matrix 3, featuring 468 occurrences, representing 1.61% of the dataset.

To systematically assess the influence of these matrices, we initiated our analysis by creating a validation split consisting of 2000 images using only Matrix 1. All our models have been tested on this single validation split.

Then, we created the different training sets: First, we formed a base training from NuScenes set comprising 5000 images. Within this set, 2886 images corresponded to Matrix 1 such as 2, while 2114 images were associated with Matrix 2 (no image used Matrix 3). Since NuScenes is a sequential dataset, we made sure that all the images from the validation and the training set come from different sequences, to avoid training our model on images too similar to the ones used for our validation. Then, to see the effect of the absence of Matrix 2 images, we made a filtered training set containing only the 2886 images belonging to Matrix 1 while all images associated with Matrix 2 were omitted from this subset. Finally, we created a normalized training set to see the potential influence of normalization techniques. Here, all 2886 images associated with Matrix 1 were retained, while the 2114 images linked to Matrix 2 underwent normalization procedures.

Table 2: Calibration Matrices of the Training and Validation Sets.

Dataset	Matrix 1	Matrix 2	Matrix 3	Total
NS3	2886	2114	-	5000
NS3 Filt	2886	-	-	2886
NS3 Norm	2886 + *2114(n)*	-	-	5000
NS3 GTA	2886	2114	2500	7500
NS3 GTA Norm	2886 + *2500(n)*	2114	-	7500
NS3 VAL	2000	-	-	2000

Figure 2: Image of the NuScenes[2] dataset using the Matrix 1.

Having established the effect of different matrices within the NuScenes dataset, we then proceeded to investigate the impact of dataset combinations on 3D monocular object detection. To achieve this, we incorporated another dataset, GTA[5], a fully synthetic dataset that is visually distinct from NuScenes while also presenting images with a different camera calibration matrix. Initially, we examined the effects of a simple combination without any further changes, by combining the base NuScenes dataset with 2500 images sourced from GTA, adhering to its native calibration matrix (Matrix 4). Then, we combined the base NuScenes dataset with a normalized version of GTA. In this scenario, 2500 images from GTA underwent normalization procedures to align with the calibration Matrix 1.

3.2 Model Configuration & Evaluation Method

For our 3D monocular object detection tests, we employed a homemade version of YOLOv7 modified to do 3D Monocular Object Detection, which we called MYv7. We used a modified method of [23] to adapt YOLOv7 from 2D to 3D monocular object detection. Previous observations we made seemed to indicate that, dataset combination yields inferior results compared to training without any form of dataset augmentation. However, these initial findings also suggested that dataset combination could reach its maximum accuracy at a higher epoch than regular model training and that this maximum was greater than the regular model's.

Table 3: Effect of different matrices within the same dataset on 2D and 3D object detection metrics on the car class.

Model	Epochs	P	R	mAP @0.5	mAP @0.95	Depth Err.	CS	DS	OS
NS3	250	0.687	0.785	0.748	**0.445**	0.0465	**0.935**	0.869	**0.953**
NS3 Filt	250	**0.699**	0.774	**0.749**	0.442	**0.0462**	0.931	**0.871**	0.949
NS3 Norm	250	0.652	**0.789**	0.737	0.435	0.0482	0.643	0.745	0.596
NS3	1000	0.68	**0.856**	0.794	**0.534**	0.0391	**0.948**	0.89	**0.978**
NS3 Filt	1000	0.736	0.828	**0.797**	0.529	**0.0379**	0.946	**0.891**	0.975
NS3 Norm	1000	**0.779**	0.8	**0.797**	0.531	0.0388	0.637	0.805	0.773
NS3	2000	0.795	0.808	0.782	**0.548**	0.0366	**0.954**	0.9	0.984
NS3 Filt	2000	**0.801**	0.804	0.782	0.545	**0.0353**	0.953	**0.901**	**0.985**
NS3 Norm	2000	0.782	**0.821**	**0.788**	**0.548**	0.0371	0.643	0.83	0.842
NS3	4000	0.843	0.776	**0.768**	**0.558**	0.036	**0.959**	0.909	**0.988**
NS3 Filt	4000	**0.857**	0.768	0.764	0.55	**0.0333**	0.958	**0.911**	0.987
NS3 Norm	4000	0.827	**0.792**	0.767	0.557	0.0352	0.65	0.85	0.88
NS3	6000	0.862	0.762	**0.756**	**0.555**	0.0351	**0.96**	0.914	**0.991**
NS3 Filt	6000	**0.889**	0.74	0.748	0.547	**0.033**	**0.96**	**0.916**	0.987
NS3 Norm	6000	0.869	**0.765**	0.755	**0.555**	0.0357	0.657	0.857	0.894
NS3	MAX	0.856	**0.766**	**0.765**	**0.557**	0.0358	0.959	0.91	0.988
NS3 Filt	MAX	**0.877**	0.748	0.747	0.547	**0.0334**	**0.96**	**0.916**	**0.987**
NS3 Norm	MAX	0.868	0.765	0.754	0.554	0.0355	0.658	0.858	0.896

To ensure a fair comparison among different models, we decided to evaluate them at their peak performance, determined by the maximum accuracy they could attain regardless of epoch. However, achieving this pinnacle necessitated extensive training durations, with model maximums usually reached after 5000-6000 epochs. Consequently, we had to reduce the dataset size and employ a smaller model variant, specifically the Tiny model. While this inherently caps the performance potential compared to larger models, even with these changes the training process still extends over three months. This means that replicating the experiment with the entire NuScenes dataset or with heavier models is not practical.

The evaluation itself is done using the usual 2D metrics (Precision, Recall, and Average Precision (AP) at IoU thresholds of 0.5 and 0.95) combined with further metrics tailored for each specific 3D estimation, these metrics are the Depth Error, the Center offset & Dimension Score defined by [24] and the Orientation Score. This has a two-fold use: firstly, this grants us insights into how each 3D estimation is affected by dataset filtering/normalization/combination, and it allows us to tailor the model learning to focus on a specific metric if needed. We assess Depth Error using metrics such as Absolute Relative Error (Abs Rel), Squared Relative Error (SRE), Root Mean Square Error (RMSE), and logarithmic RMSE (log RMSE).

4 EXPERIMENTAL RESULTS

4.1 Different Matrices within the same dataset

As we can see in Table 3, filtering the dataset increases the accuracy of both depth and dimension estimations. This improvement can be attributed to eliminating contradictions introduced by having varying focal lengths within the same dataset. Since focal length particularly affects the depth and dimensions estimations, it is logical that these two metrics are the most improved ones by the filtering. The improved depth estimations are further confirmed by the use of RMSE metrics in Table 4.

Conversely, the non-filtered regular dataset outperforms the filtered version in mAP@95 and center position & orientation estimations. The superior performance in mAP@95 can be attributed to the larger quantity of data, as additional properly labeled data invariably benefits 2D object detection tasks. Additionally, the enhanced CS and OS metrics can be attributed to the importance of data volume outweighing the impact of differing focal lengths on these specific estimations.

However, the normalized dataset demonstrates poor performance across all metrics except Recall and depth estimation. Despite removing the different focal lengths while keeping additional data, the distortions resulting from normalization and reprojections lead to

Table 4: Effect of different matrices within the same dataset on depth estimation results on the car class.

Method	Depth RMSE									
Epochs	250	500	1000	1500	2000	3000	4000	5000	6000	MAX
NS3	2.89	2.77	2.46	2.34	**2.22**	2.19	2.17	2.11	2.05	2.13
NS3 Filt	**2.81**	2.93	**2.37**	**2.31**	2.23	**2.06**	**2.01**	**1.92**	**1.96**	**2.01**
NS3 Norm	2.97	**2.68**	2.5	2.33	2.26	2.17	2.09	2.05	2.06	2.04

Table 5: Effect of dataset combination on depth estimation results on the car class.

Method	**Depth RMSE**									
Epochs	250	500	1000	1500	2000	3000	4000	5000	6000	MAX
NS3	2.89	2.77	2.46	2.34	**2.22**	2.19	2.17	2.11	**2.05**	2.13
NS3 GTA	**2.88**	**2.71**	**2.41**	**2.28**	2.26	**2.16**	**2.04**	**2.06**	2.09	**2.06**
NS3 GTA Norm	2.97	**2.71**	2.58	2.48	2.4	2.3	2.27	2.32	2.3	2.27

a significant decrease in prediction accuracy across all metrics. The decreased depth error seen in Table 4 does show that getting rid of the different focal lengths does help the depth estimation. However, the loss in image quality due to normalization cannot be compensated for by removing contradictory information.

The qualitative results given by Figure 3 further confirm these observations: Compared to the left image (NS3 model), the center image (NS3 Filt model) has more accurate bounding box sizes and positions, however, the predictions made by the base model on the left have more accurate orientation estimation.

4.2 Dataset Combination

Based on the results in Table 6, incorporating an additional dataset with its own focal length and matrix did not consistently lead to a complete decrease in 3D metrics. This unexpected outcome suggests that the accuracy of the synthetic 3D position data, despite introducing contradictory focal length information, may partially compensate for such discrepancies. Utilizing a dataset with precise ground truth compared to the image may allow the accuracy of the additional data to offset the negative impact of conflicting focal lengths and matrices. As we can see in Table 5, the depth estimations in the combined models were often better than the ones made by the regular model.

Further deviating from expectations, introducing the new synthetic dataset appears to have had a detrimental effect on 2D metrics, contrary to the usual anticipation of higher results with additional information. This unexpected observation is likely attributed to the visual appearance of the synthetic dataset. Despite its photorealism, it remains too far from real-life images to significantly enhance 2D object detection performance.

Moreover, the normalization process once again results in excessive distortion, hindering the attainment of satisfactory results. While applied to synthetic data, normalization yielded improved mAP@0.5 results, the substantial decrease in other 3D metrics outweighs this improvement. Consequently, the overall impact of normalization on synthetic data appears unfavorable, underscoring the importance of considering the trade-offs between data preprocessing techniques and resultant performance metrics.

Another result of the dataset combination that we can see from the qualitative evaluation in Figure 3 is the difference in labeling between datasets: in the GTA dataset, even cut-off objects are labeled, while they are not in the NuScenes dataset. This means that once these datasets are combined, our NS3 GTA model can detect cut-off cars using information from the GTA dataset, while these are not part of the NuScenes label, which means that the quantitative evaluation considers these detections as erroneous. This effectively means that even if the labels themselves are correct, both datasets must have similar criteria for object labelization to avoid contradicting each other.

Table 6: Effect of dataset combination on 2D and 3D object detection metrics on the car class.

Model	**Epochs**	**P**	**R**	**mAP @0.5**	**mAP @0.95**	**Depth Err.**	**CS**	**DS**	**OS**
NS3	250	**0.687**	0.785	0.748	0.445	**0.0465**	0.935	0.869	**0.953**
NS3 GTA	250	0.679	**0.798**	**0.749**	**0.447**	0.0472	**0.937**	**0.87**	0.95
NS3 GTA N	250	0.628	0.78	0.721	0.421	0.0491	0.643	0.745	0.609
NS3	1000	0.68	**0.856**	**0.794**	**0.534**	0.0391	0.948	0.89	**0.978**
NS3 GTA	1000	0.7	0.85	**0.794**	0.532	**0.0386**	**0.949**	**0.891**	0.976
NS3 GTA N	1000	**0.744**	0.817	**0.794**	0.521	0.0427	0.622	0.79	0.763
NS3	2000	**0.795**	0.808	0.782	**0.548**	0.0366	**0.954**	**0.9**	0.984
NS3 GTA	2000	0.772	0.821	0.785	**0.548**	**0.0363**	**0.954**	**0.9**	**0.985**
NS3 GTA N	2000	0.757	**0.824**	**0.796**	0.544	0.0405	0.633	0.815	0.825
NS3	4000	0.843	0.776	0.768	**0.558**	0.036	**0.959**	**0.909**	**0.988**
NS3 GTA	4000	**0.861**	0.762	0.764	0.55	**0.0347**	**0.959**	0.907	0.987
NS3 GTA N	4000	0.821	**0.788**	**0.783**	0.556	0.0393	0.643	0.836	0.867
NS3	6000	0.862	0.762	0.756	0.555	0.0351	0.96	0.914	0.991
NS3 GTA	6000	0.865	0.766	0.755	0.549	0.035	0.959	0.912	0.989
NS3 GTA N	6000	0.835	0.778	0.774	0.556	0.0395	0.65	0.846	0.879
NS3	MAX	0.856	0.766	0.765	**0.557**	0.0358	0.959	0.91	0.988
NS3 GTA	MAX	**0.873**	0.756	0.755	0.55	**0.0351**	**0.96**	**0.912**	**0.989**
NS3 GTA N	MAX	0.839	**0.771**	**0.773**	0.551	0.0392	0.654	0.845	0.886

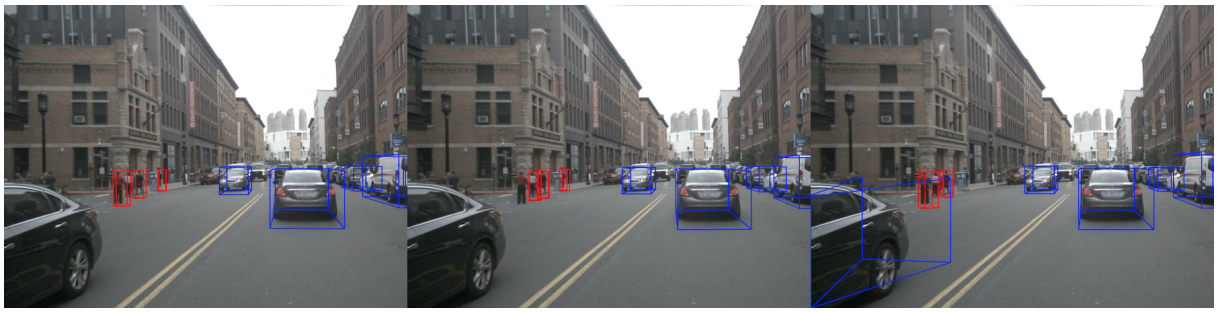

Figure 3: Comparison of 3D predictions on the Nuscenes[2] dataset. Left: NS3 model, Center: NS3 Filt model, Right: NS3 GTA model.

5 DISCUSSION

In this study, we explored the complexities surrounding the presence of several camera calibration matrices within the same dataset, as well as dataset combination and its impact on 3D monocular object detection.

Firstly, our analysis focused on the comparative performance of filtered and non-filtered datasets within the context of 3D monocular object detection. Filtering the dataset, which aimed to remove contradictions introduced by different focal lengths within the same dataset, yielded positive effects on depth error and dimension estimations as expected. However, our results show that the non-filtered regular dataset often outperformed its filtered counterpart in metrics such as mAP@95 and center position & orientation estimations. This discrepancy can be attributed to the larger quantity of data within the non-filtered dataset, which benefits 2D object detection by providing additional properly labeled data.

While dataset combination offers a promising strategy for addressing the limitations inherent in individual datasets, its efficacy varies depending on several factors. Despite anticipating improved 2D detection results with the addition of new synthetic data, we observed a negative impact on 2D metrics. This outcome suggests that although synthetic datasets offer perfect annotations, their visual dissimilarity from real-world images can compromise their utility in practical applications. Using synthetic datasets in combination with real ones does not always result in increased performance, and careful consideration must be given to dataset composition to ensure alignment with the objectives of the object detection task.

Another finding is the overwhelmingly negative effect of normalization of the camera matrix through artificial distortion, both on real and synthetic datasets, whether it affected the core training data or additional data. While normalization may improve certain metrics, such as mAP@0.5, not only are these improvements inconsistent, but its adverse effects on other 3D metrics did outweigh the benefits, and models trained on normalized datasets performed poorly in most cases.

On a more positive note, incorporating additional datasets with their own focal lengths and matrices did not consistently result in a complete decrease in 3D metrics. It seems that datasets with precise ground truth compared to the image allowed the accuracy of the additional data to partially compensate for the introduction of contradictory information induced by different focal lengths.

6 CONCLUSION

We can conclude that having different camera focal lengths within a training set does not inherently decrease the performance of a 3D monocular object detection model. While filtering the dataset results in more accurate depth and dimension estimations, it is at the expense of other results as filtering gets rid of useful data. Introducing additional data from other datasets does not necessarily reduce the accuracy of the model's 3D estimations as long as this data contains precise ground truth, the visual appearance of this new data matters a lot for 2D object detection. Finally, attempting to normalize the focal length through artificial distortion just provides unreliable data for 3D estimations.

7 ACKNOWLEDGMENTS

This research is funded and supported by SEGULA Technologies. We would like to thank SEGULA Technologies for their collaboration and for allowing us to conduct this research. We would like to thank also the engineers of the Autonomous Navigation Laboratory (ANL) of IRSEEM for their support. In addition, this work was performed, in part, on computing resources provided by CRIANN (Centre Regional Informatique et d'Applications Numeriques de Normandie, Normandy, France).

8 REFERENCES

[1] A. Geiger, P. Lenz, *et al.*, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361, IEEE, 2012.

[2] H. Caesar, V. Bankiti, *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] M. Cordts, M. Omran, *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[4] M. Everingham, L. Van Gool, *et al.*, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[5] S. R. Richter, V. Vineet, *et al.*, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 102–118, Springer, 2016.

[6] F. Yang, H. Chen, *et al.*, "Geometry-guided domain generalization for monocular 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 6467–6476, 2024.

[7] M. Furst, R. Jakkamsetty, R. Schuster, and D. Stricker, "Learned fusion: 3d object detection using calibration-free transformer feature fusion," 2023.

[8] M. Zhu, L. Ge, *et al.*, "Monoedge: Monocular 3d object detection using local perspectives," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 643–652, 2023.

[9] X. Shi, Q. Ye, *et al.*, "Geometry-based distance decomposition for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15172–15181, October 2021.

[10] C. Park, H. Kim, J. Jang, and J. Paik, "Odd-m3d: Object-wise dense depth estimation for monocular 3d object detection," *IEEE Transactions on Consumer Electronics*, 2024.

[11] J. Jinrang, Z. Li, and Y. Shi, "Monouni: A unified vehicle and infrastructure-side monocular 3d object detection network with sufficient depth clues," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[13] Y. Zhang, J. Lu, and J. Zhou, "Objects are different: Flexible monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3289–3298, June 2021.

[14] T. Wang, J. Pang, and D. Lin, "Monocular 3d object detection with depth from motion," in *European Conference on Computer Vision*, pp. 386–403, Springer, 2022.

[15] Y. Liao, J. Xie, *et al.*, "Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3292–3310, 2022.

[16] A. Gaidon, Q. Wang, *et al.*, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4340–4349, 2016.

[17] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," *arXiv preprint arXiv:2001.10773*, 2020.

[18] X. Ye, M. Shu, *et al.*, "Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21341–21350, 2022.

[19] Q.-H. Pham, P. Sevestre, *et al.*, "A*3d dataset: Towards autonomous driving in challenging environments," in *2020 IEEE International conference on Robotics and Automation (ICRA)*, pp. 2267–2273, IEEE, 2020.

[20] A. Patil, S. Malla, *et al.*, "The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9552–9557, IEEE, 2019.

[21] N. Gahlert, N. Jourdan, *et al.*, "Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection," *arXiv preprint arXiv:2006.07864*, 2020.

[22] Y. Xiang, R. Mottaghi, *et al.*, "Beyond pascal: A benchmark for 3d object detection in the wild," in *IEEE winter conference on applications of computer vision*, pp. 75–82, IEEE, 2014.

[23] A. Mauri, R. Khemmar, *et al.*, "Lightweight convolutional neural network for real-time 3d object detection in road and railway environments," *Journal of Real-Time Image Processing*, vol. 19, pp. 499–516, Jun 2022.

[24] H.-N. Hu, Q.-Z. Cai, *et al.*, "Joint monocular 3d vehicle detection and tracking," 2019.