

UDC 681.7.068
EDN: OWGQXF

PACS: 42.30.Va, 42.79.Bh, 07.05.Mh

Determination of the optimal volume of synthetic data for training object detection neural networks (using YOLOv8 as an example)

A. A. Berillo

RD&P Center ORION, JSC, Moscow, 111538 Russia
E-mail: berillo.andrey@yandex.ru

Received 25.06.2026; revised 9.07.2026; accepted 06.08.2026

The problem of determining the minimum sufficient volume of synthetic data for training YOLO-format object detection neural networks is considered. An approach to generating labeled synthetic images using the developed Yolo Synthetic Data Generator software module is proposed. Three variants of the YOLOv8 model (nano, small, medium) were trained on datasets of various sizes (from 100 to 10,000 frames). Quality assessment was performed on five test videos based on the number of detected objects and frames with detections. It is shown that for the nano and small models, a dataset of ~1,000 frames is optimal, while for the medium model, ~500 frames are sufficient. A further increase in the dataset does not lead to a significant improvement in quality but increases computational costs. The obtained results can be used in the development of machine vision systems, including for ROI-based autofocusing.

Keywords: synthetic data, YOLO, object detection, neural network training, dataset size, Yolo Synthetic Data Generator, autofocus.

DOI: 10.51368/2949-561X-2026-4-83-86

For citation: Berillo A.A., Applied Physics, No. 4, 83–86 (2026).

Introduction

Modern computer vision systems have been using deep neural networks for object detection with increasing frequency. One of the key factors determining the accuracy and efficiency of such systems is the availability of an annotated training dataset. However, manual image annotation is labor-intensive and costly, especially for specialized tasks where objects are rare or difficult to access. An alternative approach is to use synthetic data generated by software tools with automatic annotation [1].

Previously, the author developed a software Yolo Synthetic Data Generator, which allows for the creation of annotated synthetic images using three-dimensional models of objects on static backgrounds, incorporated controlled randomization of position, rotation, exposure, and gamma parameters. The program is registered in

the Register of Computer Programs (certificate No. 2026667324). The previous paper [2] examined the use of neural network detection for extracting a region of interest (ROI) during the autofocusing of an optoelectronic system. The goal of this paper is to determine the optimal volume of synthetic data required to train a detection model for a given object (a sponge) on various backgrounds, aiming to ensure high accuracy with minimal computational costs.

1. Synthetic Data Generation Method

The *Yolo Synthetic Data Generator* program is implemented in C++14 using the Vulkan graphics API and the tinyGLTF, GLM, and ImGui libraries. It allows for loading 3D models in glTF/GLB formats, overlaying them onto background images (JPG) while preserving perspective and lighting based on Physically

Based Rendering (PBR), and automatically generating annotations in the YOLO format comprising normalized coordinates of bounding boxes.

Randomization of the following parameters is provided during dataset generation:

- object position within the visible area of the scene;
- rotation around three axes (range 360°);
- exposure (0.1–10.0) and gamma correction (0.1–4.0).

Seven datasets were prepared for the training purposes. They consisted of 75 different background images, for each of which 100, 500, 1000, 1500, 3000, 5000, or 10000 frames were generated, respectively. Thus, the total volume of each dataset amounted to 7500, 37500, 75000, 112500, 225000, 375000, and 750000 images. In all cases, the same 3D model of the object was used, specifically a sponge (sized 160×55 mm).

Generation was performed at a resolution of 1280×1280 pixels, subsequently followed by conversion to the format adopted for YOLOv8 training (640×640). Annotations were saved in text files containing normalized coordinates.

2. Experimental Procedure

Three YOLOv8 architecture variants were chosen for training: **nano** (the smallest, 3.2 million parameters), **small** (11.2 million), and **medium** (25.9 million). Each model was trained on each of the seven datasets with the same hyperparameters: 100 epochs, a batch size of 16, the SGD optimizer with a momentum of 0.937, and an initial learning rate of 0.01 with cosine decay. Training was performed on an NVIDIA GeForce RTX 4080 GPU (16 GB).

Quality evaluation was carried out for five test videos with a duration of approximately ~2883 frames each (about 14415 frames in total).

The videos contained the same sponge model placed under various lighting conditions and angles, which differed from the training backgrounds. The following metrics were calculated for each trained model on each video:

- total number of detected objects (total_detections);
- number of frames with at least one detection (det_frames);
- proportion of frames with detections relative to the total number of frames;
- processing time per video (time_sec).

A confidence threshold of 0.7 was used, and the IoU threshold for non-maximum suppression was set to 0.35. All measurements were performed on the same computing device.

3. Findings and Discussion

The table shows the average metrics for each model and each dataset size, averaged over the five videos. For clarity, data for video No. 4, which recorded the maximum detection values, are also presented.

Analysis of the table shows that the maximum number of detections for the **nano** and **small** models is achieved with a dataset size of **1000 frames**. With a further increase in data volume, a decrease in detections is observed, which may be attributed to overfitting on synthetic data or insufficient background variability. Regarding the **medium** model, the peak also occurs at 1000 frames; however, the values at 500 frames are close to the maximum, and at 3000 frames a secondary rise is observed (possibly due to random factors). Considering that the medium model requires more computational resources, a dataset of 500 frames can be considered optimal for it, as it provides good quality with lower training costs.

Table

Average values for five videos (with values for video No. 4 in brackets)

Dataset size (number of frames)	Model			Time, s
100	nano	442 (596)	442 (596)	48.6
	small	485 (653)	485 (653)	59.1
	medium	602 (771)	602 (771)	67.5
500	nano	476 (641)	476 (641)	48.3
	small	549 (763)	549 (763)	59.0
	medium	695 (866)	695 (866)	67.4

End of table

1000	nano	532 (760)	529 (759)	48.0
	small	561 (793)	560 (793)	58.9
	medium	618 (776)	617 (776)	67.1
1500	nano	511 (644)	506 (643)	47.9
	small	533 (658)	531 (658)	58.8
	medium	626 (744)	624 (744)	67.2
3000	nano	399 (480)	395 (480)	48.1
	small	469 (570)	467 (570)	59.0
	medium	620 (790)	617 (790)	67.2
5000	nano	306 (281)	304 (281)	48.0
	small	395 (460)	393 (460)	58.9
	medium	532 (621)	530 (621)	67.3
10000	nano	358 (441)	355 (441)	48.1
	small	448 (547)	446 (546)	59.2
	medium	550 (584)	548 (584)	66.9

The video processing time (inference) is practically independent of the dataset size and is determined only by the model architecture: nano – ~48 s, small – ~59 s, medium – ~67 s. This implies that a model should be chosen based not only on quality but also on performance requirements. For example, the nano model

provides acceptable quality (760 detections on video No. 4) with minimal time consumption, making it preferable for real-time systems, such as ROI-based autofocusing.

The figure shows the average number of detections vs. the dataset size for each model.

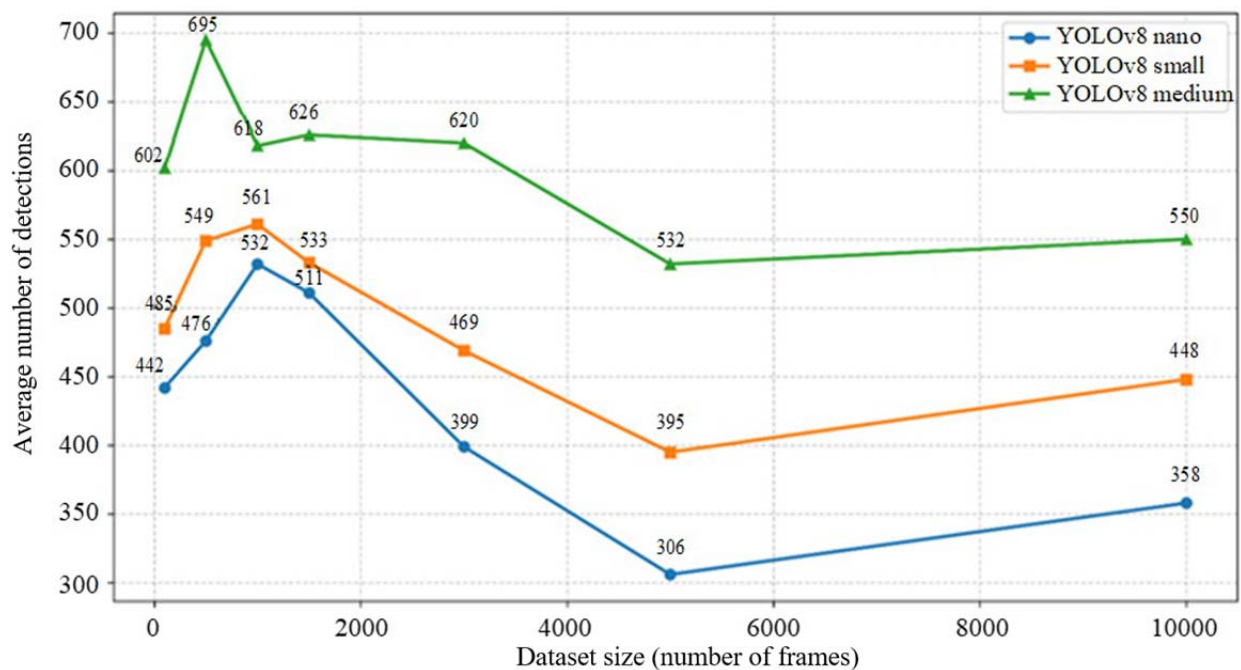


Fig. Average number of detections vs. the dataset size (averages on five videos)

The figure shows that the curves for all models have a similar pattern: rapid growth in the initial segment (up to 500–1000 frames) followed by saturation or even a decline. This

confirms the hypothesis that there is an optimal data volume, beyond which the model's generalization ability does not improve and may even deteriorate.

Discussion

The obtained results are consistent with prior studies on the use of synthetic data [3, 4], which also note that model quality ceases to grow after a certain threshold, and a further increase in the dataset may lead to overfitting on synthesis artifacts. In our case, the decline in quality at large datasets (3000–10000 frames) is likely due to the model starting to memorize specific features of backgrounds and angles that are not present in the test videos, thereby reducing its generalization ability.

It should be noted that the synthetic data were generated with limited diversity (75 backgrounds, one object). It is possible that increasing the number of backgrounds or object variations would increase the optimal data volume. Nevertheless, the obtained recommendations are justified for practical tasks similar to ours (specifically, detection of a single object on a limited set of backgrounds).

The application of the results of this paper in autofocusing systems [2] makes it possible to reduce manual annotation time and accelerate the development process. Using a synthetic dataset of 1000 frames and the YOLOv8 nano model allows for achieving high accuracy in object detection in real time, thus ensuring correct ROI extraction for focus adjustment.

Conclusion

This paper investigated the impact of synthetic data volume on the quality of training neural networks for object detection. Experiments with the YOLOv8 nano, small, and medium models using datasets of various sizes established that:

- the optimal dataset size for the nano and small models is ~ 1000 frames (with 75 backgrounds), which provides the maximum number of detections on test videos;
- the medium model shows close results already at 500 frames; however, its use is justified only when higher accuracy is required and sufficient computational resources are available;
- a further increase in data volume (above 1000–1500 frames) does not lead to quality improvement, and in some cases deteriorates it, which is a symptom of overfitting.

The practical significance of the paper lies in providing recommendations for selecting the synthetic dataset volume for single-object detection tasks on static backgrounds. In combination with the developed synthetic data generator, this allows for a significant reduction in the labor costs for the preparation of annotated samples and accelerates the development of machine vision systems, including ROI-based autofocusing systems.

Further research is aimed at expanding the variability of synthetic data (increasing the number of backgrounds, adding various objects, changing lighting conditions) and studying the influence of these factors on the optimal dataset volume.

REFERENCES

1. Nikolenko S. I., Kadurin A. A., and Arkhangelskaya E. O. Deep Learning. St. Petersburg: Piter Publ., 2018 [in Russian].
2. Berillo A. A., Applied Physics, No. 2, 82–85 (2026) [in Russian].
3. Tremblay J., Prakash A., Acuna D., et al., Training deep networks with synthetic data: Bridging the reality gap by domain randomization, arXiv:1804.06516, 2018.
4. Dwibedi D., Misra I., and Hebert M., Cut, paste and learn: Surprisingly easy synthesis for instance detection ICCV, 1301–1310 (2017).
5. Redmon J. and Farhadi A., YOLOv3: An incremental improvement, arXiv:1804.02767, 2018.

About the author

Berillo A. A., Lead Software Engineer, M.Sc., RD&P Center ORION, JSC (Moscow, 111538 Russia). E-mail: berillo.andrey@yandex.ru