Image Augmentation for Object Detection and Segmentation with Diffusion Models

Leon Useinov[a], Valeria Efimova[b] and Sergey Muravyov[c]
ITMO University, Russia

Keywords: Augmentation, Image Generation, Diffusion Models, Object Detection, Segmentation.

Abstract: Training current state-of-the-art models for object detection and segmentation requires a lot of labeled data, which can be difficult to obtain. It is especially hard, when occurrence of an object of interest in a certain required environment is rare. To solve this problem we present a train-free augmentation technique that is based on a diffusion model, pretrained on a large dataset (more than 1 million images). In order to establish the effectiveness of our method and its modifications, experiments on small datasets (less than 500 training images) with YOLOv8 are conducted. We conclude that none of the proposed versions of the diffusion-based augmentation method are universal, however, each of them may be used to improve an object detection (and segmentation) model performance in certain scenarios. The code is publicly available: github.com/PnthrLeo/diffusion-augmentation.

1 INTRODUCTION

When the time comes to implement state-of-the-art data-driven machine learning solutions for new object detection or segmentation applications (e.g., bottle defects detection (Bergmann et al., 2019), skin cancer detection (Dildar et al., 2021)), it is often hard to get enough labeled data for training. The same thing usually happens when the implemented system needs to be adapted to a novel domain (e.g., a new type of bottle, another skin color). To solve these problems several synthetic data generation approaches may be applied.

Model-free image augmentations (e.g., image translation, rotation, hue shift) can be utilized as a computationally lightweight fully-automated approach. Nevertheless, such techniques cannot be considered as a comprehensive data extension because of either limited background variations or the photo-realism lack.

Thus, image generation via 3D rendering can be used to obtain photo-realistic data with a high variety of objects' positions (Wang et al., 2019; Wood et al., 2021). Moreover, bounding boxes and segmentation masks can be automatically retrieved, since information about objects' and camera's positions is known. However, generating more diverse images requires creating or gathering more 3D models, textures, shaders and 3D environments, which may be exhausting.

Significant progress has been achieved in image synthesis via generative models (Goodfellow et al., 2020; Rombach et al., 2022; Podell et al., 2023). In comparison to 3D rendering, these models do not require creating or gathering assets to generate photo-realistic images. Thereby, it is a natural thought to leverage them for data augmentation.

Most of existing model-based methods for image augmentation are trained only on target datasets (datasets for subsequent augmentation) (Xu et al., 2023; Yang et al., 2022) without exploitation of existing large image datasets (more than 1 million images, for example, LAION-5B (Schuhmann et al., 2022)), therefore, lack creativity. Furthermore, most of them are designed for classification purposes.

The rest of the model-based methods either require additional training (Zhang et al., 2023b; Zhang et al., 2023c) or they are focused on augmentation of big datasets (more than 100000 images) (Xie et al., 2023; Zhao et al., 2023) and target datasets consisting of mainstream object classes (e.g., sofa, train, cat) (Ge et al., 2022).

Therefore, we propose a new train-free model-based augmentation approach for object detection and

[a] https://orcid.org/0009-0002-5648-4027
[b] https://orcid.org/0000-0002-5309-2207
[c] https://orcid.org/0000-0002-4251-1744

Useinov, L., Efimova, V. and Muravyov, S.
Image Augmentation for Object Detection and Segmentation with Diffusion Models.
DOI: 10.5220/0012474500003660
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 812-820
ISBN: 978-989-758-679-8; ISSN: 2184-4321

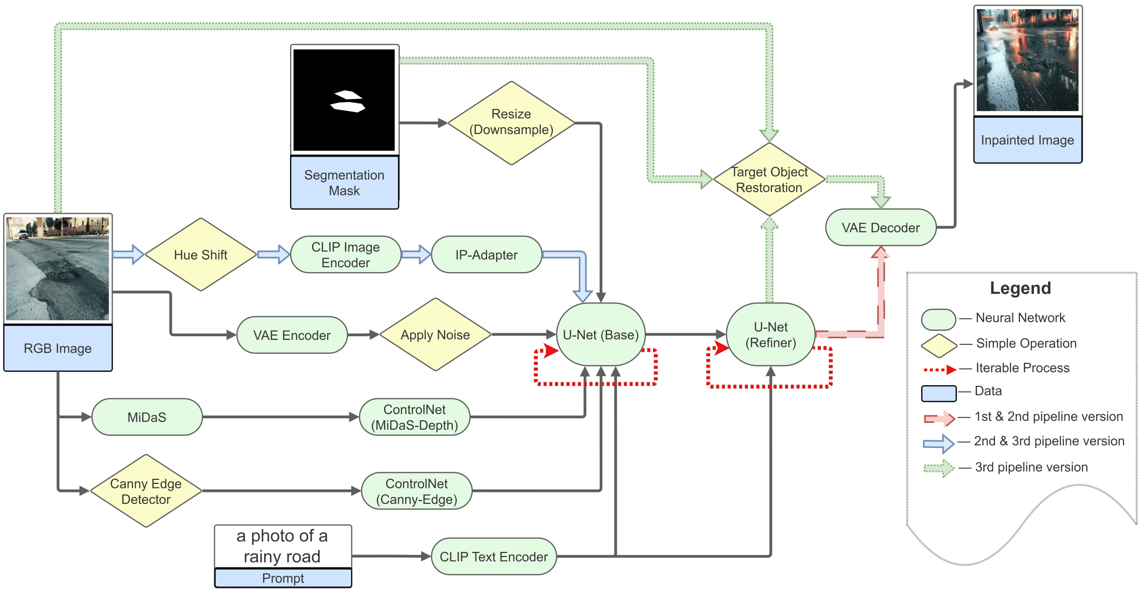

Figure 1: Illustration of the proposed image augmentation method for object detection and segmentation tasks.

segmentation tasks, which is aimed to solve data lack problem on small datasets (less than 500 images) with non-mainstream object classes (e.g., bottle defects, printed circuit board defects, road potholes). We believe that our method can be inspirational for future research in the model-based augmentation field.

2 RELATED WORKS

2.1 Model-Free Image Augmentation

Model-free augmentations include: geometrical transformations (e.g., translation, rotation, flip), color image transformations (e.g., hue shift, brightness shift), image blurring, image masking (e.g., Random Erasing (Zhong et al., 2020), Grid Mask (Chen et al., 2020)), image mixing (e.g., PuzzleMix (Kim et al., 2020), GridMix (Baek et al., 2021), Simple CutPas (Ghiasi et al., 2021), Continuous CutPas (Xu et al., 2021)). In order to perform these augmentations, no data-driven model is required, resulting in low computational cost. In addition, consequences of using these methods are predictable. In other words, if a trained model should have an additional property (e.g., be robust for input image mirroring or shifts) then, possibly, one of the augmentation methods can be used in order to obtain it (e.g., horizontal or vertical flipping). Moreover, policy-based algorithms (e.g., Faster AA (Hataya et al., 2020), RandAugment (Cubuk et al., 2020), Adversarial AA (Zhang et al., 2019), SPA (Takase et al., 2021)) can be employed to automatically find optimal data-level, class-level or instance-level combinations of model-free augmentation methods with corresponding hyperparameters.

Most model-free augmentation methods can be applied directly to object detection and segmentation tasks. However, these methods are either limited in background variations or lack photorealism.

2.2 3D Rendering

An alternative approach to increase the amount of training data is 3D rendering (e.g., CAMERA25 dataset (Wang et al., 2019), Face Synthetics dataset (Wood et al., 2021), and others (Gaidon et al., 2016; Džijan et al., 2023; Rajpal et al., 2023)). Due to progress in computer graphics research, it is possible to render photorealistic images that may be exploited as a real dataset replacement (Wood et al., 2021). In addition, since spatial information for all objects and a virtual camera is known, it is possible to automatically generate object detection and segmentation labels. However, rendering is a computationally demanding process and may require a large amount of manual work to obtain enough 3D models, textures, shaders and 3D environments.

2.3 Model-Based Image Augmentation with GANs

Having been introduced in 2014, GANs (Generative Adversarial Networks) (Goodfellow et al., 2020) were shown as a relatively good framework for image generation, thus, initiating the line of GAN-based image augmentation methods (Xu et al., 2023). Most of

Table 1: Object detection and segmentation results with YOLOv8n on MVTec AD Bottle dataset. ”B” — boxes, ”M”— masks, ”inp” — our inpainting method, ”std aug” — default YOLOv8 augmentations, v2 and v3 — the second and the third versions of our framework respectively.

Dataset	Precision		Recall		mAP50		mAP50-95	
	B	M	B	M	B	M	B	M
w/o inp, w/o std aug	0.813±0.035	0.791±0.035	0.535±0.024	0.576±0.007	0.663±0.009	0.695±0.009	0.434±0.004	0.463±0.005
with inp, w/o std aug	0.700±0.072	0.698±0.072	0.570±0.046	0.605±0.045	0.644±0.016	0.659±0.020	0.391±0.009	0.415±0.019
with inp, w/o std aug v2	0.742±0.024	0.759±0.024	0.564±0.014	0.573±0.021	0.676±0.023	0.682±0.025	0.437±0.019	0.458±0.019
with inp, w/o std aug v3	0.758±0.038	0.786±0.015	0.541±0.023	0.548±0.026	0.663±0.031	0.684±0.031	0.421±0.030	0.453±0.029
w/o inp, with std aug	0.849±0.022	0.849±0.022	0.773±0.019	0.773±0.019	**0.839±0.017**	**0.843±0.017**	0.671±0.014	0.609±0.010
with inp, with std aug	0.858±0.030	**0.870±0.022**	0.741±0.014	0.744±0.010	0.822±0.016	0.819±0.011	0.648±0.011	0.577±0.008
with inp, with std aug v2	0.847±0.016	0.849±0.014	**0.776±0.015**	**0.776±0.011**	0.835±0.010	0.837±0.004	0.664±0.012	0.603±0.008
with inp, with std aug v3	**0.861±0.011**	0.862±0.013	0.760±0.014	0.768±0.008	0.836±0.009	0.835±0.004	**0.672±0.011**	**0.612±0.006**

these methods were designed for classification purposes (e.g., DAGAN (Antoniou et al., 2017), IDA-GAN (Yang and Zhou, 2021), StyleAug (Jackson et al., 2019), Shape bias (Geirhos et al., 2018), GAN-MBD (Zheng et al., 2021), StyleMix (Hong et al., 2021)). Although, there are existing successful adaptations for object detection and segmentation tasks (e.g., CycleGAN (Sandfort et al., 2019), SCIT (Xu et al., 2022), MGD-GAN (Efimova et al., 2020)).

Despite being more computationally expensive than model-free augmentations and additional model training requirement, GAN-based methods provide an opportunity to generate more photorealistic samples. Furthermore, these methods have no need in gathering any additional assets (as in case of the 3D rendering approach) and may work even if only original training data is available. However, GAN-based augmentation methods do not utilize large image datasets for training and, consequently, lack creativity.

2.4 Model-Based Image Augmentation with Diffusion Models

Diffusion models are a comparatively new trend in image generation (Croitoru et al., 2023). They have gained huge popularity since 2021 with the release of Stable Diffusion (Rombach et al., 2022). The popularity is explained by more stable training, high generation variety, and similar photorealism in comparison to GANs. In consequence, diffusion models have become an object of interest in terms of image augmentation.

As with GANs, it is obvious idea to use diffusion-based augmentation methods for classification purposes (Trabucco et al., 2023; Burg et al., 2023). However, attempts to apply diffusion models for object detection and segmentation data augmentation also exist. Is some of them a diffusion model is trained only on a target dataset and, therefore, limited in creativity (e.g., DBDA-NIS (Yu et al., 2023)). Other methods require fine-tuning of a pretrained diffusion model (EMIT-Diff (Zhang et al., 2023c), Diffusion Engine (Zhang et al., 2023b)), which is time and computationally consuming. The rest are aimed at augmentation of big datasets or datasets with common objects (Xie et al., 2023; Zhao et al., 2023; Ge et al., 2022).

To alleviate aforementioned issues, our work is targeted on adoption of pre-trained on large datasets diffusion models, without additional training requirement, for augmentation of small datasets composed of non-mainstream object classes for object detection and segmentation problems.

3 METHOD

The idea of our augmentation framework (Fig 1) is a replacement of real image backgrounds with ones generated by a diffusion model. Therefore, RePaint (Lugmayr et al., 2022) inpainting method is employed as the core of our approach. Stable Diffusion XL (Podell et al., 2023) is used as a state-of-the-art denoising diffusion probabilistic model, which is re-

Table 2: Object detection and segmentation results with YOLOv8n on PCB Defects dataset. ”B” — boxes, ”M”—masks, ”inp” — our inpainting method, ”std aug” — default YOLOv8 augmentations, v2 and v3 — the second and the third versions of our framework respectively.

Dataset	Precision		Recall		mAP50		mAP50-95	
	B	M	B	M	B	M	B	M
w/o inp, w/o std aug	0.527±0.060	0.516±0.053	0.422±0.023	0.412±0.022	0.495±0.021	0.482±0.016	0.393±0.016	0.351±0.011
with inp, w/o std aug	0.537±0.061	0.487±0.082	0.364±0.024	0.389±0.035	0.456±0.008	0.451±0.013	0.330±0.011	0.285±0.006
with inp, w/o std aug v2	0.488±0.075	0.503±0.060	0.391±0.029	0.367±0.015	0.435±0.003	0.433±0.006	0.318±0.001	0.273±0.008
with inp, w/o std aug v3	0.484±0.076	0.473±0.074	0.412±0.019	0.404±0.021	0.456±0.016	0.447±0.017	0.345±0.016	0.303±0.012
w/o inp, with std aug	**0.705**±**0.025**	0.698±0.026	0.578±0.013	0.579±0.020	0.631±0.014	0.629±0.015	**0.511**±**0.011**	**0.457**±**0.009**
with inp, with std aug	0.697±0.027	0.694±0.025	**0.655**±**0.052**	**0.647**±**0.049**	**0.656**±**0.025**	**0.649**±**0.024**	0.503±0.017	0.429±0.011
with inp, with std aug v2	0.671±0.033	0.669±0.031	0.608±0.039	0.606±0.039	0.606±0.013	0.601±0.015	0.475±0.010	0.409±0.009
with inp, with std aug v3	0.693±0.028	**0.705**±**0.031**	0.608±0.033	0.616±0.033	0.620±0.014	0.625±0.014	0.488±0.010	0.423±0.009

quired to run RePaint. The diffusion model provides following modules:

- Variational AutoEncoder (VAE) (Kingma and Welling, 2013) to map a RGB image in and out of a reduced latent space;
- U-Net (Base) (Ronneberger et al., 2015) to perform reverse diffusion process in the latent space;
- U-Net (Refiner) to add more fine details to get more photorealistic image;
- CLIP Text Encoder (Radford et al., 2021) to encode input text condition.

In the first version of our framework, in order to generate more consistent backgrounds, ControlNets (Zhang et al., 2023a) with MiDaS depth (Ranftl et al., 2020) and canny edge image (Canny, 1986) conditions are used. Each ControlNet represents an additional neural network module for Stable Diffusion XL trained for a certain image condition. Since the modules are independent, they can be applied simultaneously, weighted by corresponding coefficients, to control background content. The conditioning images are obtained by passing the RGB image through corresponding preprocessors: MiDaS, Canny edge detector.

In the second version of our framework, IP-Adapter (Ye et al., 2023) is employed to generate images that are closer to the given RGB Image. The idea is that IP-Adapter implicitly provides image information such as style, color, and textures, which can be helpful in creating more realistic backgrounds for target objects. As an image preprocessor hue shift is used to change hue value in HSL (hue, saturation, lightness) representation of the RGB image for a more diverse color palette of generated images, since usage of IP-adapter with ControlNets strongly decreases variation. The resulting image is passed through CLIP Image Encoder to generate image features for IP-Adapter.

On top of all, in the third version of our framework, a target object restoration algorithm is added to mitigate the effect of distorted target (segmented) objects after latent image decoding. The algorithm simply replaces the objects on inpainted images with their corresponding original variants, additionally blending their edges from both (original and inpainted) versions for a more realistic look.

4 EXPERIMENTS

4.1 Datasets

MVTec AD Bottle dataset (Fig 2) is a subset of MVTec dataset (Bergmann et al., 2019) that consists of 209 images for training and 83 for testing within 3 categories of defects: broken small, broken large, contamination. The training set includes only images without defects. The test set includes images with and without defects.

The original use of the dataset was supposed to be based on generative (feature extraction) models which

Table 3: Object detection and segmentation results with YOLOv8n on Potholes dataset. ”B” — boxes, ”M” — masks, ”inp” — our inpainting method, ”std aug” — default YOLOv8 augmentations, v2 and v3 — the second and the third versions of our framework respectively.

Dataset	Precision		Recall		mAP50		mAP50-95	
	B	M	B	M	B	M	B	M
w/o inp, w/o std aug	0.544±0.064	0.570±0.087	0.425±0.029	0.418±0.033	0.497±0.012	0.495±0.013	0.286±0.008	0.253±0.009
with inp, w/o std aug	0.559±0.040	0.560±0.036	0.399±0.009	0.398±0.014	0.479±0.011	0.480±0.009	0.254±0.003	0.229±0.004
with inp, w/o std aug v2	0.524±0.027	0.521±0.034	0.440±0.014	0.433±0.012	0.487±0.014	0.484±0.009	0.249±0.009	0.230±0.009
with inp, w/o std aug v3	0.535±0.047	0.555±0.047	0.449±0.042	0.432±0.043	0.500±0.015	0.491±0.012	0.272±0.006	0.246±0.006
w/o inp, with std aug	0.647±0.020	**0.674**±**0.012**	**0.572**±**0.010**	**0.556**±**0.014**	0.594±0.006	**0.600**±**0.011**	0.304±0.004	0.282±0.004
with inp, with std aug	0.660±0.014	0.654±0.016	0.554±0.009	0.555±0.007	**0.608**±**0.005**	0.592±0.005	0.319±0.004	0.283±0.002
with inp, with std aug v2	0.662±0.037	0.668±0.031	0.510±0.034	0.506±0.038	0.563±0.055	0.550±0.052	0.301±0.028	0.271±0.024
with inp, with std aug v3	**0.666**±**0.019**	0.666±0.023	0.552±0.015	0.548±0.013	0.607±0.002	0.595±0.006	**0.330**±**0.003**	**0.294**±**0.003**

had to be trained on non-defect images and then fail to generate similar images (extract similar features) when images with defects were passed. By models extension, it was possible to achieve segmentations of the defects.

Since we are using a more classic approach to detect and segment target objects such as object detectors, we should change the training dataset by including images with defects. For this purpose, we use all the original training data and part of the original test data with defects to form a new training dataset (the rest of the test data forms a new validation dataset). Since the original test dataset with defects is small, we use the CrossValidation (Bates et al., 2023) method with 4 folds on each category (each category includes ~ 20 images: ~ 15 images form train images, ~ 5 images form validation images). Finally, we get 4 training sets that consist of $\sim (209+15*3)$ images and 4 corresponding validation sets with $\sim (20+5*3)$ images.

To get an augmented version of the training sets, we generate 15 new images with our inpainting method for each defective image in a training set. With this, we obtain 4 augmented training sets that consist of $\sim (209+15*15*3) = 929$ images. To not change detection model training hyperparameters we equalize the size of training dataset without inpainting by copying original training images 15 times, also getting ~ 929 images in total for each training set.

PCB Defects (Diplom, 2023) dataset (Fig 3) consists of 332 images for training and 40 images for validation across 3 categories: dry joint, incorrect installation, and short circuit. For each training image 6 new images are generated with our augmentation method, in total — 1992 images. Therefore, the training dataset with inpainting consists of $1992+332 = 2324$ images. To not change detection model training hyperparameters we equalize the size of training dataset without inpainting by copying original training images 6 times, also getting 2324 images in total. 40 validation images are used for model evaluation.

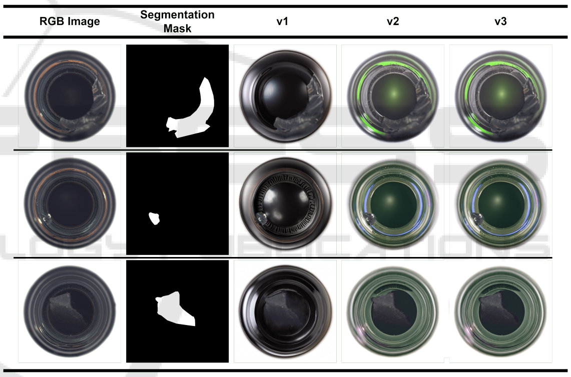

Figure 2: Original and augmented samples visualization for MVTec AD Bottle dataset. ”v1”, ”v2” and ”v3” are corresponding versions of our augmentation approach.

Potholes (Project, 2023) dataset (Fig 4) consists of 424 images for training, 124 images for validation and 60 images for test across 1 category: pothole. For each training image 6 new images are generated with our augmentation method, in total — 2544 images. Therefore, the training dataset with inpainting consists of $2544+424 = 2968$ images. To not change

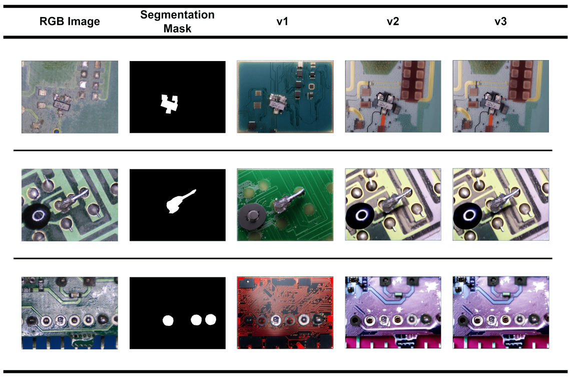

Figure 3: Original and augmented samples visualization for PCB Defects dataset. ”v1”, ”v2” and ”v3” are corresponding versions of our augmentation approach.

detection model training hyperparameters we equalize the size of training dataset without inpainting by copying original training images 6 times, also getting 2968 images in total. 124 validation images are used for model evaluation.

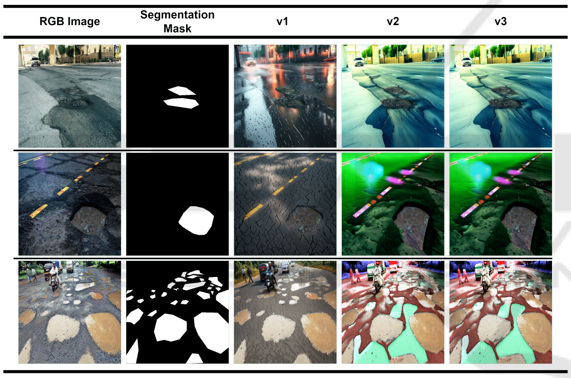

Figure 4: Original and augmented samples visualization for Potholes dataset. ”v1”, ”v2” and ”v3” are corresponding versions of our augmentation approach.

4.2 Implementation Details

All images are brought to 1024 × 1024 resolution before the inpainting algorithm: for MVTec AD Bottle and Potholes datasets it is done by bilinear interpolation upscaling, for PCB Defects dataset it is done by padding with zeros. Both ControlNets conditioning (MiDaS depth and Canny edge) weighted by 1. IP-Adapter’s conditioning weight is set to 1 and noise parameter set to 0.5. Hue shift is random for each image. For all generated samples the same negative prompt is used: “comics, cartoon, blur, text”. After inpainting, images from MVTec AD Bottle and Potholes datasets are kept in 1024 × 1024 resolution, PCB Defects’s images are unpadded.

4.3 Model for Object Detection and Segmentation

We leverage YOLOv8 (Jocher et al., 2023) for object detection and segmentation as current state-of-the-art across one stage detectors. Pre-trained YOLOv8n (nano) version is used in order to avoid overfitting and save computation time, since training datasets are small. For fine-tuning, default hyperparameters from the original repository are utilized.

Nevertheless, our augmentation methods are detector-agnostic. Therefore, they can be used without any adjustment in their pipelines or hyperparameters with any model for object detection and segmentation.

4.4 Object Detection and Segmentation Results

Results are presented in Table 1, Table 2, and Table 3. It can be seen that our augmentation methods mostly decrease performance of the models when the default model-free augmentations are not applied. Furthermore, application of the default model-free augmentations alone significantly improve performance of the models. However, it seems that the joint usage of the augmentations may lead to even better performance across several metrics:

- the third version of our augmentation method lead to major boost of precision with slight tradeoff across Recall and mAP50 metrics on MVTec AD Bottle dataset (Table 1);
- the first version of our augmentation method lead to substantial gain across Recall and mAP50 metrics with minor Precision decrease and significant mAP50-95(M) reduction on PCB Defects dataset (Table 2);
- the third version of our augmentation method lead to meaningful gain over Precision (B) and mAP50-95 with high negative impact on Recall(B) and small decline of the other metrics (Table 3).

4.5 Discussion

Absence of visible pattern in metrics distribution between different datasets and configurations may be explained by high differences in the evaluated datasets and, therefore, differences in the data, which is generated by our approach. This idea is supported by visualizations on Fig 2, Fig 3 and Fig 4. We can see that the second and the third augmentation versions were able to produce more photo-realistic results for

MVTec AD Bottle an PCB Defects datasets. At the same time, in case of augmented Potholes images completely opposite picture is shown (the first version is better). These findings, supported by quantitative results, mean that each dataset should be treated individually when choosing to apply one of the presented diffusion-based method versions.

It is worth noting, that it is difficult to predict influence of our augmentation methods in combination with existing augmentations on a detector training. This point can be supported by the quantitative results, where magnitude and direction of the impact for each metric vary based on whether the default augmentations are used or not. Thus, effects of combination of our augmentation methods with others can be a target for a future research.

In addition, it is important to say, that the current implementation of the proposed augmentations takes $35-45$ seconds to process one image on NVIDIA RTX 3090 graphics card, which make it impossible to use these methods for online augmentation. However, most of the computation time is consumed by the diffusion model itself – $20-25$ seconds. Recent works, allow to reduce a diffusion model computation to 1 seconds or less (Luo et al., 2023), which potentially might facilitate overall inference speed of our augmentation method as well.

The final thing to notice is that there is no comparison with other model-based methods in this paper. The reason for that is a requirement to generate larger augmented datasets and perform a subsequent detector training. Since this process is computationally consuming we decided to make it a theme for a future research.

5 CONCLUSION

In this work we reviewed different augmentation approaches for object detection and segmentation tasks. Next, we proposed our diffusion-based training-free method in order to solve found issues in previous works, such as lack of photorealism and computation inefficiency. Consequently, quantitative comparison results with and without suggested augmentation are shown. None of the proposed augmentation versions proved to be universal across different datasets and metrics. Nevertheless, each of them can be used in order to boost object detection and segmentation models results quality in certain scenarios.

Further research is needed in order to establish how the current framework can be modified to take into account datasets differences. Additionally, comparison and consistency with other augmentation methods should be investigated in more detail.

ACKNOWLEDGEMENTS

The research was supported by the ITMO University, project 623097 ”Development of libraries containing perspective machine learning methods”.

REFERENCES

Antoniou, A., Storkey, A., and Edwards, H. (2017). Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.

Baek, K., Bang, D., and Shim, H. (2021). Gridmix: Strong regularization through local context mapping. *Pattern Recognition*, 109:107594.

Bates, S., Hastie, T., and Tibshirani, R. (2023). Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, pages 1–12.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. (2019). Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600.

Burg, M. F., Wenzel, F., Zietlow, D., Horn, M., Makansi, O., Locatello, F., and Russell, C. (2023). A data augmentation perspective on diffusion models and retrieval. *arXiv preprint arXiv:2304.10253*.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, pages 679–698.

Chen, P., Liu, S., Zhao, H., and Jia, J. (2020). Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

Dildar, M., Akram, S., Irfan, M., Khan, H. U., Ramzan, M., Mahmood, A. R., Alsaiari, S. A., Saeed, A. H. M., Alraddadi, M. O., and Mahnashi, M. H. (2021). Skin cancer detection: a review using deep learning techniques. *International journal of environmental research and public health*, 18(10):5479.

Diplom (2023). Defects dataset. https://universe.roboflow.com/diplom-qz7q6/defects-2q87r. visited on 2023-11-22.

Džijan, M., Grbić, R., Vidović, I., and Cupec, R. (2023). Towards fully synthetic training of 3d indoor object detectors: Ablation study. *Expert Systems with Applications*, page 120723.

Efimova, V., Shalamov, V., and Filchenkov, A. (2020). Synthetic dataset generation for text recognition with generative adversarial networks. In *Twelfth International Conference on Machine Vision (ICMV 2019)*, volume 11433, pages 310–316. SPIE.

Gaidon, A., Wang, Q., Cabon, Y., and Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349.

Ge, Y., Xu, J., Zhao, B. N., Itti, L., and Vineet, V. (2022). Dall-e for detection: Language-driven context image synthesis for object detection. *arXiv preprint arXiv:2206.09592*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., and Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Hataya, R., Zdenek, J., Yoshizoe, K., and Nakayama, H. (2020). Faster autoaugment: Learning augmentation strategies using backpropagation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 1–16. Springer.

Hong, M., Choi, J., and Kim, G. (2021). Stylemix: Separating content and style for enhanced data augmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14862–14870.

Jackson, P. T., Abarghouei, A. A., Bonner, S., Breckon, T. P., and Obara, B. (2019). Style augmentation: data augmentation via style randomization. In *CVPR workshops*, volume 6, pages 10–11.

Jocher, G., Chaurasia, A., and Qiu, J. (2023). YOLO by Ultralytics.

Kim, J.-H., Choo, W., and Song, H. O. (2020). Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pages 5275–5285. PMLR.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471.

Luo, S., Tan, Y., Huang, L., Li, J., and Zhao, H. (2023). Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Project, F. (2023). Pothole detection system new dataset. https://universe.roboflow.com/final-project-iic7d/pothole-detection-system-new. visited on 2023-11-22.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rajpal, A., Cheema, N., Illgner-Fehns, K., Slusallek, P., and Jaiswal, S. (2023). High-resolution synthetic rgb-d datasets for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1188–1198.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Sandfort, V., Yan, K., Pickhardt, P. J., and Summers, R. M. (2019). Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):16884.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Takase, T., Karakida, R., and Asoh, H. (2021). Self-paced data augmentation for training neural networks. *Neurocomputing*, 442:296–306.

Trabucco, B., Doherty, K., Gurinas, M., and Salakhutdinov, R. (2023). Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., and Guibas, L. J. (2019). Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651.

Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Cashman, T. J., and Shotton, J. (2021). Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691.

Xie, J., Li, W., Li, X., Liu, Z., Ong, Y. S., and Loy, C. C. (2023). Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *arXiv preprint arXiv:2309.13042*.

Xu, M., Yoon, S., Fuentes, A., and Park, D. S. (2023). A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, page 109347.

Xu, M., Yoon, S., Fuentes, A., Yang, J., and Park, D. S. (2022). Style-consistent image translation: A novel data augmentation paradigm to improve plant disease recognition. *Frontiers in Plant Science*, 12:3361.

Xu, Z., Meng, A., Shi, Z., Yang, W., Chen, Z., and Huang, L. (2021). Continuous copy-paste for one-stage multi-object tracking and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15323–15332.

Yang, H. and Zhou, Y. (2021). Ida-gan: A novel imbalanced data augmentation gan. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8299–8305. IEEE.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., and Shen, F. (2022). Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*.

Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W. (2023). Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, X., Li, G., Lou, W., Liu, S., Wan, X., Chen, Y., and Li, H. (2023). Diffusion-based data augmentation for nuclei image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–602. Springer.

Zhang, L., Rao, A., and Agrawala, M. (2023a). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Zhang, M., Wu, J., Ren, Y., Li, M., Qin, J., Xiao, X., Liu, W., Wang, R., Zheng, M., and Ma, A. J. (2023b). Diffusionengine: Diffusion model is scalable data engine for object detection. *arXiv preprint arXiv:2309.03893*.

Zhang, X., Wang, Q., Zhang, J., and Zhong, Z. (2019). Adversarial autoaugment. *arXiv preprint arXiv:1912.11188*.

Zhang, Z., Yao, L., Wang, B., Jha, D., Keles, E., Medetalibeyoglu, A., and Bagci, U. (2023c). Emit-diff: Enhancing medical image segmentation via text-guided diffusion model. *arXiv preprint arXiv:2310.12868*.

Zhao, H., Sheng, D., Bao, J., Chen, D., Chen, D., Wen, F., Yuan, L., Liu, C., Zhou, W., Chu, Q., et al. (2023). X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. *arXiv preprint arXiv:2212.03863*.

Zheng, Z., Yu, Z., Wu, Y., Zheng, H., Zheng, B., and Lee, M. (2021). Generative adversarial network with multi-branch discriminator for imbalanced cross-species image-to-image translation. *Neural Networks*, 141:355–371.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. (2020). Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008.