Exploiting Multimodal Latent Diffusion Models for Accurate Anomaly Detection in Industry 5.0

Luigi Capogrosso[1,*], Alvise Vivenza[1,2], Andrea Chiarini[2,3], Francesco Setti[1,2] and Marco Cristani[1,2]

[1]Department of Engineering for Innovation Medicine, University of Verona, Italy
[2]QUALYCO S.r.l, Spin-off of the University of Verona, Verona, Italy
[3]Department of Management, University of Verona, Italy

Abstract

Defect detection is the task of identifying defects in production samples. Usually, defect detection classifiers are trained on ground-truth data formed by normal samples (negative data) and samples with defects (positive data), where the latter are consistently fewer than normal samples. State-of-the-art data augmentation procedures add synthetic defect data by superimposing artifacts to normal samples to mitigate problems related to unbalanced training data. These techniques often produce out-of-distribution images, resulting in systems that learn what is not a normal sample but cannot accurately identify what a defect looks like. In this paper, we show the research we are carrying out in collaboration with QUALYCO, a startup spin-off of the University of Verona, on multimodal Latent Diffusion Models (LDMs) for accurate anomaly detection in Industry 5.0. Unlike conventional image generation techniques, we work within a human feedback loop pipeline, where domain experts provide multimodal guidance to the model through text descriptions and region localization of the possible anomalies. This strategic shift enhances the interpretability of results and fosters a more robust human feedback loop, facilitating iterative improvements of the generated outputs. Remarkably, our approach operates in a zero-shot manner, avoiding time-consuming fine-tuning procedures while achieving superior performance. We demonstrate its efficacy and versatility on the challenging KSDD2 dataset, achieving state-of-the-art results.

Keywords

Diffusion Models, Anomaly Detection, Industry 5.0

1. Introduction

Surface Defect Detection (SDD) is a challenging problem in industrial scenarios, defined as the task of individuating samples containing a defect [1]. In many real-world applications, a human expert inspects every product and removes those defective pieces. Unfortunately, human experts are often inaccurate, and outputs can be inconsistent or biased. Moreover, humans are relatively slow in accomplishing this task, and their performances are subject to stress and fatigue.

Automated defect detection systems [2] can easily overcome most of these issues by learning classifiers on defective and nominal training products. The main drawback is the data collection process required to train a model effectively. Indeed, defective items (i.e., positive samples) are relatively rare compared to nominal items (i.e., negative samples). Thus, the user may need to collect massive amounts of data to have enough positive samples. Moreover, with the rise of the Industry 5.0 [3] and the transition towards flexible manufacturing processes where human operators and production line components actively collaborate, there is an increasing demand for systems that can quickly adapt to new production setups, i.e., customized products manufactured in small batches. Traditional automated systems cannot comply with these demands since data collection could easily involve the whole batch size.

Recent studies on SDD focused on limiting the impact of the labeling process by formulating the problem under the unsupervised learning paradigm [4] or training exclusively on nominal samples [5], possibly using few-shot learning strategies [6]. In both cases, the goal is to generate an accurate model of the nominal sample distribution and predict everything with a low probability score as anomalies. However, due to the limited restoration capability of these models, these approaches tend to generate many false positives, especially on datasets with complex structures or textures [7].

It is worth noting that, in industrial setups, anomalies are not generated by Gaussian processes but are the outcome of specific, often predictable, issues during the production process. Consequently, the anomalous samples are not randomly distributed outside the nominal distribution; they can be modeled as a mixture of Gaussian

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
luigi.capogrosso@univr.it (L. Capogrosso); alvise.vivenza@univr.it (A. Vivenza); andrea.chiarini@univr.it (A. Chiarini); francesco.setti@univr.it (F. Setti); marco.cristani@univr.it (M. Cristani)
0000-0002-4941-2255 (L. Capogrosso); 0000-0003-4915-5145 (A. Chiarini); 0000-0002-0015-5534 (F. Setti); 0000-0002-0523-6042 (M. Cristani)

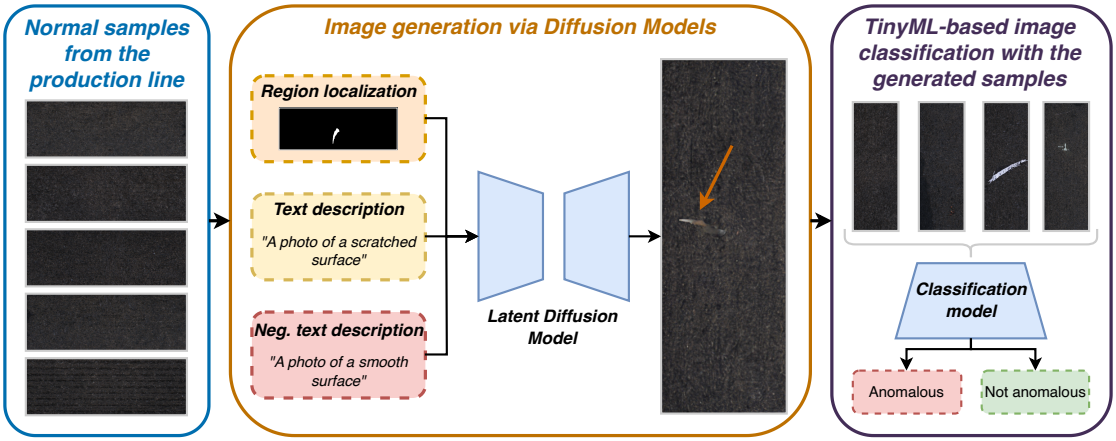

Figure 1: Our pipeline. Starting from positive samples, we leverage a Latent Diffusion Model (LDM) to synthesize novel in-distribution high-quality images of defective surfaces based on defect localization via gesture and textual prompts by a human feedback loop. Then, these synthetic images are used as anomaly samples to train a TinyML-based binary classifier directly on the production line for real-time anomaly detection.

distributions in the feature space instead. While general, unpredictable anomalies can still happen, expert operators can easily define the main problems they can expect from the manufacturing process, such as which kind of defects, in which locations, and how often they wish to appear. Thus, generative AI can represent a powerful tool for SDD, with defect image generation emerging as a promising approach to enhance detector performance.

Specifically, in this paper, we report the result of our research on Latent Diffusion Models (LDMs), a powerful class of generative models, to produce fine-grained realistic defect images that can be used as positive samples to train an anomaly detection model. We name our approach DIAG, a training-free Diffusion-based In-distribution Anomaly Generation pipeline for data augmentation in the SDD task. By leveraging pre-trained LDMs with multimodal conditioning, we can exploit domain experts' knowledge to generate plausible anomalies without needing real positive data. When using these augmented images to train an anomaly detection model, we show a notable increase in the detection performance compared to previous state-of-the-art augmentation pipelines. Specifically, this research is being carried out in collaboration with QUALYCO[1], a startup spin-off of the University of Verona. Figure 1 outlines our approach.

The main contributions of our research are as follows:

- We present a complete pipeline for training anomaly detection models based on nominal images and textual prompts. We showcase the superior outcomes achieved by utilizing generated defective samples compared to previous state-of-the-art approaches.
- We dive into spatial control approaches to enable the synthesis of defect samples incorporating regional information and exhibit enhanced controllability of the image generation through a human feedback loop pipeline, effectively utilizing domain expertise to generate more plausible in-distribution anomalies.

2. Related Work

Research on SDD has been conducted according to different setups: unsupervised approaches [8] use a mixture of unlabelled positive and negative sample images for training; supervised approaches require labeled samples in the form of binary masks representing the defects (full supervision) [9] or simply as a tag for the whole image (weak supervision) [10]. Supervised methods demonstrated superior accuracy in the identification of anomalies. Nevertheless, the effort required to provide good annotations is not always justified. Collecting positive samples can be time and resource-consuming due to the low rate of defective products generated by industrial lines.

Thus, many recent approaches adopt a "*clean*" setup, where the training set consists of only nominal samples. Two strategies can be adopted in clean setups: model fitting and image generation. Model fitting approaches aim at generating an accurate model of the nominal distribution, considering an outlier in every sample with a likelihood lower than –or a distance from the nominal prototype higher than– a predefined threshold [11].

[1]https://qualyco.com.

On the contrary, data augmentation approaches leverage generative methods to synthesize images of defects and use these images as positive samples for training a supervised model. Specifically, this work focuses on generation-based data augmentation under clean setups.

The most popular data augmentation pipeline for SDD consists of a series of random standard transformations of the input image –such as mirroring, rotations, and color changes– followed by the super-imposition of noisy patches [12].

In MemSeg [12], the pipeline for the generation of the abnormal synthetic examples is divided into three steps: *i)* a Region of Interest (ROI) indicating where the defect will be located is generated using Perlin noise and the target foreground; *ii)* the ROI is applied to a noise image to generate a noise foreground ROI; *iii)* the noise foreground ROI is super-imposed on the original image to obtain the simulated anomalous image. However, all these approaches are based on generating out-of-distribution patterns that do not faithfully represent the target-domain anomalies.

More recently, the first work that draws attention to in-distribution defect data is In&Out [13], in which we empirically show that diffusion models provide more realistic in-distribution defects. Here, we significantly improve the generation of in-distribution anomalous samples of [13], incorporating domain knowledge provided by an expert user through textual prompts and localization of salient regions in a training-free setup.

3. Methodology

3.1. Multimodal Diffusion-based image generation

LDMs [14, 15] are a class of deep latent variable models that work by modeling the joint distribution of the data over a Markovian inference process. This process consists of small perturbations of the data with a variance-preserving property [16], such that the limit distribution after the diffusion process is approximately identical to a known prior distribution. Starting with samples from the prior, a reverse diffusion process is learned by gradual denoising the sample to resemble the initial data by the end of the procedure.

We leveraged the natural ability of LDMs to incorporate multimodal conditioning in the generation process, taking inspiration from [17, 18, 19]. Specifically, we use as textual descriptions a prompt and a negative prompt, i.e., a prompt which guides the image generation "away" from its concepts of the desired output, resulting in high-quality images that comply with the given descriptions [20, 21].

In particular, we do not do full image generation to effectively enhance spatial control, opting to utilize an inpainting model, as demonstrated in [14, 18]. Given an image with a masked region, inpainting seamlessly fills it with content that harmonizes with the surrounding image. Although typically employed to eliminate undesired artifacts, the inpainting process ensures that the masked area incorporates the provided prompt, effectively merging textual and visual content.

3.2. Our proposed pipeline

To generate an anomalous image i_a, the process starts by sampling a random negative image, an anomaly description, and a mask, forming the triplet (i_n, d_a, m_a). These pieces of information will then be fed to a text-conditioned LDM to perform inpainting on image i_n using the mask m_a.

The anomaly description d_a guides the generation, filling the masked region of i_n with an anomaly that complies with the prompt. To generate images resembling real anomalous samples, domain knowledge from industrial experts is exploited, providing textual descriptions of the potential anomalies' type, shape, and spatial information.

The LDM is then conditioned on this information to inpaint plausible anomalies on defect-free samples. Formally, given pictures of defect-free (negative) samples I_n, domain experts will provide textual descriptions D_a of what different anomalies may look like. At the same time, regions where these anomalies may appear on the defect-free samples will be designated. We define this set of regions as a set of binary masks M_a of possible anomalies, shapes, and locations. The result of this operation is i_a, an anomalous version of i_n, where an anomaly has been inpainted in the masked region m_a. Due to the stochastic nature of LDMs, this process can be repeated multiple times to generate an augmented set of anomalous sample images I_a. Finally, the set I_a can be used as data augmentation for training anomaly detection models, as presented in the following section.

3.3. The anomaly detection task

We approach the anomaly detection problem as a binary classification problem, where the objective is to predict whether a sample belongs to one of two classes. Specifically, we utilized a ResNet-50 [22] backbone trained with a binary cross-entropy loss function denoted as $\mathscr{L}_{\text{BCE}}$. Mathematically, it is defined as:

$$\mathscr{L}_{\text{BCE}}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] , \tag{1}$$

where, y represents the ground truth labels, $\hat{y}$ represents the predicted probabilities, and N is the number of sam-

ples. In detail, y_i denotes the true label for sample i, which can be either 0 or 1, while $\hat{y}_i$ signifies the predicted probability that sample i belongs to class 1.

Ongoing developments aim to optimize a model through TinyML [23] techniques in order to have an ultra-efficient system that can work smoothly in real-time on a production line.

4. Experiments

4.1. Experiment setup

Datasets We use the Kolektor Surface-Defect Dataset 2 (KSDD2) [10], one of the most recent, complex, and real-world SDD datasets. This dataset comprises 246 positive and 2085 negative images in the training set and 110 positive and 894 negative images in the testing set. Positive images are images with visible defects, such as scratches, spots, and surface imperfections. Since the images have different dimensions, we standardize the dataset resolution, resizing all the images to 224×632 pixels while keeping the number of normal and anomalous samples unchanged.

Evaluation metrics The anomaly detection performance was evaluated based on Average Precision (AP), Precision, and Recall, following the evaluation protocol defined in [13].

4.2. Implementation details

In this section, we specify all the implementation details for reproducibility. All training and inferences were conducted on an NVIDIA RTX 3090 GPU.

Inpainting via Diffusion Models We use the pre-trained implementation of SDXL [21] from Diffusers as our text-conditioned LDM. Following the procedure outlined in Section 3.2, we use the negative images of KSDD2 as the set I_n. As the set of anomaly descriptions D_a, we used the prompts "`white marks on the wall`" and "`copper metal scratches`". Instead, "`smooth, plain, black, dark, shadow`" were used as a negative prompt to improve the performance further. These prompts were chosen after a series of tests, simulating the iterative process of our human feedback loop pipeline until the resulting images resembled plausible anomalies. We used the segmentation masks of positive samples in the KSDD2 dataset to simulate the domain experts' definition of plausible anomalous regions. Then, these data are fed to the pre-trained SDXL model to perform inpainting on the negative images in a training-free process, generating the set of augmented anomalous images I_a as described in Section 3.2. Finally, the generated images I_a are added to the training set, which will be used to train the anomaly detection model.

Table 1
Results between MemSeg, In&Out and DIAG when ***no*** anomalous samples are available. In **bold**, the best results. Underlined, the second best.

Model	N_{aug}	AP ↑	Precision ↑	Recall ↑
MemSeg [12]	80	.514	.733	.436
MemSeg [12]	100	.388	.633	.432
MemSeg [12]	120	.511	.683	.470
In&Out [13]	80	.556	.530	.655
In&Out [13]	100	.626	.742	.568
In&Out [13]	120	.536	.699	.534
DIAG (ours)	80	.769	.851	**.673**
DIAG (ours)	100	**.801**	.924	.664
DIAG (ours)	120	.739	**.944**	.609

ResNet-50 training and testing For a fair comparison with [13], we use the same PyTorch implementation of the ResNet-50 [22] as our anomaly detection model, in which we substitute the fully connected layers after the backbone to make it a binary classifier. The network is trained for 50 epochs with Adam [24] as an optimizer, a learning rate of 0.0001, and a batch size of 32. To maintain consistency with the training and evaluation procedures of KSDD2, our setup is the same as presented in [10, 13], where only the images and ground truth labels are used to train the model.

4.3. Quantitative results

Zero-shot data augmentation Here, we emulate the situation where ***no*** original positive samples are available in the training set. This scenario makes generating augmented positive samples necessary and restricts the users to augmentation procedures that do not rely on positive images. To do this, we build the set of augmented anomalous images I_a by generating N_{aug} augmented positive samples with different pipelines, i.e., MemSeg [12], In&Out [13] and DIAG. Then, we train the ResNet-50 model on a dataset that includes the original negative samples I_n and the augmented positive samples I_a. Finally, we evaluate the model on the original test set.

Table 1 reports the comparison between the models trained with MemSeg, In&Out, and DIAG augmented data at different values of N_{aug}. As we can see, our proposed method achieves the highest AP (.801), recorded at 100 augmented images, while also resulting in a consistently higher AP when compared to the MemSeg and In&Out pipelines. These impressive results highlight how, through domain expertise in the form of anomaly

Table 2
Results between MemSeg, In&Out and DIAG when ***all*** the anomalous samples are available. In **bold**, the best results. Underlined, the second best.

Model	N_{aug}	AP ↑	Precision ↑	Recall ↑
MemSeg [12]	80	.744	.851	.691
MemSeg [12]	100	.774	.814	.752
MemSeg [12]	120	.734	.772	.707
In&Out [13]	80	.747	.764	.734
In&Out [13]	100	.775	.868	.720
In&Out [13]	120	.782	.906	.689
DIAG (ours)	80	.869	.912	.755
DIAG (ours)	100	.911	**.978**	.800
DIAG (ours)	120	**.924**	.896	**.864**

descriptions and segmentation masks, it is possible to generate in-distribution images able to meaningfully guide an anomaly detection network, even in a complicated scenario where no real anomalous data is available.

Surprisingly, the DIAG performance with $N_{aug} = 120$ augmented images is lower than using a smaller number of augmented images. We hypothesize this is due to the stochastic nature of the LDMs image generation. While it allows the generation of various images given the same guidance, it can also lower, in some cases, the predictability of the quality of the generated samples, which sometimes may not faithfully comply with the prompt. Future works will focus on studying quality consistency in the image generation pipeline.

Full-shot data augmentation To showcase DIAG as a general data augmentation technique, we also explore the scenario where real positive samples are available in the training set. To this aim, we include all the 246 real positive samples I_p in the training set, together with the real negative images I_n and the N_{aug} augmented positive images I_a.

As we can see from Table 2, DIAG achieves the highest average AP yet (.924), surpassing the .782 set by the previous state-of-the-art data augmentation pipeline [13]. When comparing these results to the ones obtained in the "zero-shot data augmentation" scenario, it is clear how more in-distribution images improve model performance during training. This is highlighted by the improvement in performance of all the models when adding the real positive images I_p to the training set. At the same time, the inclusion of DIAG augmented images allows the model to explore the anomaly distribution further, resulting in the difference in performance between the different data augmentation pipelines.

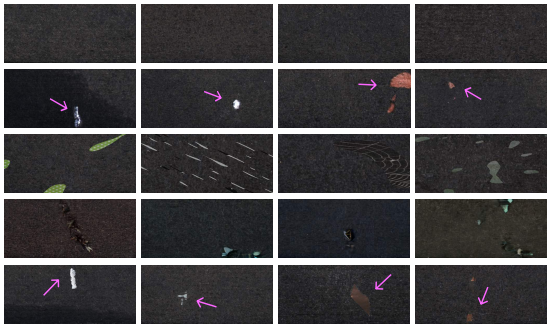

Figure 2: First row displays some negative samples from the KSDD2 dataset. The second row shows some images of positive samples from the same dataset. The third row shows the MemSeg-generated defect samples. The fourth row shows In&Out generated defect samples. Lastly, the final row showcases some images generated with DIAG. Notably, the defect images that DIAG generated are more realistic and in-distribution.

4.4. Qualitative results

The main goal of our data augmentation pipeline is to generate in-distribution synthetic positive images, meaning images that closely resemble the real ones. Figure 2 shows qualitative results. It's evident that the images produced by DIAG are markedly more realistic compared to those generated by MemSeg [12] and In&Out [13].

5. Conclusions

This work presents DIAG, a novel data augmentation pipeline that leverages visual language models to produce training-free positive images for enhancing the performance of an SDD model. We introduced domain experts in the generation pipeline, asking them to describe with textual prompts how a defect should look and where it can be localized. Then, we adopt a pre-trained LDM to generate defective images and train a binary classifier for isolating the anomalous images. We focus our experiments on the KSDD2 dataset and establish ourselves as the new state-of-the-art data augmentation pipeline, surpassing previous approaches in both the zero-shot and full-shot data augmentation scenarios with an AP of .801 and .924, respectively. These results highlight the potential of in-distribution data augmentation in the anomaly detection field, where training-free generative model pipelines such as DIAG can provide meaningful data for downstream classification, making them appealing solutions in scenarios where real anomalous data is difficult to collect or unavailable. These promising results promote further exploration across various datasets, particularly investigating how robust the image generation is compared to noisy textual prompts.

References

[1] T. Wang, Y. Chen, M. Qiao, H. Snoussi, A fast and robust convolutional neural network-based defect detection model in product quality control, The International Journal of Advanced Manufacturing Technology 94 (2018) 3465–3471.

[2] S. H. Hanzaei, A. Afshar, F. Barazandeh, Automatic detection and classification of the ceramic tiles' surface defects, Pattern Recognition 66 (2017) 174–189.

[3] P. K. R. Maddikunta, Q.-V. Pham, B. Prabadevi, N. Deepa, K. Dev, T. R. Gadekallu, R. Ruby, M. Liyanage, Industry 5.0: A survey on enabling technologies and potential applications, Journal of Industrial Information Integration 26 (2022) 100257.

[4] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, P. Gehler, Towards total recall in industrial anomaly detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14318–14328.

[5] M. Rudolph, B. Wandt, B. Rosenhahn, Same same but differnet: Semi-supervised defect detection with normalizing flows, in: Winter Conference on Applications of Computer Vision (WACV), 2021.

[6] Y. Song, T. Wang, P. Cai, S. K. Mondal, J. P. Sahoo, A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities, ACM Computing Surveys 55 (2023) 1–40.

[7] Y. Chen, Y. Ding, F. Zhao, E. Zhang, Z. Wu, L. Shao, Surface defect detection methods for industrial products: A review, Applied Sciences 11 (2021) 7657.

[8] X. Tao, D. Zhang, W. Ma, Z. Hou, Z. Lu, C. Adak, Unsupervised anomaly detection for surface defects with dual-siamese network, IEEE Transactions on Industrial Informatics 18 (2022) 7707–7717.

[9] C. Luan, R. Cui, L. Sun, Z. Lin, A siamese network utilizing image structural differences for cross-category defect detection, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020.

[10] J. Božič, D. Tabernik, D. Skočaj, Mixed supervision for surface-defect detection: From weakly to fully supervised learning, Computers in Industry 129 (2021) 103459.

[11] T. Defard, A. Setkov, A. Loesch, R. Audigier, Padim: a patch distribution modeling framework for anomaly detection and localization, in: International Conference on Pattern Recognition (ICPR), 2021.

[12] M. Yang, P. Wu, H. Feng, Memseg: A semi-supervised method for image surface defect detection using differences and commonalities, Engineering Applications of Artificial Intelligence 119 (2023) 105835.

[13] L. Capogrosso, F. Girella, F. Taioli, M. Dalla Chiara, M. Aqeel, F. Fummi, F. Setti, M. Cristani, Diffusion-based image generation for in-distribution data augmentation in surface defect detection, in: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), 2024. doi:10.5220/0012350400003660.

[14] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: International Conference on Machine Learning (ICML), 2015.

[15] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in Neural Information Processing Systems (NeurIPS) 33 (2020) 6840–6851.

[16] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, in: International Conference on Learning Representations (ICLR), 2020.

[17] J. Ho, T. Salimans, Classifier-free diffusion guidance, arXiv preprint arXiv:2207.12598 (2022).

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[19] L. Capogrosso, A. Mascolini, F. Girella, G. Skenderi, S. Gaiardelli, N. Dall'Ora, F. Ponzio, E. Fraccaroli, S. Di Cataldo, S. Vinco, et al., Neuro-symbolic empowered denoising diffusion probabilistic models for real-time anomaly detection in industry 4.0: Wild-and-crazy-idea paper, in: 2023 Forum on Specification & Design Languages (FDL), IEEE, 2023, pp. 1–4.

[20] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents. arxiv 2022, arXiv preprint arXiv:2204.06125 (2022).

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

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[23] L. Capogrosso, F. Cunico, D. S. Cheng, F. Fummi, M. Cristani, A machine learning-oriented survey on tiny machine learning, IEEE Access (2024).

[24] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).