CAGEN: CONTROLLABLE ANOMALY GENERATOR USING DIFFUSION MODEL

Bolin Jiang[1] *Yuqiu Xie*[1] *Jiawei Li*[2] *Naiqi Li*[1,*] *Yong Jiang*[1] *Shu-Tao Xia*[1]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Huawei Manufacturing

ABSTRACT

Data augmentation has been widely applied in anomaly detection, which generates synthetic anomalous data for training. However, most existing anomaly augmentation methods focus on image-level cut-and-paste techniques, resulting in less realistic synthetic results, and are restricted to a few predefined patterns. In this paper, we propose our Controllable Anomaly Generator (CAGen) for anomaly data augmentation, which can generate high-quality images, and be flexibly controlled with text prompts. Specifically, our method finetunes a ControlNet model by using binary masks and textual prompts to control the spatial localization and style of generated anomalies. To further augment the resemblance between the generated features and normal samples, we propose a fusion method that integrates the generated anomalous features with the features of normal samples. Experiments on standard anomaly detection benchmarks show that the proposed data augmentation method significantly leads to a 0.4/3.1 improvement in the AUROC/AP metric.

Index Terms— Anomaly detection, Data augmentation, Diffusion model

1. INTRODUCTION

Anomaly detection is an important task in both industrial applications and academic research. Since in real-world scenarios anomalous data are considerably less frequent than normal data, researchers have concentrated on self-supervised methods [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. These methods exclusively train on normal data and can be categorized into reconstruction based and feature distance based methods. The reconstruction based methods assume that a model trained only on normal data cannot effectively reconstruct anomalous parts. On the other hand, feature distance based approaches aim to obtain a tight representation of the normal state, with any deviation from this representation being considered anomalous.

To address the scarcity of anomalous data, both reconstruction based and feature distance based methods resort to generating synthetic anomalies [2, 3, 7]. Although these synthetic anomalies differ from real anomalies, they can still serve as supervisory signals and yield satisfactory results. Intuitively, if the synthetic anomalies follow a more similar distribution of the real anomalous patterns, they will result in more favorable detection results. This motivates us to take advantage of the recent progress of diffusion models, which can generate high-quality images and be flexibly controlled with text prompts.

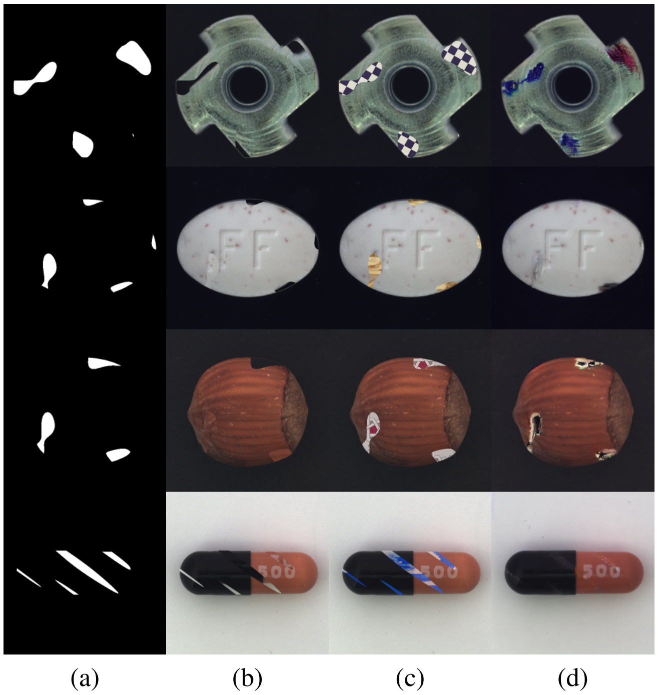


Fig. 1. The augmented anomalous images obtained through various methods. (a) presents the anomaly masks. Columns 2-4 present the images generated by (b) CutPaste [2], (c) DRAEM [3], and (d) our method.

Latent diffusion models (LDMs) like Stable Diffusion [11] have achieved remarkably impressive results in image generation. We observe that by guiding the pretrained Stable Diffusion model using textual prompts, it can comprehend the text and generate images of anomalies that closely resemble real-life instances. However, this is insufficient for augmenting datasets for anomaly detection tasks. The primary reason is that the images generated in this manner lack annotation

Corresponding author: Naiqi Li

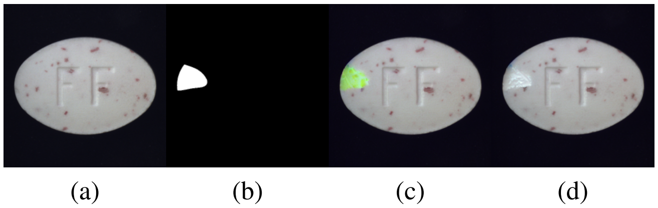


Fig. 2. Anomalous images generated through different textual prompts by CAGen. Where (a) represents the normal sample, (b) indicates the anomaly location mask, (c) corresponds to the “colored pill”, and (d) corresponds to the “scratched pill”.

information. Besides, there is a significant discrepancy between the generated images and the real samples, since the model was pretrained on a much diverse and different dataset.

In response to these limitations, we propose our two-stage method Controllable Anomaly Generator (CAGen) using diffusion model. In the first stage, we employ a fine-tuned Stable Diffusion Model via ControlNet [12], where a binary mask and text prompt are utilized to govern both the location and style of the generated anomalies. The mask and text offer the necessary annotation information for the anomalies. In the second stage, we ensure the resemblance to normal samples by fusing the generated anomalous features with the features of normal samples at the feature level. In summary, our paper has the following contributions:

1. We introduce CAGen, an innovative data augmentation methodology for anomaly detection.

2. By utilizing masks and textual prompts to dictate the location and style of anomalies, CAGen can produce high-quality anomalous images.

3. We propose Anomaly Feature Fusion, integrating the features of normal samples with the generated anomalous features.

2. RELATED WORK

2.1. Data augmentation in anomaly detection

Data augmentation is a widely adopted approach in anomaly detection to capture anomalous images as supervisory signals. Cutpaste [2] and DRAEM [3] primarily generate synthetic anomalies at the image level, while SimpleNet [7] induces anomalies at the feature level. Specifically, Cutpaste [2] fabricates anomalous images by randomly cutting a patch from an image and then pasting it to a random position. DRAEM [3], on the other hand, utilizes random Perlin noise to determine the location of the synthetic anomaly and subsequently pastes images from the DTD dataset [13] onto those positions. SimpleNet [7] generates anomalous data by adding a subtle Gaussian noise to the features of normal images. These methodologies have collectively affirmed the efficacy of data augmentation in anomaly detection.

2.2. Diffusion models

Denoising Diffusion Probabilistic Models (DDPM) [14] have demonstrated state-of-the-art performance in the realm of image generation. To address the generation time, Denoising Diffusion Implicit Models (DDIM) [15] offers improvements, reducing the required time span. Furthermore, the Stable Diffusion Model [11], by integrating VAE [16] and diffusion models, further minimizes image generation time. Concurrently, many subsequent works, such as ControlNet [12], are built upon the Stable Diffusion framework.

ControlNet [12] seeks to control image generation without compromising the generative capabilities of the pretrained Stable Diffusion Model [11]. By incorporating the ControlNet structure into the pretrained Stable Diffusion model, we obtained the final model structure. Specifically, the upper half of ControlNet replicates the top half of the Stable Diffusion U-Net, while the lower half employs a structure called “zero convolution”. By locking the weights of the pretrained Stable Diffusion model and training solely on a limited target dataset, ControlNet can achieve impressive control efficacy while maintaining robust generalization capabilities.

3. PROPOSED METHOD

In this section, we introduce our framework for generating anomalous samples. As shown in Fig. 3, the process of generating these samples is divided into two stages. The first stage is Anomaly-Guided Feature Generation. Using a designated mask combined with a textual prompt, we instruct the denoising Unet to produce anomaly features at designated locations. The second stage is Anomaly Feature Fusion, where the features of normal samples are fused with the generated defect features at the feature level, resulting in samples that closely resemble real anomalous instances.

3.1. Anomaly-Guided Feature Generation

Due to the extensive training of the pretrained Stable Diffusion model [11] on a wide range of image-text pairs, it can generate specific anomalies based on textual prompts. Typically, these anomalies can appear at various locations in the image. To utilize the generated images for the targeted anomaly detection task, they must be annotated with anomaly masks, thus necessitating the model’s ability to control the location of anomaly generation.

In this stage, we employ ControlNet [12] to fine-tune the pretrained Stable Diffusion model. ControlNet employs a binary mask and a textual prompt as conditional inputs: the mask dictates the location where the anomaly will be generated, while the textual prompt determines the style of

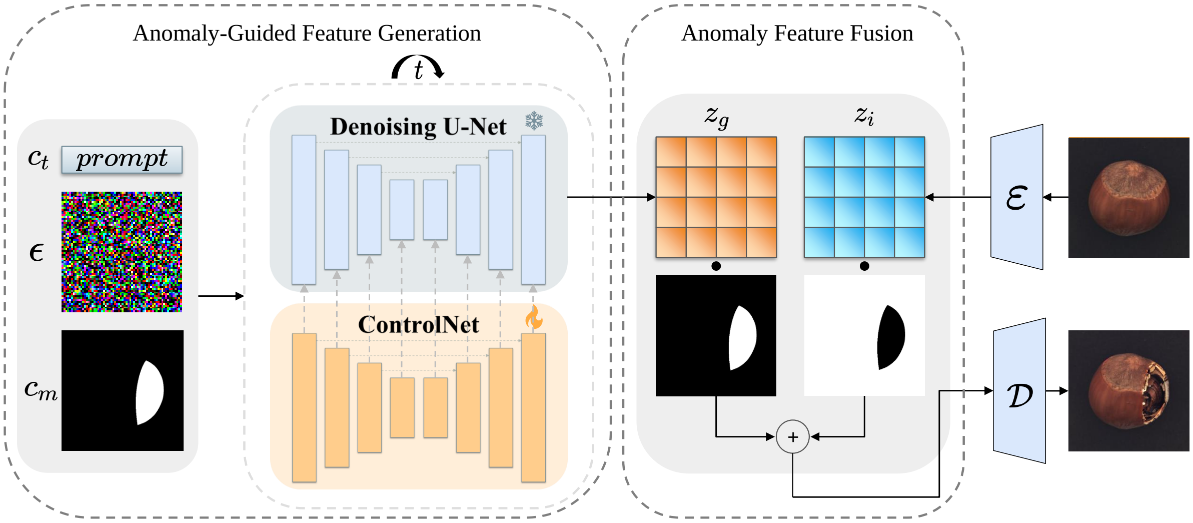


Fig. 3. The framework of our method, which consists of two stages: (i) The model takes a text prompt c_t, a positional control mask c_m, and Gaussian noise ϵ as inputs, and then produces anomalous features at the designated location after undergoing t steps of denoising. (ii) The encoder extracts features from normal samples, which are then fused at the feature level with the generated anomalous features; the decoder subsequently decodes these fused features to produce an anomalous image.

the anomalous features, enabling the production of high-quality and diverse anomalies. By training on a limited set of anomaly-mask pairs, we can guide the denoising Unet to generate anomalies at locations specified by the mask.

While training, during the forward process, given an anomalous image feature z_0, noise is progressively added to it to obtain the noise feature z_t, where t is the number of times the noise is added. In the backward process, the weights of the Denoising U-Net are frozen. Given a series of conditions, including the time step t, the text prompt c_t, and the anomaly location mask c_m, the ControlNet [12] is trained to enable the entire model ϵ_θ to predict the noise added to z_t with

$$\mathcal{L} = \mathbb{E}_{z_0,t,c_t,c_m,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_t, c_m)\|_2^2\right] \quad (1)$$

where $\mathcal{L}$ is the overall learning objective of the entire diffusion model.

3.2. Anomaly Feature Fusion

It is worth noting that the features generated in the first stage have a significant disparity from the features of normal samples, making them unsuitable for direct anomaly detection. To address this issue, we employ Anomaly Feature Fusion to integrate the generated anomalous features into the features of normal samples, thereby obtaining augmented anomalous samples. The mathematical formulation is as follows:

$$z_i = \varepsilon(i_r) \quad (2)$$

$$z_g = \epsilon_\theta(\epsilon, t, c_t, c_m), \epsilon \sim \mathcal{N}(0, 1) \quad (3)$$

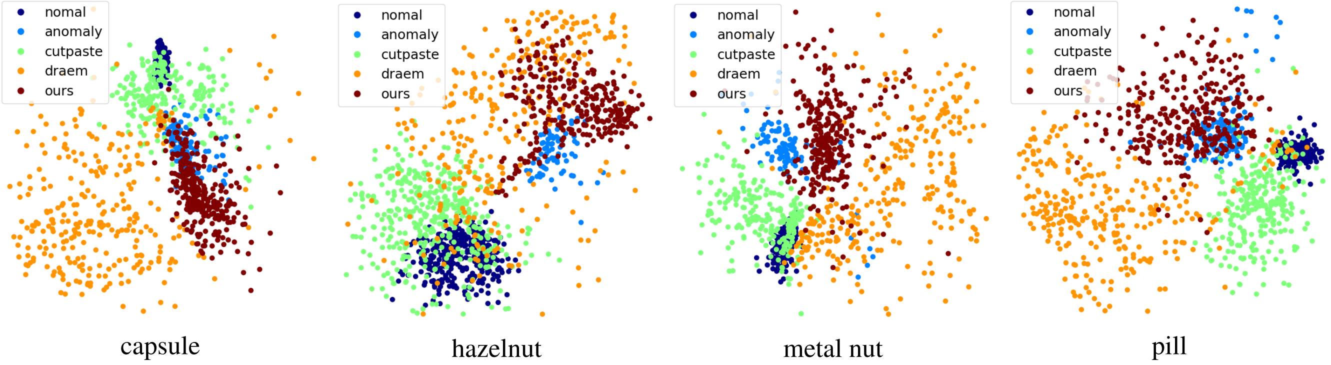
$$z_f = z_g \cdot c_m + z_i \cdot (1 - c_m) \quad (4)$$

$$i_g = \mathcal{D}(z_f) \quad (5)$$

where i_r represents the reference normal image, ε and $\mathcal{D}$ denote the encoder and decoder, z_i denotes the features of the normal image, z_g signifies the generated anomalous features, ϵ_θ stands for the denoising model, c_m stands for the conditional mask, c_t represents the textual prompt, z_f indicates the fused features, and i_g designates the final generated image.

4. EXPERIMENTS

We trained our CAGen diffusion model on the MVTec-AD [17] dataset. Specifically, we fine-tuned the pretrained Stable Diffusion [11] v1.5 model using ControlNet [12]. For the training process, we randomly selected three images from each type of defect within each category. The textual prompts for each image were simply set to "{defect type} {category}", e.g., "broken large bottle". The training was conducted over a total of 1000 epochs. For the images requiring augmentation, we first obtained the mask of the target object to be augmented. This mask was then element-wise multiplied with a randomly generated Perlin noise mask to produce the final anomaly location mask. The image to be augmented, the anomaly location mask, and the corresponding textual anomaly prompt were then fed into CAGen to produce the final anomalous image.



capsule hazelnut metal nut pill

Fig. 4. Visualization results of t-SNE for the normal samples, anomalous samples, samples generated by the Cutpaste [2], samples generated by the DRAEM [3], and samples generated by our method.

Table 1. Anomaly localization results with AUROC / AP metric on MVTec-AD [17]. The results highlighted in bold represent the best performance, while those underlined indicate an improvement compared to the DRAEM [3] method.

	Category	**Cutpaste**	**DRAEM**	**Ours**
Object	bottle	97.6 / –	99.1 / 86.5	$\underline{\textbf{99.2}}$ / $\underline{\textbf{89.8}}$
	cable	90.0 / –	94.7 / 52.4	$\underline{\textbf{95.1}}$ / $\underline{\textbf{63.0}}$
	capsule	**97.4** / –	94.3 / **49.4**	$\underline{95.6}$ / **49.4**
	hazelnut	97.3 / –	99.7 / 92.9	$\underline{\textbf{99.8}}$ / $\underline{\textbf{95.9}}$
	metal nut	93.1 / –	**99.5** / **96.3**	**99.5** / 96.1
	pill	95.7 / –	97.6 / 48.5	$\underline{\textbf{98.0}}$ / $\underline{\textbf{51.4}}$
	screw	96.7 / –	97.6 / 58.2	$\underline{\textbf{99.4}}$ / $\underline{\textbf{64.7}}$
	toothbrush	98.1 / –	98.1 / 44.7	$\underline{\textbf{98.5}}$ / $\underline{\textbf{61.9}}$
	transistor	**93.0** / –	90.9 / **50.7**	$\underline{91.5}$ / 45.7
	zipper	**99.3** / –	98.8 / 81.5	$\underline{99.0}$ / $\underline{\textbf{81.8}}$
Texture	carpet	**98.3** / –	95.5 / 53.5	$\underline{95.9}$ / $\underline{\textbf{55.9}}$
	grid	97.5 / –	**99.7** / 65.7	99.6 / $\underline{\textbf{76.1}}$
	leather	**99.5** / –	98.6 / **75.3**	$\underline{99.1}$ / 71.4
	tile	90.5 / –	99.2 / 92.3	$\underline{\textbf{99.4}}$ / $\underline{\textbf{95.8}}$
	wood	95.5 / –	**96.4** / **77.7**	96.1 / 74.8
	mean	96.0 / –	97.3 / 68.4	$\underline{\textbf{97.7}}$ / $\underline{\textbf{71.5}}$

4.1. Improvement on Anomaly Detection

We followed the setup of the anomaly detection process of DRAEM [3] and replaced 30% of the training set with anomalous images generated by CAGen. Although we trained CAGen using a small portion of the test set, the model did not directly encounter real anomalies during the anomaly detection phase. Furthermore, after removing the images used for CAGen training, we achieved an average AUROC/AP of 97.7/71.5 on MVTec-AD [17], which is essentially consistent with the full test set. Therefore, the scores we report are all based on the complete test dataset. As shown in Table 1, we achieved a **0.4/3.1** increase in the AUROC/AP metric based on the foundation of DRAEM on the MVTec-AD dataset. As demonstrated in Table 2, we achieved a **2.9/6.0** increase in the AUROC/AP metric on BTAD [18].

Table 2. Anomaly localization results with AUROC metric and AP metric on BTAD [18].

category	**DRAEM**	**Ours**
01	91.8 / 18.8	**92.7 / 19.8**
02	77.9 / 31.8	**85.5 / 45.6**
03	95.1 / 10.4	**95.4 / 13.5**
mean	88.3 / 20.3	**91.2 / 26.3**

4.2. Visualizing the Embeddings of Synthetic Images

In this experiment we utilize t-Distributed Stochastic Neighbor Embedding (t-SNE) [19] which is commonly used for visualizing high-dimensional data. We trained a binary classifier on normal images and real anomalous images by fine-tuning a pretrained ResNet-18 [20]. Subsequently, we extracted the features from this classifier and applied t-SNE visualization on normal images, real anomalous images, Cutpaste synthetic anomalous images, DRAEM synthetic anomalous images, and anomalous images generated by our method. As illustrated in Fig. 4, the anomalies generated by our method are closer to real anomalies, which is more conducive to anomaly detection.

5. CONCLUSION

In this paper, we introduce CAGen, an innovative data augmentation methodology for anomaly detection. Our approach is built upon the latest diffusion model, boasting superior generation capabilities and controllability, and allowing for the production of diverse anomalous patterns. Furthermore, we introduce the Anomaly Feature Fusion technique to blend anomalous features with normal images. Comprehensive experiments validate the effectiveness of our method.

6. REFERENCES

[1] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 475–489.

[2] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9664–9674.

[3] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8330–8339.

[4] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka, "Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 98–107.

[5] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu, "Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows," *arXiv:2111.07677*, 2021.

[6] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14318–14328.

[7] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang, "Simplenet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20402–20411.

[8] Mingqing Wang, Jiawei Li, Zhenyang Li, Chengxiao Luo, Bin Chen, Shu-Tao Xia, and Zhi Wang, "Unsupervised anomaly detection with local-sensitive vqvae and global-sensitive transformers," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 1080–1084.

[9] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia, "Unsupervised surface anomaly detection with diffusion probabilistic model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6782–6791.

[10] Jiawei Li, Chenxi Lan, Xinyi Zhang, Bolin Jiang, Yuqiu Xie, Naiqi Li, Yan Liu, Yaowei Li, Enze Huo, and Bin Chen, "Siad: Self-supervised image anomaly detection system," *arXiv:2208.04173*, 2022.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[12] Lvmin Zhang and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv:2302.05543*, 2023.

[13] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3606–3613.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[15] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," *arXiv:2010.02502*, 2020.

[16] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv:1312.6114*, 2013.

[17] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9592–9600.

[18] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti, "Vt-adl: A vision transformer network for image anomaly detection and localization," in *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*. IEEE, 2021, pp. 01–06.

[19] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and pattern Recognition*, 2016, pp. 770–778.