

# AI-Generated Anime Figure Detection

Yi'an Pei & Xiaotong Zha

Duke Kunshan University  
COMPSI 309  
May 2024

**Abstract.** Artificial intelligence-generated content (AIGC) has become pervasive across various domains, particularly with the advent of large-scale models. Text-to-image conversion, a subset of AIGC, has gained significant traction, notably in social media platforms, for its ability to automate visual design processes. In this study, we investigate the detection of AI-generated anime images, focusing on discerning between AI-generated and non-AI-generated content. We propose a novel approach that leverages YOLOv5-anime for face detection within anime images, subsequently employing a combination of convolutional neural networks (CNNs) and transformers as backbones for AI detection. Our model utilizes an ensemble architecture to integrate the outputs of these backbones, achieving enhanced performance through a weighted combination. We construct a comprehensive dataset sourced from Pixiv, a popular platform for digital artwork, and evaluate our model's performance using various metrics such as accuracy, precision, recall, specificity, and F1 score. Results demonstrate the effectiveness of our approach, with our model outperforming established architectures across all metrics. Notably, our model exhibits balanced performance in detecting AI-generated and human-generated images, indicating its potential for practical applications in identifying AI-generated content. This study contributes to the advancement of AI detection techniques and underscores the importance of responsible utilization of AIGC technologies.

**Keywords:** Machine Learning · AI-Generated Content · AI Detection · Anime Images.

## 1 Introduction

AI-generated content (AIGC), which denotes the production or creation of content through artificial intelligence systems, has exerted a significant influence across various application scenarios, particularly with the emergence of large-scale models. The content produced by generative AI models comprises text, image, audio, video, and cross-modal transformations such as text-to-image and text-to-audio[1]. Among these modalities, text-to-image conversion has garnered enormous attention in social media platforms, primarily owing to its potential to supplant human visual design. In this field, several representative models have emerged, one such model being DALL•E, which is trained using Variational Autoencoders, while another important model is VQGAN-CLIP based on

Generative Adversarial Networks (GANs). Nowadays, the diffusion model has become the core approach in text-to-image generation, with popular models like Stable Diffusion, Disco Diffusion, MidJourney, and DALL•E2 etc[22]. Users can cultivate realistic or artistic images significantly correlative with the prompt by typing the prompt containing expected image concepts, attributes, and styles. Given the rapid evolution and widespread adoption of AIGC technologies, it's imperative to critically examine its implications on creativity, ethics, and society at large, ensuring responsible and equitable utilization of these powerful tools.

Previous work in this domain has seen the utilization of familiar classification techniques prevalent in computer vision to discern between AI-generated and non-AI-generated images. In 2020, Wang et al. used Resnet50 to classify GAN-generated image with real images[20], Gagnaniello et al. enhanced on their work in 2021, deleting the down-sampling of Resnet[4]. Mandelli et al. further used multiple EfficientNet-B4 models to detect GAN-generated images[9]. Tan et al. further introduced Transformer models to this field in 2023[16]. Ricker et al. introduced these detection tech to Diffusion-generated image detection in 2024[13]. Typically, the entire image serves as the input for such classification tasks. However, we are intrigued by the possibility of exploring specific regions within an image for this purpose. Particularly, we focus on anime picture generation, where the output of anime heads tends to be of higher quality compared to other regions, which may contain errors. This raises the question of which region of the image holds the most discriminatory information. In the context of anime picture generation, focusing on the quality of the generated anime heads could potentially enhance the accuracy of AI-generated image classification. Our problem: Would focusing on the head of anime images improve the performance of AI detection?

## 2 Background

### 2.1 YOLO

You Only Look Once (YOLO) is a famous machine learning model in field of object detection which is first introduced by Redmon et al. in 2015 [12]. This model applies advanced algorithms, increasing its speed and accuracy in analyzing images, which often plays an important role in real-time object detection.

Previous models usually need to look through the original picture several times, to find the box containing objects and decide the class of objects respectively. Different from those models, YOLO does not need to scan the figure multiple times. It solves the problem in view of a single regression problem directly from image pixels when bounding box coordinates and class probabilities at the same time, which ensures the process to be fast as it finishes all its works in one single pass. To be a tradeoff, the accuracy of YOLO is usually lower, but still generally high enough to be available, making it to be an overall better model to work on.

After several years of improvement, YOLO now has many different versions and branches for various scenarios. YOLOv5-anime is one of those versions, trained and specialized for anime style. This model can highlight the parts considered as “heads” by boxes and corresponding scores, similar to what other version of YOLO can do, while it is expected to have a better performance when dealing with databases like anime images.

## 2.2 Backbone

**CNNs** Convolutional Neural Networks (CNNs) have emerged as the cornerstone of many computer vision tasks, owing to their ability to effectively extract hierarchical features from images[6]. CNNs are composed of multiple layers, including convolutional layers, pooling layers, and fully connected layers[7]. These layers collectively learn to identify patterns and features in images, enabling tasks such as object detection, classification, and segmentation. CNN architectures typically consist of a series of convolutional layers followed by pooling layers, which reduce the spatial dimensions of the feature maps while preserving important information[23]. The convolutional layers apply filters to the input image, extracting features such as edges, textures, and shapes[21]. These features are then passed through activation functions to introduce non-linearity, enhancing the model’s capacity to capture complex patterns. One of the pioneering CNN architectures is AlexNet[6], which achieved breakthrough performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. Since then, numerous CNN architectures have been proposed, each offering improvements in terms of depth, width, and computational efficiency. Examples include VGGNet[14], GoogLeNet[15], ResNet[5], and EfficientNet[18].

**Transformers** Transformers have revolutionized various fields of artificial intelligence, particularly natural language processing (NLP), but their applicability has extended to computer vision tasks as well[19]. Originally introduced for sequence-to-sequence learning tasks, transformers have demonstrated remarkable success in capturing long-range dependencies and contextual information[2]. Unlike CNNs, which rely on convolutions for feature extraction, transformers utilize self-attention mechanisms to weigh the importance of different input tokens[19]. This mechanism allows transformers to capture global dependencies in the input sequence, making them well-suited for tasks that require understanding relationships between distant elements. In computer vision, transformers have been adapted for tasks such as image classification, object detection, and image generation. Notably, models like Vision Transformer (ViT)[3] have shown competitive performance on image classification benchmarks such as ImageNet. By treating images as sequences of patches, ViT applies transformer architectures to extract features and make predictions. Additionally, transformer-based models like DALL•E[11] and CLIP[10] have demonstrated impressive capabilities in generating and understanding images by leveraging text-based prompts.

These models rely on the fusion of vision and language modalities, showcasing the versatility and power of transformer architectures in computer vision tasks.

### 3 Methods

#### 3.1 Data Source

To form a database of anime images, we collected illustrations from the website called pixiv. Pixiv is an online community and platform centered around digital artwork with mostly anime style. This dataset is regarded to be different from those of most previous researches because anime figures have no biological limitations. This provides possibility of existing "unreal" components in non-AI images, which have potential to be a misdirection for regular detection models.

All the data is downloaded from pixiv website directly through chrome extensions. Only illustrations with clear mark of "AI" or "non-AI" were collected and separated into corresponding folders. Besides, images with less than 10 likes were ignored in case there are too many extreme outliers.

#### 3.2 YOLOv5-anime

YOLOv5-anime is a variant of YOLO targeted at detecting and localizing anime-style characters within anime images and videos. This model will output the original image with highlighted boxes where containing objects considered as faces. It had a good performance at detection faces of anime characters since it is trained in similar situation.

In this project we decided to trust the model and applied a trained weight to establish a database with the help of YOLOv5-anime. All imaged from pixiv will be tested by this model to check whether there are character involved.

#### 3.3 Architecture

In our project, we aimed at combining YOLO and some of the backbones to create a better model on detecting AI-Generated anime images (see Fig. 1). Considering the common habit of focus on tags related to faces when generating AI anime pictures, we created an architecture to test the performance of dataset with, without, or even only head and made an ensemble accordingly.

For every image as an input, YOLOv5-anime will be applied. If YOLO successfully detected some faces, the image will be used to form three kinds of pictures: original, head-only and head-free. Head-only represents the cut of the highest face detected, and in the folder of head-free all faces are blocked in blank.

Afterwards, all these three kinds of data will be trained independently by different backbones. For each of these three folders, one model will be found to be

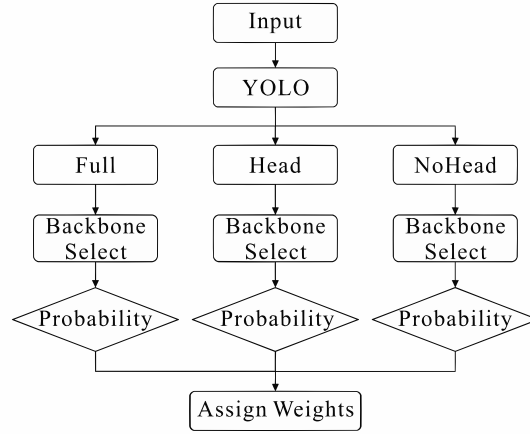


Fig. 1: Designed architecture using YOLO and Backbone Select

the best backbone, and they will be combined to give out the finally prediction of AI or not.

The combination of backbones was regarded as a classification problem, where the confidence of an image to be AI by the three models forms three features. The weights was obtained from logistic regression, attempting to output 1 for AI and 0 for human during the iterations.

### 3.4 Backbone Select

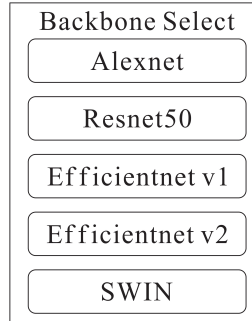


Fig. 2: The five backbones

We compared the performance of the following five backbones (see Fig. 2) to select the best for each YOLO output category.

**AlexNet** A pioneering architecture in CNNs, AlexNet’s eight-layer design[6], featuring rectified linear units (ReLUs) and overlapping max-pooling, achieved groundbreaking success in the ImageNet competition. In this study, AlexNet serves as a benchmark for evaluating subsequent architectures.

**ResNet50** Addressing the challenge of vanishing gradients, ResNet50 introduced residual connections[5], enabling training of exceptionally deep networks. Its use of residual blocks enhances feature learning and prevents overfitting, making it valuable for capturing intricate patterns.

**EfficientNet V1** Revolutionizing neural architecture design, EfficientNet V1 scales network width, depth, and resolution systematically to achieve state-of-the-art performance while maintaining computational efficiency[18]. Its balance between model complexity and cost ensures high accuracy within resource constraints.

**EfficientNet V2** An evolution of EfficientNet V1, V2 incorporates novel improvements like refined scaling methods and advanced training techniques[17]. It surpasses previous benchmarks while maintaining computational efficiency, making it ideal for cutting-edge performance within resource constraints.

**SWIN Transformer** A novel approach tailored for visual recognition tasks, the Swin Transformer divides high-resolution images into non-overlapping patches[8], facilitating efficient processing and capturing long-range dependencies. Its hierarchical design and self-attention mechanisms align well with the objectives of our research, making it a suitable choice for analyzing complex visual patterns in anime images.

### 3.5 Evaluation Metrics

**True Positives (TP)** True Positives represent the number of correctly predicted positive observations.

**True Negatives (TN)** True Negatives represent the number of correctly predicted negative observations.

**False Positives (FP)** False Positives represent the number of incorrectly predicted positive observations.

**False Negatives (FN)** False Negatives represent the number of incorrectly predicted negative observations.

**Accuracy** Accuracy is the ratio of correctly predicted observations to the total observations. It measures the overall correctness of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

**Precision** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It measures the accuracy of positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

**Recall** Recall (also known as sensitivity) is the ratio of correctly predicted positive observations to all the observations in the actual class. It measures the ability of the model to find all the relevant cases within a dataset.

$$Recall = \frac{TP}{TP + FN}$$

**Specificity** Specificity (also known as the true negative rate) measures the proportion of actual negatives that are correctly identified.

$$Specificity = \frac{TN}{TN + FP}$$

**F1 Score** F1 Score is the weighted average of Precision and Recall. It considers both false positives and false negatives. F1 Score is best if there is some sort of balance between Precision and Recall.

$$F1\_Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

## 4 Results

**Backbones** As shown in the figure 3, after weighing the pros and cons, we finally chose EfficientNet V2 to be the backbone for original full images for its good performance after several iterations. For head only dataset, EfficientNet V1 was applied for its coexistence of high accuracy and stability in low iteration. Although it seems to perform worse than EfficientNet V2 in the last several epoch, the overall effect of EfficientNet V1 in this dataset is worth considering. Furthermore, we picked EfficientNet V2 to apply to no head images for its outstanding performance of the best epoch.

**Weights** The trained weights can be seen in the following table 1. It was noticed that the head-only model weighted nearly two times of others, verifying the feasibility of testing a picture with respective of faces.

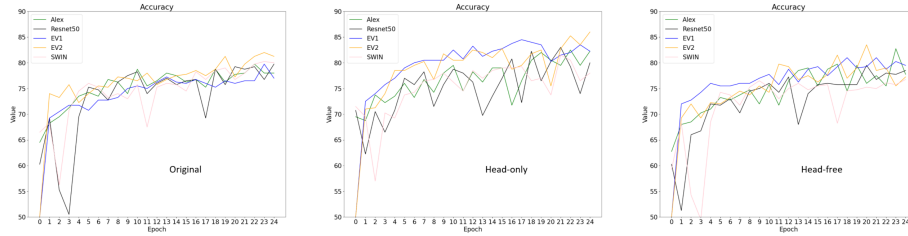


Fig. 3: Performance of different models through epochs

Table 1: Weights of three models for different type of datasets.

Dataset	Model	Weight
Original	EfficientNet V2	0.2931150682739578
Head-only	EfficientNet V1	0.4396615789134529
Head-free	EfficientNet V2	0.26722335281258935

**Model Performance** Table 2 provides a succinct overview of the performance metrics, including accuracy, precision, recall, specificity, and F1 score, achieved by each model. Notably, our model stands out with remarkable performance across all metrics, showcasing its superiority over the comparative architectures.

With an accuracy of 0.873, precision of 0.882, recall of 0.867, specificity of 0.867, and an F1 score of 0.874, our model outperforms its counterparts in terms of predictive accuracy, sensitivity, and specificity. This superior performance underscores the effectiveness and robustness of our model in accurately classifying data instances, thus highlighting its potential for practical applications in various domains.

Furthermore, the consistent and impressive performance of our model across multiple evaluation metrics reaffirms its reliability and generalizability, indicating its suitability for real-world deployment scenarios where accuracy and robustness are paramount.

Table 2: Performance Metrics of Different Models

Model	Accuracy	Precision	Recall	Specificity	F1 Score
AlexNet	0.818	0.827	0.800	0.835	0.813
ResNet50	0.815	0.820	0.803	0.826	0.811
EfficientNetV1	0.840	0.809	0.885	0.795	0.845
EfficientNetV2	0.833	0.808	0.870	0.796	0.838
SWIN	0.832	0.821	0.847	0.818	0.834
<b>Our Model</b>	<b>0.873</b>	<b>0.882</b>	<b>0.867</b>	<b>0.867</b>	<b>0.874</b>



**Model Evaluation** The evaluation of our model reveals balanced performance between AI-generated and human-generated images, indicating its reliability and practical applicability. As depicted in Figure 4, the confusion matrix illustrates the distribution of correct and incorrect classifications made by our model. The diagonal elements represent the instances where the predicted class matches the true class, while off-diagonal elements indicate misclassifications.

Furthermore, Table 3 provides a detailed overview of the performance metrics for our model across AI and human classes. With an accuracy of 0.873 for both classes, our model demonstrates robustness and consistency in its predictions. The precision, recall, specificity, and F1 score metrics further corroborate the balanced performance of our model, with comparable values achieved for both AI and human classes.

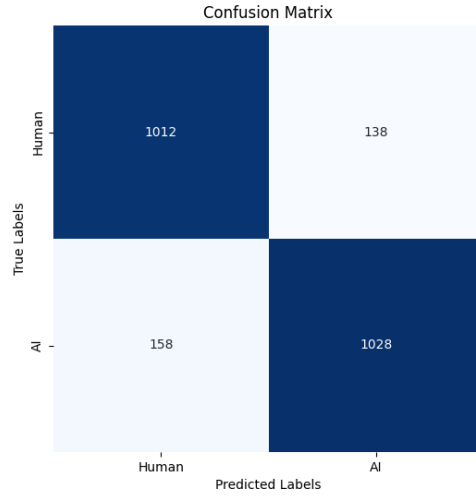


Fig. 4: Confusion matrix illustrating the performance of our model.

Table 3: Performance Metrics of Our Model

Class	Accuracy	Precision	Recall	Specificity	F1 Score
AI	0.873	0.882	0.867	0.867	0.874
Human	0.873	0.865	0.880	0.880	0.872

## 5 Conclusion

In this study, we have explored a novel approach to detecting AI-generated anime images, focusing on the specific region of anime faces. Leveraging YOLOv5-anime for face detection and a combination of convolutional neural networks (CNNs) and transformers as backbones for AI detection, we have developed an ensemble model capable of discerning between AI-generated and human-generated content with remarkable accuracy. By integrating the outputs of these backbones through a weighted combination, our model achieves superior performance compared to established architectures across a range of evaluation metrics.

Our findings not only contribute to the advancement of AI detection techniques but also underscore the importance of responsible utilization of AI-generated content. As AI technologies continue to evolve and proliferate, ensuring the authenticity and integrity of digital content becomes increasingly crucial. Our model represents a step forward in addressing this challenge, offering a reliable means of identifying AI-generated content, particularly in the context of anime images.

Moving forward, there are several avenues for further research and improvement. Refinements in model architecture, optimization of model combinations, and exploration of diverse datasets can enhance the robustness and generalizability of AI detection systems. Additionally, ongoing efforts in raising awareness about the implications of AI-generated content on creativity, ethics, and society are essential for fostering responsible and equitable utilization of AI technologies.

Overall, our study highlights the potential of AI detection methodologies in addressing emerging challenges associated with AI-generated content and underscores the importance of continued research in this domain to ensure the ethical and responsible deployment of AI technologies in society.

## 6 Contribution

Both authors contributed equally to the overall project. While both authors collaborated on various aspects of the research, Xiaotong primarily focused on the YOLO model, whereas Yi'an focused on the backbone models. Both worked on the final weight model training.

## References

1. Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S., Sun, L.: A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. arXiv preprint arXiv:2303.04226 (2023)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., Verdoliva, L.: Are gan generated images easy to detect? a critical analysis of the state-of-the-art (2021)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 770–778 (2016)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS) (2012)
7. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) (2021)
9. Mandelli, S., Bonettini, N., Bestagini, P., Tubaro, S.: Detecting gan-generated images by orthogonal training of multiple cnns (2022)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
11. Ramesh, A., Goyal, A., Krishnan, R., Chaudhari, P., Jain, V.: Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092 (2021)
12. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
13. Ricker, J., Damm, S., Holz, T., Fischer, A.: Towards the detection of diffusion model deepfakes (2024)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2015)
16. Tan, C., Zhao, Y., Wei, S., Gu, G., Wei, Y.: Learning on gradients: Generalized artifacts representation for gan-generated images detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12105–12114 (2023)
17. Tan, M., Chen, R., Pang, B., Vasudevan, V., Sandler, M., Howard, A., Le, Q.V.: Efficientnetv2: Smaller models and faster training. arXiv preprint arXiv:2104.00298 (2021)

18. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114 (2019)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
20. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now (2020)
21. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
22. Zhang, C., Zhang, C., Zhang, M., Kweon, I.S.: Text-to-image diffusion model in generative ai: A survey. arXiv preprint arXiv:2303.07909 (2023)
23. Zhou, Y., Sun, J., Zhang, W.: A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1801.10307 (2018)