American Research Journal of Computer Science and Information Technology

Volume 7, Issue 1, 63-67 Pages
Research Article | Open Access
ISSN (Online)- 2572-2921
DOI : 10.21694/2572-2921.24014

AMERICAN RESEARCH JOURNALS
An Academic Publishing House

Pretrained Diffusion Models for Image Segmentation

Gainetdionov Ainur Fanurovich
Lead ML Engineer, Ticket to the moon, Inc, Moscow, Russia.

ABSTRACT

This article discusses the use of pre-trained diffusion models for image segmentation. The study aims to increase the accuracy of segmentation and reduce required training data by using diffusion methods. During the work, models based on noise generation and removal were used, which allows for improvement in the quality of segmentation masks. The methodology includes the use of U-Net and Transformer algorithms, which contribute to the creation of high-precision segmentation masks of objects in images. The results showed that the use of pre-prepared models significantly improved the accuracy and consistency of segmentation masks compared to traditional methods. In conclusion, it is noted that the proposed approach provides higher segmentation efficiency without requiring extensive data, which opens up prospects for further technology development in various fields..

KEYWORDS: *diffusion models, image segmentation, U-Net, Transformer, segmentation accuracy.*

INTRODUCTION

In recent years, image segmentation has become one of the key tasks in the field of computer vision. This task involves dividing an image into several meaningful parts, which allows for a more precise analysis of objects and their relationships. Traditional segmentation methods, such as thresholding or boundary-based techniques, demonstrate limited accuracy when dealing with complex or heterogeneous data. As a result, the development of new methods capable of providing higher accuracy and stability in image segmentation has become increasingly important.

One of the modern solutions proposed to improve segmentation quality is the use of diffusion models. These generative-based models not only enable the creation of images but also allow for highly accurate segmentation by transforming noise and utilizing attention maps. The introduction of pretrained diffusion models offers the possibility of accelerating the training process and improving results through the use of pre-existing learned parameters. This approach shows promise for applications in various fields such as medicine, automation, and industry, where rapid and precise image analysis is required.

The relevance of this topic lies in the growing need for efficient image segmentation methods that can handle large datasets and complex images. The application of pre-trained diffusion models can significantly enhance the effectiveness of such processes, making research in this area crucial for the further development of image processing technologies.

This work aims to explore the potential of using pre-trained diffusion models for image segmentation and to evaluate their effectiveness compared to traditional methods.

Principles of Diffusion Models

In recent years, diffusion models such as Glide, DALL-E 2, Imagen, and Stable Diffusion have taken a leading role among artificial intelligence tools for image generation. Creating high-quality images based on textual descriptions is a complex task, as it requires a precise understanding of the text and the ability to generate an image that aligns with the meaning of the description. Diffusion models have proven their effectiveness in addressing this challenge, becoming an indispensable tool in the field of artificial intelligence.

In machine learning, both generative and discriminative approaches are widely used. Generative models, by their nature, are designed to create new data based on existing data. Their task is to transform noise into representative data samples, making them ideal for information synthesis. Generative models are widely applied in various fields, including advertising. For instance, in a famous Cadbury commercial, synchronization between sound and image was achieved. Generative models have also been used to create personalized images of celebrities in advertising campaigns [1].

Among the most well-known image generation models are variational autoencoders, flow-based models, generative adversarial networks, and diffusion models, which are currently at the peak of popularity.

The term "diffusion" originates from the field of nonequilibrium statistical physics and describes the process of particles moving from an area of high concentration to an area of low concentration. In the context of data generation models, this process allows the transformation of noise into representative data using specialized neural networks.

Diffusion models can operate in both unconditional image generation, where the result is a random data sample, and conditional generation, where the model is guided by additional information such as text or class labels. This enables the creation of targeted images that correspond to specific requests [2].

The process of generation using diffusion models consists of two stages: in the first stage, noise is iteratively added to the original image, eventually breaking down its structure. In the second stage, a neural network performs the reverse process—removing the noise and reconstructing the image.

In various modalities, observable data can be viewed as the result of interactions with certain hidden variables, denoted by a random variable z. To illustrate this idea, one can refer to Plato's allegory of the cave. In the allegory, people chained inside a cave see only shadows of objects projected on a wall, which are a result of three-dimensional objects passing in front of a light source. These people perceive only the projections, while the true objects remain hidden and inaccessible to them [3].

Similarly, objects in the real world can emerge as representations of higher-level abstractions that include parameters such as color, shape, or size. The phenomena we observe can be considered a three-dimensional projection of such abstractions, just as the shadows in the cave are projections of objects unseen by the people. Although these hidden objects may be inaccessible for direct perception, they can be approximately studied and inferred based on observable data [4].

From a mathematical perspective, hidden variables and observed data can be described through a joint distribution $p(x, z)$. One approach to generative modeling is to maximize the likelihood of the observed data $p(x)$. This can be achieved in two ways: by marginalizing over the hidden variables z, or by applying the chain rule of probability:

$$p(x)=\int p(x,z)dz$$

$$p(x)= p(x,z) =p(x)\ p(z|x)$$

However, directly calculating the likelihood $p(x)$ may be difficult due to the complexity of integrating over all hidden variables z or the absence of the true distribution $p(z|x)$. In such cases, the evidence lower bound (ELBO) is used, which approximates the maximization of likelihood [5].

A segmentation method is proposed based on the probabilistic noise diffusion model (DDPM), which is capable of generating uncertainty maps for the created segmentation masks. Figures 1 below provide examples of the performance of diffusion models SD2 fine-tuned for the matting task.

Fig. 1. Results of fine-tuning SD2 for the matting task.

Training the DDPM model is carried out using a pair dataset of input images and segmentation masks. Since the sampling process via the DPMS model includes stochastic components at each step, it is possible to generate different segmentation mask variants for the same input image and the same pre-trained model. The resulting segmentation variants allow for the computation of pixel-level variance maps, which quantitatively assess the uncertainty of the created mask. Moreover, averaging several segmentations improves the overall accuracy of the segmentation process [6].

Thus, diffusion models demonstrate high efficiency in generative tasks, offering several advantages such as stable training, the ability to handle large datasets, and applicability in tasks related to data privacy and synthesis.

Comparison of Modern AI-Based Image Segmentation Methods with Traditional Methods

Image segmentation methods can be broadly classified into two main categories. The first category includes traditional approaches that utilize classical computer vision algorithms. The second category is based on methods that apply artificial intelligence.

There are several major groups of image segmentation methods. Semantic segmentation assigns a class to each pixel in the image. For example, all pixels representing people in an image will be labeled with the same class, while

the background will be assigned to a different class. Instance segmentation, on the other hand, focuses on recognizing each object in an image. For instance, every person in the image will be segmented as a distinct object. Panoptic segmentation is a combination of semantic and instance approaches: it assigns a class to each pixel while also distinguishing different instances of the same class.

One of the simplest segmentation methods is thresholding. It is based on selecting a threshold value that converts a grayscale image into a binary image. A critical step here is the correct selection of the threshold value, which can be achieved using various methods such as the maximum entropy method, balanced histogram thresholding, or Otsu's method.

Adaptive Thresholding. To improve segmentation results in images with uneven lighting, adaptive thresholding is used. It divides the image into small zones, applying individual threshold values to each zone, which accounts for local lighting variations and increases segmentation accuracy.

Edge-based segmentation methods allow for detecting object contours by analyzing intensity discontinuities in the image.

Canny Algorithm. The Canny algorithm is a multi-stage process that involves Gaussian filtering for noise reduction, calculating intensity gradients, and edge thinning using non-maximum suppression. Double thresholding and edge tracking complete the process of detecting object contours.

Sobel Operator. The Sobel operator identifies high-frequency areas in an image that may correspond to object boundaries. This method is based on convolving the image with specific kernels designed to detect edge-like regions.

Clustering Methods group pixels based on their characteristics, such as color or texture, into several clusters.

K-means Clustering. This method iteratively assigns pixels to clusters by minimizing intra-cluster variance. Pixels are assigned to the nearest cluster, and cluster centers are recalculated until convergence is achieved.

Mean Shift. Mean shift is a clustering method that shifts pixels towards areas of maximum density in their neighborhood. This approach is effective for handling clusters of arbitrary shapes.

Deep Learning for Image Segmentation. With the advent of deep learning methods, image segmentation has reached a new level. The use of convolutional neural networks (CNNs) enables automatic feature extraction from images, significantly improving segmentation accuracy.

Convolutional Neural Networks (CNN). CNNs form the foundation of many modern segmentation models, extracting hierarchical features from images using convolutional filters [7].

U-Net. The U-Net architecture was designed for biomedical segmentation and combines a contracting path to capture context with an expanding path for precise object localization.

SegNet. SegNet employs an encoder-decoder architecture, allowing efficient real-time image segmentation with low memory consumption.

Fully Convolutional Networks (FCN). FCNs provide end-to-end learning for segmentation tasks by replacing fully connected layers with convolutional ones, allowing images of variable sizes to be processed.

Advanced Architectures: Mask R-CNN. Mask R-CNN extends Faster R-CNN by adding object segmentation, enabling the segmentation of each object instance within an image.

DeepLab. The DeepLab architecture incorporates atrous convolutions to improve spatial resolution, allowing for the segmentation of images with multi-scale contextual information.

PSPNet. PSPNet uses pyramid pooling to understand complex scenes, which is particularly useful when dealing with images containing a large number of objects [8].

Below, Table 1 provides a comparison between traditional segmentation methods and those based on AI algorithms.

Table 1. Comparison of traditional and modern image segmentation methods

Criterion	Traditional Segmentation Methods	AI-based Segmentation Methods (Deep Learning)
Algorithms and Approaches	Based on thresholds, gradients, and morphological operations. Examples: Canny, k-means, active contours.	Use neural networks such as CNN, U-Net, SegNet, and Mask R-CNN for automatic object detection.
Adaptability	Require parameter tuning for each image type, sensitive to changes (lighting, texture).	Adapt to various conditions and images through training on large datasets.
Accuracy and Quality	May suffer from false positives, and lower accuracy on complex images.	Provide high accuracy and detail due to automatic recognition of complex features.
Resource Requirements	Low computational requirements, can be executed on standard systems.	Require significant resources during training (GPU), but can run faster post-training.
Generalization Capability	Limited by predefined rules, perform poorly on new data without adjustment.	High generalization capability, able to apply learned knowledge to new images.
Applicability	Suitable for simple tasks where parameters can be predefined.	Suitable for complex segmentation tasks where flexibility and high accuracy are required.

Thus, modern methods demonstrate high efficiency when working with large datasets and complex scenes, making them preferable for use in AI applications.

Application of Pretrained Diffusion Models in Image Segmentation

The diffusion model, developed for image generation, employs a process of adding and removing Gaussian noise. It modifies traditional approaches by incorporating encoder-decoder structures and the U-Net architecture, which includes self-attention and cross-attention layers embedded in Transformer layers. The algorithm processes images by compressing them into a latent space and then reconstructing them, with the diffusion processes occurring within this latent space.

The U-Net architecture is represented as modular blocks utilizing ResNet and Transformer layers. It is assumed that the self-attention mechanisms within these blocks can detect objects in the image and group them to create segmentation masks without the need for textual input. Attention maps in the model capture semantic relationships and highlight object groupings, with the resolution of these maps influencing the level of segmentation detail.

For the experiments, a pretrained Stable Diffusion 2 model was used, employing its original VAE to encode both the input image and its corresponding depth map into latent space representations. The model's input layer was modified to accept not only latent noise but the concatenation of the noise with the image latent for which a segmentation mask needs to be computed.

Instead of training the entire model, only the U-Net is trained using LoRA. This approach solves the problem of catastrophic forgetting and achieves maximum quality when training on a synthetic dataset. Experiments have shown that with this method, the model generalizes well to realistic domains. Thus, the issues of dataset creation and mask accuracy are resolved, as we can render them programmatically.

This fine-tuning is achieved by optimizing the standard diffusion loss function with respect to the depth latent code. After completing the full schedule of denoising steps, we decode the resulting depth latent code back into an image using the VAE decoder. Finally, by averaging the three color channels of this decoded image, we obtain the final depth estimation.

In experiments with the COCO benchmark, baseline clustering algorithms K-Means-C and K-Means-S were used. Both methods showed significant improvement compared to previous approaches, but the suggested algorithm, which relies on the pretrained SD2, achieved the best results, significantly outperforming both K-Means variants. On the COCO-Stuff-27 test, it demonstrated a 26% increase in accuracy and a 17% improvement in mood compared to the ReCo method, while in the Cityscapes segmentation task, it achieved results that surpassed previous works.

A real dataset with matting suffers from inaccurate mask annotations caused by the ambiguity in labeling small or blurred details. Unlike previous works that use real datasets for generalization, the proposed model is trained solely on synthetic data. Synthetic data provides a cleaner gradient for denoising, and because fine-tuning SD2 on synthetic data generalizes to the real domain, we achieve a state-of-the-art solution.

Experiments on the COCO and Cityscapes datasets demonstrated the superiority of the suggested method over traditional algorithms, confirming its potential in image segmentation tasks with improved accuracy and more scores.

CONCLUSION

Thus, the proposed approach, based on generative methods of noise transformation, significantly enhances the accuracy and coherence of segmentation masks. The use of the U-Net architecture and Transformer mechanisms allows for the effective creation of attention maps, contributing to more detailed and precise object segmentation. The advantages of pre-trained models include reduced training time and improved results compared to traditional segmentation methods. An important conclusion is the models' ability to operate without real data, making them particularly promising for tasks requiring fast and accurate image analysis. Future research prospects involve expanding the application of these models across various industries and optimizing them for working with more complex visual data.

REFERENCES

1. Ho J. and others. Cascade diffusion models for high-precision image generation //Journal of Machine Learning Research. – 2022. – Vol. 23. – No. 47. – pp. 1-33.
2. Ahmad H. and others. A new method for analyzing nonlinear equations of the Cauchy reaction model with a fractional time distribution //Results in physics. – 2020. – vol. 19. – p. 103462.
3. Liu N. et al. Compositional visual generation using composable diffusion models //European Conference on Computer Vision. – Cham: Springer Nature Switzerland, 2022. – pp. 423-439.
4. Yu. Yu. et al. Methods and problems of image segmentation: an overview //Electronics. – 2023. – vol. 12. – No. 5. – p. 1199.
5. Wallab J. et al. Diffusion models for implicit image segmentation ensembles //International Conference on Medical Imaging with Deep Learning. – PMLR, 2022. – pp. 1336-1348.
6. To the auditor F. A. et al. Diffusion models in vision: a review //IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2023. – vol. 45. – No. 9. – pp. 10850-10869.

7. Malhotra P. et al. [Deleted] Deep neural networks for segmentation of medical images //Journal of Healthcare Engineering. – 2022. – T. 2022. – No. 1. – p. 9580991.

8. Minai S. et al. Image segmentation using deep Learning: an overview //IEEE transactions on pattern Analysis and machine intelligence. – 2021. – vol. 44. – No. 7. – pp. 3523-3542.

9. Guo H. et al. Acceleration of diffusion models using diffusion sampling before segmentation for segmentation of medical images //2023 20- The 1st IEEE International Symposium on Biomedical Imaging (ISBI). – IEEE, 2023. – pp. 1-5.

10. Tian J. et al. Diffuse presence and segmentation: uncontrolled zero-interval segmentation using stable diffusion //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2024. – pp. 3554-3563.

11. Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, Chunhua Shen. DiffuMask: Synthesizing Images with Pixel-level Annotations for Semantic Segmentation Using Diffusion Models. arXiv:2303.11681, 2023.

Citation: Gainetdionov Ainur Fanurovich, "Pretrained Diffusion Models for Image Segmentation", American Research Journal of Computer Science and Information Technology, Vol 7, no. 1, 2024, pp. 63-67.