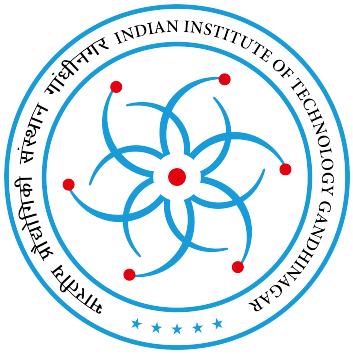
Indian Institute of Technology Gandhinagar
Computer Vision, Imaging, and Graphics (CVIG) Lab



CS 399 Project Course Report Week - 1

INDIAN INSTITUTE OF TECHNOLOGY GANDHINAGAR
Palaj, Gandhinagar - 382355

Diffusion & Segmentation

Submitted by

Guntas Singh Saran
Computer Science and Engineering (22110089)

Aadya Arora
Electrical Engineering (22110002)

Under the guidance of

Prof. Shanmuganathan Raman
Jibaben Patel Chair in Artificial Intelligence and Associate Professor
Electrical Engineering & Computer Science and Engineering, IIT Gandhinagar

Prajwal Singh
Ph.D. Scholar
Computer Science and Engineering, IIT Gandhinagar

Contents

Chapter 1

Introduction

1.1 Objective

The primary objective of this research to be able to infer 3D structure from 2D images using view-conditioned diffusion model [1] using minimal or optimal camera estimations. To better our understanding of view-based diffusion, we will explore [2, 3]. Further, we wish leverage the intersection of Segmentation with Text or Image Representation to help with this task.

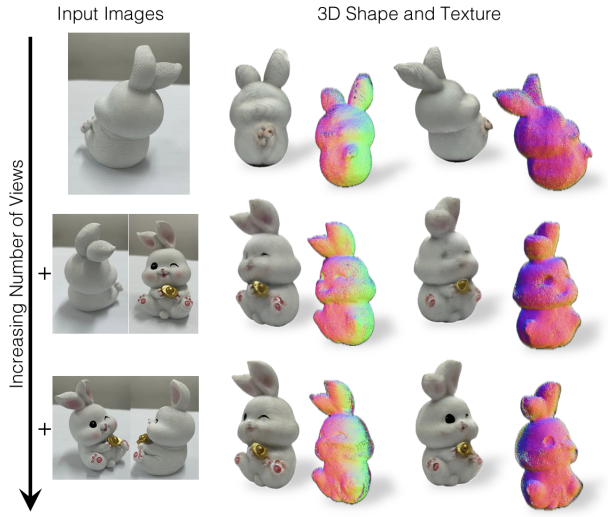


Figure 1.1: Adapted From [1].

In order to make ourselves well versed with Natural Language aspect of this project and the idea of semantic segmentation on even synthetic data, we start off with the understanding of Vision Transformers (ViT) [4] and the role that Attention Maps [5] play in their learning. Then we move onto understanding how models like CLIP [6] are able to use both the text embeddings using a transformer and image embeddings using a vision transformer to help generate captions for images. And how these can be few-shot learned catered to our task.

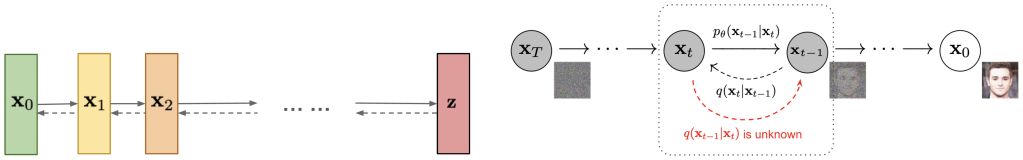
To understand segmentation, we explore models like SAM [7] and SAM-2 [8]. To study on semantic segmentation and investigate efficient learning techniques using synthetic data generated by generative models, we will refer to this [9].

Chapter 2

Literature Review

2.1 Diffusion Probabilistic Models

Diffusion Probabilistic Models (DPMs) are generative models that reverse a gradual noise-adding process to generate data. They are particularly effective in generating high-quality images with fine details.



(a) The Diffusion Process. Source [10] (b) Forward and Reverse Distributions. Source [11]

Figure 2.1: Diffusion Probabilistic Models

J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020. [11] This paper introduces the concept of DPMs and demonstrates their effectiveness in generating high-fidelity images by progressively denoising a noisy signal.

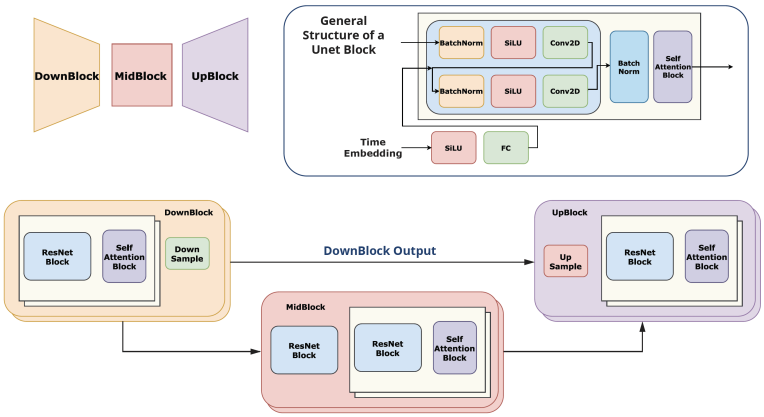


Figure 2.2: UNet Blocks used in DMs.

2.2 Vision Transformers

The paper *"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"* [4] introduces the Vision Transformer (ViT), a novel approach to image classification that adapts the

transformer architecture, originally used in natural language processing (NLP), to computer vision tasks. The key idea is to treat images as sequences of patches, similar to how sequences of words are processed in NLP.

ViT achieves competitive performance on large-scale datasets such as ImageNet, CIFAR-100, and VTAB by pre-training on large datasets and transferring the knowledge to mid-sized and small benchmarks. This approach challenges the conventional use of convolutional neural networks (CNNs) in computer vision, demonstrating that pure transformers can perform as well or better than CNN-based architectures with fewer computational resources.

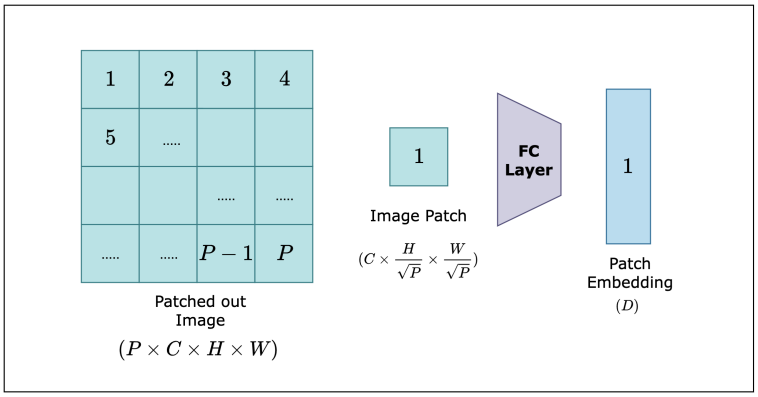


Figure 2.3: Patch Embeddings in Vision Transformers.

2.2.1 Attention Mechanism Using Patch Embeddings

In Vision Transformers, the input image is divided into fixed-size patches (e.g., 16x16 pixels), which are flattened and linearly embedded into a sequence of vectors, akin to words in NLP. These patch embeddings are passed through the transformer model, where the self-attention mechanism is employed to capture long-range dependencies between different patches.

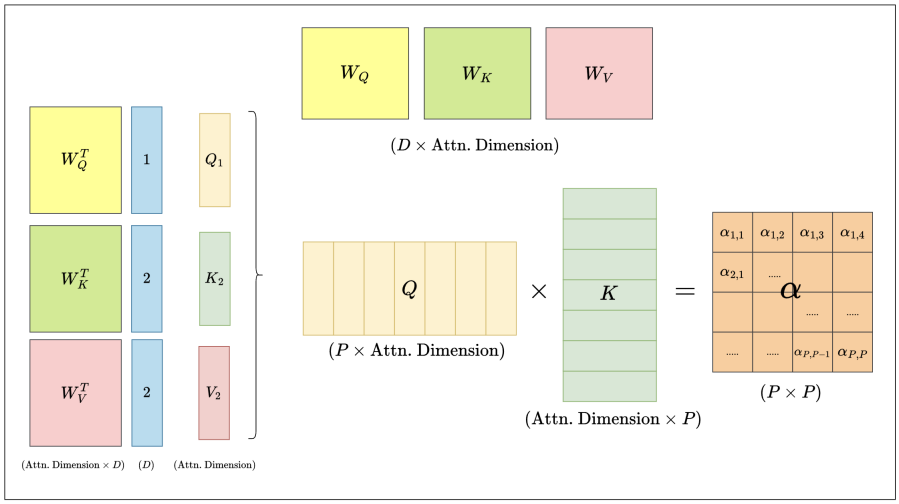


Figure 2.4: The Query, Key, and Value Matrices as Attention Maps.

2.2.1.1 Query, Key, Value Matrices

The attention mechanism operates by constructing three matrices from the input patches: **Query (Q)**, **Key (K)**, and **Value (V)** matrices. These are computed as follows:

- **Query (Q)**: Input Representation. How much is each context item relevant to input.
- **Key (K)**: Context Representation. Quantifies the relevance to the query representation.
- **Value (V)**: Context representation which will be used to add the understanding of relevant context to the input representation/query.

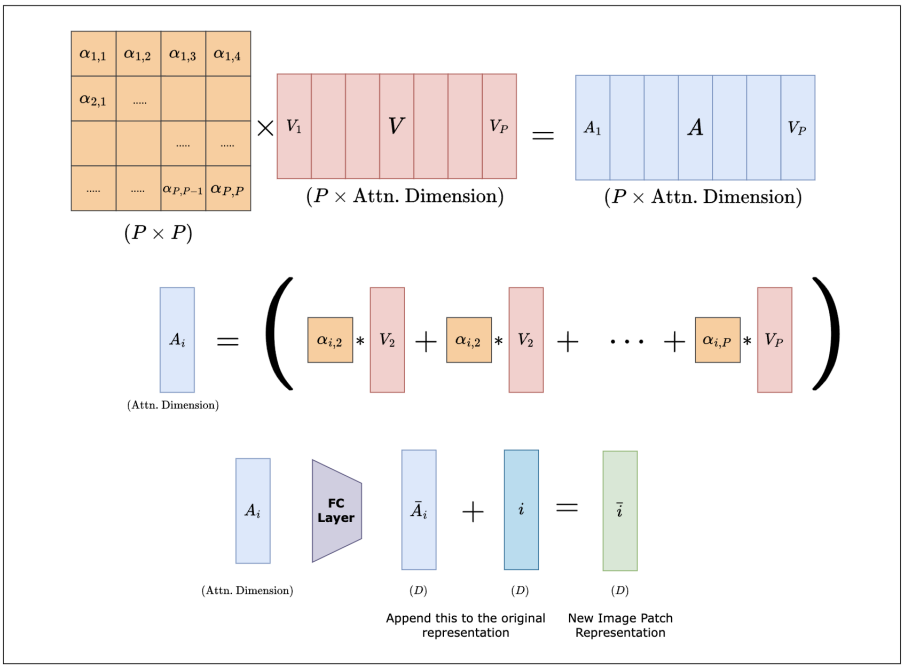


Figure 2.5: Adding the understanding of the relevant context to the input representation.

Thus, the transformer model is able to focus on different parts of the image by adjusting the attention weights based on the patch embeddings.

2.3 Constrastive Learning CLIP & Segment Anything SAM

The integration of natural language and computer vision has significantly advanced the development of visual models, with two influential approaches being the *CLIP* model and the *Segment Anything Model (SAM)*.

The paper *"Learning Transferable Visual Models From Natural Language Supervision"* [6] introduces CLIP, a model that leverages raw text about images as a supervision signal to train scalable visual models. CLIP is pre-trained on a massive dataset of 400 million image-text pairs collected from the internet. The model learns general image representations by predicting which caption corresponds to which image. This allows for zero-shot transfer to various downstream tasks without requiring task-specific training. CLIP demonstrates competitive performance across over 30 different datasets, covering tasks such as object classification, optical character recognition (OCR), and action recognition. Notably, CLIP matches the performance of ResNet-50 on ImageNet in a zero-shot setting, highlighting its generalization ability without the need for labeled data.

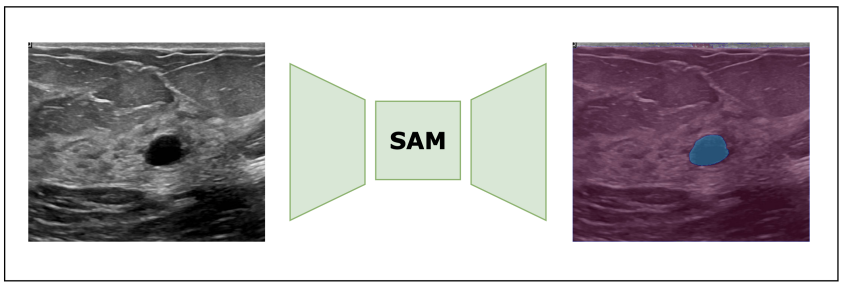


Figure 2.6: The Segment Anything Model.

The *"Segment Anything"* [7] project focuses on image segmentation by developing the largest segmentation dataset to date, with over 1 billion masks across 11 million images. SAM is designed to be *promptable*, allowing it to generalize to new image distributions and tasks in a zero-shot manner. The model's zero-shot performance is competitive with fully supervised approaches across a range of segmentation tasks. SAM's adaptability to different tasks without fine-tuning positions it as a powerful tool for segmentation in diverse contexts. The release of the model and dataset fosters further research into foundational models for computer vision.

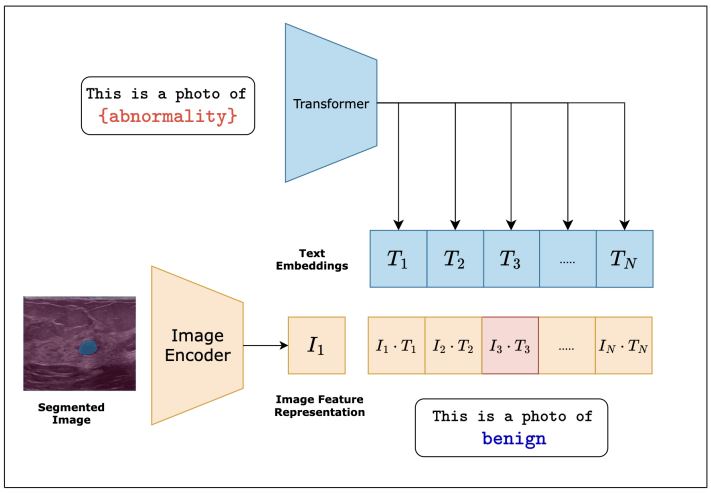


Figure 2.7: Zero-shot in CLIP gives better result with segmented image to ViT.

Both CLIP and SAM demonstrate the effectiveness of scalable pre-training and zero-shot transfer capabilities. These models emphasize the shift towards foundational models capable of generalizing across multiple vision tasks without requiring extensive, task-specific labeled datasets.

2.4 Upcoming Reviews

As we progress further in our understanding of 3D object reconstruction from 2D data, the next step will involve exploring a series of state-of-the-art papers that will significantly enhance our knowledge of neural radiance fields (NeRFs) and diffusion-based models. This includes papers like Viewset Diffusion [3], which focuses on generating consistent 3D models using diffusion methods trained with multi-view 2D data, and DiffRF [12], which extends NeRF by incorporating rendering loss to guide the generation of accurate 3D radiance fields. Both papers emphasize how the integration of 2D images can be used to construct more precise and efficient 3D models.

In addition to these, papers such as SyncDreamer [2] and Keeping Segment Mask Quality with Self-generated Masks [9] will provide valuable insights into the consistent generation of multiview images and the optimization of mask segmentation quality. These explorations will culminate in an in-depth review of the key ideas presented in *"The More You See in 2D, the More You Perceive in 3D"*.

Bibliography

[1] X. Han, Z. Gao, A. Kanazawa, S. Goel, and Y. Gandelsman, "The more you see in 2d, the more you perceive in 3d," 2024.

[2] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang, "Syncdreamer: Generating multiview-consistent images from a single-view image," 2024.

[3] S. Szymanowicz, C. Rupprecht, and A. Vedaldi, "Viewset diffusion: (0-)image-conditioned 3d generative models from 2d data," 2023.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.

[8] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," 2024.

[9] S. Mae, R. Yamada, and H. Kataoka, "Keeping segment mask quality with self-generated masks," in *ECCV 2024 Workshop The Dark Side of Generative AIs and Beyond*, 2024.

[10] L. Weng, "What are diffusion models?," *lilianweng.github.io*, Jul 2021.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[12] N. Müller, Y. Siddiqui, L. Porzi, S. R. Bulò, P. Kontschieder, and M. Nießner, "Diffrf: Rendering-guided 3d radiance field diffusion," 2023.

[13] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," 2022.

[14] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "Multidiffusion: Fusing diffusion paths for controlled image generation," 2023.