smartChef: a virtual chef that integrates segments and substitutes meal components in images

Team Members: Kuanmin Wang Yuhang Chen Rui Zhou Jingmin Xu

Mentor TA: Nathan DePiero

Date Apr 17, 2025

Introduction

What problem are you trying to solve and why?

- If you are implementing an existing paper, describe the paper's objectives and why you chose this paper.
- If you are doing something new, detail how you arrived at this topic and what motivated you.
- What kind of problem is this? Classification? Regression? Structured prediction? Reinforcement Learning? Unsupervised Learning? Etc.

Potential:

E-commerce: As the e-commerce market continues to grow, online stores need high-quality visual content to present their products and services in a way that attracts customers. However, many small entrepreneurs do not have enough time, money, or design skills to create professional advertisements. Our product provides a simple and affordable way for these small business owners to make appealing visuals. We use Mask R-CNN to automatically find and segment objects in an image, and then apply a diffusion-based inpainting model to change or replace these objects based on what the user wants. From our perspective, this is going to allow entrepreneurs to create attractive marketing images quickly and easily. It can help them save time, lower costs, and improve their online visibility, even without any design experience.

Menu visualization: To cater to the growing customer demand for customization, restaurants keep expanding their menu options. The format of combos is always a popular way to meet this kind of need. However, menu visualization becomes a serious problem for small-scale restaurants. While large chain restaurants can afford to take professional advertisement photos for each combo, smaller restaurants are struggling to deal with the high costs of numerous combo combinations. Our model offers a low-cost solution, empowering small restaurants with more flexible options. By taking a single picture of a combo, these restaurants can seamlessly replace food items with alternatives. The model could ensure that the replacement still maintains high resolution and visual authenticity. We believe our model could support local businesses and also enrich the menu visualization experience for customers.

Our approach draws inspiration from two key papers:

1. **Mask R-CNN (He et al., 2017)** – This paper introduces a powerful instance segmentation framework that detects objects in an image and generates a

high-quality segmentation mask for each. We chose Mask R-CNN because it allows us to isolate and recognize individual food items in a complex scene, which is essential for selective modification.

2. **DreamShaper Inpainting with Diffusion Models (Lykon)** – This diffusion-based model is capable of high-quality image inpainting. We use it to generate visually coherent replacements for unhealthy food, once these are masked by the segmentation module. The model enables seamless blending and realistic substitutions that match the image's context.

This is a **structured prediction** problem combined with **image generation**. Specifically:

- The segmentation and classification of food items represent a **structured prediction** task.

- The replacement of items using a diffusion model is part of a **generative modeling** task, specifically using **inpainting** techniques.

Related Work

Are you aware of any, or is there any prior work that you drew on to do your project?

- Please read and briefly summarize (no more than one paragraph) at least one paper/article/blog relevant to your topic beyond the paper you are re-implementing/novel idea you are researching.
- In this section, also include URLs to any public implementations you find of the paper you're trying to implement. Please keep this as a "living list"–if you stumble across a new implementation later down the line, add it to this list.

In addition to the two core papers, our project is based on **Mask R-CNN** (He et al., 2017) for the object segmentation task and **DreamShaper v8 Inpainting** (DreamShaper on Hugging Face) for diffusion-based image editing. Since DreamShaper v8 Inpainting is based on the Stable Diffusion Inpainting model, we also refer to the original project to help us determine the accuracy metrics.

We explored other relevant literature and public implementations to deepen our understanding. One related paper is **FoodSAM: Any Food Segmentation**. This work introduces a zero-shot framework that extends the Segment Anything Model (SAM) for robust semantic, instance, and panoptic segmentation of food images. The authors

combine coarse semantic masks with SAM's category-agnostic predictions to achieve precise semantic segmentation, closely aligning with our goal of accurately identifying and segmenting food items in complex scenes. The paper also highlights the limitations of existing segmentation models in handling diverse food images, a challenge we aim to address using fine-grained, instance-level segmentation techniques.

Public Implementations

- Mask R-CNN (Matterport implementation): https://github.com/matterport/Mask_RCNN
- Detectron2 (by Facebook AI Research): https://github.com/facebookresearch/detectron2
- DreamShaper v8 Inpainting: https://huggingface.co/Lykon/dreamshaper-8-inpainting
- Stable Diffusion model: https://huggingface.co/stabilityai/stable-diffusion-2-inpainting

Data

What data are you using (if any)?

- If you're using a standard dataset (e.g. MNIST), you can just mention that briefly. Otherwise, say something more about where your data come from (especially if there's anything interesting about how you will gather it).
- How big is it? Will you need to do significant preprocessing?

MASK R-CNN Model: The dataset we are going to fine-tune the Mask R-CNN model on is the UNIMIB2016 Food Database, which consists of 1,027 tray images with multiple foods. Each image is segmented and labeled with the corresponding food categories. It adds up to a total of 73 unique classes. We don't need a significant data preprocessing process because images are relatively consistent in content and image structure.

Inpainting: The pre-trained model we are using for inpainting is Dreamshaper 8 Inpainting. It is based on the Stable Diffusion Inpainting model that was trained on the LAION-5B dataset. The dataset contains 5,85 billion high-quality and diverse images.

Methodology

What is the architecture of your model?

- How are you training the model?
- If you are implementing an existing paper, detail what you think will be the hardest part about implementing the model here.
- If you are doing something new, justify your design. Also, note some backup ideas you may have to experiment with if you run into issues.

Our system consists of two major components: **(1) object instance segmentation**, and **(2)user prompting image inpainting using a diffusion model**. The architecture combines the strengths of Mask R-CNN and latent diffusion-based image editing to enable the identification and replacement of unhealthy food items in images.

Model Architecture

1. **Segmentation Module – Mask R-CNN**:
   We use **Mask R-CNN** as the backbone for instance-level segmentation. This model will detect and segment individual food items in an input image. We will fine-tune a pre-trained Mask R-CNN (e.g. via Detectron2) on our food dataset (UNIMIB2016) to allow the model to recognize and localize specific food types.

2. **Replacement Module – Diffusion Inpainting (DreamShaper v8)**:
   For any food items classified as "unhealthy," we mask them out and use a **latent diffusion model** (e.g., DreamShaper Inpainting) to generate healthier substitutes. The prompt used for inpainting will include context-aware instructions like “replace hamburger with salad.”

Training Strategy

- **Segmentation**: Fine-tune Mask R-CNN on the UNIMIB2016 dataset with instance-level annotations. If needed, convert or create mask annotations from bounding boxes.

- **Inpainting**: Leverage pre-trained diffusion models with prompt engineering; no training required, but extensive testing and tuning of prompts will be done.

Backup Plans: If the diffusion-based model produces low-quality or irrelevant replacements, we’ll explore **ControlNet-based guided inpainting** or fallback to a fixed set of pre-rendered healthy food templates blended using image processing techniques.

Metrics

What constitutes “success?”

- What experiments do you plan to run?
- For most of our assignments, we have looked at the accuracy of the model. Does the notion of “accuracy” apply for your project, or is some other metric more appropriate?
- If you are implementing an existing project, detail what the authors of that paper were hoping to find and how they quantified the results of their model.
- If you are doing something new, explain how you will assess your model’s performance.
- What are your base, target, and stretch goals?

We can fine-tune Mask R-CNN (Detectron2) on UNIMIB2016 with instance segmentation labels, and then evaluate **mIoU accuracy**. We can run experiments about evaluating replacement quality across different prompts by humans.

Base: Inpainted images pass human plausibility checks in >70% of cases.
We are going to use human evaluation to test this. Human evaluator needs to answer questions like “Does the edited image look real?”“Can you tell which region was modified?”. Then we are going to calculate the fooling rate to calculate how often users think the edited image is real.

Segmentation achieves ≥70% mIoU across common food types. (AP@0.7 The evaluation metric for the paper is AP@[0.5:0.95], which means Average precision across IoU 0.5–0.95.)
Target:
Inpainted images pass human plausibility checks in >80% of cases.
Segmentation achieves ≥80% mIoU across common food types. (AP@0.8)

Stretch:
Inpainted images pass human plausibility checks in >90% of cases.
Segmentation achieves ≥90% mIoU across common food types. (AP@0.9)

Ethics

Choose 2 of the following bullet points to discuss; not all questions will be relevant to all projects so try to pick questions where there's interesting engagement with your project. (Remember that there's not necessarily an ethical/unethical binary; rather, we want to encourage you to think critically about your problem setup.)

- What broader societal issues are relevant to your chosen problem space?
- Why is Deep Learning a good approach to this problem?
- What is your dataset? Are there any concerns about how it was collected, or labeled? Is it representative? What kind of underlying historical or societal biases might it contain?
- Who are the major "stakeholders" in this problem, and what are the consequences of mistakes made by your algorithm?
- How are you planning to quantify or measure error or success? What implications does your quantification have?
- Add your own: if there is an issue about your algorithm you would like to discuss or explain further, feel free to do so.

What broader societal issues are relevant to your chosen problem space?

Our project addresses broader societal issues related to small business accessibility and digital inclusion. By enabling automated visual editing of food and product images, our system empowers individuals and small businesses, particularly those lacking technical or design expertise, to create high-quality content for healthier lifestyle promotion or competitive online presence. It also supports local restaurants and entrepreneurs in overcoming the financial and logistical barriers of traditional professional marketing.

Why is Deep Learning a good approach to this problem?

Deep learning works well for this problem because it can understand complex images. We could implement or develop a more advanced system to solve the existing problems by leveraging this characteristic. Our system uses Mask R-CNN to separate and identify objects in a photo. We later use Diffusion models to help us change or replace food items in a way that looks natural and matches the style of the image. These models work together to make changes that are both accurate and realistic. It is the task that something traditional image editing tools cannot do, or take lots of effort to achieve a satisfactory solution.

Division of labor

Briefly outline who will be responsible for which part(s) of the project.

Rui Zhou: Fine-tuning the Mask R-CNN, (a healthy/unhealthy classification), applying inpainting model, system integration

Jingmin Xu: Data Preprocessing, paper writing

Yuhang Chen and Kuanmin Wang: inpainting tests, metrics evaluation and paper writing