[CS550]-MACHINE LEARNING:
SUPER RESOLUTION IMAGE GENERATION USING PROMPT AND SRGAN

Siddhi Agarwal
ID: 12141570

Sudeep Ranjan Sahoo
ID: 12141600

Konduri Naga Lakshmi Rekha
ID: 12140930

ABSTRACT

Generative Adversarial Networks (GANs) in supervised settings can generate photo-realistic corresponding output from low-definition input (SRGAN). Using the architecture presented in the SRGAN, we explore how providing a prompt affects the outcome by using modified datasets of images and human-generated prompts to see that SRGAN fundamentally learns objects, with their shape, color, and texture, and redraws them in the output rather than merely attempting to sharpen edges.

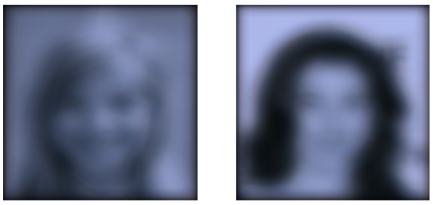
1 INTRODUCTION AND PROBLEM MOTIVATION

Introduction:
Generative Adversarial Networks (GANs) revolutionize computer vision, especially in Single Image SuperResolution (SR). SRGANs enhance low-res images, transcending edge sharpening to capture shape, color, and texture nuances. They find applications in surveillance, entertainment, and medical imaging. A pivotal aspect is the training dataset, shaping the network's ability to create precise, high-res images. SRGANs, thus, grasp and replicate object features, not just edges. This innovation holds immense potential across industries.

Previous report summary:

- In our prior tries to attain desired results, we encountered the paper titled "Photo-Realistic Single Image Super-Resolution." Our approach sought to tackle this challenge through the introduction of a user-guided enhancement system. Users would submit a low-resolution image accompanied by specific prompts outlining the desired enhancements. This composite input would then be fed into an SRGAN model with an attention mechanism. However, the implementation of the attention model proved challenging, as each epoch took 1-2 hours, making it impractical to execute such an extensive codebase with the available resources.
- We moved on to conduct research on other possibilities and conducted a literature review of three more papers titled:
  1. *Imagic: Text-Based Semantic Image Editing with Real Images* by Siddhi and Rekha
  2. *Prompt-to-Prompt Image Editing with Cross Attention Control*
  3. *Composing Text and Image for Image Retrieval - An Empirical Odyssey* by Sudeep

  After numerous experiments and implementations, we ultimately decided to implement the paper titled *Composing Text and Image for Image Retrieval - An Empirical Odyssey*. In this



paper, we explore the task of image retrieval, where the input query is specified in the form of an image along with some text describing desired modifications to the input image. In the future, we plan to decode the composite image to generate the image we require.

2 IMPLEMENATION OF TIRG

The initial task involved the implementation of the TIRG model, which presented several challenges. The model was originally developed in Python 2, posing compatibility issues with newer frameworks and advanced models like SRGAN and encoder models. Undeterred, we initiated the implementation in Python 3, carefully studying the code structure and implementation details.

To gain a deeper understanding, we cloned the repository mentioned in the paper, attempting various modifications to adapt it to our needs. However, this process was not without obstacles, as adjustments were needed to make the code compatible with modern Python environments. The initial attempts at modification involved addressing issues related to the calculation of loss and the handling of the backward pass for gradient computation. In our pursuit of understanding the model's

Figure 1: Working of the TIRG Model

intricacies, we experimented with alternative approaches, tweaking the mechanisms for loss computation and gradient backpropagation. Although we did not complete the training process due to the extensive time required for each epoch, this preliminary exploration provided valuable insights into the functioning of the TIRG model.

Building upon this foundation, we drew inspiration from the TIRG model and integrated its principles into our final approach, which focused on the composition of text and image embeddings. This incorporation of the TIRG model's insights played a crucial role in shaping our final solution, contributing to the synergy between text and image components in our project.

3 ARCHITECTURE DESIGN

- **Encoder for Downsampling Images:** Custom implementation for downsampling images, preserving essential features.
- **BERT Model for Text Embeddings:** Integration of BERT model for generating context-rich text embeddings.
- **Composition Model Based on TIRG:** Development of a composition model inspired by TIRG for combining image and text features.
- **2D Model Integration:** Fusion of downscaled images, BERT text embeddings, and composition model features for a unified 2D representation.

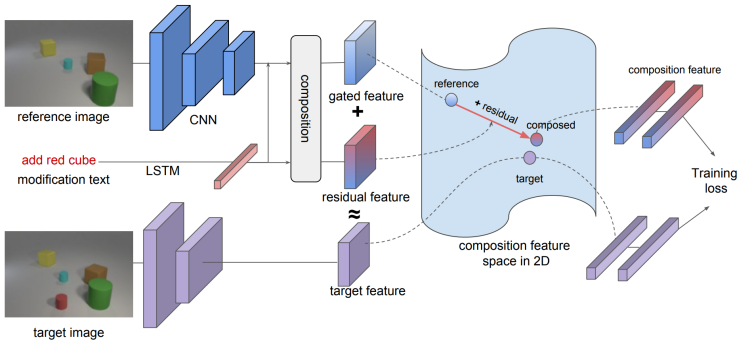


Figure 2: Architecture for the TIRG Model

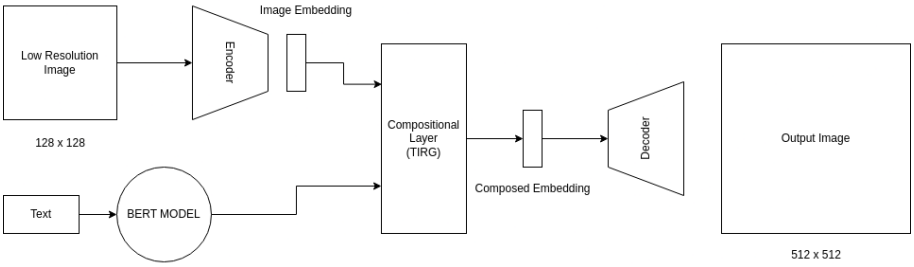


Figure 3: Architecture for our model

- **Decoder for Upsampling:** Custom decoder implementation for upsampling the final 2D representation, preserving details for downstream tasks.

4 Dataset and References to Past Papers

We are using the **InstructPix-to-Pix** dataset

- **InstructPix-to-Pix Data for Images:**
  - Employed Instruct Pix-to-Pix data for image translation, aligning generated images with corresponding instructions.
- **Decoder with Image Cropping and Noise Regularization:**
  - Enhanced the decoder with image cropping techniques to focus on specific regions of interest.
  - Introduced noise during training to implement regularization, promoting model robustness and preventing overfitting.

5 Model Selection:

- **Image Embeddings:** Utilized the TIRG model as a baseline structure for generating image embeddings, adapting and modifying it to suit the specific requirements of our project.
- **Text Embeddings:** Employed the BERT model for the generation of text embeddings, leveraging its capacity to capture contextual information and relationships within textual data.

6 Feature Enginnering:

1. **Image Processing:**

- **Cropped Image:** The cropping of the image is a strategic preprocessing step, allowing the model to focus on specific regions of interest. This is particularly useful when dealing with large or complex images.
- **Encoder for Image Encoding:** The encoder transforms the cropped image into a condensed representation, extracting high-level features that are crucial for subsequent denoising and generation processes.
- **Denoising with Generator:** The generator plays a key role in denoising the image, eliminating unwanted artifacts and enhancing the visual quality. This step is essential for refining the image and improving its clarity.

2. **Text Processing:**
   - **BERT Model for Text Embeddings:** Leveraging the BERT model for text embeddings ensures that the model captures nuanced semantic information from the associated text. BERT's bidirectional context-awareness allows for a richer understanding of the textual content.
3. **Composition of Image and Text:**
   - **Combining Image and Text Embeddings:** The fusion of image and text features in the 2D composition facilitates a holistic representation, where the strengths of both modalities contribute synergistically. This integration enhances the model's ability to comprehend and generate meaningful outputs.
4. **Regularization with Image Cropping and Noise:**
   - **Image Cropping for Focus:** The incorporation of image cropping not only aids in dimensionality reduction but also assists the model in focusing on salient visual features. This selective attention mechanism contributes to improved interpretability and efficiency.
   - **Noise Introduction for Regularization:** Introducing controlled noise during training serves as a regularization technique. This regularization effect prevents the model from overfitting to the training data, promoting generalization to unseen examples. It adds a level of robustness to the model by exposing it to various perturbations.
5. **Synergistic 2D Composition:**
   - **Harmonizing Image and Textual Information:** The 2D composition phase involves more than mere concatenation; it's about creating a synergistic blend of image and text features. This holistic representation is designed to capture nuanced relationships between visual and semantic elements, fostering a more cohesive understanding.
   - **Facilitating Cross-Modal Understanding:** By combining image and text embeddings, the model is equipped to comprehend cross-modal relationships. This enables the generation of outputs that reflect a nuanced interplay between visual and textual content, enhancing the overall expressive power of the model.
6. **Loss Interpretation for Model Refinement:**
   - **Loss as Dissimilarity Metric:** The loss calculated during image comparison serves as a dissimilarity metric. It quantifies the deviation between the generated image and the target, guiding the optimization process.
   - **Iterative Model Refinement:** Through backpropagation, the model iteratively refines its parameters based on the calculated loss. This dynamic adjustment ensures that the model learns to generate images that align more closely with the desired output, improving its overall performance over training epochs.
7. **End-to-End Training Paradigm:**
   - **Integration of Steps for Coherent Learning:** The entire process is part of an end-to-end training paradigm, where each step influences the others. This holistic approach allows the model to learn not only individual components but also the interdependencies between them, leading to a more coherent and effective learning process.

7 RESULTS AND FURTHER FINE TUNING:

1st Iteration (1 - 150 epochs):

- Implemented the model architecture with careful attention to each block and layer as defined in the initial structure.
- Observed that the initial results closely resembled the baseline output, indicating a limitation in capturing subtle nuances within the images.
- Investigated the training dynamics and discovered that the model struggled to learn intricate details, possibly due to issues in information flow across layers.
- Noted potential challenges in the skip connection between the ninth and first blocks, hypothesizing that its long-range connectivity might be disrupting the gradient flow during backpropagation.

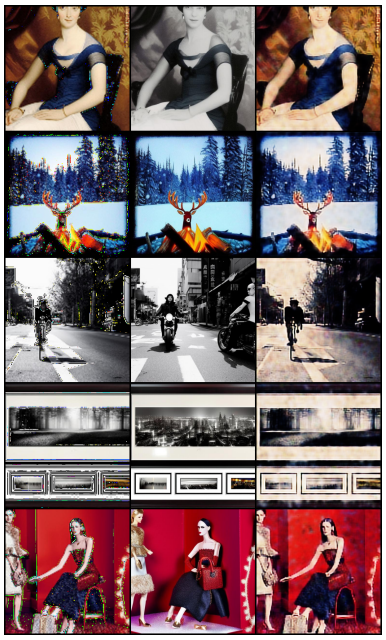


Figure 4: Result after the first 1-150 epochs

2nd Iteration (151 - 250 epochs):

- Responding to insights from the first iteration, the skip connection between the ninth and first blocks was temporarily removed to assess its impact.
- Resulted in images with increased error-proneness and noticeable noise, revealing the crucial role of the skip connection in maintaining image quality.
- Analyzed the model's internal representations and found that without the long-range skip connection, the model struggled to capture contextual information crucial for preserving fine details.
- Experimented with different loss functions and regularization techniques to mitigate the observed increase in noise, but challenges persisted.

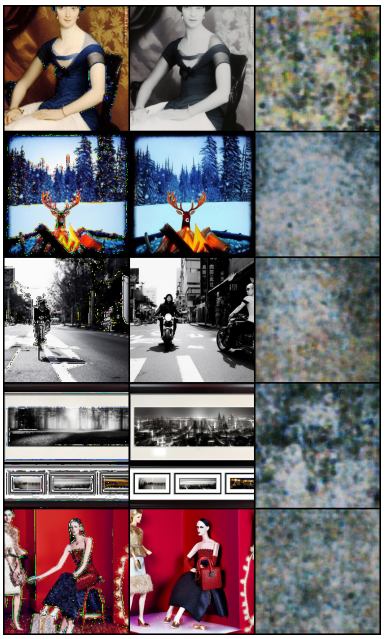


Figure 5: Result after 151 - 250 epochs. The vanishing gradient is the reason of above result and hence the model could not even generate the structure. SSIM

3rd Iteration (250 - 350 epochs):

- Building on lessons learned, reintroduced the skip connection but strategically placed it between the output of the sixth block and the final block.
- While the resulting images showed some similarity to the goal image, they exhibited an excess of noise, making it challenging to discern the desired features.
- Investigated the impact of skip connections on intermediate representations and found that the long-range skip connection introduced during this iteration contributed to an accumulation of noise in the later stages.
- Explored the use of additional regularization techniques specifically tailored to address noise amplification, experimenting with dropout and batch normalization strategies.

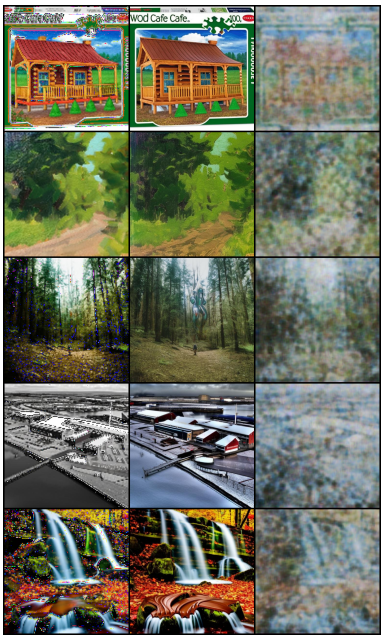


Figure 6: Result after 250-350 epochs (comparatively better than the previous set of epochs)

Resizing and Manual Refinement:

- **Resizing Images:** Recognizing the need for alignment with text embeddings, efforts were directed towards resizing images to match the dimensions of the text embedding size. This aimed at establishing a harmonious integration between visual and textual components.
- **Manual Input-Output Refinement:** To further enhance the alignment between input and output, manual adjustments were made to the sizes. This meticulous process involved fine-tuning the dimensions to ensure compatibility and improve overall model performance.

8 Scope for Improvement:

- **Refinement in Resizing Process:** Despite efforts in resizing, there is a scope for further refinement in the resizing process to optimize the alignment between image dimensions and text embeddings.
- **Automated Size Adjustment:** Consider exploring automated methods for adjusting image sizes to eliminate the manual refinement step, potentially incorporating dynamic resizing strategies during training.
- **Evaluation of Image Quality Metrics:** Implement and assess image quality metrics to quantitatively evaluate the impact of resizing and manual adjustments on the generated images, providing a more objective measure of model performance.

9 Future scope

Next Steps:

- Further Exploration: Consider additional variations in skip connections, such as utilizing attention mechanisms, and explore alternative architectural adjustments to refine the model's ability to capture long-range dependencies.
- Fine-Tuning Parameters: Iteratively fine-tune hyperparameters, including learning rates and regularization strengths, to achieve a more nuanced balance between loss minimization and perceptual image quality.

- Collaborative Analysis: Engage in collaborative discussions with domain experts to gain insights into the significance of certain features and inform adjustments to the model architecture. **Scope for Improvement:**
- **Exploration of Multiple Skips:** Investigate the potential benefits of incorporating multiple skip connections within the model architecture. Experiment with skip connections at various stages to enhance information flow and improve the model's ability to capture features across different scales.
- **Resizing Instead of Cropping:** Re-evaluate the choice of resizing images instead of cropping them. Explore the impact of resizing on preserving important details and contextual information, aiming for a balance between image size and feature retention.

10 CONTRIBUTIONS

We have distribited the work among us as of now as follows:

Sudeep: Implemented the TIRG Model. Developed final synergistic 2D composition code to generate the final output.Data extraction and fine tuning of hyperparameters.

Siddhi: Designing the final pipeline and training and implementing the model. Data extraction and hyperparameter tuning.

Rekha:: Implemented the TIRG model. Implemented the bert model for text Embeddings. Data extraction and hyperparameter tuning.