2D Image Relighting With ControlNet

Tony Xia, Danica Xiong

March 15, 2024

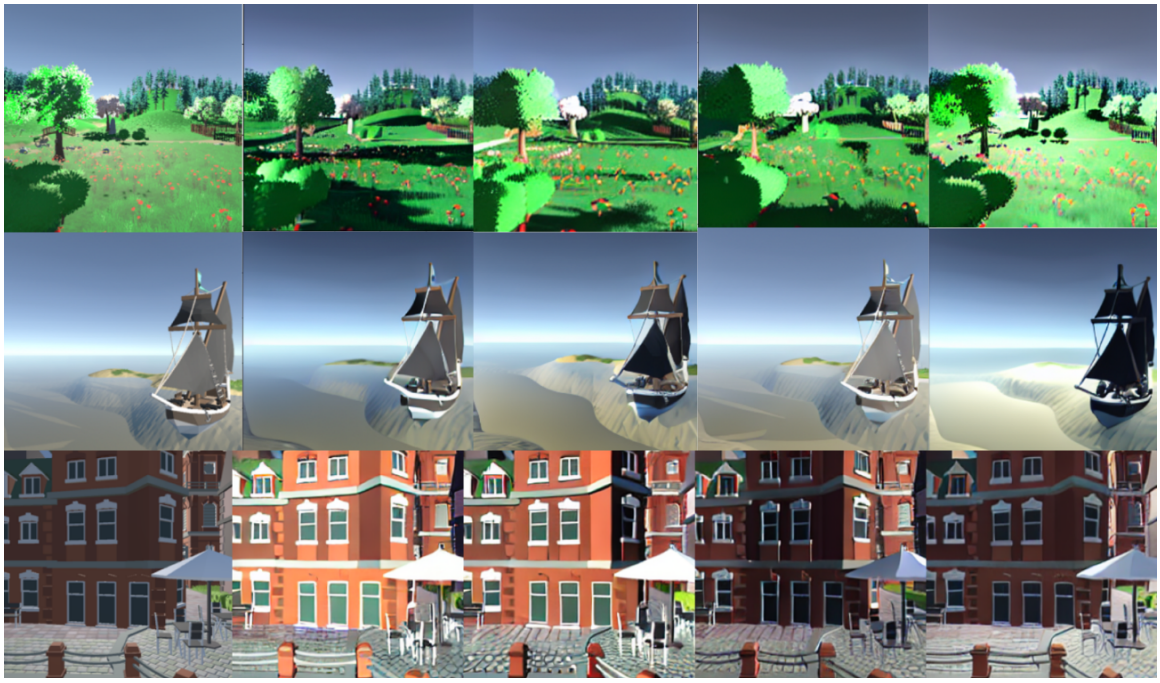

Figure 1: The first column shows the input view. The following columns each contain images generated by our ControlNet conditioned on the four primary directions: front, left, right, and back.

Abstract

Image relighting, the process of manipulating lighting conditions within images, plays a crucial role in various applications such as computer graphics, photography, and augmented reality. Traditional approaches often rely on complex 3D reconstructions or explicit light control parameters, limiting their applicability and efficiency. In this paper, we present a novel approach to image relighting utilizing ControlNet, a neural network architecture originally designed for diverse conditional controls in image generation. By integrating ControlNet into our methodology, we circumvent the need for explicit 3D reconstruction, making our approach particularly well-suited for 2D image relighting tasks. We propose training a model to learn the intricate relationship between spatial features and lighting conditions within 2D images, thus enabling accurate relighting without relying on depth information.

1 Introduction

Image relighting, the process of manipulating lighting conditions within images, is a fundamental task in computer graphics, computer vision, and related fields. The ability to alter lighting can significantly impact the visual aesthetics, mood, and realism of images, making it a valuable tool for various applications such as virtual reality, augmented reality, and digital content creation.

Traditional approaches to image relighting often rely on complex 3D reconstructions or explicit manipulation of light parameters, which can be computationally intensive and cumbersome. These methods may also require depth information or scene geometry, limiting their applicability to 2D images or requiring additional preprocessing steps.

In this paper, we present a novel approach to image relighting leveraging ControlNet, a neural network architecture initially designed for diverse conditional controls in image generation. ControlNet offers a unique framework for incorporating spatial conditioning controls into large pretrained models, enabling precise manipulation of image features based on specified conditions.

Our approach aims to overcome the limitations of traditional methods by eliminating the need for explicit 3D reconstruction and depth information. Instead, we train a model to learn the intricate relationship between spatial features and lighting conditions within 2D images, enabling accurate relighting without relying on depth data.

Through experimental validation and demonstrations, we showcase the effectiveness and versatility of our approach in achieving realistic and visually appealing image relighting results.

2 Related Work

2.0.1 Stable Diffusion and ControlNet

"The advent of diffusion models signifies a pivotal advancement within the landscape of deep generative modeling, replacing Generative Adversarial Networks to become the ..." is what I would have typed if I did not read the announcement on canvas. So here's a Haiku about Stable Diffusion and Control Net:
CONTROL NET IS COOL,
SO IS STABLE DIFFUSION,
SO IS THIS PAPER.

2.0.2 SOTA Reconstruction

Recent advancements in computer vision and graphics have focused on reconstructing the shape, surface appearances, and illumination of physical-world objects based on 2D images, such as photographs. Traditional methods have explored various approaches, including physics-based inverse rendering [3] and neural-based object reconstruction.

One approach combines neural-based object reconstruction with physics-based inverse rendering to create an accurate and efficient object reconstruction pipeline. This pipeline utilizes a neural signed distance function (SDF) based shape reconstruction to generate high-quality object shapes, followed by a neural material and lighting distillation stage to predict surface material properties and illumination conditions. Finally, PBIR is employed to refine the initial results and achieve a final reconstruction of object shape, material, and illumination.

Another research direction focuses on enriching neural networks with physical insight to improve image relighting. By employing intrinsic image decomposition [5] and a direct black box approach, researchers aim to generate relighted images that accurately capture material reflectance parameters and illumination conditions. Additional components, such as surface normal vectors and multiscale blocks, are incorporated to enhance model performance for specific relighting scenarios.

Furthermore, efforts have been made to enhance the decomposition of 3D scenes into shape, materials, and lighting components using advanced shading models and denoising techniques. By incorporating ray tracing, Monte Carlo integration [1], and multiple importance sampling, researchers aim to improve the accuracy and efficiency of inverse rendering pipelines, leading to high-quality reconstructions with reduced noise and artifacts.

2.0.3 Netizens

The internet has played a significant role in advancing image manipulation techniques, with contributions ranging from innovative algorithms to extensive datasets. One notable example is the work of

a Chinese researcher who trained a ControlNet model to condition on black and white images, where black represents zero light and white represents full light intensity. This groundbreaking approach, trained on millions of images over several months, has demonstrated remarkable results in generating visually appealing relit images.

While this approach may not adhere to physical accuracy, its success highlights the potential of leveraging massive datasets and distributed computing resources available on the internet. By harnessing the collective intelligence and computational power of online communities, researchers can explore novel approaches to image manipulation that may not be feasible within traditional academic settings.

The success of internet-driven initiatives underscores the importance of collaboration and knowledge sharing in driving scientific progress. As researchers continue to leverage the power of online communities and resources, we can expect further advancements in image manipulation techniques, paving the way for new applications and creative possibilities.

3 Method

3.1 Data Collection

In order to generate physically plausible results, we decided to train our controlnet with physically grounded data. We used data from two sources to construct two separate training datasets.

Rene: Toschi et al.[4] constructed a dataset with 20 scenes depicting a variety of objects from 50 different viewpoints under 40 different lighting conditions. The dataset provides the images along with the camera extrinsics and lighting positions, perfect for our need for data accurate lighting information.

Unity Renderings: To collect data for the study, a custom script was developed in Unity to automate the process of capturing images from different lighting conditions. The scenes used for data collection were obtained from the Unity Asset Store and included diverse environments such as a town, desert, landscape, and graveyard. Each scene provided unique visual aesthetics and served as a backdrop for studying image relighting techniques.

For the test set, for each of the 4 scenes, images were captured from three different camera locations, allowing for variations in perspective and composition. Additionally, images were captured from 20 different lighting positions per scene, covering a wide range of lighting conditions. These lighting positions included varying colors (e.g., white, blue, and orange lights), as well as different attenuation settings, perturbation distances, and point light/directional light configurations.

The following is a list of differing conditions in the scenes:

- Light position (XYZ)
- Light type (point/directional)
- Light Rotation (For directional light only)
- Light color (RGB)
- Scene (Town/Desert/Landscape/etc)
- Camera position (Random on the scene)

The custom script, named "LightMovementAndCapture" simulated various lighting scenarios by randomly positioning a light source relative to the camera and perturbing it by 50-100 units in one direction (front/back/left/right) which was used in the prompt generation. To avoid camera rotations changing the orientation of front/back/left/right, the camera was positioned at XYZ=(0,0,0) with rotation (0,0,0) on the X,Y,Z axis and the scene was translated and rotated around the camera. This technique created dramatic lighting effects, enabling the study of different relighting scenarios.

Overall, the custom script facilitated the efficient and systematic collection of data by automating the process of simulating various lighting scenarios and capturing images from different camera positions within the Unity environment.

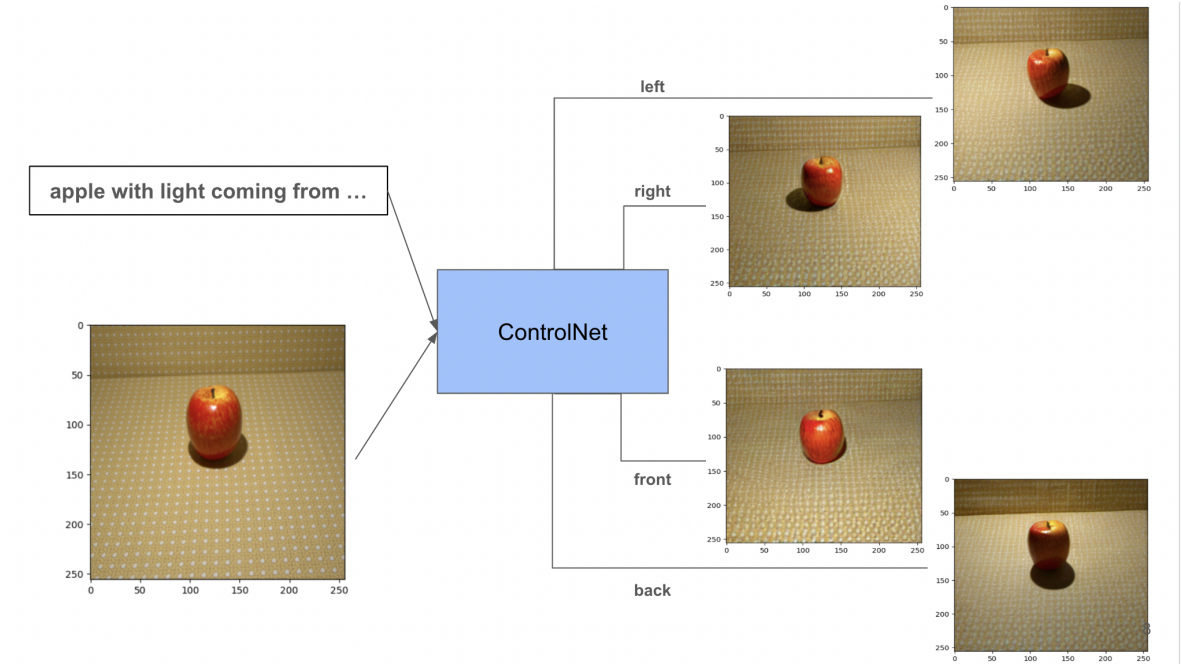

Figure 2: Illustration of the inputs and outputs

3.2 Labeling, Sampling and Prompt Generation

Rene: For each of the 20 scenes, we randomly select 3 lighting positions as the starting image. We then pair these images up with other images with the same camera extrinsics but different lighting position. This resulted in a total of $20 \times 3 \times 40 \times 50 = 120,000$ pairs of images. Each of these image pairs are associated with a text prompt derived from the second image using the following scheme: We first categorize the 40 lighting directions into 4 categories: "front", "back", "left", and "right" depending on the position of the light relative to the object. Light positions that are in between two categories are put into the category that it represents more. Then, we create the text prompt with the template: "*object* with light coming from *direction*."

Unity: Unity provides accurate physical shadows through its built-in rendering engine, which utilizes advanced rendering techniques to simulate realistic lighting and shadow effects. Here's how Unity achieves accurate physical shadows:

Real-time Shadow Mapping: Unity employs real-time shadow mapping techniques to simulate the interaction of light sources with objects in the scene. This involves projecting shadow maps from the perspective of each light source onto the scene geometry, allowing for dynamic and interactive shadow casting.

Soft Shadows: Unity supports soft shadows by incorporating techniques such as percentage closer filtering (PCF) and variance shadow mapping (VSM). These methods simulate the softening of shadows caused by the finite size of light sources and occluders, resulting in more natural and realistic shadow transitions.

For the test set, for each of the 4 scenes, images were captured from three different camera locations, allowing for variations in perspective and composition. Additionally, images were captured from 20 different lighting positions per scene, covering a wide range of lighting conditions. These lighting positions included varying colors (e.g., white, blue, and orange lights), as well as different attenuation settings, perturbation distances, and point light/directional light configurations.

3.3 Conditioning

To generate images with new lighting conditions, we condition the control net on 1) a text prompt with the new lighting information, and 2) a image showing the original object (scene) we want to relight.

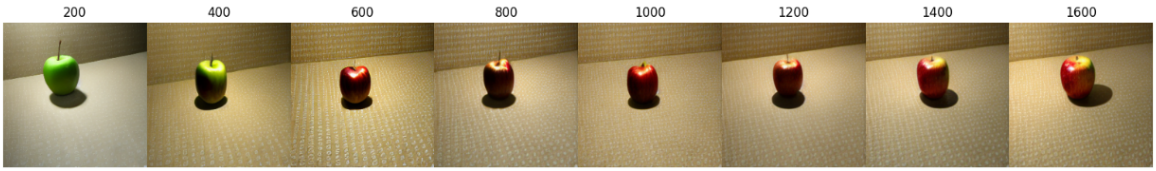

Figure 3: ControlNet generation of "apple with light coming from right" conditioned on an image of a red apple over 1600 training steps

The process is illustrated in Figure 2.

3.4 Training

We used stable diffusion 1.5 with the original control net implementation [6] for our experiments. We used a learning rate of 1e-5 and a batch size of 32. As this project is a proof of concept, we decided to crop and downscale the input images to 256×256 for faster training and sampling. The model was then trained on the rene dataset for 4 epochs. The training process took 7 hours on a single A5000 GPU with 24 GB of VRAM.

For unity scene relighting, as the dataset size is smaller, we trained the model for 80 epochs with the other hyperparameters kept the same. The process took under 7 hours.

3.4.1 Sudden Convergence

In our training, we noticed a sudden convergence to the original scene at training step 1600, as shown in Figure 3. Even though the model starts generating apples since as early as step 200, it is not able to generate the original apple faithfully until step 200. We found a similar pattern in our model trained on Unity generated images as well.

4 Results

Figure 1 shows some results generated by our model. The very left column contains the input image, while the rest of the columns correspond to images generated with lighting coming from the front, left, right, and back side of the scene. As can be seen from the shadows cast by the objects as well as the lighting on the objects surface, our approach achieves decent results.

4.1 Comparisons

We experimented with SD v1.5 without control net. Conditioning on the text prompt "apple with light coming from right", we generated 10 images. Only 1 out of the 10 images have correct lighting. Out of the other 9, 8 do not have well-defined lighting features such as shadows and highlights, and the other one have the light source coming from the opposite direction. Figure 4 shows some of the selected examples generated in our very brief "ablation study".

4.2 Discussion

Despite training on only the four primary directions, we experimented with input prompts that contain two adjacent directions. The results show that our model was able to interpolate the directions and generate expected results. This observation shows that if a continuous light conditioning method (an encoder with that encodes light position X,Y,Z) were to be implemented, the model could learn to generate images with lights from all directions. However, the quality of the generated image deteriorated, likely due to overfitting as our training dataset was not very diverse.

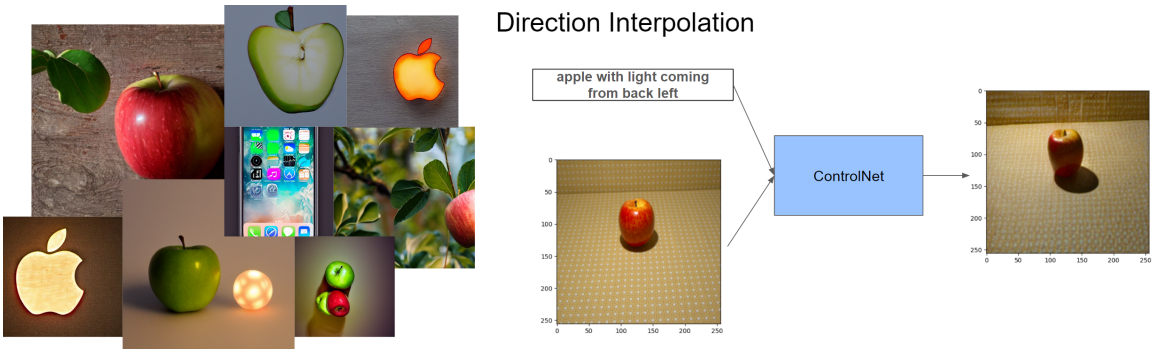

Figure 4: The left diagram shows Stable Diffusion generated images conditioned on text prompt "apple with light coming from right". The right diagram shows direction interpolation observed in our ControlNet model.

5 Limitations

5.1 Generalization to New Environments

The trained relighting model struggles to generalize to new environments or scenes that significantly differ from the training data. Adapting the model to novel environments or extending its capabilities to handle diverse lighting conditions remains a challenge. More diverse data is needed to achieve this, as well as a lot of parameter tuning.

5.2 Training Data Diversity and Quantity

The effectiveness of the relighting model heavily relies on the diversity, quantity, and quality of the training data. Limited diversity in the training dataset, especially in terms of lighting conditions and scene compositions, leads to biases and generalization issues in the model's predictions.

5.3 Physical Accuracy

Despite efforts to simulate realistic lighting conditions, the relighting model still lacks full physical accuracy due to the nature of not using any lighting simulation. Instilling some sort of physics prior should positively impact the realism and fidelity of the results.

6 Future Work

6.1 Enhanced Physical Accuracy Via Path Tracing

Future research could focus on improving the physical accuracy of the relighting model by incorporating more sophisticated physics-based lighting models and material properties. This may involve integrating techniques such as ray tracing, bidirectional reflectance distribution function (BRDF) modeling, and global illumination algorithms. The normal and depth data of the model can be retrieved with SOTA research such as Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation by Ke et al [2].

6.2 Real-Time Relighting Applications

Thanks to the speed of ControlNet, real-time is an option for our relighting scheme. Exploring real-time relighting applications for interactive experiences, such as virtual reality (VR) or augmented reality (AR), presents exciting opportunities. Developing lightweight and efficient relighting algorithms suitable for real-time rendering could enable immersive and interactive visual experiences with dynamic lighting effects.

6.3 Encoder for Precise Lighting Conditioning

Currently, our model only supports changing the lighting to the four primary directions: left, right, front, and back. Alternatively, we could train an encoder to encode the X,Y,Z positions of the light source into the input space of controlnet and pass this directly into controlnet. This encoder could be trained end-to-end along with the controlnet. This allows the user for more granular lighting control.

6.4 User-Centric Design

Integrating user-centric design principles into the development of relighting interfaces could improve usability and accessibility. Sliders that define X,Y,Z positions could be an option to allow users to adjust images on the fly and view their changes quickly. Conducting user studies and gathering feedback from artists, designers, and content creators can inform the design of intuitive and user-friendly interfaces for specifying desired lighting conditions and manipulating relit images.

6.5 Reflective Material

Future work could focus on incorporating reflective materials into our image relighting framework. While our current approach assumes Lambertian materials, the inclusion of reflective materials would significantly enhance the realism of relighted images, particularly in scenes with reflective surfaces such as water or glass. This could leverage ControlNet's generative behaviour to create realistic-looking reflections with varying lighting.

7 Conclusion

In this paper, we have presented a novel approach to image relighting leveraging ControlNet. Our approach offers a decent solution to the challenging task of image relighting by circumventing the need for explicit 3D reconstruction and depth information, making it particularly suitable for 2D image relighting tasks that are close to trained images.

Overall, this was a great learning experience to try stable diffusion and model training for Control Net. Throughout the project, we learned a lot of things that would aid us in our future endeavors.

7.1 Reflection - Tony

We've delved into the intricacies of diffusion models and control net on a theoretical level, deepening our understanding significantly. In addition, the iterative process of training and encountering setbacks with our self-collected dataset underscored the crucial role of size, diversity, and quality of the dataset in effectively training a model. Lastly, this project has acquainted us with essential tools and frameworks like the Unity rendering engine and PyTorch Lightning, which are widely utilized in the field.

Aside from technical lessons, we also learned the lesson that any project that involves machine learning is a marathon as opposed to a sprint. Time management is key to the intelligence of the model and the sanity of the model trainer. Moreover, care and patience should be present at every moment, as a single careless mistake can result in hours of wasted effort.

7.2 Reflection - Danica

I personally come from a complete computer graphics background (path tracing/inverse rendering/spherical harmonics w/ prof. Ramamoorthi) and have only recently started learning about AI after coming to Stanford. This course has taught me the intricacies of AI and its applications to the forefront of graphics research. This was my 2nd time training a model and my 1st time using Pytorch. We spent such a long time debugging, dealing with Stanford's computing clusters, and debugging Pytorch lightning. We also delved into ControlNet's and Stable diffusion's code to figure out what was going on (turns out we needed more VRAM). This was also my first time dealing with such a large dataset and writing scripts to generate the prompts we needed. We trained for 5-12 hours at a time and made many adjustments to the models and our training data to get the time and output down. It was really exciting trying to figure out how to adjust our pairing and prompt generation to create the correct

output, and trying to explain why we got the output we got (ie. texture mapping and why it didn't work for differing many viewpoints). I'm really proud of the fact that we came up with the idea of heavily skewed lighting + a single directional lighting prompt to achieve emergent multi-directional lighting. Furthermore, I've learned about overfitting, analysis, and batch methods in class but this has allowed me to put it all to work which was an amazing experience. I can't wait to work more with AI in graphics, especially Gaussian Splatting. Thank you for the amazing quarter!

8 References

[1] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. *Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising*. 2022. arXiv: 2206.03380 [cs.GR].

[2] Bingxin Ke et al. *Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation*. 2023. arXiv: 2312.02145 [cs.CV].

[3] Cheng Sun et al. *Neural-PBIR Reconstruction of Shape, Material, and Illumination*. 2024. arXiv: 2304.13445 [cs.CV].

[4] Marco Toschi et al. "ReLight My NeRF: A Dataset for Novel View Synthesis and Relighting of Real World Objects". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 20762–20772.

[5] Amirsaeed Yazdani, Tiantong Guo, and Vishal Monga. *Physically Inspired Dense Fusion Networks for Relighting*. 2021. arXiv: 2105.02209 [cs.CV].

[6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 3836–3847.