

Towards Controllable Image Relighting

Maria Pilligua

July 17, 2026

Abstract

Relighting is the task of modifying the illumination conditions of a scene while preserving its semantic and structural content. In this work, we investigate different strategies for achieving controllable image relighting, reviewing recent advances in diffusion models and image-to-image transformer architectures. While diffusion-based methods have shown strong generative capabilities in other tasks, we identify key limitations in their controllability and computational cost. Motivated by this, we propose a novel transformer-based framework that introduces lighting control via a prompt-based conditioning mechanism. Specifically, we evaluate two architectures, DAT and Restormer, both trained on a synthetic relighting dataset and subsequently fine-tuned on real-world images to enhance generalization. Our approach achieves competitive performance while enabling intuitive and fine-grained control over lighting directions and intensities. These results highlight the potential of transformer-based models, when properly conditioned, as a powerful tool for controllable scene relighting.

Keywords: Relighting, Diffusion, Transformer, Latent-Space, Controllable, Image2Image, StableDiffusion, Restormer

1 INTRODUCTION

Relighting is a fundamental task in computer vision and computer graphics, enabling applications in photography, film production, augmented reality, and virtual environments. The goal is to modify the lighting conditions of a scene while preserving its structure, appearance, and semantics. Traditional approaches to relighting often rely on *intrinsic image decomposition*, where an image is separated into albedo and shading components. These methods produce physically consistent outputs but require explicit modeling of lighting behavior and are limited to specific domains.

Recent advances in generative models have opened new possibilities for more flexible and data-driven relighting. In particular, diffusion-based models and transformer-based image-to-image architectures have shown strong performance in image synthesis and manipulation tasks. Diffusion models offer powerful generation capabilities and can be guided using techniques such as textual prompts or structural conditions. However, they often lack fine-grained control and struggle to isolate lighting changes from other visual attributes. On the other hand, transformer-based image-to-image models are also a competitive alternative for image enhancement and restoration, and usually they

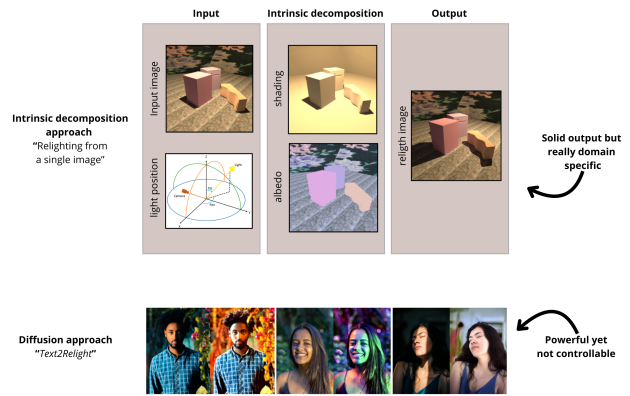


Fig. 1: Comparison between intrinsic decomposition-based relighting and generative approaches. The former provides structured and domain-specific outputs, while the latter enable more flexible and data-driven relighting.

allow more direct control in pixel space and show potential for structured conditioning.

This project aims to explore novel techniques for controllable and realistic relighting by leveraging recent advances in both diffusion models and image to image transformers. We investigate how different conditioning mechanisms, such as image prompts, edge maps, or lighting parameters, can be incorporated into these type of architectures to achieve high-quality relighting while preserving scene consistency. Our goal is to identify effective strategies for light-aware generation and contribute to the development of more controllable relighting models aligned with current trends in

- Contact E-mail: 1638307@uab.cat
- Supervised by: Javier Vazquez-Corral (Computer Science)
- Academic Year 2024/25

the literature.

This work will be structured like the following: First we will discuss the objectives and the scope of this project. Then, we will show the planning strategy we will follow to organize the work. Next, we will review the context and state of the art of existing relighting models. We will continue with the most important tests we have computed during the duration of the project, and finally, we will discuss the best model and its qualitative and quantitative results.

2 OBJECTIVES

The primary objectives of this project are:

1. Study different approaches to do relighting
2. Design an pipeline that integrates:
 - (a) A model to generate a relit image.
 - (b) Mechanisms to control relighting based on user-specified light source positions or other lighting parameters.
3. Train the proposed architecture on a diverse dataset to ensure robustness and generalization.
4. Evaluate the generalization capability of the model by testing it on unseen images.

Through this research, we aim to bridge the gap between controllability and generalization in relighting models, contributing a novel method that enhances realism and usability in real-world applications.

3 RELATED WORK

Controllable relighting has been approached from several directions in the literature, spanning both physically motivated methods and data-driven generative models. Early techniques rely on intrinsic decomposition to disentangle lighting components from scene content, while more recent advances leverage powerful generative architectures such as diffusion models and transformers to learn relighting from data. In this section, we review the most relevant approaches across these categories, highlighting their mechanisms, strengths, and limitations in the context of our goal: precise and intuitive lighting manipulation in real-world images.

3.1 Intrinsic Decomposition Methods

One branch of relighting research focuses on physically grounded approaches, such as intrinsic decomposition, which explicitly separates images into albedo and shading. The key idea behind them is to separate the part of the image that needs to be changed from the part that needs to be kept. For instance, Yang et al. [1] and StyleGAN-based methods [2] utilize this decomposition to manipulate shading while keeping the albedo intact. Other work, like StyLitGAN, extends this decomposition to include gloss maps, enabling

finer control over reflections on non-lambertian surfaces. These methods often leverage U-Net style architectures enhanced with CNNs to reconstruct relit images. While they offer precise control over lighting effects, their reliance on domain-specific datasets limits their ability to generalize to diverse, uncontrolled real-world scenes.

3.2 Generative Methods

3.2.1 Image-to-Image Transformer Models

Image-to-image transformers encompass all of the deep learning models that use transformer architectures to solve image-to-image translation tasks. These tasks involve taking an input image and generating a new image with some desired transformation, e.g., low-light enhancement. Transformers were first introduced for NLP tasks, but they have been adapted for vision with models like ViT (Vision Transformer). Image-to-image transformers typically follow an encoder-decoder architecture similar to U-Net but replace convolutional layers with self-attention blocks.

There are a lot of image-to-image transformer models. One of the most relevant over the last few years has been Restormer [3], an all in one restoration model. It is widely regarded as a leading model in multiple tasks, including de-raining, deblurring, denoising, and dehazing. In addition to general-purpose models, several transformer variants have been proposed for more specialized applications. For instance, Retinexformer [4] and LLFormer [5] are tailored for low-light enhancement, while DAT [6] focuses on high-fidelity image super-resolution.

3.2.2 Diffusion Models and Latent Space Manipulation

Several recent works have focused on adapting Stable Diffusion (SD) to the relighting task by fine-tuning or extending its capabilities. Some approaches retrain the model on relighting-specific datasets, such as ICLight [7], while others introduce new conditioning modalities. For instance, Text2Relight [8] enables relighting based on natural language descriptions, aiming to adjust illumination while preserving semantic content.

A major challenge in these approaches is maintaining the structural integrity of the input scene, as diffusion models tend to hallucinate or distort details when given loosely defined conditions. To mitigate this, methods like ControlNet [9] and IP-Adapter [10] inject stronger structural guidance. ControlNet incorporates spatial priors such as depth or edge maps, while IP-Adapter aligns visual embeddings from frozen vision encoders (e.g., CLIP) with the text conditioning space of SD, enabling image-based control.

Beyond conditioning, IP-Adapter also operates at the latent level, bridging to a broader class of methods that steer diffusion generation via latent space manipulation. Concept Sliders [11], for example, use LoRA modules to disentangle semantic attributes and allow continuous control over them. SliderSpace [12] expands on this idea by identifying unsupervised latent directions aligned with CLIP features,

enabling intuitive editing without task-specific training.

4 DATASETS

A central requirement for developing and evaluating relighting models is access to datasets that contain the **same scene captured under multiple lighting conditions**. Such datasets allow the model to learn lighting-dependent variations while preserving geometry, texture, and object identity. During my work, I have worked with the following datasets:

- **Real Scene Relighting (RSR)** The **RSR** dataset, proposed by Yixiong Yang et al. in [1], provides around 80 real-world tabletop scenes. Each scene is photographed under 9 distinct lighting conditions and from 8 viewpoints (rotated in-plane up to 180°). This makes it ideal for studying structured, multi-illumination relighting in real, complex scenes with textured surfaces and cluttered arrangements.
- **Synthetic ISR** In addition to the real images, the authors also released a synthetic dataset following the same structure as RSR but with much more data. It was rendered using a physically-based graphics engine. This version maintains the same layout and lighting configurations as the real set but includes perfect ground truth for intrinsic components such as albedo and shading. The ISR has 7801 scenes, each one under 10 different light conditions. Each scene has between 3 and 10 nonoverlapping objects which are randomly selected from various categories of the ShapeNet dataset [13]. In addition, they pay special attention to the roughness of shapes and introducing variability in walls and floors patterns.

4.1 Other Datasets Considered

Several existing datasets were considered for training, but were ultimately excluded due to limitations that prevent them from meeting the specific requirements of our model. Some of these may be suitable for future extensions of this work.

- **Multi-Illumination Images in the Wild (MIW)** [14] contains real-world indoor and outdoor scenes photographed under varying lighting conditions. The dataset uses a servomotor to rotate a light source during capture, resulting in a large number of images under different illuminations. While MIW is extensive and well-suited in terms of scale, the nature of the lighting variation, primarily rotational changes rather than positional, does not align with the type of control we aim to learn. As illustrated in Figure 2, MIW corresponds to type (b), where light rotates around a fixed position, whereas our objective requires datasets that involve changes in the physical position of the light source (type a).
- **VIDIT** offers 390 synthetically generated scenes created with Unreal Engine, each rendered under 40 dis-

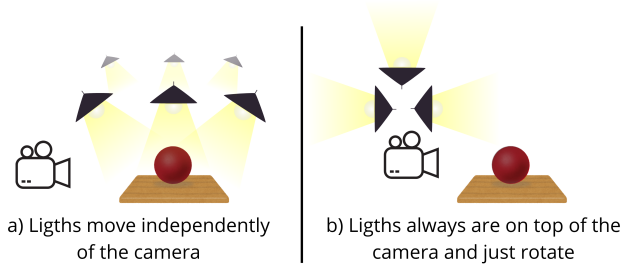


Fig. 2: Comparison of two common approaches for capturing datasets with varying lighting conditions: (a) changing the physical position of the light source around the scene, enabling diverse illumination angles and intensities; (b) rotating a fixed-position light source, which varies the direction of illumination but does not change its spatial location. Our work focuses on datasets following approach (a) to achieve full relighting control.

tinct lighting conditions, resulting in a total of 15,600 images. These conditions span 5 color temperatures (2500K to 6500K) and 8 discrete light directions (N, NE, E, SE, S, SW, W, NW). Although this dataset is well-structured and diverse, integrating it into our pipeline requires a proper mapping between its coordinate system and those used in the RSR and ISR datasets. As such, we leave this dataset for future work, where a translation mechanism can be developed to align illumination representations across datasets.

5 METHODOLOGY AND TIMELINE

To efficiently manage the development of this project, we will follow an agile-inspired approach that combines elements of Scrum and Kanban. Scrum provides a structured yet flexible framework where work is divided into iterative cycles, allowing for continuous improvement through weekly progress evaluations with the tutor. Meanwhile, Kanban will help visualize the workflow, ensuring a smooth transition between different project stages and avoiding bottlenecks.

The project will be organized into weekly sprints, during which specific tasks will be tackled and reviewed. Each sprint will conclude with a progress review and feedback session during the weekly meeting with the tutor, ensuring alignment with the project objectives. A Kanban board (Notion) will be used to track ongoing, completed, and pending tasks at different stages of development.

5.1 Gantt Chart

The Gantt chart showed in figure 3 provides the detailed visualization of the project’s timeline, outlining the main tasks and their respective durations. The project is structured into distinct phases, ensuring a smooth progression from research and preliminary tests to training, evaluation, and final documentation.

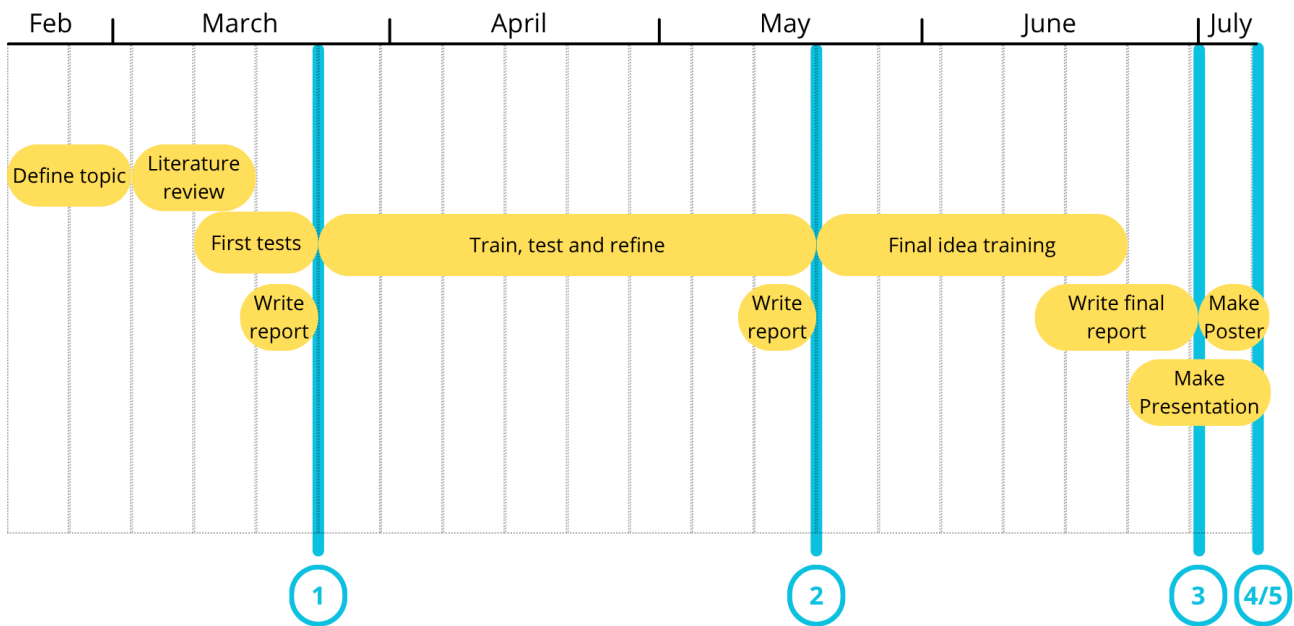


Fig. 3: Gantt Chart explaining the overall structure followed during the project. Notice that as it is a more research-based project. There are not that many subtasks to divide the executing as we don't know from the beginning which approach will give better results and what we will spend more time doing. The marks in blue indicate the submission deadlines explained in section 5.1.2.

5.1.1 Project Phases and Tasks

The primary tasks outlined in the Gantt chart include:

1. **Define the Topic** – During this time we finished defining the topic and the scope of our project. We actually decided to change the original topic of the work from low-light enhancement using diffusion to controlled relighting.
2. **Literature Review** – A comprehensive review of relevant papers, including state-of-the-art techniques in diffusion-based relighting and latent space manipulation.
3. **First Tests and Ideas** – Initial experiments and feasibility tests to assess potential approaches for relighting.
4. **Write First Report** – Summarizing the literature review and initial experimental results.
5. **Train, Test, and Refine** – Iteratively implementing and training different approaches and refining our approach based on performance and feasibility.
6. **Write Second Report** – Writing a report of mid-term findings and refinements.
7. **Final Model Training** – Conducting the last round of training using the final approach.
8. **Write Final Report** – Preparing the complete documentation of the research, methodology, experiments, and results.
9. **Make Poster** – Designing a poster summarizing key findings for the final presentation.

10. **Prepare Presentation** – Structuring and practicing the oral defense of the project.

5.1.2 Submission Deadlines

The following deadlines, marked in blue in the Gantt chart, represent the official submission dates for various deliverables:

1. **First Report:** March 23rd
2. **Second Report:** May 18th
3. **Final Submission:** June 30th
4. **Presentation:** July 4th
5. **Poster Submission:** July 6th

6 WORK DONE

Over the past months, I have conducted a series of experiments to explore the feasibility of performing relighting using both diffusion-based models and state-of-the-art transformer-based image-to-image architectures. The objective has been to evaluate how lighting information can be effectively introduced and used to condition these models.

6.1 Diffusion-based relighting

In this section we will explore some of the most relevant tests done using diffusion models. We have explored multiple ways of conditioning diffusion models in order to do relighting.



Fig. 4: Results of concept sliders trained for the relighting task. The prompt to generate the image was "Image of a table with bright light illuminating it". The numbers on top of the image shows the value given to the concept slider for each generation. Notice how the shadow below the table changes as we increase the concept value. However, the illumination of the whole scene seems a bit inconsistent.

6.1.1 Stable Diffusion

Stable Diffusion [15] is a class of text-to-image generative models based on latent diffusion. It operates by gradually denoising a latent representation of an image conditioned on a text prompt, producing high-quality and semantically coherent outputs. Unlike pixel-space diffusion models, Stable Diffusion performs the diffusion process in a compressed latent space learned by an autoencoder, significantly reducing computational cost while maintaining visual fidelity. It has been widely adopted for various generative tasks due to its flexibility, efficiency, and ability to incorporate different forms of conditioning, such as text, images, and structural cues.

6.1.2 Concept Sliders

Concept Sliders [11] proposes to train a LoRA module over SD parameters to allow fine-grained, interpretable control over specific image attributes during the generation process. In their paper [11] this is mostly used to modify images using semantic concepts. See figure 5 to view an example of what changes this approach can make.



Fig. 5: Example shown in concept sliders teaser. This model is able to change only a concept in each generated image.

To train this LoRA module you simply need a few examples of the paired images with different amounts of the "attribute" you want to play with. In my case, to start with a simple test, I trained using 5 pairs of images from the RSR dataset, which all go from illumination A to illumination B. Their training process optimizes the LORA applied in both the negative and positive directions. Basically, they are changing the loss to the one shown in eq. 1. Where $\epsilon_{\theta+}$ represents the application of positive LoRA and $\epsilon_{\theta-}$ the negative case. Ideally, this slider would allow us to pass smoothly from one lighting to the other.

$$\|\epsilon_{\theta-}(x_t^A, \cdot, t) - \epsilon\|^2 + \|\epsilon_{\theta+}(x_t^B, \cdot, t) - \epsilon\|^2 \quad (1)$$

This has the effect of causing the LORA to align to a direction that causes the visual effect of A in the negative direction and B in the positive direction.

However, for this to work, the concept of "relighting an image" had to be present in the latent space of stable diffusion, that is, there had to be a direction where you could potentially move to only change the position of the light source.

After careful training of this method we got the results obtained in Figure 4. As you can see, these are not really promising results. Even though it seems like the model is learning something about the light (notice the change in the shadow made by the table). It is obvious that the model doesn't understand what we are asking for as the shading on top of the table remains constant and the whole shading in general seems a bit off.

6.1.3 IP-Adapter

Next, we decided to test conditioning the generation process with the desired lighting using an IP-Adapter [10]

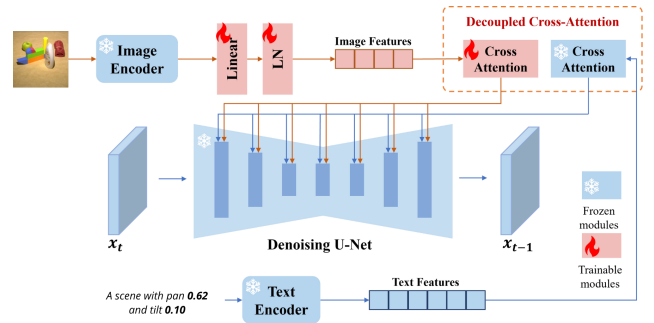


Fig. 6: Architecture of IP-Adapter with decoupled cross-attention. Only the modules in red are trained; the rest of the pre-trained text-to-image model remains frozen.

The IP-Adapter is a lightweight conditioning module that enables the use of image prompts in pre-trained diffusion models. This is achieved by aligning the image embeddings, extracted from a frozen vision encoder like CLIP, with the text embeddings used by the diffusion model. This allows an input image to be used as a conditioning signal without requiring paired data; training only needs image-caption pairs. See the architecture on Figure 6.

Figure 7 shows various types of conditioning IP-Adapters can support, including single or multiple image prompts, and image-text combinations.



Fig. 7: Example generations using IP-Adapter: top to bottom: conditioning with one image, conditioning with two images, and conditioning with image+text.

Our goal was to use an image of the same scene under different lighting conditions as a prompt to guide the generation toward a target illumination. See Figure 8.

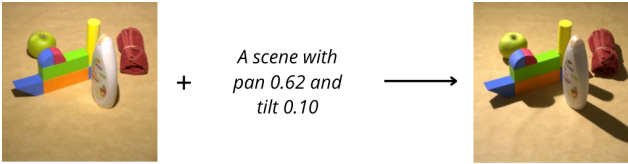


Fig. 8: Relighting objective: conditioning on an input image with different lighting and a target light description.

To this end, we trained two IP-Adapters, one for Stable Diffusion 1.5 and another for Stable Diffusion XL. We used the Real Scene Relighting Dataset from the Deep Intrinsic Relighting paper explained in section 4.

The results, shown in Figure 9, indicate that the IP-Adapter alone was insufficient to control the generation effectively. The model often failed to preserve the structure of the original scene, introducing distortions or altering semantic content. As such, it was difficult to evaluate whether the relighting task was successful, since the underlying scene had changed substantially. While the XL version produced slightly more coherent outputs than SD 1.5, neither model offered the desired level of control.

Additionally, we conducted an exploratory experiment by conditioning the model on a shadow map, following the idea presented in [8], to see whether more explicit lighting cues could improve performance. However, as expected it did not work successfully as it was not trained for this specific task.

6.1.4 ControlNet

Given that the IP-Adapter alone was unable to preserve the scene structure during generation, we explored the use of ControlNet [9] to add stronger structural constraints.

As introduced in the original paper, ControlNet is a neural network architecture designed to control diffusion models by introducing additional conditions. It works by duplicating the neural network blocks into two copies: a "locked" copy, which preserves the original model, and a "trainable" copy, which learns the new condition. See fig-

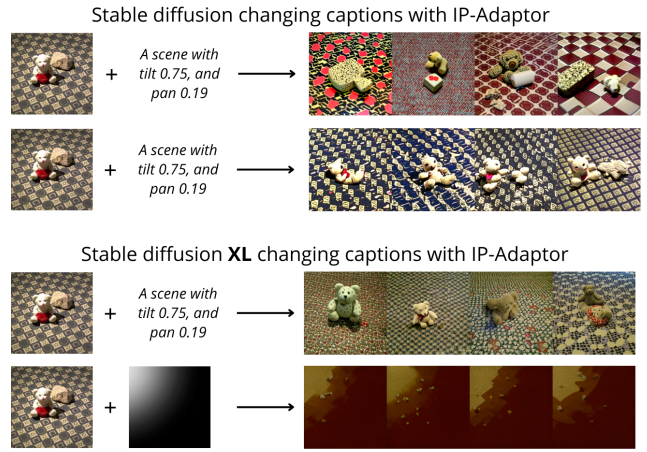


Fig. 9: Qualitative results from IP-Adapter training, conditioned on the desired lighting via text and an image of the same scene under different illumination.

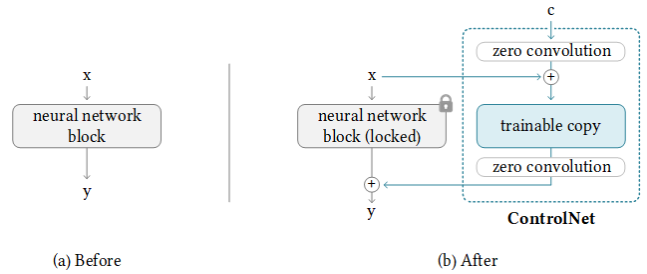


Fig. 10: Simplified architecture of ControlNet. The "trainable" branch is initialized with zero convolutions and progressively learns to integrate conditioning information.

ure 10 to view the simplified architecture. This setup allows ControlNet to be fine-tuned on small datasets of image pairs without damaging the original, production-ready diffusion model. A key component is the "zero convolution", a 1×1 convolution with both weights and biases initialized to zero. Before training, these layers output only zeros, ensuring that ControlNet introduces no distortion to the model until it begins learning.

An advantage of using ControlNet is that there are many production-ready pretrained ControlNet models tailored for different types of input, such as edge maps, depth, pose, and normals. So, our first step was to check if any of these modules helped to conserve our image. We selected two conditioning strategies that seemed most relevant to our task: *Canny edges* and *image normals*.

Canny Edges.

Figure 11 shows results obtained using ControlNet conditioned on Canny edges. The model faithfully followed the edge maps and successfully maintained the overall structure of the original image. However, this approach presented two key limitations:

1. Canny edges do not capture object color or texture information.
2. Edges often encode shadows—features that should ideally be removed in a relighting task.

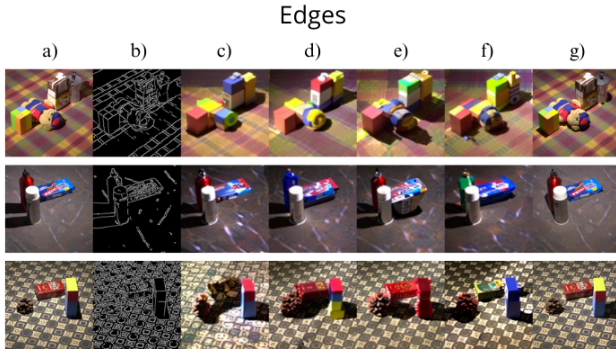


Fig. 11: Relighting results using IP-Adapter and ControlNet conditioned on Canny edges. a) Original input, b) extracted edge map, c–f) outputs with different seeds, g) ground truth.

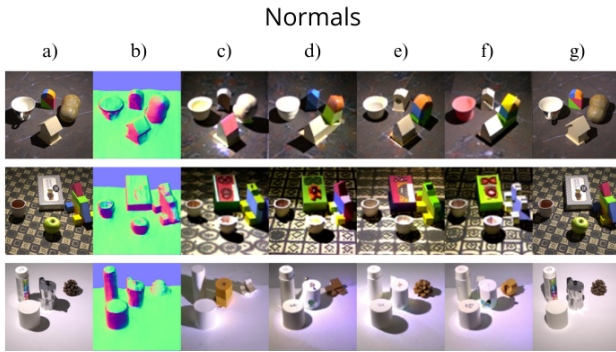


Fig. 12: Relighting results using IP-Adapter and ControlNet conditioned on normals. a) Input image, b) MiDaS-generated normals, c–f) generations with different seeds, g) ground truth.

As a result, while the output images preserved spatial structure, they lacked appearance accuracy and were not suitable for light-specific modifications.

Image Normals.

Next, we tested a ControlNet model conditioned on surface normals. In Figure 12, you can see the results of these tests. The normals are extracted using a model called Midas [16]. The extracted normals are good, but again, using this type of conditioning there is one major problem. Similar to edges, normals do not retain color or texture information, and that can be seen in the results generated. Additionally, patterned surfaces, such as the textured “floors” used in the dataset, were lost. However, although it has these disadvantages, the model does generate an image with a similar overall structure.

Albedo Conditioning.

After testing several pretrained ControlNet models, we thought that, for our task, the most suitable model would be one that conditions generation on image albedo maps. As explained in the introduction, the concept of albedo comes from intrinsic image decomposition, a technique used in computer vision and graphics to separate an image into its albedo (intrinsic color) and shading (illumination) components. Albedo maps retain only the diffuse reflectance of surfaces, removing lighting effects such as shadows and

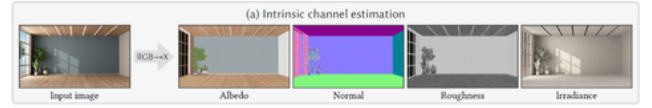


Fig. 13: Intrinsic decomposition results using RGBX. Output includes albedo, normals, roughness, and irradiance components.

highlights. This property makes them ideal for our task, albedo maps retain all the information we would like the model to preserve, and they remove all information we would like the model to ignore. A rough example of albedo and shading decomposition is shown in Fig. 1.

Unfortunately, no ControlNet variant exists that supports albedo inputs. Therefore, we trained a custom ControlNet using albedo maps as the conditioning signal. For this, we required a dataset containing: (a) albedo maps, (b) image captions, and (c) a wide range of visual variability to support generalization. We constructed this dataset by selecting a subset of the Laion-Aesthetic dataset [17] (used to train the original Stable Diffusion) and extracting albedo maps using the RGBX model [18]—a recent method developed by Adobe for intrinsic decomposition. This model is able to decompose an image into albedo, normals, roughness, and irradiance (shading); see Fig. 13 for an example of this decomposition.

However, the RGBX model itself does not always produce clean or consistent albedo maps. Figure 14 shows examples of albedo maps extracted from Laion-Aesthetic. To avoid training with incorrect albedo maps, we applied a two-step filtering process: (1) we discarded albedo maps with low gradient magnitude (i.e., flat color), and (2) we removed those dominated by gray values, which are typically over-smoothed. In Figure 14, red boxes highlight samples that would be excluded during this filtering stage.

While this filtering improved training quality, it does not address issues during inference: if RGBX produces a poor albedo at test time, the model cannot compensate for it.

Figure 15 shows the results obtained using our custom ControlNet trained on RGBX-generated albedo maps. Compared to previous approaches, the outputs are significantly more consistent with the original image and better reflect the desired lighting conditions. However, the effectiveness of this pipeline is heavily dependent on the quality of the extracted albedo maps.

6.1.5 Comparison between models

In Figure 16, we show an abstract diagram of what each of these methods does.

1. **Concept Sliders:** This method implies training a Lora for each of the layers of SD and at the end just adding the parameters together. Then this is trained using a few paired image prompts and weighing how much LoRA parameters weigh during the addition. This way ideally, you achieve a way to modify a single concept of an image. The idea was to train using relighting images from A to B. Where A and B are two positions of



Fig. 14: Examples of albedo maps extracted by RGBX from Laion-Aesthetic. Red boxes mark filtered-out samples.

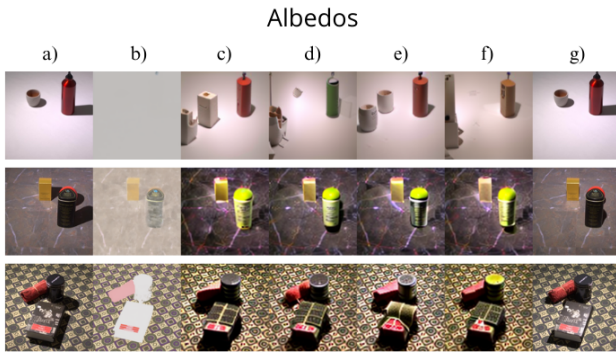


Fig. 15: Results of relighting using IP-Adapter and a custom ControlNet trained on albedo maps.

the light.

2. **IP-Adapter:** This method trains only the cross-attention modules of Stable Diffusion. Image prompts are encoded using CLIP, and a cross-attention module aligns these embeddings with the model’s latent space. Unlike Concept Sliders, IP-Adapter accepts both an image prompt and a textual prompt describing the desired lighting. Training was conducted on the RSR dataset with textual prompts specifying target illumination.
3. **ControlNet:** ControlNet duplicates the Stable Diffusion model and trains it with a large number of input-output pairs, approximately 10,000 examples, for a conditioning task. In this work, it was trained to preserve the scene’s structural integrity by conditioning on inputs such as Canny edges, normal maps, or albedo maps. Albedo maps provided the best performance in maintaining structure. The final idea was that ControlNet enforced scene consistency, while IP-Adapter was responsible for relighting via textual conditioning, thus separating structure preservation from lighting modification.

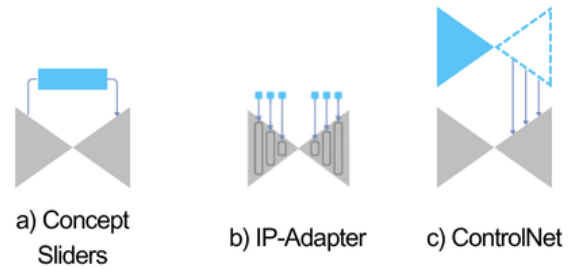


Fig. 16: Overview illustrating the training parameters (in blue) of the different conditioning approaches. a) Concept Sliders learns LoRA parameters to adjust specific image attributes via slider control. b) IP-Adapter aligns image embeddings with the latent space of SD by retraining the cross-attention modules. c) ControlNet makes a copy of the whole model and adds it to the original model parameters.

6.1.6 Problems

After extensive experimentation with Concept Sliders, IP-Adapter and ControlNet, we encountered several persistent limitations that hindered our ability to perform high-quality relighting. A summary of the main issues is as follows:

1. **Concept Sliders** was not able to understand the overall scene. This resulted in poor results as the model didn’t manage to move from A to B in a realistic manner. In addition, during inference, we only input the text prompt and the value of the slider. Therefore, we don’t have control over what is generated directly. We would need to introduce some kind of control mechanism if we wanted to force the generation to change a given scene and not completely generate a new scene from scratch.
2. **IP-Adapter** failed to preserve the structure and semantics of the original scene. Due to the significant changes in image content, it was not possible to reliably assess the quality of the relighting itself.
3. **IP-Adapter plus ControlNet with edge conditioning** preserved structural layout but lacked information about texture and color. Furthermore, edge maps retained shadow information, which is undesirable for relighting tasks.
4. **IP-Adapter plus ControlNet with normals** similarly failed to retain appearance and texture, and did not provide enough cues to guide lighting direction.
5. **IP-Adapter plus ControlNet with albedo maps** yielded the most promising results. However, it relied on the RGBX model heavily for intrinsic decomposition, which did not consistently produce high-quality albedo maps, especially at inference time, when no filtering was possible.

These limitations led us to explore alternative modeling paradigms. In particular, we turned to transformer-based *image-to-image* architectures, which offer more direct control over pixel-level transformations and are well-suited for learning dense mappings such as relighting.

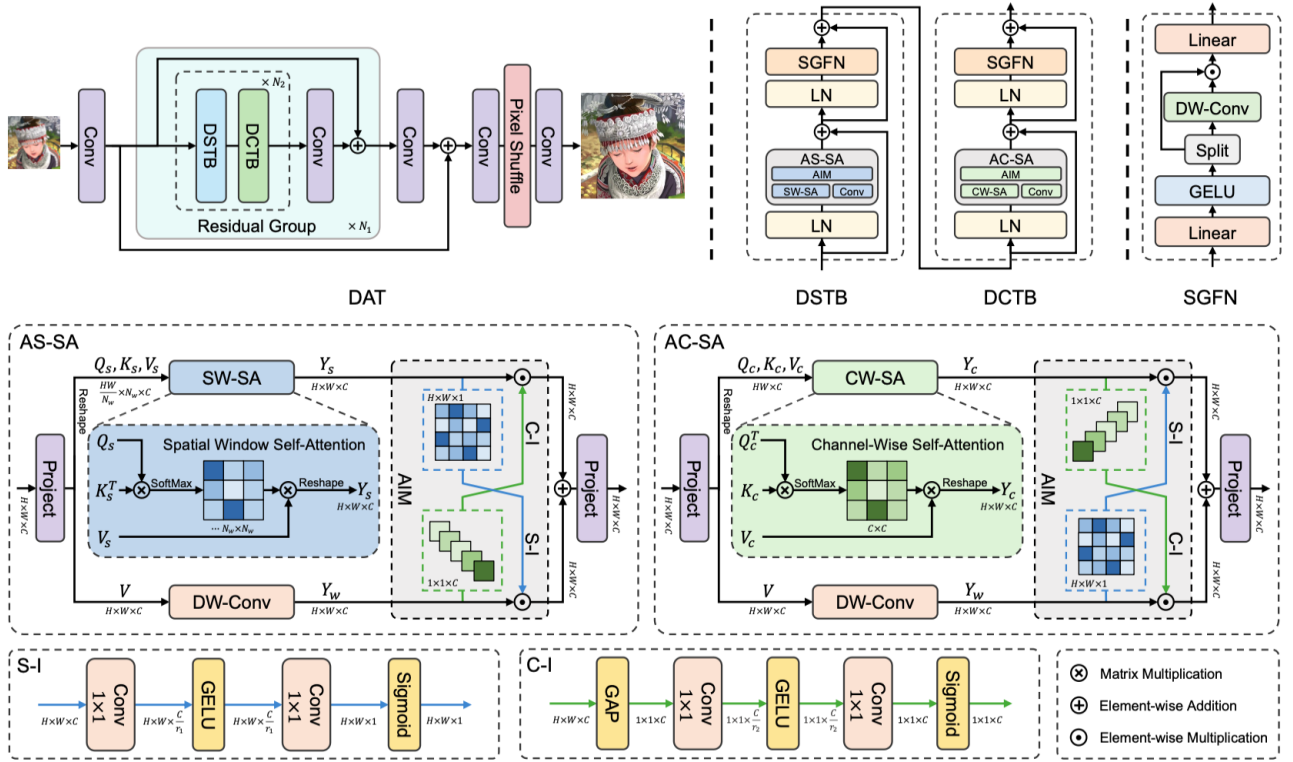


Fig. 17: Illustration of DAT architecture. Top-left: overall simplified structure; others: internal modules. DATB blocks include DSTB (dual spatial transformer block) and DCTB (dual channel transformer block).

6.2 Transformer-based relighting

To explore transformer-based alternatives, we reviewed recent state-of-the-art image-to-image architectures and selected two models that seem to be good based on their results in their corresponding fields.

- On one side we have **Dual Aggregation Transformer (DAT)** [6], although originally proposed for image super-resolution, DAT uses an interesting dual attention mechanisms that operates across spatial and channel dimensions, which, suggest strong potential for controllable image transformation tasks.
- On the other side we have **Restormer** [3] which is one of the best models for image to image tasks until now. It has been widely used by the community due to its great performance and for being a really efficient architecture. Initially, it was designed for denoising, but it has also been trained for deblurring, dehazing and deraining.

6.2.1 DAT

The DAT architecture is composed of multiple *residual groups*, each containing both spatial and channel attention blocks. While the original implementation uses six such groups, we reduced this to three due to memory constraints. Figure 17 presents the basic structure of DAT.

6.2.2 Initial Conditioning Strategy.

The main challenge in adapting DAT for relighting is how to introduce lighting information. Our first approach is illustrated in Figure 18. We encoded the desired lighting configuration, defined by pan and tilt angles, using a multi-layer perceptron (MLP), producing a feature vector. This vector was then broadcast spatially to match the resolution of the image features and concatenated to the feature map output of DAT’s first convolutional layer.

We tried different amounts for the hyperparameter C . The best results were obtained using a C of 180 which actually makes the input have 360 channels 180 describing the input and 180 describing the light prompt.

The model was trained over two days using a batch size of three—the maximum permitted by memory constraints. Qualitative results are shown in Figure 20. Interestingly, the model appears to attempt shadow removal (e.g., row 1) and to preserve partial lighting patterns in some cases (e.g., row 2). However, results were inconsistent, and the model struggled to clearly separate which aspects of the image to preserve and which to modify.

This led us to hypothesize that the light embedding introduced via simple concatenation was not sufficiently emphasized by the network. To address this, we looked for alternative conditioning mechanisms.

6.2.3 PromptIR-based Conditioning.

PromptIR proposes a unified image restoration model guided by prompt embeddings, which encode degradation-

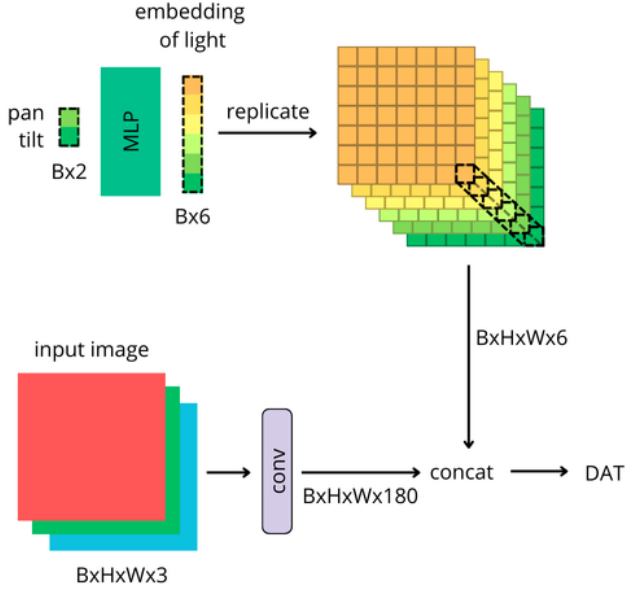


Fig. 18: Initial conditioning method: MLP encodes lighting angles into a spatially replicated embedding concatenated with early feature maps.

specific information.

What is interesting about this paper is not the model itself but the key idea behind the conditioning with prompts.

Figure 19 illustrates how we adapt this mechanism for our purpose. The prompt module takes pan and tilt values as input, which are transformed by an MLP and then replicated to produce a tensor of shape $\hat{H} \times \hat{W} \times 6$. This tensor is processed through a CNN to generate the prompt components.

Next, following the PromptIR approach, global average pooling (GAP) is applied to these prompt components, followed by a 1×1 convolution. The resulting vector is passed through a softmax function to produce dynamic weights. These weights are used to re-weight the original prompt components, implementing a dynamic weighting scheme.

The final output, referred to as the light prompt, is incorporated into the cross-attention mechanism of both DAT and Restormer models. Notice that $\hat{H}$ and $\hat{W}$ depend on where we are going to add the final embedding.

6.3 Final Results

6.3.1 Quantitative Results

To measure the quality of the predicted results we computed the PSNR (Peak To Noise Ratio) and SSIM (Structural Similarity) of all the possible pairs of images between the validation set. That is, for each scene, we make the model predict the relighted image for each of the other possible light sources we have as ground truth. From all the possible combinations we can make, we have a total of 6480 pairs of images in the validation set.

In Figure 21, we show the mean PSNR and SSIM of images going from one light direction to another. The axis we

show are sorted based on similarity. That is, we put images where fewer changes should be made closer together. On the diagonal, we can see the model has learned to identify when the image should not be changed, and we obtain a PSNR of 100 (it was actually infinity but clipped to 100) and SSIM of 1.0. Notice that SSIM goes from 0 to 1, so this means, the two compared images are exactly the same. As for the other combinations. There is not a huge difference between going from one particular direction or another. However, the middle squares seem to be darker indicating a slightly smaller PSNR and SSIM. This can be due to the intuitive idea that images that require more changes to arrive to the correct illumination have a tendency to be harder for the model.

However, the model seems to perform pretty well in all cases with a mean PSNR of 28.664 and a mean SSIM of 0.785 over the whole validation set.

6.3.2 Qualitative Results

Figure 23 presents qualitative results of the best-performing model at various training epochs. The model trained for 2000 epochs (second column) achieves the highest overall performance. With additional training, the model appears to overfit the fine-tuning dataset and the fixed light positions, resulting in rough transitions rather than smooth lighting changes. Conversely, earlier epochs exhibit a slight decline in performance, indicating that 2000 epochs provide an optimal balance between memorization and generalization.

The figure includes two examples from the training set. The model successfully learns to reproduce images with discrete light positions while also generalizing to unseen light positions, enabling smooth transitions between known lighting conditions.

The validation results involve completely unseen images with backgrounds different from those in the training set. The model effectively relights these scenes and accurately predicts the corresponding shadows.

Lastly, we show out of domain images taken with an smartphone. The model is able to correctly predict the image with the mug and the image with the two color boxes. However, it seems to have difficulties when relighting the octopus image as it creates some unwanted artifacts in the logotype of the laptop. We believe these artifacts are due to a lack of training data with non-Lambertian effects.

6.4 Conclusions and Continuation Lines

In this report, we have presented the progress made toward developing a model for controllable and realistic image relighting. We began by evaluating the performance of diffusion-based methods, incorporating popular control techniques such as Concept Sliders, IP-Adapter, and ControlNet. While these approaches provided partial improvements, they suffered from limitations including poor preservation of scene structure, entanglement of content and lighting, and reliance on unreliable conditioning inputs like automatically extracted albedo maps.

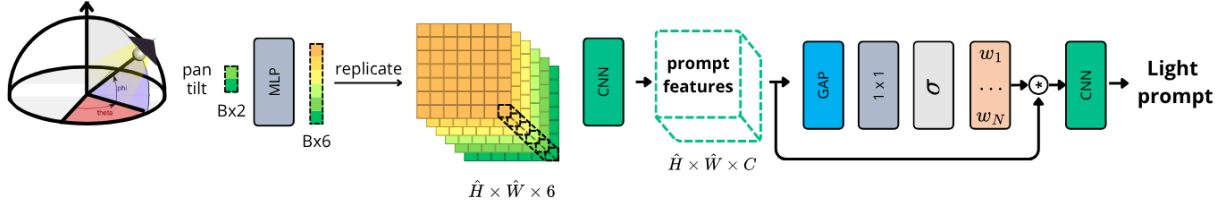


Fig. 19: PromptIR style of getting the light prompt to condition the transformer model. It takes pan and tilt as input and outputs a vector of size $\hat{H} \times \hat{W} \times C$ representing this lighting prompt.



Fig. 20: Qualitative results using DAT with light embedding via early concatenation. a) Input, b) Prediction, c) Ground truth.

In light of these challenges, we shifted our focus toward transformer-based image-to-image architectures, specifically, we explored the Dual Aggregation Transformer (DAT) and Restormer. The first tests using DAT showed promise but still seemed to have trouble performing the full relighting process correctly. We hypothesized it could be due to the way we were introducing the light information so we decided to change it to a more advanced conditioning scheme inspired by PromptIR to integrate lighting cues more effectively.

However, the results still were not as good as expected, and we decided to change the transformer model to Restormer. Training this model using the PromptIR style to introduce the information gave really good results.

Therefore, we have successfully achieved all the objectives of the project. We have studied different approaches of doing image relighting. We have designed an architecture that conforms 'a model to generate a relit image' and 'mechanisms to control relighting based on user specified light source positions or other lighting parameters'. We have trained this architecture on multiple datasets and we have evaluated it in unseen images, even out of domain images, to test its generalization capabilities.

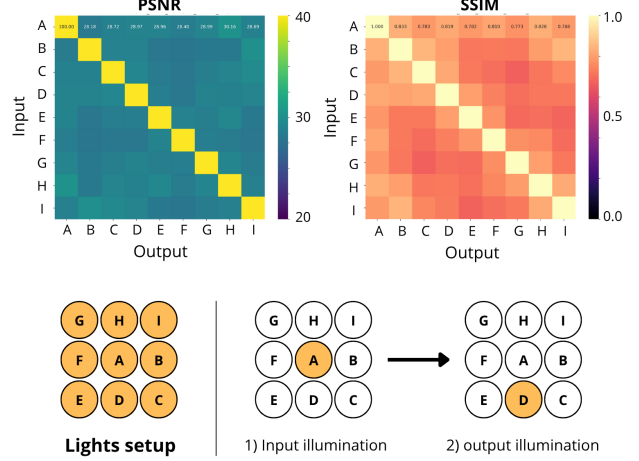


Fig. 21: Quantitative evaluation results. The matrices display the mean PSNR and SSIM scores for all scene transitions in the validation set, where each cell corresponds to the performance of relighting from one illumination condition to another. Overall, the matrix is relatively homogeneous, with an average PSNR of 28.664 and SSIM of 0.785, indicating consistent performance across lighting pairs. The bottom left grid illustrates the spatial arrangement of the light sources. Notice that for each image, only one of the A-I bulbs is active. These bulbs are directly on top of the object, so having light A active means the object has a light coming from the top, B means the light is slightly to the right, and so on. The bottom right corner shows an example of what a specific matrix entry means, cell [A, D] represents the average PSNR and SSIM obtained when transforming scenes from lighting configuration A to D.

As future work this model can be tested on more datasets like VIDIT, explained in section 4. And other image-to-image transformer models could be tested to compare the performance of different architectures. To achieve better results using diffusion models I believe there is a need for a bigger dataset with paired albedo/image, so another line of work may be to create a big and variate dataset to be able to train a diffusion model from scratch.

This has been a really challenging project, as many of the initial attempts did not yield promising results. In addition, working with such a new topic to me as diffusion, I felt lost in many moments and did not know how to continue. Special mention to David Serrano-Lozano (and obviously my tutor, Javier Vazquez) for the great help during the realization of this work.

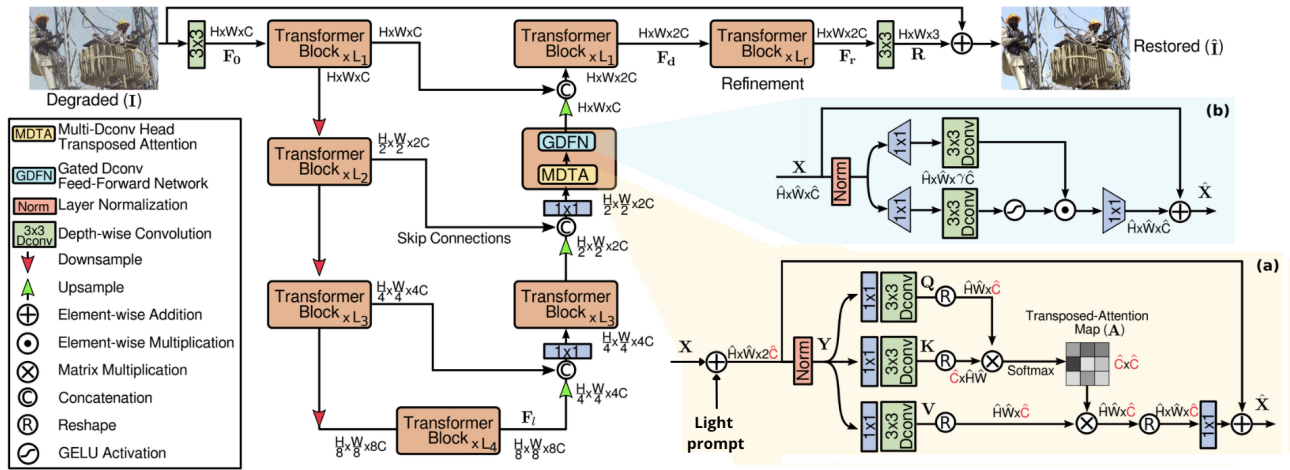


Fig. 22: Final architecture of the best-performing model. The base architecture is the model "Restormer," and we introduce the light prompt inside the cross-attention module. See figure 19 to view how the light prompt is generated.

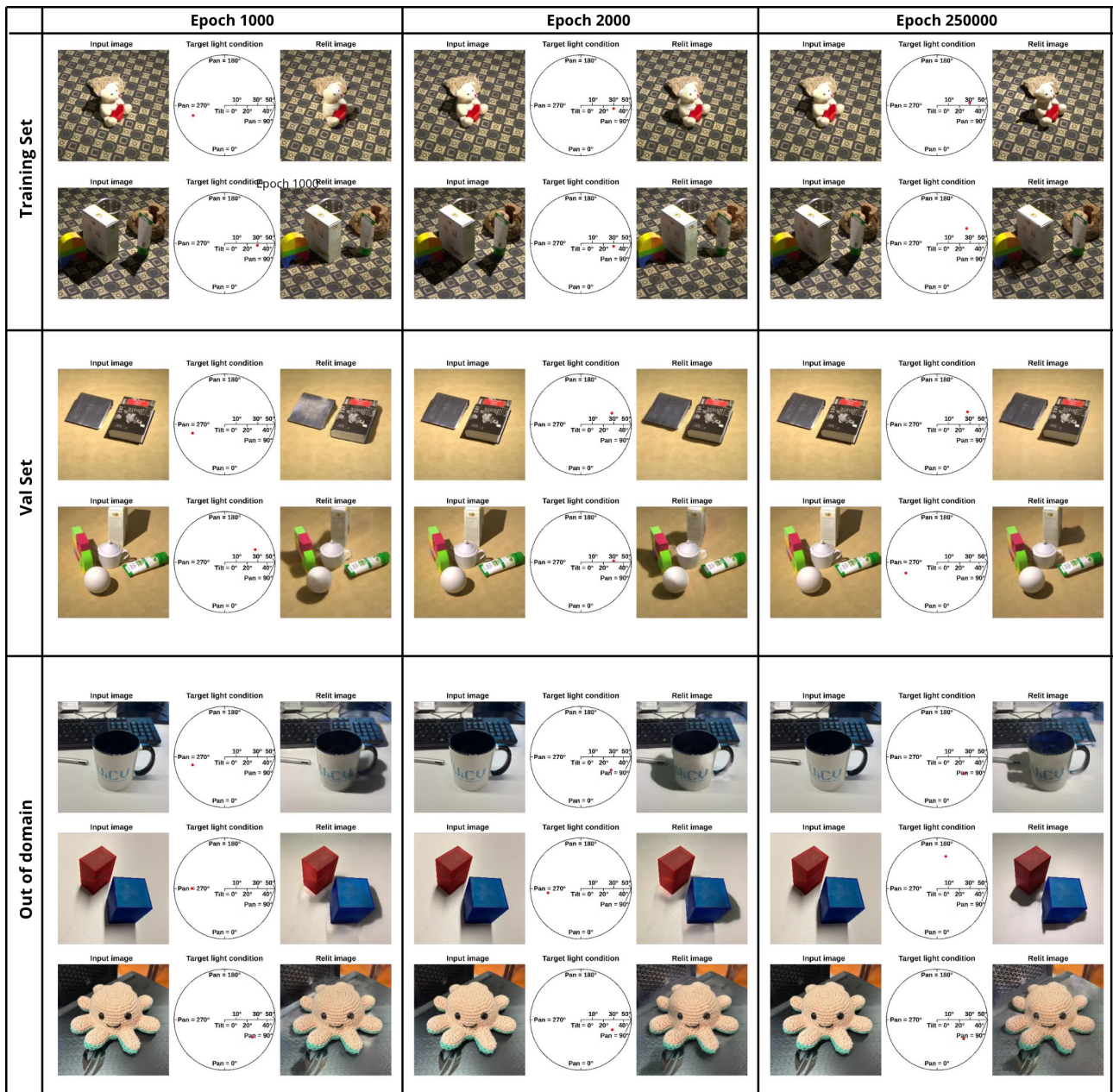


Fig. 23: Please view the following link to view the results with movement. The image shows qualitative results of the best performing model in the training set, validation set and totally out of domain images taken with an Iphone 14.

REFERENCES

- [1] Y. Yang, H. A. Sial, R. Baldrich, and M. Vanrell, "Relighting from a single image: Datasets and deep intrinsic-based architecture," 2024. [Online]. Available: <https://arxiv.org/abs/2409.18770>
- [2] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2019. [Online]. Available: <https://arxiv.org/abs/1812.04948>
- [3] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," 2022. [Online]. Available: <https://arxiv.org/abs/2111.09881>
- [4] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," 2023. [Online]. Available: <https://arxiv.org/abs/2303.06705>
- [5] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, and T. Lu, "Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method," 2022. [Online]. Available: <https://arxiv.org/abs/2212.11548>
- [6] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," 2023. [Online]. Available: <https://arxiv.org/abs/2308.03364>
- [7] L. Zhang, A. Rao, and M. Agrawala, "Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=u1cQYxRI1H>
- [8] J. Cha, M. Ren, K. K. Singh, H. Zhang, Y. Hold-Geoffroy, S. Yoon, H. Jung, J. S. Yoon, and S. Baek, "Text2relight: Creative portrait relighting with text guidance," 2024. [Online]. Available: <https://arxiv.org/abs/2412.13734>
- [9] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023. [Online]. Available: <https://arxiv.org/abs/2302.05543>
- [10] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023.
- [11] R. Gandikota, J. Materzyńska, T. Zhou, A. Torralba, and D. Bau, "Concept sliders: Lora adapters for precise control in diffusion models," *arXiv preprint arXiv:2311.12092*, 2023.
- [12] R. Gandikota, Z. Wu, R. Zhang, D. Bau, E. Shechtman, and N. Kolkin, "Sliderspace: Decomposing the visual capabilities of diffusion models," 2025. [Online]. Available: <https://arxiv.org/abs/2502.01639>
- [13] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," 2015. [Online]. Available: <https://arxiv.org/abs/1512.03012>
- [14] L. Murmann, M. Gharbi, M. Aittala, and F. Durand, "A multi-illumination dataset of indoor object appearance," in *2019 IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.
- [15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. [Online]. Available: <https://arxiv.org/abs/2112.10752>
- [16] R. Birkel, D. Wofk, and M. Müller, "Midas v3.1 – a model zoo for robust monocular relative depth estimation," 2023. [Online]. Available: <https://arxiv.org/abs/2307.14460>
- [17] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: An open large-scale dataset for training next generation image-text models," 2022. [Online]. Available: <https://arxiv.org/abs/2210.08402>
- [18] Z. Zeng, V. Deschaintre, I. Georgiev, Y. Hold-Geoffroy, Y. Hu, F. Luan, L.-Q. Yan, and M. Hašan, "Rgbx: Image decomposition and synthesis using material- and lighting-aware diffusion models," in *ACM SIGGRAPH 2024 Conference Papers*, ser. SIGGRAPH '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: <https://doi.org/10.1145/3641519.3657445>