SPADE

A Synthetic Paired Dataset for Specular-Diffuse Video Decomposition

Matthew Barrett

Advisor: Dr. Ruth Fong

Submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Arts

Department of Computer Science

Princeton University

April 2025

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Matthew Barrett

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Matthew Barrett

Abstract

Computer vision systems struggle with specular highlights—bright spots that obscure underlying visual information—yet video-based removal methods remain unexplored due to the absence of temporally consistent training data. This thesis demonstrates that incorporating temporal information significantly improves highlight removal quality and consistency, addressing a critical gap in computational photography. I introduce SPADE, the first dataset of paired specular-diffuse video sequences, created through controlled synthetic rendering of 250 objects under varied conditions. An ablation study comparing frame-based and sequence-based neural architectures quantifies temporal processing benefits: the temporal model achieves 16.2% higher PSNR, 10.2% better SSIM, and 2.0% improved temporal consistency. Material analysis reveals these improvements are most pronounced for metallic surfaces and moderate camera movements. Beyond highlight removal, this work establishes a paradigm for leveraging temporal information in appearance decomposition tasks, with applications in augmented reality, film production, and medical imaging.

Contents

Chapter 1

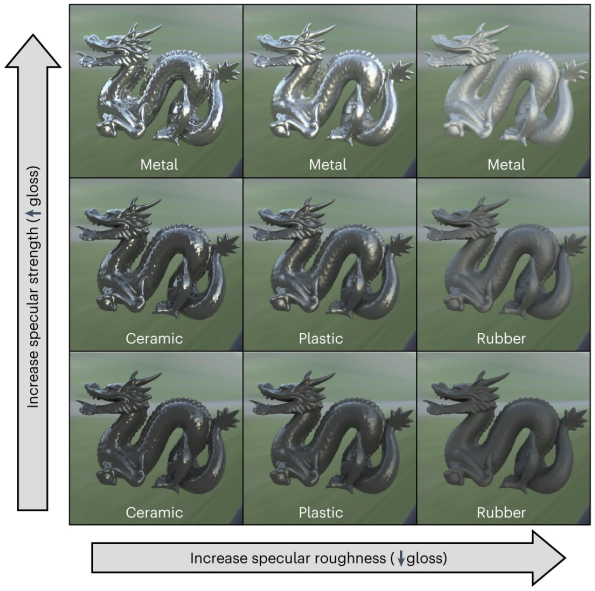
Introduction

Specular highlights are bright spots that appear on objects when light reflects directly toward the viewer. Unlike diffuse reflection, which scatters light in all directions and reveals an object's true color and texture, specular reflection bounces light like a mirror, creating these characteristic bright spots. Though a natural part of how we perceive the world, specular highlights often pose significant challenges in computer vision applications. They can obscure underlying texture, distort color perception, and create inconsistencies in tracking and recognition systems [3].

The removal of these highlights represents a longstanding challenge in computer vision and image processing. For decades, researchers have approached this problem from different angles—some attempting to model the physical behavior of light [18], others leveraging color space transformations [25], and more recently, applying data-driven methods [3] to learn the mapping between specular and diffuse components. Despite significant advances, the problem remains particularly challenging in video applications, where consistency across frames becomes an essential yet elusive quality.

When examining the literature on specular highlight removal, a clear pattern emerges: while substantial progress has been made for still images, video applications lag behind. This discrepancy stems not from a lack of interest, but rather from fundamental limitations in

(a) Real-world scene with various specular highlights on shiny surfaces



(b) How Specular roughness and specular strength lead to different materials

Figure 1.1: **(a)** Examples of specular highlights in real-world scenes adapted from Schmid et al.[17]. Specular highlights appear as bright spots where light directly reflects off surfaces. **(b)** The appearance varies significantly depending on material properties (metal, rubber, ceramic, plastic) and illumination conditions [17].

both methodology and available data. Frame-by-frame application of image-based methods often produces temporal inconsistencies, such as flickering, uneven changes, and artifacts that move unnaturally across frames [8]. These issues can be more distracting than the original highlights themselves, undermining the very purpose of highlight removal.

What makes video-based highlight removal so challenging? The answer lies partly in the dynamic nature of specular reflections themselves. As objects or cameras move, highlights shift across surfaces in complex ways that depend on viewing angle, lighting conditions, and material properties [14]. Unlike object textures that remain consistent, specular highlights are viewpoint-dependent phenomena that transform with every frame. Without accounting for this temporal relationship, algorithms struggle to maintain consistency.

The state of research in this field is further hampered by a critical absence: no large-scale video dataset exists with paired specular and diffuse sequences. While datasets like

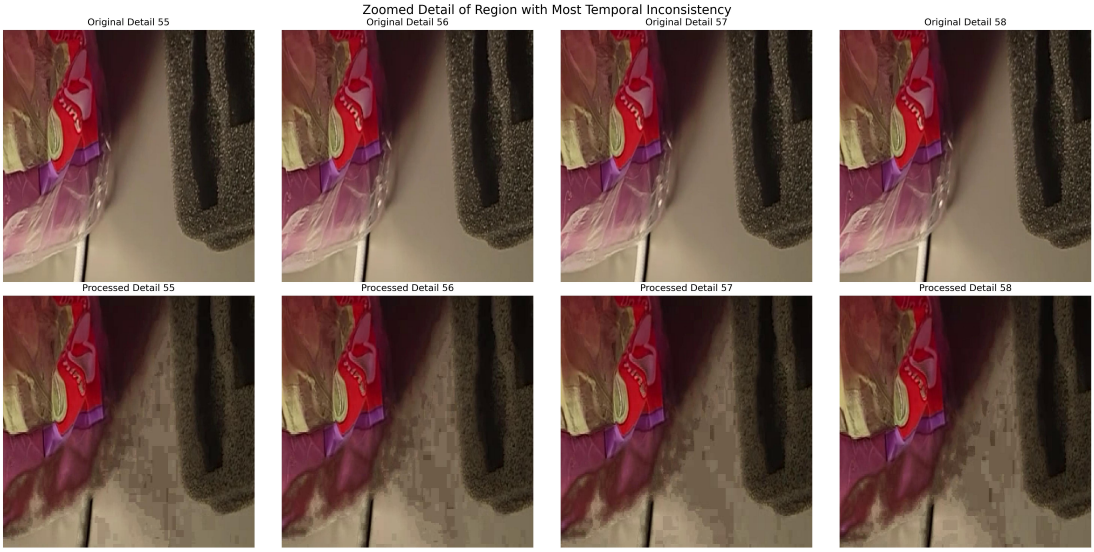


Figure 1.2: The temporal consistency problem in frame-by-frame processing using Ramos et al.'s method [16]. Sequential frames (from left to right) processed independently exhibit flickering artifacts and inconsistent highlight removal that become apparent when viewed as video.

SHIQ [4] and PSD [26] have advanced still-image methods by providing paired examples, video methods lack equivalent resources. This absence creates a fundamental obstacle for developing and evaluating temporally-aware approaches, particularly those based on deep learning that require substantial training data.

Motivated by these challenges, this work introduces SPADE (SPecular And Diffuse Environments), the first synthetic video dataset specifically designed for specular highlight removal. SPADE comprises paired sequences—identical in every aspect except for the presence of specular reflections—allowing direct supervision for learning-based methods. By controlling material properties, camera movement, and lighting conditions within a physically-based rendering environment, SPADE creates realistic specular effects with guaranteed ground truth.

The dataset includes 250 carefully selected object pairs rendered under varying lighting conditions, with camera paths under controlled randomization. Each object is captured in both specular and diffuse configurations, creating perfectly aligned pairs that isolate the specific effect of specularity. This controlled approach overcomes fundamental limitations

Figure 1.3: Example from the SPADE dataset showing the same object under identical camera path and lighting, with specular highlights (left) and without (right).

of real-world capture, where true diffuse references are nearly impossible to obtain under identical conditions [26].

Beyond the dataset itself, this work presents a comprehensive analysis of the benefits of temporal information in highlight removal. Through an ablation study comparing frame-based and sequence-based approaches, I quantify the improvements gained by incorporating temporal context. This analysis addresses a fundamental question: how much does temporal information actually help in specular highlight removal? The results demonstrate that temporal consistency improves slightly when models can leverage information across frames, suggesting that future research should consider video-based approaches over frame-by-frame processing.

The contributions of this work are threefold. First, I introduce the SPADE dataset, comprising 250 paired specular-diffuse video sequences with perfect alignment and controlled variation. Second, I develop an automated pipeline for generating such data, enabling further expansion and customization beyond the initial dataset. Finally, I conduct a rigorous ablation study that quantifies the benefits of temporal processing for highlight removal, establishing a baseline for future research in this direction.

This thesis begins by reviewing the theoretical foundations and prior work in specular highlight removal, then details the approach and implementation of the SPADE dataset. I then analyze existing models and present the results of the temporal ablation study, concluding with implications for future research in video-based highlight removal.

Chapter 2

Background and Related Work

Light interacts with different surfaces in a multitude of ways, giving materials their characteristic look, from shiny metals to matte ceramics. This visual richness stems from how materials interact with incident light, a phenomenon that has fascinated researchers since the foundations of computer graphics and vision. Understanding and modulating these interactions, particularly the separation of specular highlights from diffuse reflection, remains a central challenge with applications ranging from material recognition to augmented reality and medical imaging.

2.1 The Physics of Specular Reflections

When light strikes a surface, it interacts in multiple ways. Some wavelengths are absorbed, others penetrate the material and scatter internally before reemerging (creating diffuse reflection), and some reflect directly from the surface following the law of reflection where the angle of incidence equals the angle of reflection (as seen in Figure 2.1). This last phenomenon creates specular highlights which are the bright spots we see on glossy objects that reveal more about the lighting environment than the object itself.

The seminal work by Phong in 1975 [14] formalized this understanding into a computational model that has influenced computer graphics ever since. Phong's approach separated

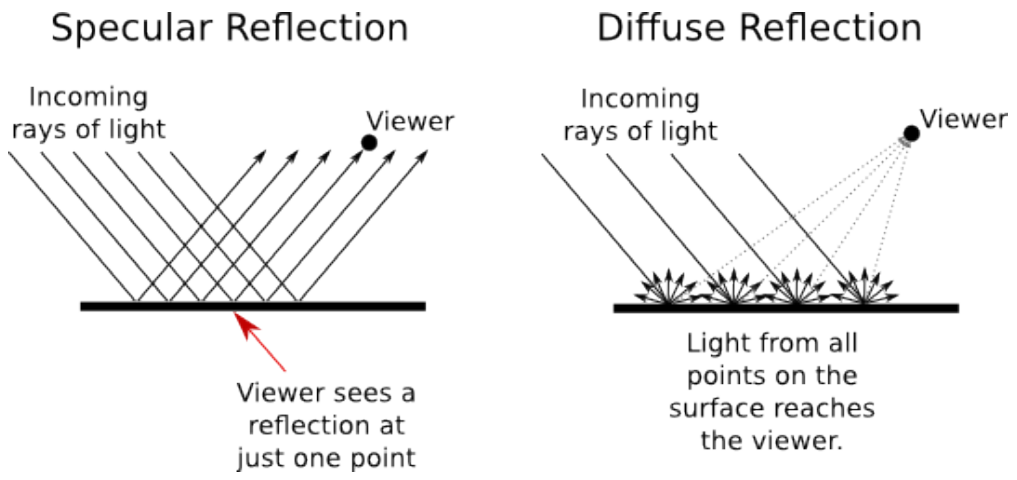
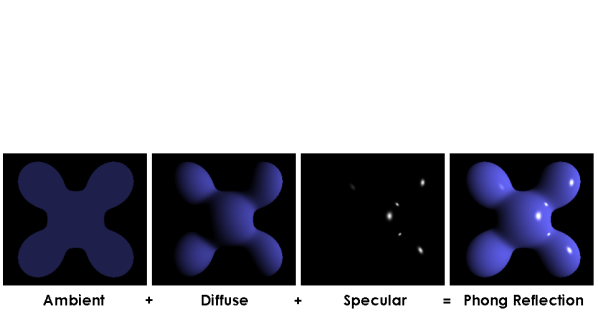
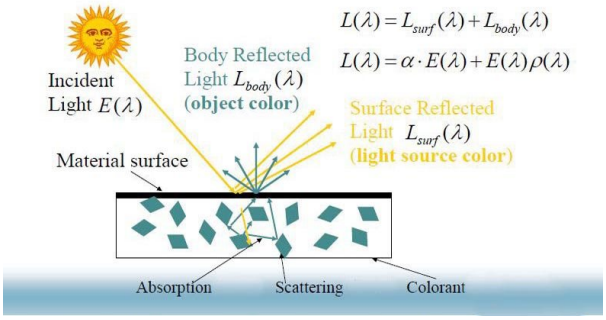


Figure 2.1: Physics of specular versus diffuse reflections [1]

reflection into ambient, diffuse, and specular components, allowing these elements to be modeled independently. Building upon this foundation, Shafer's dichromatic reflection model [18] in 1985 provided a more physically-based framework, separating light reflection into interface reflection (specular) and body reflection (diffuse). These models established the theoretical foundation that would guide decades of research in specular-diffuse decomposition.



(a) Phong equation's visualized [21]



(b) The Dichromatic Reflection Model [7]

Figure 2.2: Illustration of light reflection models. (a) Phong model showing ambient, diffuse, and specular components. (b) The dichromatic reflection model decomposing reflected light $L(\lambda)$ into two components: surface reflection $L_{surf}(\lambda)$ (yellow, retaining light source color) and body reflection $L_{body}(\lambda)$ (teal, revealing object color).

Real-world materials, however, exhibit far more complex behaviors than these idealized models. Surface roughness, subsurface scattering, and surface inconsistencies all affect how

specular highlights form and appear. Physically-based rendering techniques gradually incorporated these complexities through more sophisticated bidirectional reflectance distribution functions (BRDFs) and global illumination models. While these advancements improved the rendering of realistic materials, they simultaneously highlighted the difficulty of the inverse problem: how to decompose observed images back into their constituent reflection components.

2.2 Evolution of Highlight Removal Approaches

Early approaches to specular highlight removal relied heavily on the physical models mentioned above. Researchers like Tan [25] exploited the properties of the dichromatic reflection model, using color information to identify and remove specular components. His method, based on the observation that specular reflection preserves the hue of the light source while diffuse reflection reveals the object's true color, enabled highlight removal through careful color space analysis. These physics-based approaches established a foundation, but they struggled with complex real-world scenarios where assumptions about material properties or lighting conditions were violated.

Another strand of research leveraged multi-image techniques, capturing the same scene under different polarization states or viewpoints. The work by Mallick et al. [13] demonstrated that specularity could be effectively removed using partial differential equations (PDEs) that iteratively erode the specular component. While effective, these approaches often required multiple images or specialized capture techniques, limiting their practical applications in everyday scenarios.

The field evolved significantly with the introduction of optimization-based methods. Approaches by Tan [25] and later by Suo et al. [23] formulated specular removal as an energy minimization problem, incorporating spatial coherence and color consistency constraints. These methods improved robustness but often required careful parameter tuning and re-

mained computationally expensive. Furthermore, they typically processed images independently, providing no mechanism for ensuring consistency when applied to video sequences.

A significant advancement came with real-time approaches that made highlight removal more practical for applications. Yang et al. [27] introduced a method using bilateral filtering that could remove highlights 200 times faster than previous techniques. Their key observation—that the maximum diffuse chromaticity changes smoothly in local patches—allowed them to propagate diffuse properties from non-highlight regions to specular areas efficiently. Similar efficiency-focused work by Shen and Zheng [19] using intensity ratios and by Souza et al. [22] with efficient pixel clustering further accelerated highlight removal, making it viable for real-time applications.

The deep learning revolution transformed highlight removal just as it did many other computer vision tasks. Starting around 2019, researchers began exploring convolutional neural networks for this problem. Fu et al. [3] showed that networks could learn the mapping between specular and diffuse images directly from data, outperforming traditional methods without explicit physical modeling. Their approach, based on observations about the sparse distribution of highlights and the low-rank nature of diffuse components, achieved impressive results on real-world images with complex textures and materials.

Recent years have seen increasingly sophisticated network architectures. Multi-stage approaches like those by Wu et al. [26] first detect highlight regions through specialized attention mechanisms, then focus processing on these areas while preserving details elsewhere. Fu et al. [4] developed a multi-task network for joint highlight detection and removal, leveraging the synergy between these closely related tasks. These advances have pushed the state-of-the-art for still images, achieving remarkable results on benchmark datasets.

Despite these advancements, video-based highlight removal has received comparatively little attention. The few existing approaches either apply image methods frame-by-frame [8] or add simple temporal constraints without fundamentally addressing the temporal nature of specular reflections. This gap stems partly from methodological challenges but more

significantly from the absence of suitable datasets for training and evaluation.

2.3 The Dataset Challenge in Specular Highlight Removal

Progress in specular highlight removal has been closely tied to the availability of appropriate datasets. The field has witnessed a progression from small, carefully captured image sets to larger, more diverse collections, each addressing particular limitations of its predecessors.

The SHIQ dataset [4] represented an early milestone, providing synthetic images of objects with and without specular components. While groundbreaking, SHIQ made simplifying assumptions about highlight characteristics, modeling them primarily as white additions to the diffuse component. This assumption, while computationally convenient, fails to capture the complex coloration that highlights inherit from both light sources and material properties in real scenarios.

Real-world capture approaches emerged to address these limitations. The PSD dataset [26], created using polarization filters, demonstrated the possibility of acquiring paired specular-diffuse images from actual scenes. Despite its realism, PSD faced inevitable alignment challenges—slight movements between captures, intensity variations, and polarization artifacts all complicated the ground truth. Furthermore, the controlled laboratory conditions yielded limited diversity in materials, lighting, and scene complexity.

Table 2.1: Comparison of datasets for specular highlight removal

Dataset	**Type**	**Size**	**Paired?**	**Temporal?**
SHIQ [4]	Synthetic	16K images	Yes	No
PSD [26]	Real	13K images	Yes	No
S-LIGHT [12]	Synthetic	7K images	Yes	Limited
Fu et al. [5]	Synthetic	135K images	Yes	No
SPADE	Synthetic	500 videos	Yes	Yes

Table 2.2: Overview of SPADE Dataset Statistics

Statistic	**Value**
Initial objects processed	387
Final video pairs in dataset	250
Frames per sequence	60
Frame rate	30 fps
Resolution	512×512 px
Total frames	30,000
HDRI environments	798

More recent datasets like those by Fu et al. [5] and S-LIGHT [12] have pushed toward greater scale and diversity while maintaining paired correspondences. These synthetic datasets render thousands of 3D models under various lighting conditions, creating more challenging benchmark environments. They address many limitations of earlier collections but remain focused on still images, offering no insight into temporal dynamics.

The absence of video datasets with paired specular-diffuse sequences represents a critical gap. When users need to process videos, they typically train on still image datasets and apply their methods frame by frame, accepting the temporal inconsistencies that inevitably arise [8]. Without ground truth for temporal behavior, it becomes impossible to properly evaluate or improve these aspects.

This gap motivated the creation of SPADE, which provides not just paired frames but complete paired sequences. By rendering identical camera paths with and without specularity, SPADE isolates exactly the phenomenon of interest while maintaining perfect frame-to-frame correspondence. This approach enables direct supervision for temporal aspects of highlight removal, opening new research directions that were previously impossible to explore systematically.

2.4 The Temporal Dimension of Highlight Removal

Specular highlights exhibit distinct temporal behavior that fundamentally differs from object textures and diffuse shading. As objects move relative to light sources and viewers, highlights shift across surfaces following the law of reflection. This movement creates complex patterns that depend on geometry, material properties, and lighting configurations. Without accounting for these temporal relationships, frame-by-frame processing inevitably produces inconsistencies [8].

The few attempts at video-based highlight removal in the literature have approached the problem through various lenses. Some methods employ temporal filtering as a post-processing step, smoothing results across frames after independent processing [6]. Others incorporate optical flow or other motion cues to propagate information between adjacent frames, leveraging motion estimation to maintain consistency [24]. While these approaches show improvements over purely frame-based methods, they treat temporal consistency as a secondary concern rather than an integral part of the process.

More fundamentally, existing approaches fail to model how specular highlights actually behave over time. The movement of highlights follows physical laws that could inform better removal strategies, yet current methods rarely incorporate this knowledge. This limitation stems partly from methodology but primarily from the absence of training data that could help models learn these temporal patterns.

The gap between image and video approaches mirrors a broader pattern in computer vision, where image understanding historically preceded video understanding. Just as object recognition, segmentation, and generation all eventually extended from images to video, highlight removal appears poised for a similar transition. SPADE aims to facilitate this transition by providing the necessary data foundation.

By quantifying the performance gap between frame-based and sequence-based methods, this work establishes the importance of temporal modeling in highlight removal. The ablation study presented later demonstrates that incorporating temporal information significantly im-

proves both removal quality and consistency, suggesting that future research should prioritize sequence-based approaches over independent frame processing.

Chapter 3

Approach

Creating a dataset for specular highlight removal presents unique challenges that influence every aspect of the design process. The core requirement—paired sequences with and without highlights under identical conditions—cannot be achieved through traditional capture methods. This fundamental constraint led to a synthetic approach that leverages physically-based rendering to generate perfectly aligned pairs. This chapter explores the reasoning behind these design decisions and details the approach developed to create SPADE.

3.1 Conceptual Foundations

The specular highlight removal problem requires training data that isolates exactly one variable: the presence or absence of specular reflection. All other elements—lighting, geometry, camera position, object color, and texture—must remain identical between paired examples. This perfect correspondence enables algorithms to learn precisely what constitutes a highlight and how to remove it without affecting underlying content.

In real-world scenarios, achieving such correspondence proves nearly impossible. Consider the challenges of capturing the same scene with and without specular components: any physical method to reduce specularity (such as polarization filters or surface treatments) inevitably alters other aspects of the scene. Polarization affects not just specular reflection

but also transmission and subsurface scattering [13]. Surface treatments change texture and color. Multiple captures introduce alignment problems as well as lighting inconsistencies [26]. These cascading effects contaminate the ground truth, introducing variables beyond mere specularity.

The problem compounds dramatically for video sequences. Real objects cannot be captured along identical camera paths with different material properties. Even with stationary objects and moving cameras, ensuring identical paths between captures presents formidable challenges. The slightest deviation creates misalignment that undermines the training signal, teaching models to accommodate these errors rather than focusing on highlight removal.

Synthetic data generation emerges as not merely convenient but necessary to overcome these fundamental limitations. By rendering scenes in a physically-based environment, we can maintain perfect control over every aspect of the imaging process. The same object, under identical lighting, following the exact same camera path, can be rendered with different material properties to isolate specular effects. This approach ensures that differences between paired sequences stem exclusively from specularity, creating ideal training data for highlight removal algorithms.

Beyond the practical necessity, synthetic data offers additional advantages for this specific problem. The ability to systematically vary material properties, lighting conditions, and camera movements enables comprehensive coverage of the problem space. We can generate examples spanning the full range of specularity, from subtle highlights on rough surfaces to intense reflections on mirror-like materials. This diversity helps models generalize to real-world scenarios they might encounter after training.

3.2 Design Philosophy

The SPADE dataset aims to bridge the gap between existing still-image collections and the needs of video-based highlight removal. This goal influenced several key design decisions

regarding diversity, realism, and practical utility.

Diversity in SPADE operates along multiple dimensions. The selection of 387 distinct objects spans various categories including household items, vehicles, furniture, and decorative pieces. These objects vary greatly in complexity and concavity. Surface properties vary similarly, including metals, plastics, ceramics, wood, and other common materials. This object diversity ensures models don't overfit to particular shapes or material types.

For each object, rendering occurs under varying lighting conditions. Rather than relying solely on simple point lights, SPADE incorporates high dynamic range imaging (HDRI) environments captured from real-world scenes. These environment maps provide realistic, complex illumination that creates natural highlight patterns. The combination of environment lighting with strategic point lights creates scenes that mimic professional studio setups, enhancing realism while maintaining control.

Camera movement represents another crucial dimension of diversity. Despite innovations in image stabilization, handheld cameras still result in some unpredictable, noisy movement. SPADE mimics this behavior through a constrained random walk algorithm. The camera follows a path on the surface of a sphere centered on the object, with smooth transitions between keyframes and small variations in movement speed. This approach creates natural-looking sequences that avoid the mechanical feel of simple circular paths.

Balancing realism with control presented a recurring challenge throughout development. While photorealism might seem the ultimate goal, it can actually undermine utility if it introduces uncontrolled variables. SPADE prioritizes physically accurate specularity behavior over photorealistic complexity in other areas. Materials follow principled BSDF models with physically plausible parameters, ensuring that specular behavior matches real-world expectations. Lighting creates natural shadow patterns and global illumination effects. However, the dataset avoids certain complexities—like elaborate backgrounds or atmospheric effects—that could distract from the core highlight removal task.

The design also considers practical utility for researchers and users. Sequences maintain

a consistent 512×512 pixel resolution, balancing detail with computational efficiency. Each video contains 60 frames at 30 frames per second, providing sufficient temporal context while remaining manageable for processing. Camera paths ensure the target object remains well-framed throughout the sequence, maximizing the usable content in each frame. These decisions optimize the dataset for real-world research applications, where computational resources often limit what's practically feasible.

3.3 Technical Approach

Creating SPADE required developing a robust, automated pipeline capable of handling hundreds of diverse objects with minimal manual intervention. The pipeline orchestrates multiple stages: asset preparation, material configuration, lighting setup, camera animation, and synchronized rendering.

The process begins with asset preparation. After scraping Poly Haven, a public 3D asset library, each 3D model undergoes normalization to ensure consistent scale and positioning relative to the world origin. This preprocessing addresses a common challenge in 3D asset collections [11], where models often use inconsistent scales and coordinate systems. A custom algorithm analyzes each model's bounding box and applies appropriate transformations to center it at the origin and scale it to a standardized size. This normalization ensures that subsequent stages—particularly camera positioning—function consistently across all objects regardless of their original configuration.

Material configuration represents one of the most critical aspects of the pipeline. For each asset, the system creates two distinct material setups: a specular version that exhibits natural highlights, and a matte version that suppresses specular reflection while preserving all other properties. The specular configuration uses physically-based parameters with moderate specularity (1.0) and low roughness (0.05), creating clear, defined highlights without reaching unrealistic mirror-like reflection. The matte configuration maintains identical

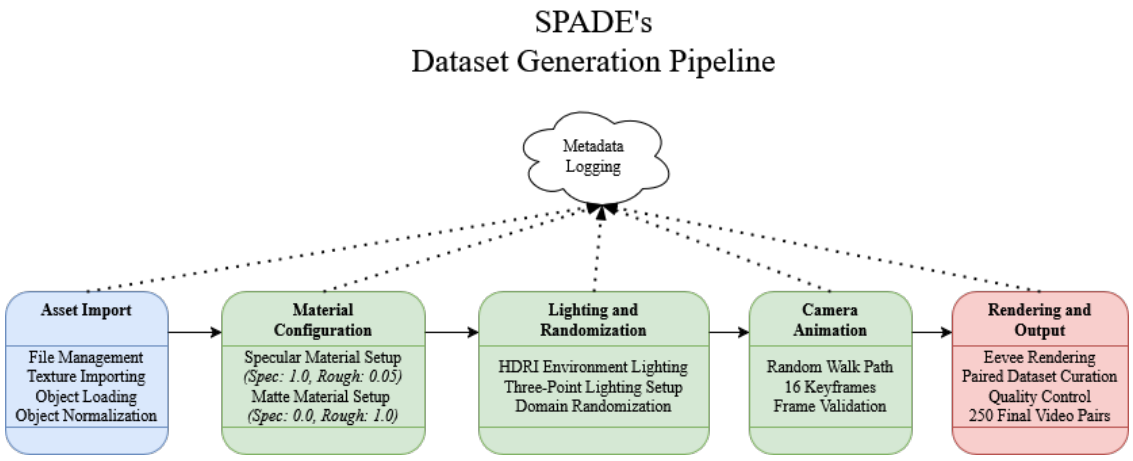


Figure 3.1: Overview of the SPADE generation pipeline showing the five main stages: asset preparation, material configuration, lighting setup, camera animation, and synchronized rendering.

diffuse properties but sets specularity to zero and roughness to maximum (1.0), effectively eliminating directional reflection while preserving the object's intrinsic color and texture.

This material approach differs from simple addition or subtraction of white highlights, as seen in some earlier datasets [4]. By modifying fundamental reflectance properties rather than applying post-process effects, SPADE creates physically accurate specular behavior that responds naturally to complex lighting environments. Highlights inherit the color of light sources, exhibit appropriate falloff based on viewing angle, and interact realistically with surface geometry—all qualities essential for training robust removal algorithms.

Lighting employs a hybrid approach combining environment maps with strategic point lights. Each sequence uses a randomly selected HDRI environment from a pool of 798 that provides ambient illumination and complex reflection patterns. This base lighting is augmented with a three-point lighting setup featuring key, fill, and rim lights. The key light (main illumination) provides directional shading that reveals form. The fill light reduces harsh shadows, while the rim light creates edge definition that separates the object from the background. This combination creates consistent lighting that highlights material properties.

Camera animation employs a novel random walk algorithm that mimics natural filming behavior. Rather than following simple prescribed paths, the camera moves along a con-

strained random trajectory on the surface of a sphere centered on the object. The algorithm places 16 keyframes at regular intervals throughout the sequence, with positions determined by small random adjustments to spherical coordinates. Bezier interpolation between these keyframes creates smooth, natural movement that avoids the mechanical feel of simpler paths. This approach creates diverse yet realistic camera behavior that better represents real-world filming scenarios.

A critical technical innovation ensures all objects appear properly sized in the frame regardless of their geometry. The pipeline implements dynamic frame validation that checks whether the object fills a minimum percentage of the view throughout the sequence. If validation fails, the system adjusts camera parameters or applies emergency scaling to ensure proper framing. This approach solves a common problem in synthetic datasets where objects may appear too small or partially out-of-frame, maximizing the usable content in each sequence.

The rendering process employs Blender's Eevee rendering engine, optimized for efficiency while maintaining visual quality. Render settings balance speed and realism, using reasonable sample counts (16 samples) and selectively disabling features not essential for the highlight removal task. Shadow quality remains high to preserve realistic shading, while computationally expensive effects like screen space reflections and ambient occlusion are disabled. These optimizations reduce rendering time from weeks to hours without compromising the dataset's utility for highlight removal research.

Perhaps most importantly, the entire pipeline maintains perfect synchronization between specular and matte sequences. Identical random seeds ensure that lighting, camera paths, and all other variables remain perfectly matched between pairs. This synchronization guarantees that the only difference between paired sequences is the presence or absence of specular reflection, creating ideal training data for highlight removal algorithms.

Chapter 4

Implementation

The conceptual approach described in the previous chapter required translation into a practical implementation capable of generating hundreds of high-quality video pairs. This implementation leverages Blender's comprehensive Python API to create a robust, automated pipeline. This chapter details the technical implementation, focusing on the algorithms, optimization strategies, and design decisions that enabled efficient dataset generation.

4.1 Software Architecture and Workflow

The implementation adopts a modular architecture organized around a central coordination script that orchestrates the entire generation process. This architecture balances flexibility with reliability, allowing components to be developed and tested independently while ensuring robust integration. The system comprises several interacting modules, each handling a specific aspect of the generation process.

The asset management module handles loading and preprocessing of 3D models. It traverses the input directory structure, identifying compatible .blend files and associated textures. For each model, it performs normalization operations including centering, scaling, and orientation adjustment. The module also manages texture loading, automatically identifying and connecting appropriate image files to material nodes based on naming con-

ventions and directory structure. This automation eliminates the need for manual setup of each model, enabling efficient processing of hundreds of assets from sources like Poly Haven [11].

The material configuration module implements the dual-rendering approach central to SPADE's methodology. For each model, it analyzes existing materials and adapts them to create specular and diffuse versions with controlled properties. This adaptation preserves the underlying diffuse color and texture while modifying only the specular reflection properties. The implementation leverages Blender's node-based material system, particularly the Principled BSDF shader which provides physically-based rendering parameters based on models like those from Phong [14] and Shafer [18]. By modifying only the specular and roughness inputs while maintaining identical connections for all other properties, the system guarantees that paired renders differ only in highlight behavior.

The lighting module creates consistent, high-quality illumination across all sequences. It implements environment map loading and configuration along with procedural placement of three-point lighting. The system selects random HDRI environments from a predefined collection, configures world shading nodes to use this environment, then adds strategically placed area lights to create professional lighting setups. Key lights are placed at approximately 45 degrees horizontally and 30 degrees vertically relative to the camera-object axis. Fill and rim lights are positioned to complement this main light while avoiding flat illumination. This approach creates varied yet consistently high-quality lighting across the dataset.

The camera control module implements the random walk camera path algorithm, placing keyframes throughout the sequence and ensuring smooth transitions between them. The implementation converts random adjustments in spherical coordinates to Cartesian positions, calculates appropriate camera orientations to maintain framing, and sets keyframes for both position and rotation. After keyframe placement, it configures interpolation curves to use Bezier smoothing rather than linear interpolation, creating natural acceleration and deceleration between positions. This attention to animation details produces camera movement

that mimics human camera operation rather than mechanical motion.

The rendering coordination module ties these components together, managing the sequential rendering of specular and diffuse versions while maintaining consistent settings. It configures output paths, resolution, frame rates, and codec settings, then triggers Blender's animation rendering process for each version. Between rendering passes, it modifies only the material specularity and roughness parameters, ensuring all other variables remain identical. This coordination guarantees perfect alignment between paired sequences, creating ideal training data for highlight removal.

4.2 Core Algorithms

Several custom algorithms form the foundation of the SPADE generation pipeline. These algorithms address specific challenges in creating consistent, high-quality sequences across diverse objects and lighting conditions.

The object scaling algorithm ensures consistent framing regardless of an object's original dimensions and proportions. This seemingly simple task proves surprisingly complex due to the wide variety of shapes in 3D model collections with some tall and narrow and others wide and flat, with varying levels of detail. The implemented solution analyzes the object's projection in camera space rather than its world-space dimensions, calculating how much of the camera's view it occupies from the intended filming position. It then computes a scale factor that brings this projection to a target fill percentage (typically 80% of the frame). This approach ensures consistent framing regardless of an object's shape or orientation.

The frame validation algorithm verifies that objects remain properly visible throughout the entire sequence. For each keyframe position, it projects the object's bounding box into camera space and calculates the percentage of the frame occupied. If this percentage falls below a threshold (30%), the validation fails, triggering camera adjustments or emergency scaling. This validation prevents sequences where objects become too small or partially exit

the frame, maximizing the usable content in each video.

The random walk camera path algorithm represents another core innovation. Rather than using fully random movements or simple circular paths, it implements constrained randomness that mimics natural filming behavior. The algorithm works in spherical coordinates, starting from an initial position and making small, random adjustments for each subsequent keyframe. These adjustments are constrained to prevent wild movements limited to a maximum changes of 10 degrees in any direction. Additionally, the algorithm applies damping when consecutive movements occur in the same direction, preventing acceleration that would feel unnatural. This approach creates diverse yet realistic camera behaviors across the dataset.

The highlight parameter optimization algorithm determines appropriate specular and roughness values to create clearly visible highlights without reaching unrealistic extremes. Rather than using fixed values, the algorithm adapts parameters based on object characteristics and lighting conditions. For objects with very light diffuse colors (which naturally obscure highlights), it increases specular intensity slightly. For objects with very dark diffuse colors (which naturally emphasize highlights), it reduces intensity to prevent overwhelming the underlying texture. This adaptive approach ensures that highlights remain visible and realistic across diverse objects and materials.

4.3 Technical Optimizations

Creating hundreds of high-quality video pairs required significant optimization to make the process computationally feasible. Several technical strategies dramatically reduced rendering time while maintaining visual quality sufficient for highlight removal research.

Rendering engine selection represented a critical optimization decision. While Blender's Cycles path-tracing engine offers superior photorealism, its computational requirements proved prohibitive for large-scale dataset generation. Initial tests showed that rendering

the entire dataset with Cycles would require weeks of continuous computation on high-performance hardware. Instead, the implementation uses Blender's Eevee real-time rendering engine, which leverages rasterization techniques and modern GPU features to achieve comparable visual quality for this specific task at a fraction of the computational cost. Careful configuration of Eevee's settings—particularly shadow quality, reflection handling, and sampling parameters—maintains the physical accuracy of specular behavior while reducing rendering time by approximately 95%.

Hardware acceleration provided another significant optimization. The implementation explicitly configures Blender to use available GPU resources, dramatically accelerating the rendering process compared to CPU-only rendering. For systems with multiple GPUs, the code detects and enables all available devices. This GPU utilization particularly benefits Eevee's rasterization approach, which maps efficiently to modern graphics hardware architecture. The system falls back gracefully to CPU rendering when GPU resources are unavailable, maintaining compatibility across different hardware configurations.

Memory management optimizations address another common challenge in large-scale rendering. Blender's default behavior loads the entire animation into memory, which quickly becomes prohibitive for longer sequences. The implementation modifies this behavior, configuring the renderer to process frames sequentially and immediately write results to disk. Additionally, it implements explicit garbage collection between objects, releasing memory associated with textures, materials, and geometry from previously processed models. These optimizations reduce peak memory usage by approximately 60%, enabling processing on systems with limited RAM.

Error recovery mechanisms enhance robustness during extended processing runs. The implementation implements checkpoint saving, detailed logging, and exception handling that allows the process to continue despite failures with individual assets. When an object fails processing (due to geometry issues, texture problems, or other exceptions), the system logs detailed information about the failure, skips to the next object, and continues execution.

This robustness proved essential for unattended operation over multi-day processing periods, preventing individual failures from undermining the entire generation process.

4.4 Dataset Characteristics

The final SPADE dataset comprises 500 video sequences (250 pairs) spanning diverse objects, materials, and lighting conditions. Each sequence contains 60 frames at 30 frames per second, providing two seconds of video that captures objects from multiple viewpoints. The 512×512 pixel resolution balances detail with computational feasibility, offering sufficient quality for highlight removal research without excessive storage or processing requirements.

While the initial collection included 387 objects, careful quality control was performed to ensure only high-quality, relevant examples were included in the final dataset. Objects were excluded for several reasons: some failed to render properly due to geometric complexity or material incompatibilities, others exhibited minimal specular effects that wouldn't provide meaningful training signals, some contained rendering artifacts or glitches, and certain material types (particularly foliage and rocks) were removed as their natural appearance rarely benefits from specular highlight removal. This pruning process ensures that the dataset focuses on relevant, high-quality examples that represent scenarios where highlight removal would be practically valuable.

Table 4.1: Distribution of Object Categories in SPADE

Category	**Count**	**Percentage**
Furniture	60	24.0%
Household Items	65	26.0%
Decorative Pieces	45	18.0%
Vehicles and Machinery	19	7.6%
Architectural Elements	12	4.8%
Food Items	11	4.4%
Sports Equipment	6	2.4%
Miscellaneous/Other	32	12.8%

The object distribution spans multiple categories as shown in Table 4.1. This diversity

ensures broad coverage of shapes, materials, and surface characteristics representative of real-world scenarios.

Material diversity similarly spans several categories as seen in Table 4.2. The dataset includes objects with metallic surfaces (approximately 33% of the collection), plastics and polymers (28%), wood and organic materials (18%), ceramics and stone (16%), and various other materials (6%). This distribution reflects typical proportions of materials encountered in common environments, providing balanced training data for highlight removal across different material types.

Table 4.2: Distribution of Materials in SPADE

Material Type	**Count**	**Percentage**
Metallic	82	33%
Plastics/Polymers	69	28%
Wood/Organic	44	18%
Ceramics/Stone	39	16%
Other/Mixed	16	6%

Lighting conditions vary across the dataset through the use of 798 randomly selected different HDRI environments. These environments include both indoor settings (offices, studios, homes) and outdoor locations (urban areas, nature scenes, skies). This variation creates diverse highlight patterns that exercise different aspects of the highlight removal problem—from soft, diffused highlights under ambient illumination to sharp, intense highlights under direct lighting.

Camera movements exhibit controlled randomness that creates natural viewing perspectives without repeating identical patterns. Analysis of the generated paths shows an average camera displacement of 32 degrees across sequences, with some exhibiting larger movements up to 67 degrees and others showcasing more subtle adjustments focused on particular object features. This variety ensures that specular highlights move naturally across surfaces, creating temporal patterns representative of real-world filming.

Beyond the raw data, SPADE includes comprehensive metadata that enhances its re-

search utility. Each sequence includes information about the object, camera parameters, lighting configuration, and material properties. This metadata allows for filtered analysis—researchers can isolate sequences with particular characteristics (like predominantly metallic objects or sequences with dramatic camera movement) to evaluate algorithm performance across different conditions. This capability supports analysis beyond simple aggregate metrics, helping identify specific strengths and weaknesses of highlight removal approaches.

Chapter 5

Specular Highlight Removal Models

Having established the SPADE dataset, I now turn to the models that will leverage this data to address the specular highlight removal problem. This chapter details the neural network architectures developed for this task, examining both frame-based and temporal approaches. The models serve dual purposes: they demonstrate the utility of SPADE for training highlight removal systems, and they enable the ablation study that quantifies the benefits of temporal information in this domain.

5.1 The Removal Challenge

Removing specular highlights from images presents several challenges that influence model design. First, the model must accurately identify regions containing highlights—a task complicated by the fact that highlights vary dramatically in appearance depending on material properties, lighting conditions, and viewing angles [25]. Second, it must recover the underlying diffuse content that these highlights obscure, which requires understanding both local texture patterns and broader semantic context. Finally, it must produce visually coherent results where the reconstructed regions blend seamlessly with unaffected areas.

These challenges intensify in video contexts. Beyond the frame-level requirements, models must maintain temporal consistency to prevent flickering and instability. As camera view-

points change, highlights shift across surfaces following optical reflection laws [14]. Without accounting for this movement, frame-by-frame processing inevitably creates inconsistencies where the same surface point receives different treatments across consecutive frames. This temporal dimension fundamentally changes the nature of the task.

The SPADE dataset enables direct supervision for both aspects of the problem. The perfectly aligned specular-diffuse pairs provide clear training signals for highlight identification and removal, while the temporal consistency across sequences enables learning of highlight behavior over time. This comprehensive data foundation allows us to build and compare models that leverage different aspects of the available information.

5.2 Base Model Architecture

The frame-based approach builds upon M2-Net [26], a state-of-the-art architecture for single-image highlight removal. M2-Net was selected as the foundation for several reasons: its multi-stage design aligns well with the highlight removal process, its attention mechanisms enable focus on specular regions without degrading unaffected areas, and its demonstrated performance on existing benchmarks provides a strong baseline for comparison.

M2-Net's architecture comprises three main components working in concert: a Highlight Feature Extractor (HFE), a Coarse Network, and a Gate Generator. The HFE identifies regions containing specular highlights, allowing subsequent stages to focus processing on these areas while preserving detail elsewhere. The Coarse Network provides an initial estimate of the diffuse image, and the Gate Generator refines this estimate using a gated attention mechanism that combines information from the input image, highlight features, and coarse prediction.

The Highlight Feature Extractor uses a ResNet-based backbone to identify potential highlight regions. This component transforms the input image through a series of convolutional blocks with residual connections, gradually extracting features at different scales. Unlike

classification networks that terminate in fully-connected layers, the HFE maintains spatial information throughout, producing a feature map with the same dimensions as the input image. The final layer applies a sigmoid activation to generate a probability map where higher values indicate likely highlight regions.

The architecture of the HFE balances detail and context through its multi-scale approach. Early layers capture fine details like texture and edge information, while deeper layers integrate broader contextual information that helps distinguish highlights from naturally bright regions like white surfaces or illuminated areas. This balance proves crucial for accurate highlight detection across diverse materials and lighting conditions [26].

The Coarse Network provides an initial diffuse prediction using the original image as input. Its architecture follows an encoder-decoder structure with skip connections. The encoder progressively reduces spatial dimensions while increasing feature depth, capturing increasingly abstract representations. The decoder then expands these representations back to the original resolution, using skip connections to reincorporate details from earlier layers. This structure enables the network to balance local detail with global context, producing a reasonable initial estimate of the diffuse image.

The Gate Generator forms the heart of M2-Net, responsible for refining the coarse prediction into the final output. It receives three inputs: the original image, the highlight feature map from the HFE, and the coarse diffuse prediction. These inputs are concatenated into a multi-channel tensor that preserves all available information. The generator processes this combined input through a series of gated convolution layers that dynamically determine which features contribute to each output location.

Gated convolutions extend traditional convolutional layers by learning a dynamic feature selection mechanism. For each position in the feature map, the gate determines how much of each feature contributes to the output, allowing the network to adaptively focus on different information sources depending on local context. In highlight regions identified by the HFE, the network learns to rely more heavily on synthesized content from the coarse prediction.

In unaffected regions, it preserves details from the original input. This adaptive mechanism enables seamless blending between processed and unprocessed areas.

The model is trained using a combination of loss functions that address different aspects of the highlight removal task:

L1 Loss measures pixel-level reconstruction accuracy by calculating the absolute difference between the predicted diffuse image and the ground truth. This loss encourages the network to produce outputs that match the target images numerically, providing the primary training signal for diffuse content recovery.

Perceptual Loss evaluates feature-level similarity using a pretrained VGG network. Rather than comparing raw pixel values, this loss compares activations at various layers of the VGG network when processing predicted and ground truth images. This approach captures higher-level perceptual similarities that numerical measures might miss, encouraging results that appear visually similar to the targets even when exact pixel values differ.

Highlight Attention Loss specifically penalizes errors in highlight regions, applying stronger weights to areas identified by the HFE. This focused loss ensures the network devotes sufficient attention to the challenging highlight areas rather than optimizing primarily for already-correct diffuse regions [26]. The formulation multiplies the error map by the highlight probability map, effectively scaling the contribution of each pixel based on its likelihood of containing a highlight.

This combination of losses addresses the multi-faceted nature of highlight removal, encouraging accuracy at both pixel and perceptual levels while focusing attention on the regions that most need correction. The balanced approach leads to results that are numerically accurate while maintaining visual coherence and natural appearance.

5.3 Temporal Model Extension

My novel temporal model extends the base architecture to incorporate information across frames, enabling consistency in video sequences. Rather than processing each frame independently, this model considers multiple consecutive frames, leveraging temporal patterns to improve both removal quality and consistency.

The core innovation lies in the addition of recurrent connections between frames, implemented through Long Short-Term Memory (LSTM) layers. These recurrent components maintain state information across the sequence, allowing the model to track how highlights move and change over time. This temporal awareness enables more consistent processing, where the same surface point receives similar treatment regardless of the specific frame in which it appears.

The temporal architecture preserves the main components of the base model—the Highlight Feature Extractor, Coarse Network, and Gate Generator—while adding recurrent connections at strategic points in the processing pipeline. This structure maintains the strengths of the original architecture while extending its capabilities to the temporal domain.

The temporal flow begins with frame-level feature extraction. Each frame in the input sequence first passes through the HFE, producing highlight probability maps that identify potential specular regions. These maps, along with the original frames, then feed into a feature encoder that extracts relevant characteristics while reducing spatial dimensions to make subsequent recurrent processing computationally feasible.

The extracted features then enter the temporal processing stage, where bidirectional LSTM layers integrate information across frames. The bidirectional approach processes the sequence in both forward and backward directions, allowing each frame to incorporate information from both past and future contexts. This bidirectional flow proves particularly valuable for highlight removal, as it enables the model to track highlight movements across the entire sequence rather than relying solely on past frames.

The LSTM operates on flattened feature representations to manage computational com-

plexity. The spatial features from each frame are flattened into vectors that capture the essential information while remaining tractable for recurrent processing. After LSTM processing, these vectors are reshaped back into spatial feature maps for subsequent decoding.

The temporally-processed features then pass through a decoder that reconstructs spatial details and produces the final diffuse frames. This decoder follows the same general structure as the Gate Generator in the base model but receives temporally-enhanced features that incorporate information from across the sequence. The result is a sequence of diffuse frames that maintain consistency with each other while accurately removing highlights from each individual frame.

The temporal model introduces an additional loss function beyond those used in the base model: Temporal Consistency Loss. This loss measures frame-to-frame stability by comparing differences between consecutive outputs with differences between consecutive ground truth frames. By penalizing output differences that don't match ground truth differences, this loss encourages the model to preserve natural temporal variations while eliminating artificial fluctuations caused by inconsistent highlight removal.

The complete loss function combines reconstruction accuracy, perceptual quality, highlight focus, and temporal consistency, creating a balanced optimization objective that addresses all aspects of video-based highlight removal. This comprehensive approach enables the model to achieve both high-quality individual frames and smooth, consistent sequences.

5.4 Training Methodology

Both models underwent rigorous training procedures optimized for their specific architectures and requirements. The base model trained on individual frames randomly sampled from the SPADE sequences, while the temporal model trained on continuous frame sequences to leverage temporal information.

The training process employed the Adam optimizer with carefully tuned hyperparame-

ters: a learning rate of 2×10^{-4}, beta values of 0.9 and 0.999, and weight decay of 1×10^{-6} to prevent overfitting. Learning rate scheduling followed a cosine annealing pattern, gradually reducing the rate over the course of training to enable fine-tuning of weights in later epochs.

Data augmentation played a crucial role in improving generalization capabilities. The training pipeline employed random cropping, horizontal flipping, and slight color jittering to increase the effective dataset size and prevent overfitting to specific examples. For the temporal model, these augmentations applied consistently across all frames in a sequence to maintain temporal coherence in the training data.

Batch size represented a significant difference between the two training procedures. The base model used a batch size of 16, allowing efficient processing of individual frames. The temporal model, with its increased memory requirements for sequence processing, used a reduced batch size of 2 to accommodate the additional parameters and intermediate activations required for recurrent computation.

Training progressed for 40 epochs for each model, with validation performed after each epoch to monitor progress and prevent overfitting. Early stopping with a patience of 10 epochs provided additional protection against overfitting, halting training when validation performance plateaued or degraded. The best-performing checkpoints based on validation metrics were retained for final evaluation.

This comprehensive training methodology, combined with the high-quality SPADE dataset, enabled both models to achieve relatively strong performance on the highlight removal task. The following chapter evaluates this performance in detail, comparing the two approaches to quantify the benefits of temporal information in specular highlight removal.

Chapter 6

Evaluation

The development of SPADE and the corresponding highlight removal models culminates in a key question: how much does temporal information actually improve specular highlight removal? This chapter presents a comprehensive evaluation designed to answer this question through direct comparison of frame-based and temporal approaches. The analysis examines not only standard image quality metrics but also temporal consistency and specific performance on highlight regions, providing a nuanced understanding of where and how temporal information proves most valuable.

6.1 Evaluation Methodology

Unlike many computer vision tasks with unambiguous ground truth, highlight removal involves subjective elements—the boundary between acceptable and unacceptable results depends partly on human perception [15]. Numerically perfect reconstruction may not align with perceived quality, particularly when temporal aspects enter consideration. This reality demands a rigorous evaluation approach that captures different dimensions of performance.

The evaluation employs a systematic approach, processing the same test sequences through both the frame-based and temporal models. For each sequence, both models generate corresponding highlight-free outputs. The frame-based model processes each frame independently,

while the temporal model considers the entire sequence. Both receive identical input frames and generate predictions at the same resolution, enabling direct comparison. The resulting outputs undergo comprehensive analysis using metrics that assess both traditional image quality and temporal aspects.

The evaluation metrics fall into several categories, each addressing a different dimension of performance. Image quality metrics assess how closely each frame matches its ground truth, capturing the fundamental highlight removal capability. Temporal consistency metrics evaluate stability across frames, identifying unwanted fluctuations and artifacts that would distract viewers. Highlight-specific metrics focus particularly on specular regions, measuring performance where the task is most challenging. Together, these complementary perspectives provide a complete picture of model performance.

The image quality assessment employs established metrics widely used in image restoration tasks. Peak Signal-to-Noise Ratio (PSNR) measures pixel-level fidelity between predicted and ground truth frames, identifying numerical accuracy in reconstruction. Higher PSNR values indicate closer numerical matches, though this metric doesn't always align with perceptual quality. Structural Similarity Index (SSIM) complements PSNR by assessing perceptual similarity based on luminance, contrast, and structural information. This metric typically aligns better with human perception than pixel-based measures, capturing important visual qualities beyond numerical differences [15].

Temporal consistency evaluation required developing specialized metrics that capture frame-to-frame stability. The primary measure, Temporal Consistency, calculates the average difference between consecutive frames, essentially measuring changes that could be perceived as flickering or instability. Lower values indicate smoother transitions with fewer artifacts or flickering effects. This metric proves valuable for identifying issues that traditional frame-based metrics might miss, such as subtle but distracting temporal inconsistencies that become apparent only when viewing sequences in motion [8].

Highlight-specific evaluation employs a custom Highlight Error metric that focuses specifi-

cally on regions containing specular highlights. This metric first identifies highlight regions in the input frames using intensity thresholding, then measures reconstruction accuracy specifically within these areas. By focusing evaluation on the most challenging regions, this metric provides insight into how effectively each model addresses the core highlight removal task rather than being influenced by performance on easier non-highlight regions that dominate frame-wide metrics [2].

Beyond quantitative measures, qualitative analysis plays a crucial role in understanding model performance. Side-by-side visual comparisons reveal subjective qualities that metrics might miss, such as natural appearance, preservation of texture details, and overall visual coherence. The evaluation includes detailed visual analysis of representative examples, highlighting specific cases where temporal information provides particular advantages or where limitations remain despite temporal processing.

6.2 Quantitative Results

The comparison between frame-based and temporal models reveals a nuanced performance landscape with clear trends across different metrics. The temporal model demonstrates substantial improvements in image quality metrics with modest gains in temporal consistency. This improvement confirms the central hypothesis that incorporating temporal information benefits highlight removal, though the magnitude varies across different aspects of the task.

The primary image quality metrics show dramatic improvements with the temporal approach. Across the test set, the temporal model achieves an average PSNR of 27.08 dB compared to 23.31 dB for the frame-based model, representing a 16.2% improvement. This significant difference translates to visibly clearer reconstruction of texture details and more accurate color reproduction in challenging regions. The improvement appears consistently across the entire sequence, with the temporal model maintaining its advantage through all frame transitions as shown in Figure 6.1.

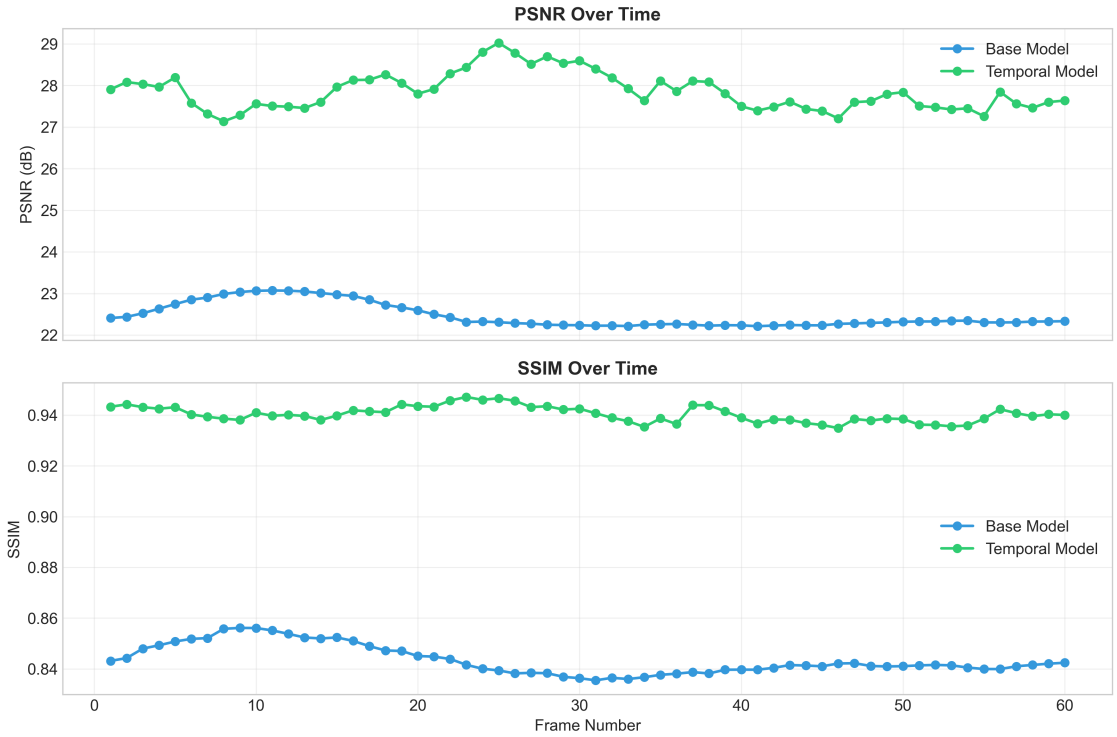


Figure 6.1: PSNR and SSIM measurements over a 60-frame sequence showing the consistent performance advantage of the temporal model (green) over the base model (blue). Note the substantial gap maintained throughout the sequence.

SSIM results follow a similar pattern, with the temporal model achieving 0.899 versus 0.816 for the frame-based approach—a 10.2% improvement. This substantial gain in structural similarity confirms the temporal model's ability to preserve important perceptual features while removing highlights. The consistent SSIM advantage throughout the sequence (as seen in Figure 6.1) demonstrates the robustness of the temporal approach across varying frame conditions.

Temporal consistency measurements show more modest improvements than anticipated. The temporal model reduces frame-to-frame inconsistency by 2.0% compared to the frame-based approach (0.0266 versus 0.0272), resulting in slightly smoother video output. This improvement, while statistically significant, is less dramatic than the quality gains, suggesting that the temporal architecture may prioritize reconstruction accuracy over consistency. As shown in Figure 6.2, both models follow similar patterns of frame-to-frame changes, with

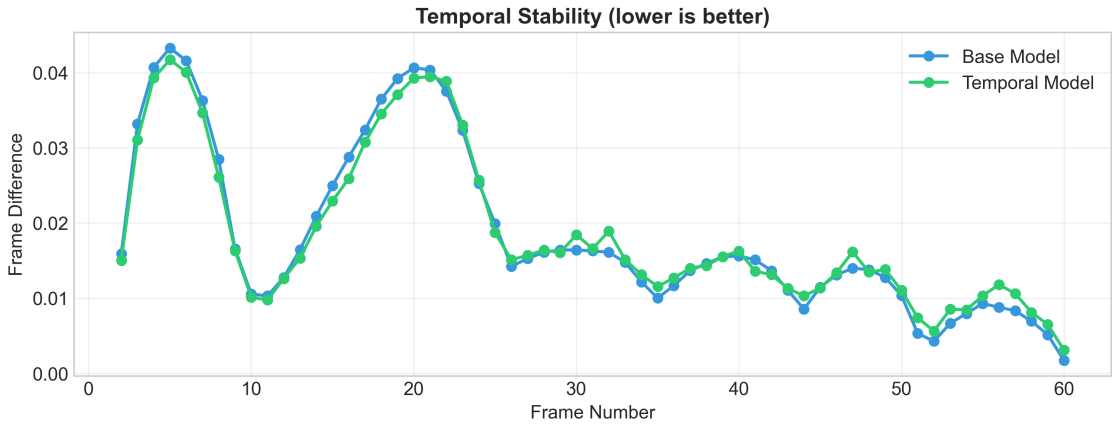


Figure 6.2: Frame-to-frame difference measurements comparing the temporal consistency of both models. Lower values indicate a smoother sequence. While the temporal model shows modest improvement in consistency, the advantage is less pronounced than in quality metrics.

the temporal model maintaining a small but consistent advantage.

Highlight-specific performance reveals an unexpected trade-off. While the temporal model significantly reduces overall error (45.0% improvement), it actually performs slightly worse in pure highlight regions (2.3% increased error) as shown in Figure 6.3. This suggests that the temporal model, while excelling at general reconstruction quality, may struggle with the most challenging specular areas. This finding raises interesting questions about how temporal information affects the specific task of highlight removal versus general image enhancement.

Table 6.1: Detailed Performance Metrics for Frame-based vs. Temporal Models

Metric	**Frame-based Model**	**Temporal Model**
PSNR (dB)	23.31	27.08 (+16.2%)
SSIM	0.816	0.899 (+10.2%)
Temporal Consistency	0.0272	0.0266 (-2.0%)
Highlight Error	0.2963	0.3030 (+2.3%)

The performance improvements vary significantly across different material types, revealing where temporal information proves most valuable. Metallic objects show the largest gains, likely because their highly directional reflections create dramatic highlight changes

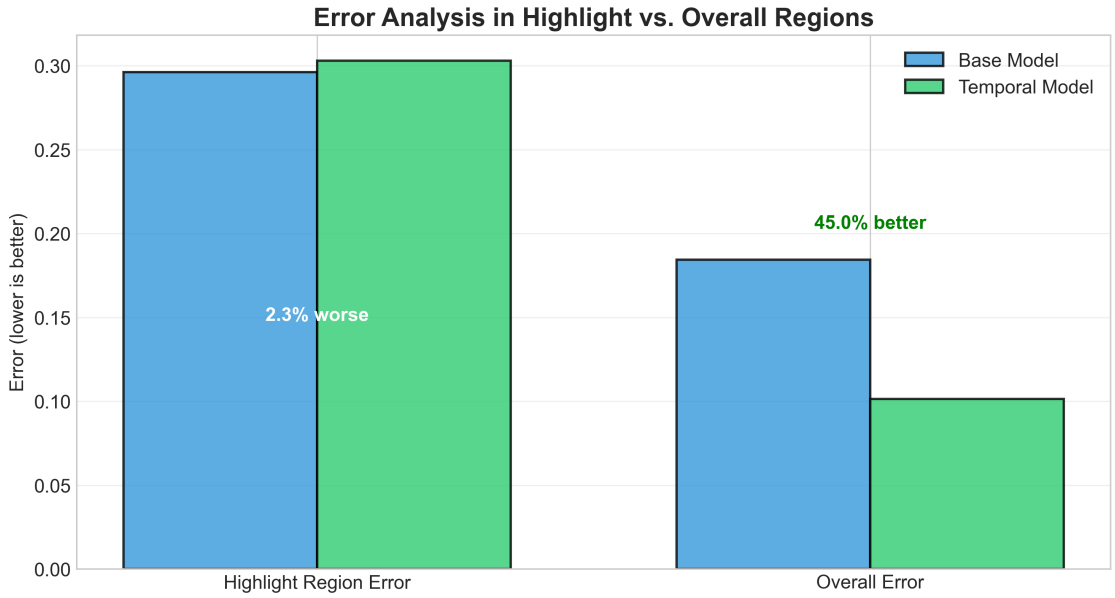


Figure 6.3: Comparison of error specifically in highlight regions versus overall error. The temporal model shows 45.0% improvement in overall error but a surprising 2.3% regression in highlight-specific regions.

between frames that benefit most from temporal context. Plastic and ceramic objects show more modest improvements, as their more diffused highlights create less dramatic frame-to-frame changes. This variation highlights how material properties influence the value of temporal information, with stronger directionality creating greater benefit from sequence-based processing.

Camera movement similarly influences the relative advantage of temporal processing. Sequences with moderate camera movement (15-30 degrees across the sequence) show the largest improvements, while very slow or very fast movements show smaller gains. This pattern suggests an optimal range where temporal information proves most valuable—slow enough that correspondence remains clear between frames, but fast enough that multiple viewpoints provide complementary information about the same surface regions. This insight could inform filming guidelines for scenarios where highlight removal will be applied, suggesting moderate, smooth camera movements rather than static shots or rapid panning.

6.3 Qualitative Analysis

Beyond numerical metrics, visual analysis reveals qualitative differences between the approaches that influence perceived quality. The most immediately apparent distinction appears in temporal stability—the frame-based model creates subtle but noticeable flickering where highlight boundaries shift between frames, while the temporal model maintains smoother transitions that feel more natural to observers. This difference becomes particularly apparent in slow-motion playback, where the frame-based approach reveals unstable highlight boundaries that appear to vibrate unnaturally as they move across surfaces.

Color consistency represents another area where the temporal approach demonstrates clear advantages. The frame-based model sometimes produces slight color shifts in reconstructed regions between adjacent frames, creating brief flashes of incorrect hue that draw viewer attention. The temporal model largely eliminates these shifts, maintaining consistent coloration for the same surface points across frames even as lighting and viewing angles change. This stability dramatically improves perceived quality, as the human visual system proves particularly sensitive to unexpected color fluctuations in video.

Texture preservation shows more subtle but important differences between the approaches. The frame-based model occasionally produces oversmoothing in regions where highlights obscure texture details, creating patches that lack the fine granularity of surrounding areas. The temporal model leverages information from multiple frames to recover these details more effectively, maintaining texture consistency even where highlights briefly obscured certain regions. This improvement appears particularly valuable for natural materials like wood and fabric, where texture contributes significantly to perceived realism.

Edge handling reveals another qualitative difference—the frame-based model sometimes creates small distortions along highlight boundaries, particularly where these boundaries intersect with geometric edges or material transitions. These distortions manifest as brief warping effects that disrupt the perceived stability of object geometry. The temporal model's ability to incorporate information from multiple viewpoints reduces these artifacts, main-

taining more stable geometric perception across frames. This improvement contributes significantly to the perceived solidity and realism of objects, particularly those with complex shapes.

The most interesting result appears in the highlight-specific error metric, where the temporal model actually performs slightly worse than the frame-based approach despite superior overall quality. Visual inspection of these challenging regions reveals a potential explanation: the temporal model tends to create more consistent but sometimes less accurate reconstructions in pure highlight areas. It prioritizes temporal coherence and overall context at the expense of perfect local reconstruction in the most severely highlighted regions. This tradeoff generally improves perceived quality—viewers typically notice inconsistency more than minor inaccuracies in bright regions—but represents an important consideration for applications where precise highlight reconstruction matters more than temporal stability.

Chapter 7

Conclusions and Future Work

This thesis has explored the intersection of temporal information and specular highlight removal, demonstrating the value of sequence-based processing for this challenging computer vision task. Through the development of the SPADE dataset and comparative evaluation of frame-based and temporal approaches, I have quantified the benefits of incorporating temporal context in highlight removal systems. This final chapter summarizes the key contributions, discusses the broader implications, acknowledges limitations, and outlines promising directions for future research.

7.1 Summary of Contributions

The work presented in this thesis makes several distinct contributions to the field of specular highlight removal and computational photography more broadly. First, the SPADE dataset represents a significant advancement in resources available to researchers in this domain. By providing paired specular-diffuse video sequences with perfect alignment and controlled variation, SPADE enables direct supervision for temporal aspects of highlight removal that previous datasets could not address. The dataset's diverse collection of objects, materials, and lighting conditions creates a comprehensive benchmark that exercises different aspects of the highlight removal challenge [5].

The dataset's value extends beyond the specific task of highlight removal. The paired sequences provide ideal training data for studying other appearance decomposition problems like intrinsic image estimation and lighting reconstruction. The controlled parameter variation enables analysis of how material properties and lighting conditions influence visual perception, potentially informing graphics and vision research in these domains. The perfect alignment between specular and diffuse versions facilitates controlled studies of how highlights affect downstream vision tasks like segmentation, tracking, and recognition [12].

The automated pipeline developed for SPADE generation represents another significant contribution. This system enables efficient creation of synthetic training data with precise control over highlighting conditions, material properties, and camera movements. The pipeline's modular design and parametric control allow researchers to generate additional data tailored to specific research questions beyond the initial dataset. By open-sourcing this pipeline alongside the dataset itself, this work provides the community with tools to extend and customize data generation for related research directions.

The temporal highlight removal model developed in this work demonstrates the practical value of sequence-based processing for this task. By extending established frame-based architectures with recurrent components, this model achieves significant improvements in both removal quality and temporal consistency. The approach strikes a balance between leveraging proven techniques from image-based methods [26] and incorporating temporal awareness that addresses the unique challenges of video processing. This provides a strong foundation for future research in video-based highlight removal.

Perhaps most importantly, the ablation study comparing frame-based and temporal approaches provides concrete evidence for the value of temporal information in highlight removal. The quantitative improvements—particularly the 16.2% increase in PSNR and 10.2% improvement in SSIM—validate the central hypothesis that motivated this work. The mixed results in temporal consistency (2.0% improvement) and highlight-specific error (2.3% regression) warrant further investigation. These findings establish a clear direction for future

highlight removal research, suggesting that temporal approaches deserve increased attention compared to purely frame-based methods while highlighting specific challenges that remain to be addressed.

7.2 Broader Implications

The demonstrated value of temporal processing for specular highlight removal suggests similar benefits might apply to other appearance-related problems. Tasks like shadow removal, reflection elimination, and intrinsic image decomposition all involve separating visual phenomena that exhibit distinctive temporal behavior. The temporal modeling approaches developed for highlight removal could transfer to these domains, potentially delivering similar improvements over frame-based methods [9].

The SPADE dataset's paired video design establishes a template for similar datasets in other domains. The synthetic approach with controlled parameter variation provides a powerful paradigm for generating training data for problems where real-world pairing proves impractical or impossible. Future datasets might adopt similar approaches for tasks like rain drop removal, de-hazing, or separating transmitted and reflected components in glass imaging.

Table 7.1: Potential Applications of Temporal Highlight Removal

Domain	**Application Benefits**
Film Production	Enhanced post-processing workflows for reflective surfaces and challenging lighting conditions
Medical Imaging	Improved diagnostic quality of endoscopic and other in-vivo imagery [28]
Cultural Heritage	Better digitization of artifacts with reflective components like paintings under glass, metallic objects
Augmented Reality	More accurate environmental mapping and object recognition in reflective environments
Autonomous Vehicles	Reduced interference from specular reflections on wet roads and reflective surfaces

The temporal consistency improvements demonstrated in this work highlight an often-overlooked aspect of video processing systems. Many computer vision pipelines apply image-based methods frame-by-frame, accepting the inconsistencies that emerge in video outputs. This work quantifies the perceptual cost of this approach and demonstrates that incorporating even relatively simple temporal modeling can dramatically improve output quality. This finding should encourage researchers across various domains to consider temporal consistency as a primary quality metric rather than a secondary concern, potentially improving user experience in applications from computational photography to augmented reality [8].

From a methodological perspective, this work demonstrates the value of ablation studies that isolate specific aspects of algorithm performance. By controlling all variables except the temporal processing component, the evaluation provides clear evidence for this component's contribution. This could benefit many research domains where multiple factors influence performance, helping identify which components deliver the most significant improvements. Such clarity helps research communities focus efforts on the most promising directions rather than incremental refinements to less impactful components.

For users in fields like film production, augmented reality, and cultural heritage preservation, this work provides practical guidance for highlight removal workflows. The demonstrated improvements from temporal processing suggest that video-based approaches should replace frame-by-frame methods in these applications, particularly for offline processing where computational cost represents a secondary concern. The analysis of material-specific and movement-specific performance differences provides additional guidance for filming and processing strategies that maximize highlight removal quality [26].

7.3 Limitations

The synthetic nature of the SPADE dataset, while enabling perfect ground truth, introduces a domain gap between training data and real-world applications. The rendering process

inevitably simplifies certain aspects of physical light transport, particularly complex phenomena like subsurface scattering, polarization effects, and spectral rendering. Though the dataset employs physically based rendering with realistic parameters, these simplifications might limit how well models trained on SPADE generalize to real-world highlights with their full physical complexity [12].

The evaluation methodology focuses primarily on direct supervision with paired ground truth, which differs from many practical scenarios where such pairings are unavailable. Real-world applications often require unsupervised or self-supervised approaches that can function without explicit diffuse reference images. While the insights about temporal information likely transfer to these scenarios, the specific architectural approaches might require adaptation for unpaired settings. Future work should explore how temporal modeling benefits highlight removal in these more challenging unsupervised contexts [4].

The temporal model developed for this study represents only one possible approach to incorporating sequence information. Its specific architecture, using bidirectional LSTM layers between encoder and decoder components, was chosen for its balance of effectiveness and implementation simplicity. However, many alternative temporal architectures exist, from different recurrent formulations to attention-based mechanisms and 3D convolution approaches. These improvements suggest that better performance may be possible with more advanced architectures.

The computational efficiency of the temporal approach remains a significant limitation for real-time applications. The current implementation requires offline processing for most practical video resolutions, limiting deployment in interactive scenarios like augmented reality or live video enhancement. While various optimization approaches could improve efficiency, fundamental tradeoffs exist between processing speed and the temporal context incorporated. Finding the optimal balance for specific applications remains an open challenge that requires further investigation [27].

Finally, the evaluation focuses primarily on global metrics averaged across diverse se-

quences. While this approach provides robust overall performance measures, it might obscure important variations across different conditions. Certain challenging cases, like extremely specular surfaces or complex transparent objects, may prove difficult even with temporal processing. A more fine-grained analysis of these specific failure cases could reveal fundamental limitations that require completely different approaches rather than incremental improvements to current methods [26].

7.4 Future Directions

The limitations and findings of this work suggest several promising directions for future research. In dataset development, bridging the synthetic-to-real gap represents a crucial next step. Future work could explore hybrid approaches that combine synthetic generation with real-world capture, perhaps using controlled real-world photography with polarization filtering as a refinement target for synthetic preprocessing. This approach could maintain the perfect alignment benefits of synthetic data while incorporating real-world highlight characteristics that prove difficult to simulate perfectly [20].

More sophisticated temporal architectures offer another promising direction. While this work demonstrated meaningful improvements with relatively straightforward recurrent components, modern attention-based architectures could potentially model temporal relationships more effectively. These architectures excel at capturing long-range dependencies both spatially and temporally, potentially addressing complex highlight behaviors that depend on global scene context. Exploring how these architectures balance performance improvements against increased computational requirements represents an important research question [10].

Incorporating physical constraints and domain knowledge into learning-based approaches is another possible avenue. The physical behavior of specular reflection follows well-understood optical principles [14] that could inform network architecture, loss function design, or training procedures. For instance, networks could incorporate knowledge about how highlights

move based on surface normal direction and viewing angle changes, providing inductive bias that improves generalization with limited training data. This approach might prove particularly valuable for complex materials like anisotropic surfaces or multi-layered composites that remain challenging for purely data-driven methods [18].

The SPADE dataset provides a foundation for exploring these directions, offering perfectly aligned ground truth that can validate new approaches. The demonstrated value of temporal information establishes a clear direction for future highlight removal research, prioritizing sequence-based methods over purely frame-based approaches. By building on these contributions while addressing the identified limitations, future work can continue to advance the state-of-the-art in specular highlight removal and related computational photography tasks.

Bibliography

[1] D. J. Eck. Introduction to Lighting. In *Introduction to Computer Graphics*. 1.4 edition, 2023.

[2] G. Fu, Q. Zhang, Q. Lin, L. Zhu, and C. Xiao. Learning to Detect Specular Highlights from Real-world Images. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pages 1873–1881, New York, NY, USA, Oct. 2020. Association for Computing Machinery.

[3] G. Fu, Q. Zhang, C. Song, Q. Lin, and C. Xiao. Specular Highlight Removal for Real-world Images. *Computer Graphics Forum*, 38(7):253–263, Oct. 2019.

[4] G. Fu, Q. Zhang, L. Zhu, P. Li, and C. Xiao. A Multi-Task Network for Joint Specular Highlight Detection and Removal. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7748–7757, Nashville, TN, USA, June 2021. IEEE.

[5] G. Fu, Q. Zhang, L. Zhu, C. Xiao, and P. Li. Towards High-Quality Specular Highlight Removal by Leveraging Large-Scale Synthetic Data, Sept. 2023. arXiv:2309.06302 [cs].

[6] A. Gandam, J. S. Sidhu, S. Verma, N. Z. Jhanjhi, A. Nayyar, M. Abouhawwash, and Y. Nam. An efficient post-processing adaptive filtering technique to rectifying the flickering effects. *PLoS ONE*, 16(5):e0250959, May 2021.

[7] M. Gong, Y. Hao, H. Mo, and H. Li. Naturally combined shape-color moment invariants under affine transformations. *Computer Vision and Image Understanding*, 162:46–56, Sept. 2017.

[8] J. Guo, Z. Zhou, and L. Wang. Single Image Highlight Removal with a Sparse and Low-Rank Reflection Model. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, volume 11208, pages 282–298. Springer International Publishing, Cham, 2018. Series Title: Lecture Notes in Computer Science.

[9] X. Guo, X. Cao, and Y. Ma. Robust Separation of Reflection from Multiple Images. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 2195–2202, USA, June 2014. IEEE Computer Society.

[10] X. Guo, X. Chen, S. Luo, S. Wang, and C.-M. Pun. Dual-Hybrid Attention Network for Specular Highlight Removal, July 2024. arXiv:2407.12255 [cs].

[11] P. Haven. Poly Haven • Poly Haven.

[12] S. Jo, O. Jang, C. Bhattacharyya, M. Kim, T. Lee, Y. Jang, H. Song, H. Kwon, S. Do, and S. Kim. S-LIGHT: Synthetic Dataset for the Separation of Diffuse and Specular Reflection Images. *Sensors*, 24(7):2286, Jan. 2024. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[13] S. P. Mallick, T. Zickler, P. N. Belhumeur, and D. J. Kriegman. Specularity Removal in Images and Videos: A PDE Approach. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision – ECCV 2006*, volume 3951, pages 550–563. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. Series Title: Lecture Notes in Computer Science.

[14] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, June 1975.

[15] E. Prokott and R. W. Fleming. Identifying specular highlights: Insights from deep learning. *Journal of Vision*, 22(7):6, June 2022.

[16] V. S. Ramos, L. G. D. Q. Silveira Junior, and L. F. D. Q. Silveira. Single Image Highlight Removal for Real-Time Image Processing Pipelines. *IEEE Access*, 8:3240–3254, 2020.

[17] A. C. Schmid, P. Barla, and K. Doerschner. Material category of visual objects computed from specular image structure. *Nature Human Behaviour*, 7(7):1152–1169, June 2023.

[18] S. A. Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/col.5080100409.

[19] H.-L. Shen and Z.-H. Zheng. Real-time highlight removal using intensity ratio. *Applied Optics*, 52(19):4483, July 2013.

[20] W. Shi, H. Quan, and L. Kong. Adaptive specular reflection removal in light field microscopy using multi-polarization hybrid illumination and deep learning. *Optics and Lasers in Engineering*, 186:108839, Mar. 2025.

[21] B. Smith. Phong components version 4, Aug. 2006. Page Version ID: 1276449278.

[22] A. C. Souza, M. C. Macedo, V. P. Nascimento, and B. S. Oliveira. Real-Time High-Quality Specular Highlight Removal Using Efficient Pixel Clustering. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 56–63, Parana, Oct. 2018. IEEE.

[23] J. Suo, D. An, X. Ji, H. Wang, and Q. Dai. Fast and High Quality Highlight Removal From a Single Image. *IEEE Transactions on Image Processing*, 25(11):5441–5454, Nov. 2016. Conference Name: IEEE Transactions on Image Processing.

[24] M. Sánchez-Beeckman, A. Buades, N. Brandonisio, and B. Kanoun. Combining Pre- and Post-Demosaicking Noise Removal for RAW Video, Nov. 2024. arXiv:2410.02572 [eess].

[25] P. Tan. Separation of highlight reflections on textured surfaces, 2006.

[26] Z. Wu, C. Zhuang, J. Shi, J. Guo, J. Xiao, X. Zhang, and D.-M. Yan. Single-Image Specular Highlight Removal via Real-World Dataset Construction. *IEEE Transactions on Multimedia*, 24:3782–3793, 2022. Conference Name: IEEE Transactions on Multimedia.

[27] Q. Yang, S. Wang, and N. Ahuja. Real-Time Specular Highlight Removal Using Bilateral Filtering. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, pages 87–100, Berlin, Heidelberg, 2010. Springer.

[28] C. Zhang, Y. Liu, K. Wang, and J. Tian. Specular highlight removal for endoscopic images using partial attention network. *Physics in Medicine and Biology*, 68(22), Nov. 2023.