# Bridging Machine Learning and Computational Photography to Bring Professional Quality into Casual Photos and Videos

*Cecilia Zhang*

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2021-1
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-1.html

January 14, 2021

# Bridging Machine Learning and Computational Photography to Bring Professional Quality into Casual Photos and Videos

by

Xuaner Zhang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ren Ng, Chair
Professor Alexei A. Efros
Professor Martin S. Banks

Fall 2020

Bridging Machine Learning and Computational Photography to Bring Professional Quality into Casual Photos and Videos

Abstract

Bridging Machine Learning and Computational Photography to Bring Professional Quality into Casual Photos and Videos

by

Xuaner Zhang

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Ren Ng, Chair

Having a compact, casual pocket camera always within reach is a delight. It opens the opportunity to capture spontaneous moments and casual events. While users appreciate the convenience of mobile experience, their crave for visual quality of the professionals is hard to achieve. Because of hardware limitations and a lack of control over suboptimal conditions in the environment, casual photos and videos suffer from noise, lack of sharpness, unflattering lighting, wrong focus, distracting obstructions, etc. The desires are eager to make cameras see as our human visual system does, to understand the world and produce photographs that are perceptually pleasing and meaningful. Professional studio photography and cinematography have made the best attempts delivering high-quality photos and videos by incorporating intricate hardware and gathering professional crew. Casual imaging, on the other hand, is still nowhere close.

In this thesis, I argue that it is key for a camera to understand the semantics of the scene – the *context* – presented in its viewfinder in order to intelligently capture and process sensor data. The approach to bring in such contextual information is through machine learning. Thankfully, modern mobile cameras are integrated with fast image processors and even dedicated machine learning chips to drive the development of computational capacities. Machine-learning-driven computational photography algorithms are lifted to great practicality more than ever before. Throughout the thesis, I discuss the challenges of causal imaging and how its quality can benefit from professional photography and cinematography principles. The thesis focuses on the quality enhancement from three aspects – perceptual, lighting and focus. We propose a number of learning-based methods to lift these limitations to produce unprecedented results, and show a potential direction that integrates machine learning and imaging systems to enhance casual photos and videos towards the quality of the professionals.

To mom, dad and Yoo who have kept me smiling
and whom I will never get tired of photographing

# Contents

# List of Figures

# List of Tables

## Acknowledgments

My PhD is nowhere close to being easy, but because of these people, I have gained enough support and strength to get through and accomplish this journey.

I feel lucky to have the people I truthfully respect to be on my committee, Ren Ng, Alexei Efros and Martin Banks. They have guided me from the first class I took from them, to my thesis proposal, to my dissertation day, and I believe to my future everyday. My understanding of human perception originates from the vision science class taught by Marty. Learning about how human brain 'sees' the visual world fascinates me, and in retrospect, greatly contributes to the goal of my thesis to bring such visual signal processing power to a camera system. My influence from Alyosha started before I met him. Some of his early works in image editing and composite were what brought me into computational photography. He lets me see the joy of manipulating every pixel, and inspires me to look beyond a single image to exploit patterns and correlations among millions of photos. Plus, he shows me what passion feels like from his teaching, presentations and even normal conversations. I also thank Jonathan Regan Kelly for being on my qualification committee, sharing his insights on systems for image and video processing.

My gratitude for Ren is immeasurably more than a single paragraph. He patiently took me on track when I was young and naïve with research. He guided me through Research 101 to being an independent researcher, and have been encouraging me to think broad and bold. Ren also taught me all the secret sauce for public speaking and academic presentation. I can only enjoy standing on the stage of SIGGRAPH because of him. Outside of work, Ren, Yi, Lana and Reya show me the importance of family and how it can make a person so warm and caring. I can never have asked for a more inspiring advisor.

A significant part of the thesis involves collaborations with industrial labs from Adobe, Intel, Facebook and Google. I am grateful for Qifeng Chen and Vladln Koltun for letting me explore raw sensor data and lending me a Sony A7 camera with a Sony FE 24-240mm zoom lens for experiments. I thank Kevin Matzen for the support on my video defocus paper even after the internship ends. I had some of the most fulfilled time interning at Gcam with David E. Jacobs, Jonathan Barron and Yun-Ta Tsai. I am grateful for David being a mentor and friend, giving me the freedom of doing research and advise me on my professional path. Of course, I will never forget us staying up all night revising each word of the paper before the SIGGRAPH deadline. I thank Jon for always bringing intellectual insights and joy into conversations. It is such a bless to work with him. I am also grateful for Yun-Ta giving me the opportunity to work with the Light Stage, one of the most fascinating technology I have seen. There are other people who have contributed to this thesis work and they are all indispensable for what I achieved today. I thank Joon-Young Lee for taking me as an intern when I was a junior student and patiently guiding me through research, and Kalyan Sunkavalli and Zhaowen Wang for supporting me over the ups and downs of my first research project. I am thankful for Dillon Yao porting my giant, unoptimized system to the Scanner framework, for Vivien Nguyen iterating with me on the

defocus GUI, which is the key for our final demo. And for Yoo Zhang helping me with my adventures of data collection.

I thank the people from my undergraduate, Rice University, where Don Johnson taught me the first lessons of sampling theorem and how he managed to use signal processing to authenticate Van Gogh painting. I thank Ashok Veeraraghavan for bringing me into research and encouraging me to pursue a research path. I wholeheartedly thank Jason Holloway, Vivek Boominathan and Adithya Pediredla who act as mentors and friends, and have bared my endless and stupid questions and spent their time teaching me about image sensors, optics and camera systems.

I spent the most time of my graduate school with my lab-mates, some of the brightest people I have ever met: Pratul Srinivasan, Ben Mildenhall, Grace Kuo, Matthew Tancik, Utkarsh Singhal, Tim Brooks and Vivien Nguyen. The weekly boba time we had together, the GPU heat we bore together, the piano recital from Ben we have all enjoyed together, the dumplings and boba cookies we made together, and the refilled snacks, green tea shots, grumpy cat..., all these are remembered and cherished deep in my heart. In particular, for Vivien, for inspiring me to see how photography coincides with arts, and for making me feel I finally have a sister.

My PhD would never have been completed without the administrative help from Shirley Salanio, Audrey Sillers, Jean Nguyen, Shirley, Audrey, Jean, Susanne Kauer and Angela Waxman. They spent tremendous efforts to make us able to concentrate on our research. I especially thank Audrey for her listening and support, carrying me through the toughest time of my PhD.

I thank the senior students who encouraged me as people who have been there. I thank Tinghui Zhou, Junyan Zhu and Biye Jiang for being inspiring mentors and friends from the visit day. A fun fact is that Biye and I took an 8:30am Introduction to Korean class together and we once studied Korean in VCL (Visual Computing Lab). I thank Lingqi Yan for his insights in rendering and material modeling, and Daniel Seita for his encouragement and feedback for my blogs[1], which I enjoy writing and have decided to carry on like he does[2].

My life at Berkeley has been delightful because of Jenny Huang. I enjoy our time hunting for new dessert places, hiking to the inspiration point to oversee Berkeley and making chicken paella for Christmas. I will never forget all the late-night chats and snacks on Soda 5th floor. Outside of Berkeley, I had the luck to have Holly Liang accompanying me all the way from Houston to the Bay area. She was with me the first day I entered this unfamiliar country 9 years ago, and have gone through the joy and tears with me, generously offering me love, trust, and letting me see the positive side of life. I also have enjoyed the genuine companionship with friends like Spencer Kent, Xuan Luo, Xiuming Zhang, Charles He, Yi Hua, Yoko Li, Julie Hao, Julia Xu, Lantao Yu, Zhixin Shu, Mandy Xia, Shutong Zhang, Zhengqi Yang,... the people I have met at different ages of my life, and have helped me discover, define and refine myself.

[1] https://people.eecs.berkeley.edu/~cecilia77/#blog

[2] https://danieltakeshi.github.io

The last few words go to my parents and my newly married husband Yoo Zhang. Mom and Dad have been the inspiration throughout my life. They are mentors who advise and friends who listen, being open-minded and always granting me trust and respect for my own decisions. Mom taught me how to love by showing her love to her students, to her families and to everyone around her. Dad's rigorous scholarship even shows in his scratch paper, and has influenced me to do so in my own research. Yoo and I just stepped into our decade anniversary since we were together in high school. There are only a handful of decades in one's lifetime, and I am lucky to have gone through one and look forward to the future ones to come. Yoo has accepted every side of me, and supported me with his full heart. I hope to do the same for him, and cannot wait to witness his every highlight moment in his career of filmmaking. Not to mention he contributed to the cinematography section of this thesis.

This thesis is also for my grandma and my uncle, who I am sure are happy to see what I have accomplished from heaven. My love and memory of you will never pass away.

Looking back, I am lucky to have received the love and support from all these people in my life. I am also proud of myself for the perseverance, belief and growth I have made. If I were to go backwards in time and choose again whether to pursue a PhD. The answer is a definite yes, and I hope I could have met the same people and even encountered the same challenges. Every single piece of what I experienced in my PhD made me the person I am today, and all these are my treasure for a lifetime.

# Chapter 1

# Introduction

When I walked from the office building to my car the other day, I took three photos with my phone within the 10-minute walk: a light yellow vintage van on Bancroft avenue, bright orange and red cotton-like clouds lit by the golden hour, and blooming flowers grown on a roof shelf on campus (see Figure 1.1). American landscape photographer Robert Adams once said " No place is boring, if you've had a good night's sleep and have a pocket full of unexposed film." Put in today's context, we may replace "a pocket full of unexposed film " with "a pocket phone". Those three photos I took on my way to my car are spontaneous and quick snapshots, enabled by modern compact imaging system that internally accomplishes all the fundamental hardware and software image processing pipeline such as stabilization, autofocus, auto-exposure, auto-white-balance, etc. Consumer-grade cameras have shrunk their sizes to fit in our pockets, accompanied by advanced optics and electronics to preserve imaging quality. Artificial Intelligent (AI) efforts in the post-processing pipeline even enable single-snap features that reach or even surpass physical limits, such as high dynamic range from a burst of photos [60], hand-held low-light imaging [109] and synthetic shallow depth of field for portrait photography [170]. Cloud storage and high speed telecommunication made it easy for us to instantly upload and share photos and videos with our families and other communities. We, as visual beings, use cameras to record not only significant historical moments, but also our daily lives. Casual imaging, in other words, has democratized visual content creation and documentation.

## 1.1 Casual Imaging

In this dissertation, I use the term *casual imaging* to refer to using consumer-grade cameras for photography and videography, as opposed to professional imaging as studio photography and cinematography.

Since Philippe Kahn took the first cell-phone photo of his new-born daughter in 1997, the way we communicate and perceive the world has been tied with casual imaging – everyone with a phone has a camera within reach, and photo enthusiasts have DSLR's or

Figure 1.1: Photographs taken with my smartphone on a random day when I walked from my office building to my car.


Figure 1.2: Computational photography on smartphones features developed at Google Research and deployed on Google Pixel smartphones. A) HDR+ using burst photography [60]. B) Night sight for low-light imaging [109]. C) Portrait mode with synthetic shallow depth of field [170].

better. Up until the year of 2020, over 1.43 trillion of photos have been taken worldwide, and 500 hours of video is uploaded every minute.

Casual Imaging offers unique advantages for both people holding the camera and people facing the camera. It provides content creators (*i.e.*, photographers and videographers) with the convenience and accessibility to capture spontaneous life moments and documentaries when point-and-shoot is preferred [11]. The size of casual imaging devices also benefits non-intrusive documentary filmmaking by placing the camera in a visually unobtrusive position, especially when the subjects are not used to facing large cameras and lighting setups.

While casual imaging has these practical benefits, achieving high image and video quality on these casual devices is extremely challenging and has long been a research question.

To state more clearly, I define "quality" from aspects of both perceptual quality and semantic quality. Perceptual quality is inherent to the camera optical system, electronics and signal processing software, referring to image resolution, noise level, geometric distortion, dynamic range, etc. Perceptual quality can be measured quantitatively using conventional reference-based metrics such as PSNR, SSIM [180] or neural-network-based perceptual metric such as the LPIPS [195]. Semantic quality, on the other hand, is inherent to the scene content and the clarity of storytelling. Focusing point, field of view, scene composition, video cuts from one shot to another all affect how a viewer interprets a certain image or a video and thus its semantic quality. Semantic quality is challenging to evaluate and measure because it is subjective and oftentimes by virtue of artistic choices.

Producing images with high perceptual and semantic quality is challenging. Professional photographers and cinematographers achieve so by using advanced imaging equipment, scripted storyline as guidance, and a constructed environment. But on the other hand, casual imaging is carried out in natural environments by average users who record daily lives. Challenges regarding casual imaging arise from: 1) limitations in optical and hardware systems, 2) inability to modify existing environments such as lighting and 3) real-time autofocusing latency for spontaneous events. The goal of this thesis is to design algorithms and software systems that combat these challenges to enhance perceptual and semantic quality of casually taken images and videos.

## 1.2 Challenges of Casual Imaging

The quality of casual images and videos are yet comparable with those taken under a professional context. Challenges primarily come from optical and hardware limitation of casual devices (*e.g.* image sensors and lenses), imperfect shooting environment, and the nature of uncertainty that makes real-time autofocusing fundamentally difficult.

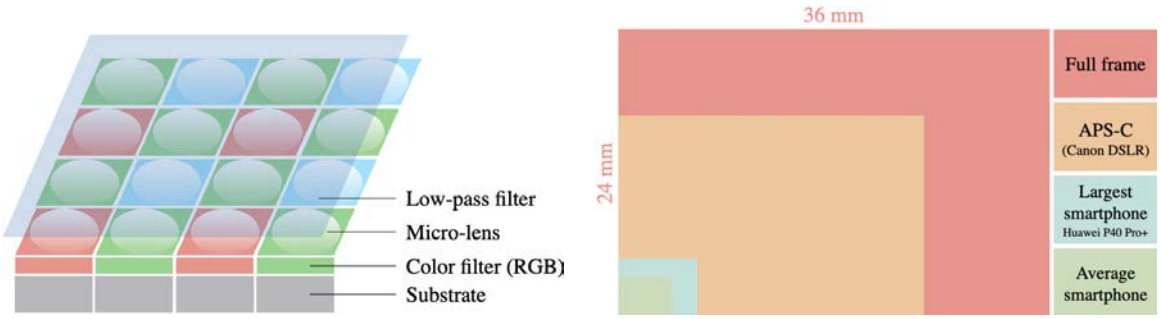



Figure 1.3: **Left**: A simple diagram of Bayer sensor used in most CMOS sensors. Incident light is focused by the micro-lenses and filtered by the green, red and blur color filters. Low-pass filter, or Anti-aliasing filter (AA filter) alleviates moiré or other high-frequency artifacts, but undesirably blurred out details. Many modern cameras have thus left the AA filter out. **Right**: Image sensor size comparison. Even the largest smartphone sensor is small in size, compared to conventional large cameras (*e.g.* DSLR). This causes great challenges to produce good quality imagery on small cameras (*e.g.* smartphones).

### 1.2.1 Image Sensors

The basic structure of image sensors consist of a mosaic filter for sensing color information and a microlens array to condense the incident light on each pixel (silicon surface) [119], see Figure 1.3. The majority of digital cameras uses CMOS image sensors these days for its practical advantages such as low power consumption and camera-on-chip integration. Major performance characteristics of image sensors include quantum efficiency, sensor noise, dynamic range and spatial resolution. Understanding these characteristics provides intuition in the parameters when evaluating imaging systems, and enables physically-based simulation for statistical modeling. A comprehensive overview of CMOS sensors can be found in [27].

The biggest limitation of image sensors on mobile devices is the sensor size. While most smartphone cameras as of today produce mega-pixel images, high pixel count is not equivalent to high quality because of the physically limited sensor size. With the same pixel count, the best smartphone cameras use a sensor with a crop factor (the sensor's diagonal size compared to a full-frame 35 mm sensor) that ranges from $5$ to $6.5$, while mainstream DSLR have a $1.5$ crop factor, indicating each pixel area in a smartphone camera is $1/16$ to $1/11$ that of a common DSLR (see Figure 1.3). With the same exposure time, much less light goes into each pixel and thus less accurately light is measured. One dominant sensor noise type is photon shot noise. Photon shot noise measures the variance of the number of photons arriving when exposure varies from pixel to pixel. It follows a Poisson distribution due to its independent occurrence. If on average $\lambda$ photons are detected in an interval of time, according to properties of Poisson distribution, its mean and

variance, and the standard deviation follow:

$$\mu = \lambda, \quad \sigma^2 = \lambda, \quad \sigma = \sqrt{\lambda}$$

The signal-to-noise ratio (SNR) is measured by the ratio of mean pixel value and the standard deviation of pixel value, which is thus

$$SNR = \frac{\mu}{\sigma} = \frac{\lambda}{\sqrt{\lambda}} = \sqrt{\lambda}$$

This indicates shot noise scales as square root of number of photons. A $16\times$ decrease in area will lower the SNR by $4\times$ or $-12$dB. To achieve a similar quality imagery, the camera either opens up aperture to let more light in, or increase the sensitivity (ISO) of the sensor. The former is difficult for compact cameras like smartphones, for the same reason as the sensor being physically limited. Increasing ISO will induce dominance of read noise, and has little effect under dim or low light environment.

Quanta Image Sensor (QIS) [37] is envisioned to be the next paradigm of image sensor. In a QIS, photoelectrons are counted one by one as a binary measurement in jots. This indicates low read noise in a QIS. The readout speed can reach 1000fps to allow burst frame processing for extreme low-light or high dynamic range imaging. Though image reconstruction from raw QIS data is still in its early research stage to be made practical, and how to make the best use of its temporal and spatial resolution also remains open questions.

### 1.2.2 Camera Lenses

Similar to the physical limitation of image sensor size, camera lenses on casual devices are constrained in size, and thus the maximum focal length as well as the maximum entrance pupil, also called the aperture (seen from an axial point on the object side).

Focal length limits the optical zoom power of smartphones, and thus the ability to capture objects far-away. The average telephoto lens on smartphones today has a focal length of 56 mm, while large cameras can mount a zoom lens that ranges from 24 mm to 240 mm, or even 400 mm. All these are 35 mm equivalent numbers. To achieve the same level of zoom, smartphones can use digital zoom. Conventionally, digital zoom is achieved by various interpolation methods to upsample the image to higher resolutions or by converting between spatial and frequency domains from a signal processing perspective [10, 128]. Recently, learning-based methods on super-resolution leverage the correlation among local and global patches as well as image semantics to recover the missing high frequency details in more faithful ways [187]. On smartphones, a widely used approach is to resolve details using multiple images taken within a short amount of time – a burst sequence [183]. This is often operated in the raw space to preserve the maximum information.

The most powerful optical system on a smartphone is deployed on the Huawei P40 Pro, which uses a folded optic periscope system that achieves $10\times$ optical zoom. However, the

zoom levels are discrete and sparse. For example, a $5\times$ zoom is still achieved by digital zoom.

Aperture controls the area over which light can pass through the camera lens, often measured as a F-number $(N)$, which is the ratio between the the focal length and the diameter of the aperture, denoted as

$$N = \frac{f}{A} \tag{1.1}$$

With the same F-number, a smartphone camera and a DSLR receive a drastically different amount of light because of the aperture size. Widening up the aperture also requires attention to lens design as a bad design could worsen aberration distortion [52] such as lens-flare artifacts [71].

Another physical limitation induced by aperture is depth of field, which is the range of distance over which objects appear in sharp focus. Although a f/2 aperture often associates with a photo with shallow depth of field, the defocus blur size may appear just noticeable on a smartphone device. The depth of field is determined by the size of the circle of confusion (CoC), which depends on the sensing medium, reproduction medium, viewing distance, human vision and other various aspects. With a thin lens model, the depth of field can be written as

$$D \approx \frac{2NCS_o^2}{f^2} \tag{1.2}$$

Where $N$ is the F-number, $C$ refers to the CoC, $S_o$ the object distance and $f$ the focal length. The full derivation can be found in Section 2.1. Using the same F-number (same $N$), taking a photo of the same object (same $S_o$) and maintaining the same field of view ($f$ shrinks, along with $C$), the depth of field increases linearly with decreasing sensor size. This is one reason small casual devices are not able to render aesthetically pleasing defocus blur. It is also worth noticing that the depth of field is quadratic with focusing distance $S_0$, which explains when we use our phone to take a photo of a close-up object, the defocus blur appears more salient, though still hardly noticeable compared to that of large cameras.

The defocus blur is proportional to the object depth by a normalized scalar [63] (see Section 2.1 for details), which has the same geometry with disparity in a stereo setup. This is intuitive by viewing the two ends of our eyes (or the lens aperture) as two pinhole camera. Each point in the object space projects to the images formed by the two cameras.

In order to simulate more perceptually-appealing defocus blur on a smartphone, one needs to know the depth (inverse disparity) of each pixel in the object space. But even with perfect depth information, the rendering is always an approximation to the physical phenomenon because out-of-focus region receives contribution from pixels that are occluded, which is hard to acquire from a single fixed viewpoint.

### 1.2.3 Lighting

Lighting, either natural or artificial, refers to how the light sources are positioned relative to the subjects in space, time and direction. As modeled in computer graphics, the irradiance (*i.e.*, optical power) received at a single point to be the integral of radiance from all directions (*e.g.* a hemisphere over the point) weighted by their angles between the surface normal:

$$E = \int L(\omega) \cos\theta d\omega \tag{1.3}$$

where $L(\omega)$ is the radiance in the direction $\omega$ and $\theta$ is the angle between $\omega$ and surface normal.

The incoming radiance with respect to the sensor or film, is the outgoing radiance from points in the scene. The rendering equation tells us that given a scene point $\mathbf{p}$ and a direction of interest $\omega_o$, its outgoing radiance is formulated as:

$$L_o(\mathbf{p}, \omega_o) = L_e(\mathbf{p}, \omega_o) + \int f_r(\mathbf{p}, \omega_i \rightarrow \omega_o) L_i(\mathbf{p}, \omega_i) \cos\theta d\omega_i \tag{1.4}$$

where $L_e(\mathbf{p}, \omega_o)$ is the emitted radiance (*e.g.* if $\mathbf{p}$ is on a light source), $f_r$ is the scattering function that characterizes how light interacts with the object surface. $L_i$ is the relative incoming radiance at point $\mathbf{p}$.

In photography, the sensor records the irradiance ($E$) at each sample location. The integral is the summed radiance affected by the scattering function $f_r$ and thus the properties of the light sources and the subjects (*i.e.*, material and geometry) they interact with. The position and quality of light can affect how an object is depicted, from clarity to emotion. While there is no single formula of lighting that works for all scenarios, there is a lighting vocabulary (derived from studio photography) that breaks down the infinite possibilities of light into manageable topics for discussion [78]. For example, the quality of a light describes primarily its hardness and softness; the density of the shadow is determined by the amount of bounce and fill light in the scene. An intimate family scene may dictate the use of sunset and sunrise lighting to render a warmer tone; strong directional lighting exaggerates details and 3D shape, forming crisp and harsh shadows; diffuse lighting smooths out surface roughness and renders a softer looking. All these lighting styles can be constructed in a controlled environment with auxiliary lighting equipment.

Casual imaging only uses available lighting from the environment. While this preserves the real world in an unaltered way, it suffers from the lighting conditions being suboptimal for photography. It is not practical to always wait for the optimal lighting that happens a few hours of a day. It is also not intuitive for an average user to decide the best camera position given existing lighting. Casual imaging is constrained by what lighting is available at the moment of capture – no matter if it is portrait or landscape, the camera passively uses existing lighting, which can oftentimes be unflattering. Figure 1.4 illustrates a few common lighting problems in casual photography.

(A) Backlit (B) Low light (C) Hard light (D) Lens flare (E) Cast shadow

Figure 1.4: When the subject is lit from the back (A), it is hard to set a single exposure to cover the wide dynamic range, and thus the background easily goes saturated and the subject appears darker than desired. Under low-light (B), the image suffers from strong shot noise. Midday sunlight (C) is harsh and strong, leaving high contrast facial shadows with hard edges, making the photo look distracting. When strong light is within the field of view of the camera, lens flare (the green circle and light streaks) caused by imperfect optics and dust will be visible in the image. In (E), close objects (the window frame) cast shadows and form an unpleasing look.

### 1.2.4 Focus

Conventional camera autofocus systems are classified into two buckets: contrast-detection autofocus (CDAF) and phase-detection (PDAF). CDAF is slower, seeking focus by aiming to maximize image contrast as the lens focus is changed; it performs poorly for video because the "focus seeking" behavior is visible in the recorded video. Phase detection can be much faster, and is based on separately detecting and comparing light passing through different parts of the lens aperture. This was achieved in SLR cameras by reflecting light onto PDAF units that each comprised microlens atop multiple pixels [49]. More recently, dual pixel autofocus (DPAF) begins to dominate the consumer cameras. Unlike PDAF where only a small portion of the pixels are responsible for autofocus, DPAF has each pixel on the CMOS image sensor attached with two photodiodes, often accomplished by placing two lenses on-top each pixel, and the two sides can operate separately or jointly. Each pixel is therefore calculating phase differences and focus, greatly improving the focus speed and accuracy.

Despite these physical autofocus systems, autofocus mistakes remain common and inevitable in casual videography, because the focus of each frame is "locked in" as it is being shot. A full solution is impossible because the autofocus algorithm would have to predict the future (*e.g.* where actions go, who picks up the next dialogue line, etc.) at every frame to correctly determine what to focus on, or transition focus to. Because there is no movie script as in cinematography, memorable moments and decisive actions occur unpredictably and can be easily missed.

Smartphones visually suffer less from autofocus errors because the small lenses cause essentially everything to be in focus at the same time. This is undesirable as videos often appear cluttered and lens defocus quality is low; the ability of using focus to guide the viewer's gaze is lost. We either sacrifice shallow DOF, as in smartphone videos; or we struggle to deliver accurate focus, as in videos from larger cameras.

### 1.2.5 Other Challenges

The challenges in optical limitations (sensors and lenses), lighting environment and focus are addressed in this thesis, but they do not cover all the obstacles in casual imaging. Traditional camera white balance problems – the accuracy of rendering neutral colors – are unsolved under low light and complicated mixed lighting conditions. One example is correctly rendering the skin tone of a person lit by neon light. Another aspect that requires considerable professional experiences and is hard to achieve in casual imaging is composition and framing, which determines the arrangement of the visual elements in an image. There are rules of composition such as the rule of thirds, S-shapes, contrasting/complementing colors, visual rhythm, leading lines, etc., but having a thorough understanding of these rules to decide what to use for a particular image requires practice and relevant education. Not to mention composition is sometimes often of an artistic choice that differentiates between good works and masterpieces. Related to the problem of composition, obstructions in the environment are difficult to be identified and effectively removed. Imagine a tourist sight with strangers, the camera system needs to first identify who are the strangers and then remove them accompanied by filling the removed regions with contents. Another common and similarly challenging scenario is removing powerlines that appear thin structured and extremely distracting.

There are additional challenges for casual imaging on smartphones. Smartphone optics suffer from geometric aberration that produces lens flare that appears as light spots with color artifacts; high quality action shot is challenging on handheld device because the shot needs to be sharp (low shutter speed) and clean (low noise); long exposure are difficult to achieve on handheld devices where optical and sensor stabilization may not suffice to account for large hand motion.

## 1.3 The Importance of Context

The eyes look, but the brain sees. A camera can gain intelligence if it understands what it sees. A key contribution of this thesis is to identify and bring such contextual information about the scene to casual camera systems. The thesis proposes solutions to limitations in sensors, camera lenses, lighting and focus uncertainty under the framework of machine learning, showing the effectiveness of learning contextual and semantic signals from large-scale dataset. These contextual and semantic signals, encapsulated as *context*

in the following, includes but is not limited to pixel and patch correlations among natural images, the statistics of human faces and perceptual attention recognition.

To put it more concretely, if a camera 'detects' a familiar face among a crowd, it could set focus to that familiar person while blurring out the rest of scenes. If a camera 'recognizes' distracting obstructions in the scene, it could suggest the photographer reframing or removing the obstructions. If a camera 'knows' what a flower petal looks like, it could recover the details of a low-quality image of a flower petal. The capability of 'detecting', 'recognizing' and 'knowing' suggest a context-aware imaging system, which affords semantics to the streaming frames or captured imagery. Casual photographers will benefit from such an imaging system when they are limited by the device and environment.

Such context-aware imaging systems behave similarly to how humans perceive. Human perception operates in a collaborative manner between our visual system and brain. Because each individual sees the world in a different way as basic as colors [44], it may bring confusion to defining 'physical reality'. On the more positive other side, because our brain adapts to the environment continuously, it helps us see with context to understand easily and quickly. For example, when we see through a fence, our brain filters out the obstruction (the fence) and observes lucidly what's on the other side; when we see a person under colorful neon light at night, we are still able to infer his/her ethnicity even though the illumination has severely distorted the skin tone, because the human visual system is so sensitive to faces that parts of our brain are dedicated for face processing [79]; we are good at focusing on objects we can easily understand [134], and thus we naturally respond to spontaneous events we care about. These capabilities are based on our knowledge and understanding of the surroundings. My goal in this thesis is to amend a imaging system with a similar 'brain power' as humans do.

Similar to how the human brain develops in response to new information and data points, one approach to build a system with semantic understanding is learning patterns and features from big data. Recent breakthroughs in computing power (*i.e.*, GPU) and the emergence of large-scale datasets have enabled using machine learning, in particular deep neural networks, to automatically and effectively extract semantic features from data. Since the first deep neural network, known as the AlexNet [94], revolutionizes performance on classifying ImageNet dataset [20] in 2012, the field of deep learning has grown rapidly. Various computer vision tasks have leveraged semantic and high-level image understanding, such as object recognition, object detection and semantic segmentation, to name a few.

## 1.4 Challenges of Learning the Context From Data

The challenges of applying learning-based methods to computational photography problems are 1) formulating the problem into a machine learning framework, and 2) collecting labeled training datasets at scale.

Unlike image classification with a clear problem formulation (*i.e.*, image in, label(s) out) and has data with categorical labels that can be collected through crowdsourcing, many computational photography problems are not straightforward to fit into machine learning frameworks with clear definitions of input and output pairs. What is the ground truth for an algorithm that enhances portrait lighting? How to define and evaluate a good tone mapping method? What is the label for training an autofocusing model? Even if one is given a well-defined problem, it remains a challenge how to acquire a dataset at scale with sufficient diversity for training. Because the output of computational photography problems are almost always images, the labels are dense at pixel level and often at the same resolution as the input. For example, the label for image denoising is an image that has a higher signal-to-noise ratio (SNR) than that of the input. Capturing pairs of images with low and high SNR, however, requires great labor and domain knowledge in photography.

An alternative approach for real data collection is to use synthetic data, for example, simulating sensor noise to learn an image denoising model. The advantage is that we can use internet-scale images for simulation, but the downside is that synthesis is always an approximation of the real physical process, and thus the distribution gap [1] between training and testing data degrades the performance.

## 1.5 Applying Machine Learning to Enable Context-aware Casual Imaging

In view of the challenges of applying machine learning to casual imaging, the following chapters (Chapter 3, 4, 5) propose learning-based methods to the problems of superresolution [200], reflection removal [198], facial lighting editing [196] and casual video autofocus [199], by breaking them down into tractable sub-problems, and by collecting and processing tailored datasets.

Chapter 3 brings machine learning into raw sensor space (Section 3.1) and learns correlation among raw sensor patches, formulating super-resolution as a joint task that performs image processing pipeline (camera ISP) and upsampling. The other problem mentioned in Chapter 3 is reflection removal (Section 3.2), which can be naturally formulated as a supervised machine learning problem. We demonstrate that it is necessary to use a hybrid collection of synthetic, which assures diversity and scale, and real data, which brings in natural image statistics, to obtain good removal quality. Chapter 4 uses a data-driven method to understand human faces and lighting in order to fix and improve portrait lighting. The key is to break down the problem into sub-problems – in this case, categorize portrait shadow into two types: foreign and facial shadows based on the source of shadows – to make the problem manageable and well-defined to fit into a machine learning framework. Similarly, Chapter 5 decomposes the problem of smartphone autofocus into

[1] Transfer learning [126] and domain adaptation [41] are ongoing research directions to improve model generalization.

two parts: rendering synthetic shallow depth-of-field and identifying salient focus. The former uses machine learning to infer depth from a single image by learning the correlation between object appearance and their relative depth with the camera; the latter learns about scene semantics to predict the salient region that is more likely to be a meaningful focus.

The goal to integrate machine learning is to extract *context* that is useful for downstream imaging tasks, minimizing the compromises on image quality in conventional casual imaging. This thesis shows that with this approach, casual imaging can produce visual quality of professionals that is sharp, visually pleasing and semantically meaningful.

## 1.6 Dissertation Road Map

The organization of this dissertation is as follows:
**Chapter 2** briefly overviews the lens geometry, talks about professional photography and cinematography and their differences from casual imaging. Studio lighting techniques and film crew responsibilities are briefly described. This chapter then introduces *contextual information* that has been missing in conventional casual imaging, and how machine learning can bring it back by extracting useful information from large-scale data collections.
**Chapter 3, 4, and 5** dive into the technical details of making casual imaging cinematic from the three aspects – perceptual, lighting and focusing quality. These techniques use machine learning to embed contextual information – object statistics, face semantics and scene descriptors into features that can then be used to guide each task. In particular, **Chapter 3** focuses on image clarity and removing distracting signals from an image such as reflections to enhance image perceptual quality. **Chapter 4** talks about the challenges and importance of handling lighting and shadow for casual portrait photography, leveraging studio lighting principles to manipulate and enhance unflattering portrait lighting. **Chapter 5** proposes a prototype of real-time autofocusing for casual videography on smartphones, simulating shallow depth of field to render refocusable videos and bringing in video saliency to guide where and when to shift focuses.

If the readers prefer, a visual version of the roadmap is available in Figure 1.5.

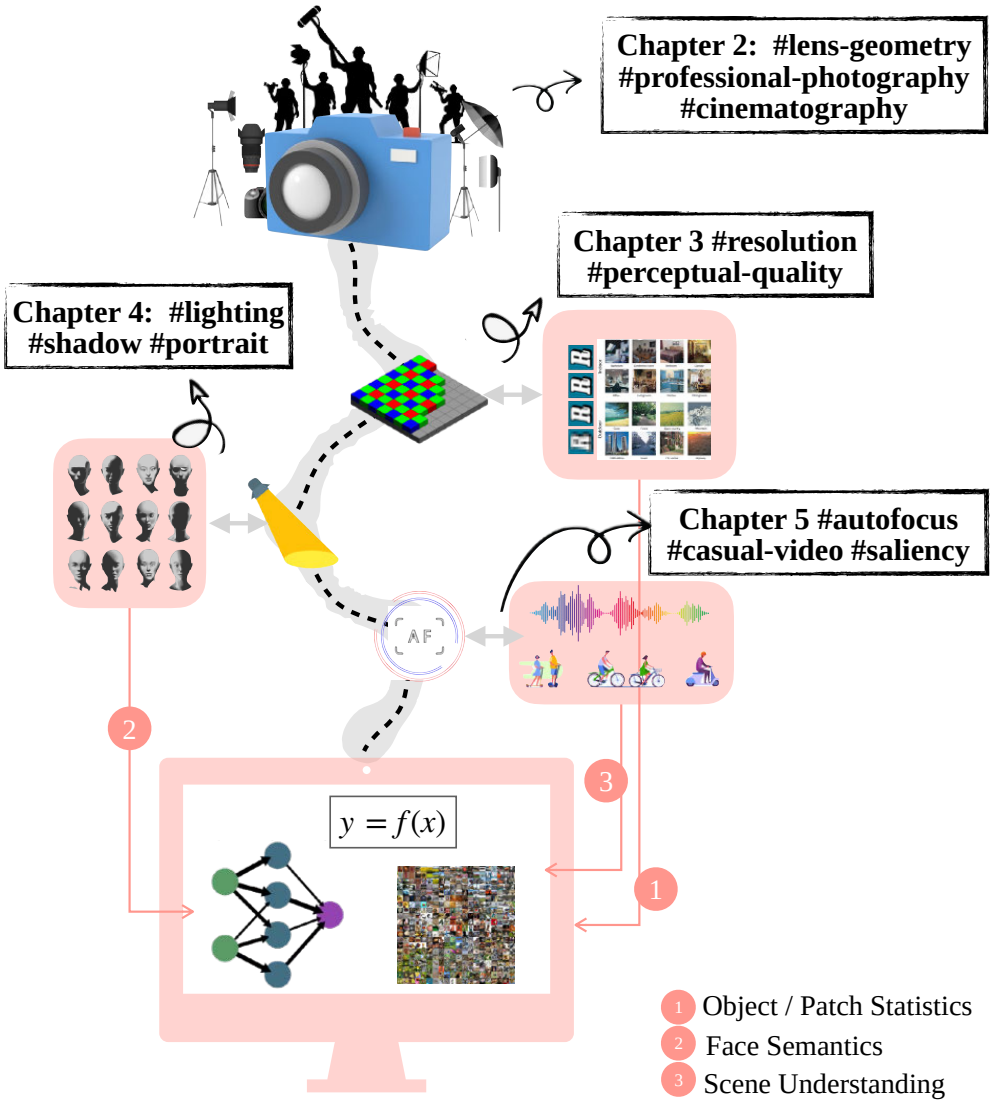



Figure 1.5: The dissertation roadmap. Chapter 2 starts with an overview in lens geometry, and continues with professional techniques in photography and videography (cinematography). The importance of context is introduced as key to making casual imaging cinematic. The thesis continues with a deep dive into the technical details of improving casual imaging from the three aspects of perceptual, lighting and focusing quality. Contextual information is extracted from different aspects accordingly: object statistics, face semantics and scene understanding.

# Chapter 2

# Background

The primary goal of building computational tools for casual imaging is to produce images of high quality like the ones produced by professional photographers and cinematographers. Professionals use better cameras and lenses, and more importantly, they manipulate and control the environment to be supportive of their photo-shooting plan or a movie script. An iconic example is the opening scene of "The Godfather" [182], where Gordon Willis, the cinematographer, sets up low-key overhead lighting to establish a dark and insidious tone that aligns with the story.

While professional photography and cinematography largely attribute to artistic and subjective choices, there are principles in this creative process that are commonly adopted – on camera settings, lens choices, lighting, composition, focusing and other visual elements. These principles are grounded on the understanding of context, which professionals take full control over but casual users know little : who is the subject, what is the occasion and how does the narrative go. To achieve high perceptual and semantic quality images and videos, one needs to understand and sets an expectation on the context. This chapter is decomposed into a brief overview of professional photography and cinematography, followed by how large-scale dataset and machine learning can transform professional principles into contextual signal to guide casual imaging.

## 2.1 Lens Geometry

Lens geometry explains many of the camera features mentioned in Chapter 1 such as depth of field and defocus blur. To simplify the analysis, I briefly overview its geometric formation using a thin-lens model (see Figure 2.1). The Gaussian lens formula [146] tells us

$$\frac{1}{S_o} + \frac{1}{S_i} = \frac{1}{f} \tag{2.1}$$

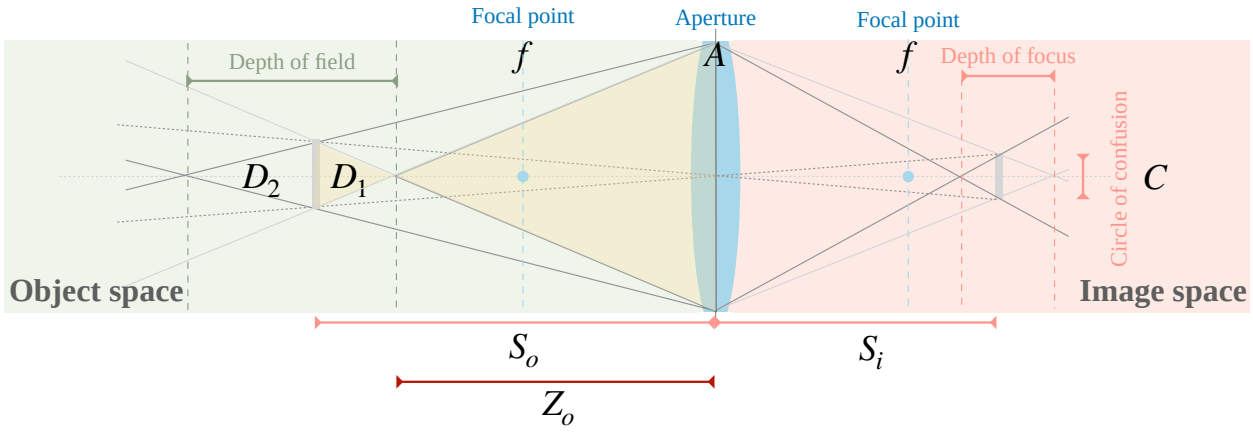



Figure 2.1: A thin lens diagram that demonstrates the geometric relationship between the circle of confusion (C) and depth of field $(D_1 + D_2)$. From the analysis we show depth of field increase linearly with decreasing sensor size.

The depth of field of this thin lens system is $D = D_1 + D_2$ (note that $D_1 \neq D_2$), and can be derived by similar triangles marked in yellow. We have

$$\frac{D_1 * f}{CS_o} = \frac{S_o - D_1}{A}$$
$$\frac{D_1 * f}{CS_o} = \frac{S_o - D_1}{\frac{f}{N}}$$
$$D_1 = \frac{NCS_o^2}{f^2 + NCS_o}$$

Similarly, $D_2 = \frac{NCS_o^2}{f^2 - NCS_o}$. The complete depth of field is thus

$$D = D_1 + D_2 = \frac{2NCS_o^2 f^2}{f^4 - N^2C^2S_o^2}$$

When circle of confusion is small relative to the aperture, which is often the case, we can simplify the depth of field further to

$$D = \frac{2NCS_o^2}{f^2} \tag{2.2}$$

Equation 2.2 shows how depth of field is correlated with the F-number, circle of confusion, object distance and the focal length.

With the geometry illustrated in Figure 2.1, we can also derive the formulation of the size of blur with respect to the depth of the object in the scene.

For an object at distance $Z_o$, it's projected circle of confusion $C_o$ on the image plane can be written as

$$C_o = |A(\frac{S_i}{S_o})(1 - \frac{Z_o}{S_o})| \tag{2.3}$$

The sign of the numerical value (without $|\cdot|$) of $C_o$ depends on whether the object is located in front of or behind the focal plane. As similarly derived in [63], if we express the $S_o/Z_o$ as the relative depth $d$, equation 2.3 can be written as

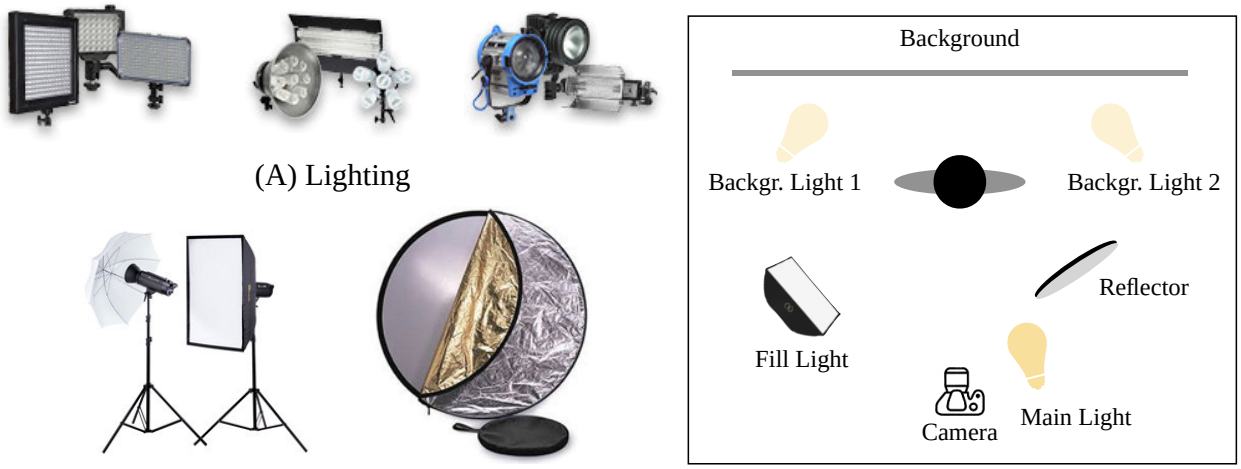
$$C_o = |A(\frac{S_i}{S_o})(1 - \frac{1}{d})| \tag{2.4}$$

This indicates that the defocus blur is proportional to the object depth by a normalized scalar, and that it is necessary to have scene depth to synthesize defocus blur.

## 2.2 Professional Photography

I use the term 'professional' to describe people who take photos for sessions of time with auxiliary equipment (*e.g.* artificial lighting), either in a studio or outdoors. Apart from high-end cameras and lenses that assist professional photographers obtain high quality signal on the film or sensor, carefully controlled lighting is another crucial factor that distinguishes professional from casual photography. In professional photography, lighting is adjusted to establish certain clarity and tone, from the aspects of direction, quality (softness), temperature and intensity (brightness).

Studio photography constructs a virtual environment with full control over lighting and placement of the subject. Because there is no existing lighting to comply with, the space for creativity is broad and possibilities are almost endless. Figure 2.2 illustrates the basic lighting equipment used in studio portrait photography, including various types of lighting, diffusers and reflectors. The most common studio photography is portrait photography, where the photographer spends a few hours optimizing the position and quality of lighting that aligns best with the subject's identity and appearance. The goal is to achieve the balance of lighting and shadow that establish the desired tone and emotion. There are two major types of shadows on human faces – cast shadow caused by occlusion, and attached shadow caused by geometry change (the cosine between the lighting direction and surface normal increases), see Figure 2.3 for an illustration. A slight change in the lighting direction changes how shadow appears on the subject's face. Because the human perception system uses shadows, especially cast shadows to infer object shape and 3D geometry [135, 114], portrait shadow is essential for 3D perception of faces. The artistic uses of portrait lighting and shadow are inspired from classical portrait paintings. One of the most iconic lighting types is Rembrandt lighting, named after the famous Dutch painter Rembrandt Harmenszoon van Rijn. Rembrandt lighting, together with loop lighting, split lighting, butterfly lighting and others belong to the category of dramatic lighting, which




(A) Lighting (B) Diffusors (C) Reflectors (D) Typical Studio Lighting Setup

Figure 2.2: A, B and C illustrate various types of lighting equipment used in studio photography. Diffusers are used to evenly spread the light. Reflectors are used to lighten up shadows, or to reflect colored light onto the subject. A typical lighting setup – Three Point Lighting – is layout in (D), where a main light, two background lights are used, together with a diffused fill light to compensate for the darker region of the face.

is common in studio photography but not in outdoor photography, because the latter requires a harmonization between artificial and existing lighting.

Outdoor or natural lighting photography uses the sun and skylight as the main light source. It follows similar lighting principles and uses similar lighting equipment as in studio photography. The major difference in outdoor photography is its constraints from the existing environment – time of day and weather. For example, mid-day sun casts strong and directional sunlight where shadows appear high-contrast and hard, while cloudy day gives diffuse and soft lighting that is often cooler. Depending on the desired portrait appearance, photographers may use diffusers to soften harsh lighting, or use reflectors and artificial lighting to add intensity and change dominant lighting color/temperature. Without any additional equipment, natural lighting can be ideal as it is, but only during specific time of day and weather conditions. Many photographers are willing to to wait for the right moments to come, such as the golden hour, which is the period of daytime shortly after sunrise or before sunset, or the sunset after a storm where partial sunlight streams through the cloudy sky and creates dramatic contrast.

Another important aspect in professional photography is the strength and quality of catch light – the reflection of light sources in the subject's eyes [51]. Humans are naturally inclined to make eye contacts, and thus eyes are the first thing a person notices when showing a photo or painting a portrait. Catch light is often overlooked in casual photography, partly because it requires additional equipment such as reflectors, and also careful


(A) Different portrait shadow types (B) Catch light

Figure 2.3: Two major shadow types on human faces are cast shadow and attached shadow. Cast shadow is caused by occlusions that block the main light; the shadow region is mostly lit only by global illumination. Attached shadow is caused by change of angle between the surface normal and lighting direction, as illustrated in (A): as the side of face curves inward, there is a gradual decrease of light that lands on the surface. Catch light (B) is crucial for professional photography because it makes the subject look more lively. It is often overlooked by casual photography, partly because catch light often requires intentional placement of both the light source and subject.

placement of the light source and subject – all is difficult for casual imaging.

## 2.3 Cinematography

Cinematography is the artistic process of telling a story in motion through a lens. If the goal of professional photography is to produce a still image at its best quality possible, the goal of cinematography is to make each frame of the quality to be cohesive and supportive of the story, even if each frame is not perfect on its own. Cinematography emphasizes live-action issues and techniques; it needs to account for the subject movements – how lighting changes along with the movements and when focus shifts from the current to the next focusing point. Getting these right requires a dedicated crew (see Figure 2.4 for a mini-version of a film set) to perform staging, framing and lighting design to plan ahead where to place the camera, how to set the lighting and where the actors move for every single shot.

The gap between cinematography and casual videography is largely due to the amount of contextual understanding and control over the scene. Cinematography happens on a set – an environment established with collaborative efforts between the director and the

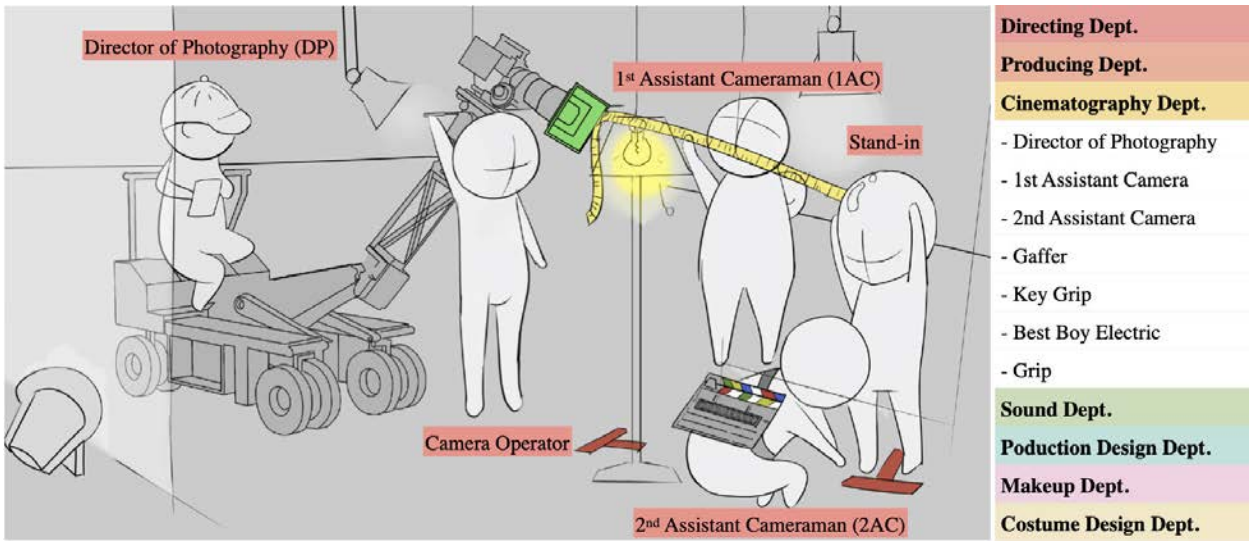



Figure 2.4: (Left) An illustration of a film crew on set. The director of photography (DP) makes decision of the lighting and camera according to the movie script. The 1st assistance cameraman (1AC) measures the focus distance for the shot, and the 2nd assistant cameraman (2AC) marks the position on the ground to guide the actor. On the right is a list of key crew in the cinematography department and the key departments for a feature film.

crew. Actors can then rehearse the scene, staging and framing can be altered, and props can be redressed to take best advantage of the lighting design. The camera setups and actor movements are then marked for correct blocking; occasionally even the eyeline of the actors are marked to avoid crossing the line [1] and having a reverse angle. During the actual shooting, precise focus is guaranteed because the focusing points have been marked on the camera lens according to the blocking. In case there is a focus error on one camera, there are on average seven other shots for any single movement for the editor to later choose from. These shots cover different angle or field of view of the scene, including but not limited to full shot, medium shot, close-up shots, over the shoulder shot and high/low angle shots. There are movies where focus is compromised for other purposes such as emotion continuity. Les Misérables (2012) is one example where the audience can observe focus errors in close-up shots because the director pursues the integrity of a complete musical piece and actor emotion, which inevitably induce misfocus when the movement is relatively large. In movies with improvisational acting, such as Coherence (2014), the cinematographer must try to anticipate what the actors will do; focus is occasionally wrong and delayed. These are of course not standard cinematography, and require the crew to

[1] Cinematography terminology, meaning two characters in a scene should maintain the same left/right relationship to one another. When the camera passes over the invisible axis connecting the two subjects, it is called "crossing the line".

be even more experienced and capable to produce cinema quality shots.

Cinema lighting is design in a unique way that accounts for motion. Lighting is designed layer by layer with a base light as foundation, similar in spirit to painting. Each light source may or may not use a 'flag' – a black cloth supported by a C-stand [2] – to crop a portion of its light path so that they do not interfere. Sometimes the use of the 'flag' is to make a light path that aligns well with the subject motion. There are exceptions where the film pursues naturalism to minimize the altering of the environment lighting. The Revenant (2016) is one of the movies to pursue such realism by using natural lighting. As a result, the movie is shot for only few hours per day due to natural lighting being undesirable most of the time. Barry Lyndon (1975) is another film that is shot with candle lights using a specialized Zeiss lens with a f/0.7 aperture that is originally developed for NASA's Apollo missions.

Hardware is another significant difference between casual videography and cinematography. Cinema camera is big in size to support efficient heat dissipation for long-time shooting, to balance the heavy cinema lens built to be durable even in the harshest weather conditions, and to act as a hardware hub for live-streaming to multiple monitors (*e.g.* for the director, cinematographer, etc.), sound systems and various types of stabilizers. Different crew members in the camera department are in charge of operating the camera, controlling the dolly and pulling the focus. There is hardly any hardware aspect that is compromised for storytelling on a film set.

Cinematography is a highly customized and elaborated process that pursues perfection. None of its designs, setups or crew is available under a casual setting – no auxiliary hardware, no script, no expectation of subject movement, and no amelioration to unsatisfying lighting conditions. Casual videography is extremely challenging, because it is an one-man job to capture spontaneous actions in life, and often on a hand-held device.

[2] Cinematography terminology, referring to a stand that is primarily used to position light modifiers, such as silks, nets, or flags, in front of light sources.

# Chapter 3

# Learning to Enhance Perceptual Quality

In this chapter, I aim to demonstrate how machine learning can improve the perceptual quality of images through the tasks of super-resolution that learns patch statistics and reflection removal that exploits natural image priors. The former shows that when applying machine learning to digital zoom, it is beneficial to operate on real, RAW sensor data. Existing learning-based super-resolution methods do not use real sensor data, instead operating on processed RGB images. We show that these approaches forfeit detail and accuracy that can be gained by operating on raw data, particularly when zooming in on distant objects. The key barrier to using real sensor data for training is that ground-truth high-resolution imagery is missing. We show how to obtain such ground-truth data via optical zoom and contribute a dataset, SR-RAW, for real-world computational zoom. The latter part of the chapter presents one of the first learning-based approaches to separating reflection from a single image. The model learns and uses priors of natural images – low-level and high-level image features extracted from a convolutional neural network. We also collect one of the first real-world datasets for reflection removal, which is used to demonstrate the necessity of training with a hybrid collection of real and synthetic data for this task.

## 3.1 Learning from Raw Sensor Data for Super-resolution

### 3.1.1 Introduction

Zoom functionality is a necessity for mobile phones and cameras today. People zoom onto distant subjects such as wild animals and sports players in their captured images to view the subject in more detail. Most smartphones are equipped with at least two cameras at different zoom levels, indicating the importance of high-quality zoom functionality for the consumer camera market.

Optical zoom is an optimal choice for image zoom and can preserve high image quality, but zoom lenses are usually expensive and bulky. Alternatively, we can conveniently use

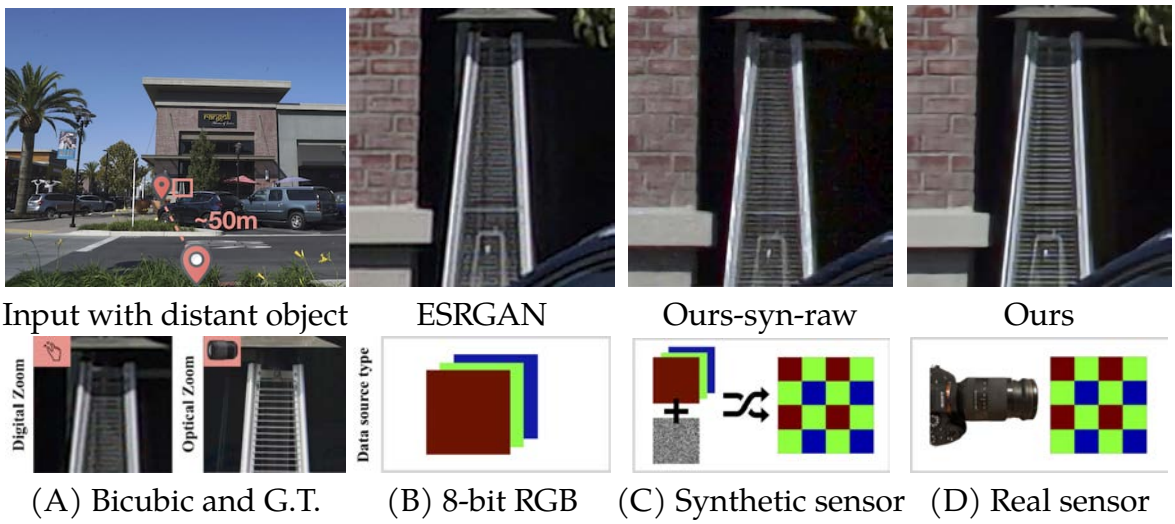



Figure 3.1: Our model (D) trained with real raw sensor data achieves better 4X computational zoom. We compare zoomed output against (B) ESRGAN [178], representative of state-of-the-art learning-based super-resolution methods, which operate on processed 8-bit RGB input, and (C) our model trained on synthetic sensor data. In (A), digital zoom via bicubic upsampling is the naïve baseline and optical zoom serves as the reference ground truth. Our output is artifact-free and preserves detail even for challenging regions such as the high-frequency grillwork.

digital zoom with a standard lens. However, digital zoom simply upsamples a cropped region of the camera sensor input, producing blurry output. It remains a challenge to obtain high-quality images for distant objects without expensive optical equipment.

We propose to improve the quality of super-resolution by starting with real raw sensor data. Recently, single-image super-resolution has progressed with deep models and learned image priors from large-scale datasets [12, 75, 85, 86, 96, 99, 110, 140, 201]. However, these methods are constrained in the following two respects. First, they approach computational zoom under a synthetic setup where the input image is a downsampled version of the high-resolution image, indirectly reducing the noise level in the input. In practice, regions of distant objects often contain more noise as fewer photons enter the aperture during the exposure time. Second, most existing methods start with an 8-bit RGB image that has been processed by the camera's image signal processor (ISP), which trades off high-frequency signal in higher-bit raw sensor data for other objectives (*e.g.* noise reduction).

In this work, we raise the possibility to apply machine learning to computational zoom that uses real raw sensor data as input. The fundamental challenge is obtaining ground truth for this task: low-resolution raw sensor data with corresponding high-resolution im-

ages. One approach is to synthesize sensor data from 8-bit RGB images that are passed through some synthetic noise model [46]. However, noise from a real sensor [166] can be very challenging to model and is not modeled well by any current work that synthesizes sensor data for training. The reason is that sensor noise comes from a variety of sources, exhibiting color cross-talk and effects of micro-geometry and micro-optics close to the sensor surface. We find that while a model trained on synthetic sensor data works better than using 8-bit RGB data (*e.g.* compare (B) and (C) in Figure 3.1), the model trained on real raw sensor data performs best (*e.g.* compare (C) and (D) in Figure 3.1).

To enable learning from real raw sensor data for better computational zoom, we propose to capture real data with a zoom lens [89], where the lens can move physically further from the image sensor to gather photons from a narrower solid angle for optical magnification. We build SR-RAW, the first dataset used for real-world computational zoom. SR-RAW contains ground-truth high-resolution images taken with high optical zoom levels. During training, an 8-bit image taken with a longer focal length serves as the ground truth for the higher-bit (*e.g.* 12-14 bit) raw sensor image taken with a shorter focal length.

During training, SR-RAW brings up a new challenge: the source and target images are not perfectly aligned as they are taken with different camera configurations that cause mild perspective change. Furthermore, preprocessing introduces ambiguity in alignment between low- and high-resolution images. Mildly misaligned input-output image pairs make pixel-wise loss functions unsuitable for training. We thus introduce a novel contextual bilateral loss (CoBi) that is robust to such mild misalignment. CoBi draws inspiration from the recently proposed contextual loss (CX) [116]. A direct application of CX to our task yields strong artifacts because CX doesn't take spatial structure into account. To address this, CoBi prioritizes local features while also allowing for global search when features are not aligned.

In brief, we "Zoom to Learn" – collecting a dataset with ground-truth high-resolution images obtained via optical zoom, to "Learn to Zoom" – training a deep model that achieves better computational zoom. To evaluate our approach, we compare against existing super-resolution methods and also against an identical model to ours, but trained on synthetic sensor data obtained via a standard synthetic sensor approximation. Image quality is measured by distortion metrics such as SSIM, PSNR, and a learned perceptual metric. We also collect human judgments to validate the consistency of the generated images with human perception. Results show that real raw sensor data contains useful image signal for recovering high-fidelity super-resolved images. Our contributions can be summarized as follows [1]:

- We demonstrate the utility of using real high-bit sensor data for computational zoom, rather than processed 8-bit RGB images or synthetic sensor models.
- We introduce a new dataset, SR-RAW, the first dataset for super-resolution from raw data, with optical ground truth. SR-RAW is taken with a zoom lens. Images taken

[1] Project webpage at: `https://ceciliavision.github.io/project-pages/project-zoom.html`

with long focal length serve as optical ground truth for images taken with shorter focal length.

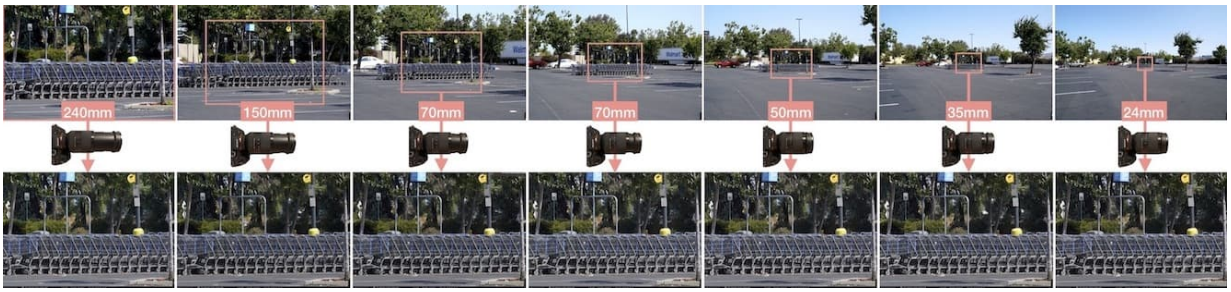
- We propose a novel contextual bilateral loss (CoBi) that handles slightly misaligned image pairs. CoBi considers local contextual similarities with weighted spatial awareness.

### 3.1.2 Background

**Image Super-resolution.** Image super-resolution has advanced from traditional filtering to learning-based methods. The goal is to reconstruct a high-resolution image from a low-resolution RGB image. Traditional approaches include filtering-based techniques such as bicubic upsampling and edge-preserving filtering [105]. These filtering methods usually produce overly smooth texture in the output high-resolution image. Several approaches use patch matching to search for similar patches in a training dataset or in the image itself [39, 47, 69]. Recently, deep neural networks have been applied to super-resolution, trained with a variety of losses [23, 75, 86].

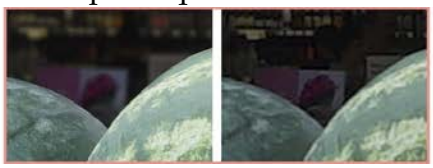
Many recent super-resolution approaches are based on generative adversarial networks. SRGAN [99] is an image super-resolution approach that applies a GAN to generate high-resolution images. The loss used in SRGAN combines a deep feature matching loss and an adversarial loss. Lai *et al*. . [96] propose the Laplacian Pyramid Super-Resolution Network to progressively predict the residual of high-frequency details of a lower-resolution image in a coarse-to-fine image pyramid. Wang *et al*. . [178] propose ESRGAN, which enhances image super-resolution with a Relativistic GAN [76] that estimates how much one image is relatively more realistic than another. Wang *et al*. . [179] study class-conditioned image super-resolution and propose SFT-GAN that is trained with a GAN loss and a perceptual loss. Most existing super-resolution models take a synthetic low-resolution RGB image (usually downsampled from a high-resolution image) as input. In contrast, we obtain real low-resolution images taken with shorter focal lengths and use optically zoomed images as ground truth.

**Image Processing with Raw Data.** Prior works have used raw sensor data to enhance image processing tasks. Farsiua *et al*. . [31] propose a maximum a posteriori technique for joint multi-frame demosaicing and super-resolution estimation with raw sensor data. Gharbi *et al*. . [46] train a deep neural network for joint demosaicing and denoising. Zhou *et al*. . [205] address joint demosaicing, denoising, and super-resolution. These methods use synthetic Bayer mosaics. Similarly, Mildenhall *et al*. . [117] synthesize raw burst sequences for denoising. Chen *et al*. . [13] present a learning-based image processing pipeline for extreme low-light photography using raw sensor data. DeepISP is an end-to-end deep learning model that enhances the traditional camera image signal processing pipeline [145]. Similarly, we operate on raw sensor data and propose a method to




(A) Example sequence from SR-RAW


(B1) Perspective



(B2) Depth-of-field

(B3) Resolution

Figure 3.2: Example sequence from SR-RAW and three sources of misalignment in data capture and preprocessing. The unavoidable misalignment motivates our proposed loss.

super-resolve images by jointly optimizing for the camera image processing pipeline and super-resolution from raw sensor data.

### 3.1.3 Dataset With Optical Zoom Sequences

To enable training with real raw sensor data for computational zoom, we collect a diverse dataset, SR-RAW, that contains raw sensor data and ground-truth high-resolution images taken with a zoom lens at various zoom levels. For data preprocessing, we align the captured images with different zoom levels via field of view (FOV) matching and geometric transformation. The SR-RAW dataset enables training an end-to-end model that jointly performs demosaicing, denoising, and super-resolution on raw sensor data. Training on real sensor data differentiates our framework from existing image super-resolution algorithms that operate on low-bit RGB images.

**Data Capture with a Zoom Lens** We use a 24-240 mm zoom lens to collect pairs of RAW images with different levels of optical zoom. Each pair of images forms an input-output pair for training a model: the short-focal-length raw sensor image is used as input and the long-focal-length RGB image is regarded as the ground-truth for super-resolution. For example, the RGB image taken with a 70mm focal length serves as the 2X zoom ground truth for the raw sensor data taken with a 35mm focal length. In practice, we collect 7 images under 7 optical zoom settings per scene for data collection efficiency. Every pair of images from the 7-image sequence forms a data pair for training a particular zoom model.

In total, we collect 500 sequences in indoor and outdoor scenes. ISO ranges from 100 to 400. One example sequence is shown in Figure 3.2A.

During data capture, camera settings are important. First, depth of field (DOF) changes with focal length and it is not practical to adjust aperture size for each focal length to make DOF identical. We choose a small aperture size (at least f/20) to minimize the DOF difference (still noticeable in Figure 3.2 B2), using a tripod to capture indoor scenes with a long exposure time. Second, we use the same exposure time for all images in a sequence so that noise level is not affected by focal length change. But we still observe noticeable illumination variations due to shutter and physical pupil being mechanical and involving action variation. This color variation is another motivation for us to avoid using pixel-to-pixel losses for training. Third, although perspective does not change with focal length, there exists slight variation (length of the lens) in the center of projection when the lens zooms in and out, generating noticeable perspective change between objects at different depths (Figure 3.2 B1). Sony FE 24-240mm, the lens we use, requires a distance of at least 56.4 meters from the subject to have less than one-pixel perspective shift between objects that are 5 meters apart. Therefore, we avoid capturing very close objects but allow for such perspective shifts in our dataset.

**Data Preprocessing** For a pair of training images, we denote the low-resolution image by RGB-L and its sensor data by RAW-L. For high-resolution ground truth we use RGB-H and RAW-H. We first match the field of view (FOV) between RAW-L and RGB-H. Alignment is then computed between RGB-L and RGB-H to account for slight camera movement caused by manually zooming the camera to adjust focal lengths. We apply a Euclidean motion model that allows image rotation and translation via enhanced correlation coefficient minimization [29]. During training, RAW-L with matched FOV is fed into the network as input; its ground truth target is RGB-H that is aligned and has the same FOV with RAW-L. A scale offset is applied to the image if the optical zoom does not perfectly match the target zoom ratio. For example, an offset of 1.07 is applied to the target image if we use (35mm, 150mm) to train a 4X zoom model.

**Misalignment Analysis** Misalignment is unavoidable during data capture and can hardly be eliminated by the preprocessing step. Since we capture data with different focal lengths, misalignment is inherently caused by the perspective changes as described in Section 3.1.3. Furthermore, when aligning images with different resolutions, sharp edges in the high-resolution image cannot be exactly aligned with blurry edges in the low-resolution image (Figure 3.2 B3). The described misalignment in SR-RAW usually causes 40-80 pixel shifts in an 8-megapixel image pair.


(A) Bicubic (B) Train with CX (C) Train with CoBi (D) Ground truth

Figure 3.3: Training with the contextual loss (CX) results in periodic artifacts as shown on the flat wall in (B). These artifacts are caused by inappropriate feature matching between source and target images, which does not take spatial location into account. In contrast, training with the proposed contextual bilateral loss (CoBi) leads to cleaner and better results, as shown in (C).

### 3.1.4 Contextual Bilateral Loss

When using SR-RAW for training, we find that pixel-to-pixel losses such as $L_1$ and $L_2$ generate blurred images due to misalignment in the training data (Section 3.1.3). On the other hand, the recently proposed Contextual Loss (CX) [116] for unaligned data is also unsatisfactory as it only considers features but not their spatial location in the image. For a brief review, the contextual loss was proposed to train with unaligned data pairs. It treats the source image $P$ as a collection of feature points $p_i{}_{i=1}^{N}$ and the target image $Q$ as a set of feature points $q_j{}_{j=1}^{M}$. For each source image feature $p$, it searches for the nearest neighbor (NN) feature match $q$ such that $q = \arg\min_q \mathbb{D}(p, q_j)_{j=1}^{M}$ under some distance measure $\mathbb{D}(p, q)$. Given input image $P$ and its target $Q$, the contextual loss tries to minimize the summed distance of all matched feature pairs, formulated as

$$\mathrm{CX}(P, Q) = \frac{1}{N} \sum_{i}^{N} \min_{j=1,\dots,M} (\mathbb{D}_{p_i, q_j}). \tag{3.1}$$

We find that training with the contextual loss yields images that suffer from significant artifacts, demonstrated in Figure 3.3. We hypothesize that these artifacts are caused by inaccurate feature matching in the contextual loss. We thus analyze the percentage of features that are matched uniquely (i.e., bijectively). The percentage of target features matched with a unique source feature is only 43.7%, much less than the ideal percentage of 100%.

In order to train our model appropriately, we need to design an image similarity measure applicable to image pairs with mild misalignment. Inspired by the edge-preserving bilateral filter [167], we integrate the spatial pixel coordinates and pixel-level RGB infor-

mation into the image features. Our Contextual Bilateral loss (CoBi) is defined as

$$\mathrm{CoBi}(P, Q) = \frac{1}{N}\sum_{i}^{N} \min_{j=1,\ldots,M} (\mathbb{D}_{p_i,q_j} + w_s \mathbb{D}'_{p_i,q_j}), \tag{3.2}$$

where $\mathbb{D}'_{p_i,q_j} = \|(x_i, y_i) - (x_j, y_j)\|_2$. $(x_i, y_i)$ and $(x_j, y_j)$ are spatial coordinates of features $p_i$ and $q_j$, respectively, and $w_s$ denotes the weight of spatial awareness for nearest neighbor search. $w_s$ enables CoBi to be flexible to the amount of misalignment in the training dataset. The average number of one-to-one feature matches for our model trained with CoBi increases from 43.7% to 93.9%.

We experiment with different feature spaces for CoBi and conclude that a combination of RGB image patches and pre-trained perceptual features leads to the best performance. In particular, we use pretrained VGG-19 features [153] and select 'conv1_2', 'conv2_2', and 'conv3_2' as our deep features, shown to be successful for image synthesis and enhancement [15, 198]. Cosine distance is used to measure feature similarity. Our final loss function is defined as

$$\mathrm{CoBi}_{\mathrm{RGB}}(P, Q, n) + \lambda \mathrm{CoBi}_{\mathrm{VGG}}(P, Q), \tag{3.3}$$

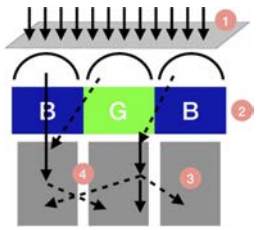
where we use $n \times n$ RGB patches as features for $\mathrm{CoBi}_{\mathrm{RGB}}$, and $n$ should be larger for the 8X zoom (optimal $n = 15$) than the 4X zoom model (optimal $n = 10$).

### 3.1.5 Experimental Setup

We use images from SR-RAW to train a 4X model and an 8X model. We pack each $2 \times 2$ block in the raw Bayer mosaic into 4 channels as input for our model. The packing reduces the spatial resolution of the image by a factor of two in width and height, without any loss of signal. We subtract the black level and then normalize the data to $[0, 1]$. White balance is read from EXIF metadata and applied to the network output as post-processing for comparison against ground truth. We adopt a 16-layer ResNet architecture [61] followed by $\log_2 N + 1$ up-convolution layers where $N$ is the zoom factor.

We split 500 sequences in SR-RAW into training, validation, and test sets with a ratio of 80:10:10, so that there are 400 sequences for training, 50 for validation, and 50 for testing. For a 4X zoom model, we get 3 input-output pairs per sequence for training, and for an 8X zoom model, we get 1 pair per sequence. Each pair contains a full-resolution (8-megapixel) Bayer mosaic image and its corresponding full-resolution optically zoomed RGB image. We randomly crop $64 \times 64$ patches from a full-resolution Bayer mosaic as input for training.

We first compare our approach to existing super-resolution methods that operate on processed RGB images. Then we conduct controlled experiments on our model variants trained on different source data types. All comparisons are tested on the 50 held-out test sequences from SR-RAW.




| Features | Syn | Real |
|---|---|---|
| 1. AA Filter | No | Yes/No |
| 2. Bit Depth | 8 | 12-14 |
| 3. Crosstalk | No | Yes |
| 4. Fill Factor | 100% | <100% |

Table 3.1: A range of sensor characteristics exist in real sensor data, but are not accurately reflected in synthesized sensor data. Each of the features listed in the table corresponds to its numbered label on the illustration, indicating the challenge to model realistic synthetic sensor data.

**Baselines.** We choose a few representative super-resolution (SR) methods for comparisons: SRGAN [99], a GAN-based SR model; SRResnet [99] and LapSRN [96], which demonstrate different network architectures for SR; a model by Johnson *et al.* . [75] that adopts perceptual losses; and finally ESRGAN [178], the winner of the most recent Perceptual SR Challenge PIRM [9].

For all baselines except [75], we use public pretrained models; we first try to fine-tune their models on SR-RAW, adopting the standard setup in the literature: for each image, the input is the downsampled (bicubic) version of the target high-resolution image. However, we notice little difference in average performance ($<\pm 0.04$ for SSIM, $<\pm 0.05$ for PSNR, and $<\pm 0.025$ for LPIPS) in comparison to the pretrained models without fine-tuning, and thus we directly use the models without fine-tuning for comparisons. For baseline methods without pretrained models, we train their models from scratch on SR-RAW.

**Controlled Experiments on Our Model.** For comparison, we also train a copy of our model ("Ours-png") using 8-bit processed RGB images to evaluate the benefits of having real raw sensor data. Different from the synthetic setup described in Section 3.1.5, instead of using downsampled RGB image as input, we use the RGB image taken with a shorter focal length as input. The RGB image taken with a longer focal length serves as the ground truth.

To test whether synthesized raw data can replace real sensor data for training, we adopt the standard sensor synthesis model described by Gharbi et al. [46] to generate synthetic Bayer mosaics from 8-bit RGB images. In brief, we retain one color channel per pixel according to the Bayer mosaic pattern from a white-balanced, gamma-corrected sRGB image, and introduce Gaussian noise with random variance. We train a copy of our model on these synthetic sensor data ("Ours-syn-raw") and test on real sensor data that is white-balanced and gamma-corrected.

### 3.1.6 Results

**Quantitative Evaluation** To quantitatively evaluate the presented approach, we use the standard SSIM and PSNR metrics, as well as the recently proposed learned perceptual metric LPIPS [195], which measures perceptual image similarity using a pretrained deep network. Although there is mild misalignment in the input-output image pairs in SR-RAW (see Section 3.1.3), this misalignment exists across all methods and thus the comparisons are fair.

The results are reported in Table 3.2. They indicate that existing super-resolution models do not perform well on real low-resolution images that require digital zoom in practice. These models are trained under a synthetic setting where input images (usually downsampled) are clean and only contain 8-bit signal. GAN-based methods often generate noisy artifacts and lead to low PSNR and SSIM scores. Bicubic upsampling and SRResnet produce blurry results and get a low score in LPIPS. Our model, trained on high-bit real raw data and supervised by optically zoomed images, can effectively recover high-fidelity visual information with 4X and 8X computational zoom.

In Table 3.3, we show evaluations on our model trained with two different strategies. "Ours-png" is our model trained on processed RGB images. By accessing real low-resolution data taken by a short focal length, the model learns to better handle noise, but its super-resolution power is limited by the low-bit image source. "Ours-syn-raw" is our model trained on synthetic Bayer images. While the model gets access to raw sensor data during test time, it is limited by the domain gap between synthetic and real sensor data. We illustrate in Figure 3.1 that a range of real sensor features are not reflected in a synthetic sensor model. Anti-aliasing filter (AA filter) exists in selected camera models. Synthetic sensor data is generated from 8-bit images while real sensor data contains high-bit signals. Inter-sensor crosstalk and sensor fill factor introduce noise into the color filter array and can be hardly parameterized by a simple noise model [186]. The synthetic sensor model is insufficient to represent these complicated noise patterns.

**Qualitative Results** We show qualitative comparisons in Figure 3.4 against baseline methods, and in Figure 3.5 against our model variants trained with different data. Most input images contain objects that are far from the viewpoint and require computational zoom in practice. Ground truth is obtained using a zoom lens with 4X optical zoom. In Figure 3.4, baseline methods fail to separate contents from the noise; it appears that their performance is limited by only having access to 8-bit signals in color images, especially in "Stripe", which contains high-frequency details. Text in "Parking" appear noisy in all baseline results, while our model generates a clean and discernible output image. In Figure 3.5, the model trained on synthetic sensor data produces jagged edges in "Mario" and "Poster," and demosaic color artifacts in "Pattern." Our model, trained on real sensor data with SR-RAW, can generate a clean demosaiced image with high image fidelity.

| | 4X | | | 8X | | |
|---|---|---|---|---|---|---|
| | SSIM↑ | PSNR↑ | LPIPS↓ | SSIM↑ | PSNR↑ | LPIPS↓ |
| Bicubic | 0.615 | 20.15 | 0.344 | 0.488 | 14.71 | 0.525 |
| SRGAN [99] | 0.384 | 20.31 | 0.260 | 0.393 | 19.23 | 0.395 |
| SRResnet [99] | 0.683 | 23.13 | 0.364 | 0.633 | 19.48 | 0.416 |
| LapSRN [96] | 0.632 | 21.01 | 0.324 | 0.539 | 17.55 | 0.525 |
| Johnson *et al*. . [75] | 0.354 | 18.83 | 0.270 | 0.421 | 18.18 | 0.394 |
| ESRGAN [178] | 0.603 | 22.12 | 0.311 | 0.662 | 20.68 | 0.416 |
| Ours | **0.781** | **26.88** | **0.190** | **0.779** | **24.73** | **0.311** |

Table 3.2: Our model, trained with raw sensor data, performs better computational zoom than baseline methods, as measured by multiple metrics. Note that a lower LPIPS score indicates better image quality.

| | 4X | | | 8X | | |
|---|---|---|---|---|---|---|
| | SSIM↑ | PSNR↑ | LPIPS↓ | SSIM↑ | PSNR↑ | LPIPS↓ |
| Ours-png | 0.589 | 22.34 | 0.305 | 0.638 | 21.21 | 0.584 |
| Ours-syn-raw | 0.677 | 23.98 | 0.231 | 0.643 | 22.02 | 0.473 |
| Ours | **0.781** | **26.88** | **0.190** | **0.779** | **24.73** | **0.311** |

Table 3.3: Controlled experiments on our model, demonstrating the importance of using real sensor data.

**Perceptual Experiments** We also evaluate the perceptual quality of our generated images by conducting a perceptual experiment on Amazon Mechanical Turk. In each task, we compare our model against a baseline on 100 4X-zoomed images (50 test images from SR-RAW and additional 50 images taken without ground truth). We conduct blind randomized A/B testing against LapSRN, Johnson *et al*. ., ESRGAN, and our model trained on synthetic sensor data. We show the participants both results side by side, in random left-right order. The original low-resolution image is also presented for reference. We ask the question: "A and B are two versions of the high-resolution image of the given low-resolution image. Which image (A or B) has better image quality?" In total, 50 workers participated in the experiment. The results, listed in Table 3.4, indicate that our model produces images that are seen as more realistic in a significant majority of blind pairwise comparisons.


Figure 3.4: Our 4x zoom results show better perceptual performance in super-resolving distant objects against baseline methods that are trained under a synthetic setting and applied to processed RGB images. From left to right are: input, ground truth, Johnson *et al*. [75], SRResnet [99], ESRGAN [178], LapSRN [96] and ours.

| | Preference rate |
|---|---|
| Ours>Syn-raw | 80.6% |
| Ours>ESRGAN [178] | 83.4% |
| Ours>LapSRN [96] | 88.5% |
| Ours>Johnson *et al*. . [75] | 92.1% |

Table 3.4: Perceptual experiments show that our results are preferred over baseline methods by a large margin.

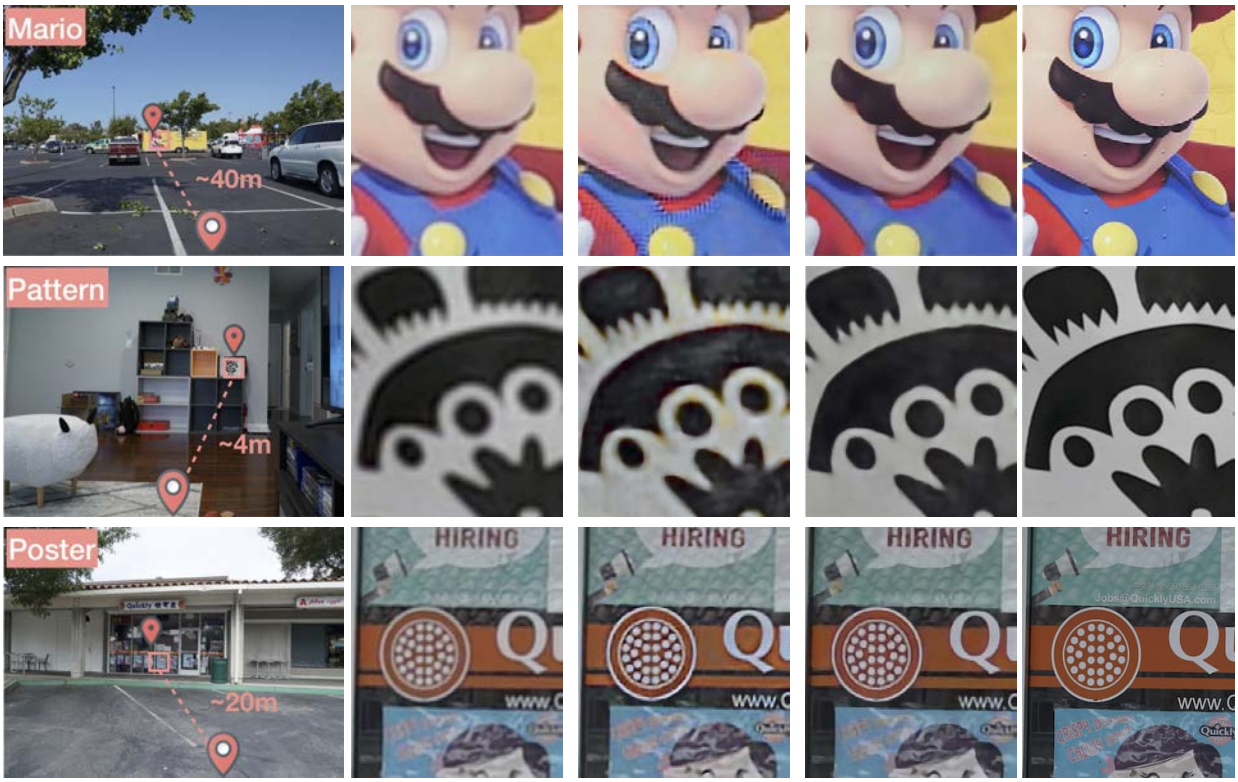



Figure 3.5: The model trained on synthetic sensor data produces artifacts such as jagged edges in "Mario" and "Poster" and color aberrations in "Pattern", while our model, trained on real sensor data, produces clean and high-quality zoomed images.

### 3.1.7 Generalization to Other Sensors

Different image sensors have different structural noise patterns in their Bayer mosaics (See Figure 3.1). Our model, trained on one type of Bayer mosaic, may not perform as well when applied to a Bayer mosaic from another device (*e.g.* iPhoneX). To explore the potential of generalization to other sensors, we capture 50 additional iPhoneX-DSLR data pairs in outdoor environments. We fine-tune our model with only 5000 iterations to adapt our model to the iPhoneX sensor. A qualitative result is shown in Figure 3.6. The results indicate that our pretrained model can be generalized to another sensor by fine-tuning the model on a small dataset captured with that sensor, and also indicate that input-output data pairs can come from different devices, suggesting the application of our method to devices with limited optical zoom power.


Figure 3.6: Our model can adapt to input data from a different sensor after fine-tuning on a small dataset.

### 3.1.8 Discussion

We have demonstrated the effectiveness of using real raw sensor data for computational zoom. Images are directly super-resolved from raw sensor data via a learned deep model that performs joint ISP and super-resolution. Our approach absorbs useful signal from the raw data and produces higher-fidelity results than models trained on processed RGB images or synthetic sensor data. Other downstream applications such as image recognition may similarly benefit from using raw sensor.

While we have shown model generalization capability through lightweight additional data collection and fine tuning, the model performance and data scale are not addressed in this work. Exploring the possibility of learning a generative model (reversely from JPEG to RAW) to produce sensor-specific noise may enable training with internet-scale dataset.

## 3.2 Learning to Remove Reflection From an Image

### 3.2.1 Introduction

Reflection from windows and glasses is ubiquitous in the real world, but it is usually undesirable in photographs. Users often want to extract the hidden clean transmission image by removing reflection from an image. For example, we may have been tempted to take photos through an aquarium glass or skyscraper windows, but reflection can often damage the image quality. Removing reflection from a single image allows us to recover visual content with better perceptibility. Thus, separating the reflection layer and transmission layer from an image — the *reflection separation problem* — is an active research area in computer vision.

Let $I \in \mathbb{R}^{m \times n \times 3}$ be the input image with reflection. $I$ can be approximately modeled as the sum of the transmission layer $T$ and the reflection layer $R$: $I = T + R$. Our goal is to recover the transmission layer $T$ given $I$, which is an ill-posed problem without additional constraints or priors.

As the reflection separation problem is ill-posed, prior works often require additional input images and hard-crafted priors. A line of previous research uses multiple images as input or requires explicit user guidance [56, 156, 185]. Multiple images, however, are not always available in practice, and user guidance is inconvenient and error-prone. Recent researchers proposed methods for reflection removal from a single image [149, 107], but these approaches rely on hand-crafted priors such as ghost cues and relative smoothness which may not generalize to all images with reflection. More recently, CEILNet [30] uses a deep neural network to train a model with low-level losses on color and edges, but this approach does not directly enable the model to learn high-level semantics which can be highly useful for reflection removal. Low-level information is insufficient for reflection separation when there is color ambiguity or the model needs to "recognize" objects in the image. For example, in Figure 3.7, our model trained with perceptual losses may have learned the representations of lamps and faces, and thus correctly removes them from the input image, while CEILNet fails to do so.

In this paper, we present a fully convolutional network with perceptual losses that encode both low-level and high-level image information. Our network takes a single image as input and directly synthesizes two images: the reflection layer and the transmission layer. We further propose a novel exclusion loss that effectively enforces the separation of transmission and reflection at pixel level. To thoroughly evaluate and train different approaches, we build a dataset that contains real-world images and the ground-truth transmission images. Our dataset covers diverse natural environments including indoor and outdoor scenes. We also use this real-world dataset to compare our approach quantitatively to previous methods. In summary, our main contributions are [2]:

[2]Project webpage at: `eecs.berkeley.edu/~cecilia77/project-pages/reflection.html`


Figure 3.7: Results by CEILNet [30] and our approach on real-world images. The top row shows a real image from the CEILNet dataset with a window reflecting a poster of a human face; the bottom row shows an image taken by ourselves, with a lamp as the reflection. From left to right: the input images, CEILNet results and our results. Note that our approach trained to learn both low-level and high-level image statistics successfully removes the reflection layers of the face and lamp, while CEILNet does not.

- We propose to use a deep neural network with perceptual losses for single image reflection separation. We impose perceptual supervision through two losses with different levels of image information: a feature loss from a visual perception network, and an adversarial loss to refine the output transmission layer.
- We propose a carefully designed exclusion loss that emphasizes independence of the layers to be separated in the gradient domain.
- We build a dataset of real-world images for reflection removal with corresponding ground-truth transmission layers. This new dataset enables quantitative evaluation and comparisons among our approach and existing algorithms.
- Our extensive experiments on real data and synthetic data indicate that our method outperforms state-of-the-art methods in SSIM, PSNR, and a perceptual user study on Amazon Mechanical Turk. Our trained model on reflection separation can be directly applied to two other image enhancement tasks, flare removal and dehazing.

(a) Input (b) Without $L_{\text{feat}}$ (c) Without $L_{\text{adv}}$ (d) Without $L_{\text{excl}}$ (e) Complete model

Figure 3.8: Visual comparisons on the three perceptual loss functions, evaluated on a real-world image. In (b), we replace $L_{\text{feat}}$ with image space $L^1$ loss and observed overly-smooth output. (c) shows artifacts of color degradation and noticeable residuals without $L_{\text{adv}}$. In (d), the lack of $L_{\text{excl}}$ makes the predicted transmission have undesired reflection residuals. Our complete model in (e) is able to produce better and cleaner prediction.

### 3.2.2 Related Work

**Multiple-image methods.** As the reflection separation problem is ill-posed, most previous work tackles this problem with multiple input images. These multi-image approaches often use motion cues to separate the transmission and reflection layers [185, 56, 106, 158, 141, 40, 163, 57]. The motion cues are either inferred from calibrated cameras, or motion parallax that assumes the background and reflection objects have greatly different motion fields. Some other multi-image approaches include the use of flash and no-flash image pairs to improve the flash image with reflection removed [2]. Schechner *et al*. [144] use a sequence of images with different focus settings to separate layers with depth estimation. Kong *et al*. [91] exploit physical properties of polarization and use multiple polarized images taken with angular filters to find the optimal separation. More recently, Han and Sim [57] tackle the glass reflection removal problem with multiple glass images, assuming that the gradient field in background image is almost constant while the gradient field in reflection varies much more. Although multiple-image methods have shown promising performance in removing reflection, capturing multiple images is sometimes impossible, for example, these methods can not be applied to existing or legacy photographs.

**Single-image methods.** Another line of work considers using a single image with predefined priors. A widely used prior is the natural image gradient sparsity [102, 101] to find minimum edges and corners for layer decomposition. The gradient sparsity prior is also explored together with optimal and minimum user assistance to better guide the ill-posed separation problem [100, 156]. A recent work by Arvanitopoulos *et al*. [4] uses the gradient sparsity constraint, combined with a data fidelity term in the Laplacian space to suppress reflection. However, all these approaches rely on low-level heuristics and are limited in cases where a high-level understanding of the image is needed.

Another prior for reflection separation is that the reflection layer is often out of focus and appears smooth. This is explicitly formulated into an optimization objective by Li and Brown [107], in which they penalize large reflection gradients. Although the assumption of relative smoothness is valid, their formulation can break down when the reflection layer has high contrast. Wan *et al*. [172] propose a variation of this smoothness prior where depth of field is used as guidance for edge labeling and layer separation. Additionally, Shih *et al*. [149] focus on a subset of the problem where reflection has ghost effects, and use estimated convolution kernel to optimize for reflection removal.

Fan *et al*. [30] recently propose a deep learning network, the Cascaded Edge and Image Learning Network (CEILNet), for reflection removal. They formulate reflection removal as an edge simplification task and learn an intermediate edge map to guide layer separation. CEILNet is trained purely with a low-level loss that combines the differences in color space and gradient domain. The main difference between CEILNet and ours is that they did not explicitly utilize perceptual information during training.

**Benchmark datasets.** A benchmark dataset by Wan *et al*. [171] was proposed recently for reflection removal. The authors collected 1500 real images of 40 scenes in a controlled lab environment by imaging pairs of daily objects and postcards, as well as 100 scenes in natural outdoor environments with three different pieces of glasses. However, the dataset has not been released publicly yet at the time of submission. In order to evaluate among different models quantitatively on real-world images, we collect a dataset of 110 real images with ground truth in natural scene environments.

### 3.2.3 Overview

Given an image $I \in [0,1]^{m\times n\times 3}$ with unwanted reflection, our approach decomposes $I$ into two layers: a transmission layer $f_T(I;\theta)$ and a reflection layer $f_R(I;\theta)$ using a single network $f(I;\theta) = (f_T(I;\theta), f_R(I;\theta))$, where $\theta$ is the network weights. We train the network $f$ on a dataset $\mathcal{D} = \{(I, T, R)\}$ where $I$ is the input image, $T$ is the transmission layer of $I$, and $R$ is the reflection layer of $I$.

Our loss function contains three terms: a feature loss $L_{\text{feat}}$ by comparing the images in feature space, and an adversarial loss $L_{\text{adv}}$ for realistic image refinement, an exclusion loss $L_{\text{excl}}$ that enforces separation of the transmission and reflection layers in the gradient domain. Our overall loss function is

$$L(\theta) = w_1 L_{\text{feat}}(\theta) + w_2 L_{\text{adv}}(\theta) + w_3 L_{\text{excl}}(\theta), \tag{3.4}$$

where we set $w_1 = 0.1$, $w_2 = 0.01$ and $w_3 = 1$ to balance the weight of each term.

An ideal model for reflection separation should be able to understand contents in an image. To train our network $f$ with semantic understanding of the input image, we form hypercolumn features [59] by extracting features from a VGG-19 [154] network pre-trained on the ImageNet dataset [139]. The benefit of using hypercolumn features is that

the input is augmented with useful features that abstract visual perception of a large dataset such as ImageNet. The hypercolumn feature at a given pixel location is a stack of activation units across selected layers of a network at that location. Here, we sampled the layers 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2', and 'conv5_2' in the pre-trained VGG-19 network. The hypercolumn feature has 1472 dimensions in total. We concatenate the input image $I$ with its hypercolumn features as the augmented input for $f$.

Our network $f$ is a fully convolutional network that has a similar network architecture to the context aggregation network [190, 15]. Our network has a large receptive field of $513 \times 513$ to effectively aggregate global image information. The first layer of $f$ is a $1 \times 1$ convolution to reduce feature dimension (1472+3) to 64. The following 8 layers are $3 \times 3$ dilated convolutions. The dilation rate varies from 1 to 128. All the intermediate layers have 64 feature channels. For the last layer we use a linear transformation to synthesize 2 images in the RGB color space.

We evaluate different methods on the publicly available synthetic and real images from the CEILNet dataset[30] and the real-world dataset we collected. We compare our method to the state-of-the-art reflection removal approach CEILNet [30], an optimization based approach [107], and Pix2pix [73], a general framework for image translation.

### 3.2.4 Training a Reflection Removal Model

**Feature loss** We use a feature loss to measure the difference between our predicted transmission layer and the ground-truth transmission in feature space. As the aforementioned observation in Figure 3.7 shows, semantic reasoning about the scene would benefit the task of reflection removal. A feature loss that combines low-level and high-level features from a perception network would serve our purpose. Feature loss has also been successfully applied to other tasks such as image synthesis and style transfer [15, 43, 98, 75].

Here, we compute the feature loss by feeding the predicted image layer and the ground truth through a pre-trained VGG-19 network $\Phi$. We compute the $L^1$ difference between $\Phi(f_T(I;\theta)$ and $\Phi(T)$ in selected feature layers:

$$L_{\text{feat}}(\theta) = \sum_{(I,T)\in\mathcal{D}} \sum_{l} \lambda_l \|\Phi_l(T) - \Phi_l(f_T(I;\theta))\|_1, \tag{3.5}$$

where $\Phi_l$ indicates the layer $l$ in the VGG-19 network. The weights $\{\lambda_l\}$ are used to balance different terms in the loss function. We select the layers 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2', and 'conv5_2' in the VGG-19 network.

**Adversarial loss** During the course of our research, we find that transmission image can suffer from unrealistic color degradation and undesirable subtle residuals without an adversarial loss. We adopted the conditional GAN [73] for our model. Our generator would be $f_T(I;\theta)$. The architecture of our discriminator, denoted as $D$, has 4 layers and 64 feature channels wide. The discriminator tries to discriminate between patches in the real

transmission images and patches given by $f_T(I;\theta)$ conditioned on $I$. The goal is to let the network $D$ learn a suitable loss function for further refining layer separation, and to push the predicted transmission layers toward the domain of real reflection-free images.

Loss for the discriminator $D$ is:

$$\sum_{(I,T)\in\mathcal{D}} \log D(I, f_T(I;\theta)) - \log D(I,T), \tag{3.6}$$

where $D(I,x)$ outputs the probability that $x$ is a natural transmission image given the input image $I$. Then our adversarial loss is:

$$L_{\text{adv}}(\theta) = \sum_{I\in\mathcal{D}} -\log D(I, f_T(I;\theta)). \tag{3.7}$$

We optimize over $-\log D(I, f_T(I;\theta))$ instead of $\log\left(1 - D(I, f_T(I;\theta))\right)$ for better gradient performance [50].

**Exclusion loss** We further propose an exclusion loss in the gradient domain to better separate the reflection and transmission layers. We explore the relationship between the two layers through analysis of the edges in the two layers. Our key observation is that the edges of the transmission and the reflection layers are unlikely to overlap. An edge in $I$ should be caused by either $T$ or $R$, but not both. Thus we minimize the correlation between the predicted transmission and reflection layers in the gradient domain. We formulate the exclusion loss as the product of normalized gradient fields of the two layers at multiple spatial resolutions :

$$L_{\text{excl}}(\theta) = \sum_{I\in\mathcal{D}}\sum_{n=1}^{N} \|\Psi(f_T^{\downarrow n}(I;\theta), f_R^{\downarrow n}(I;\theta))\|_F, \tag{3.8}$$

$$\Psi(T,R) = \tanh(\lambda_T|\nabla T|) \odot \tanh(\lambda_R|\nabla R|), \tag{3.9}$$

where $\lambda_T$ and $\lambda_R$ are normalization factors, $\|\cdot\|_F$ is the Frobenius norm, $\odot$ denotes element-wise multiplication, and $n$ is the image downsampling factor: the images $f_T$ and $f_R$ are downsampled by a factor of $2^{n-1}$ with bilinear interpolation. We set $N = 3$, $\lambda_T = \sqrt{\frac{\|\nabla R\|_F}{\|\nabla T\|_F}}$, and $\lambda_R = \sqrt{\frac{\|\nabla T\|_F}{\|\nabla R\|_F}}$ in our experiments.

Note that the normalization factors $\lambda_T$ and $\lambda_R$ are critical in Equation 3.9, since the transmission and reflection layers may contain unbalanced gradient magnitudes. The reflection layer can be either blurred with low intensity and thus consists of small gradients, or it could reflect very bright light and composes brightest spots in the image, which produces high contrast reflection and thus large gradients. A scale discrepancy between $|\nabla T|$ and $|\nabla R|$ would cause unbalanced updates to the two layer predictions. We observe that without proper normalization factors, the network would suppress the layer with a

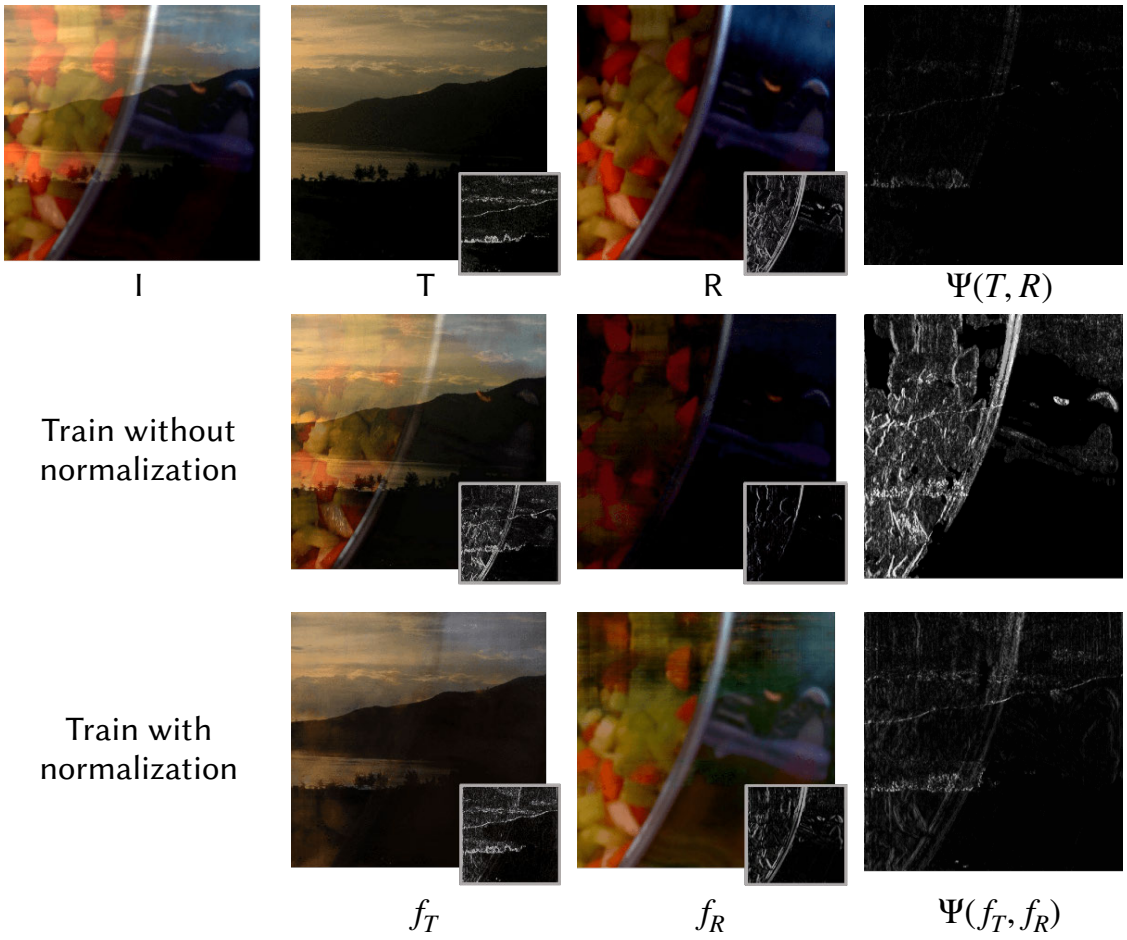



Figure 3.9: Visual comparisons of training with and without gradient normalization. In the middle two columns, the small window at the right bottom corner of each image shows the gradient magnitude of each image. In the rightmost column, $\Psi$ denotes the normalized gradient product formulated in Equation 3.9. The first row left to right shows: input, ground truth transmission $T$, ground truth reflection $R$, and $\Psi$. $\Psi(T, R)$ is close to zeros indicating that the gradient fields of $T$ and $R$ are not correlated. The middle row shows results trained with no normalization in the gradient fields. We observe that the reflection prediction trained without normalization is heavily suppressed. Bottom row shows results trained with gradient normalization with better reflection separation.

smaller gradient update rate to close to zero. A visual comparison of results with and without normalization is shown in Figure 3.9.

$L_{\text{excl}}$ is effective in separating the transmission and reflection layers at the pixel level. If we disable $L_{\text{excl}}$ in our model, some residual reflection may remain visible in the output transmission image, as shown in Figure 3.8 (d).

**Implementation** Given the ground-truth reflection layer $R$, we can further constrain the prediction $f_R(I;\theta)$ with $R$. Reflection layer is usually not in focus and thus blurry. We simply add a $L^1$ loss in color space to constrain $f_R(I;\theta)$:

$$L_R(\theta) = \sum_{(I,R)\in\mathcal{D}} \|f_R(I;\theta) - R\|_1. \tag{3.10}$$

We train the network $f$ by minimizing $(L+L_{\text{R}})$ on synthetic and real data jointly. Note that we disable $L_{\text{R}}$ when training on a real-world image as it is difficult to estimate $R$ precisely. We tried computing $R = I - T$ but $R$ sometimes contains significant artifacts because $I = R + T$ may not hold when $I$ is overexposed.

For the training data, we use 5000 synthetic images and extract 500 image patches from 90 real-world training images with random resolutions between 256p and 480p. To further augment the data, we randomly resize image patches while keeping the original aspect ratio. We train for 250 epochs with batch size 1 on an Nvidia Titan X GPU and weights are updated using the Adam optimizer [87] with a fixed learning rate of $10^{-4}$.

### 3.2.5 Reflection Dataset Collection

**Synthetic data** To create synthetic images with reflection, we choose 5000 random pairs of images from Flickr: one outdoor image and one indoor image for each pair. We use an image (either indoor or outdoor) as the transmission layer and the other image as the reflection layer. We assume the transmission and reflection layers locate on different focal planes so that the two layers exhibit noticeable different blurriness. This is a valid assumption in real-life photography, where the object of interest (e.g. artwork through museum windows) is often in the transmission layer and is set to be in focus. In addition, reflection could be intentionally blurred by shooting with a wide aperture. We use this assumption to create a synthetic dataset, by applying a Gaussian smoothing kernel with a random kernel size in the range of 3 to 17 pixels to the reflection image.

Our image composition approach is similar to the one proposed by Fan *et al.* . [30], but our forward model has the following differences. We remove gamma correction from the images and operate in linear space to better approximate the physical formation of images. Instead of fixing the intensity decay on $R$, we apply variation to the intensity decay since we observe that reflection in real images could have comparable or higher intensity level than the transmission layer. We apply slight vignette centered at random position in

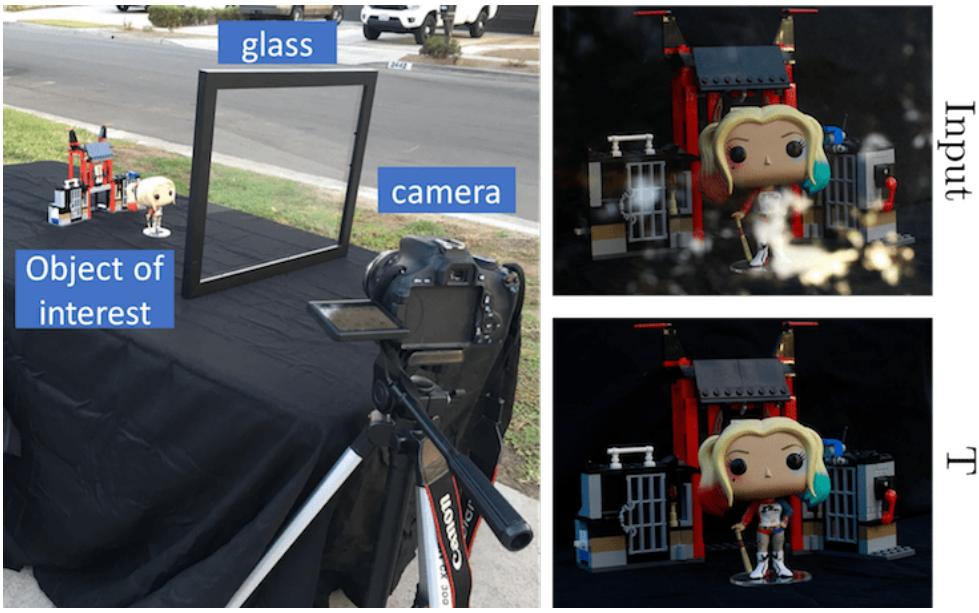



Figure 3.10: Real data collection setup and captured images. We capture two images with and without the glass with same camera settings in a static scene. Right column from top to bottom: captured image with reflection and the ground-truth transmission image $T$.

the reflection layer, which simulates the scenario when camera views the reflection from oblique angles.

**Real data** At the time of developing this work, there is no publicly available benchmark with ground-truth transmission to evaluate different reflection removal approaches on real data. We collected a dataset of 110 real image pairs: image with reflection and its corresponding ground-truth transmission image. The images with reflection were taken with a Canon 600D camera on a tripod with a portable glass in front of the camera. The ground-truth transmission layer was captured when the portable glass was removed. Each image pair was taken with the same exposure setting. Our setup for data capture is shown in Figure 3.10. We captured the dataset with the following considerations:

- Environments: indoor and outdoor
- Lighting conditions: skylight, sunlight, and incandescent
- Camera viewing angles: front view and oblique view
- Camera apertures (affecting the reflection blurriness): $f/2.0$ — $f/16$

| | Synthetic | | Real | |
|---|---|---|---|---|
| Method | SSIM | PSNR | SSIM | PSNR |
| Input | 0.689 | 15.09 | 0.697 | 17.66 |
| Pix2pix [73] | 0.583 | 14.47 | 0.648 | 16.92 |
| Li and Brown [107] | 0.742 | 15.30 | 0.750 | 18.29 |
| CEILNet [30] | 0.826 | 20.47 | 0.762 | 19.04 |
| Ours | **0.853** | **22.63** | **0.821** | **21.30** |

Table 3.5: Quantitative comparison results among our method and 3 other previous methods. We evaluated on synthetic data provided by CEILNet [30], and our real image test set. We also provide a trivial baseline that takes the input image as the result transmission image.

| | Preference rate |
|---|---|
| Ours>CEILNet [30] | 84.2% |
| Ours>Li and Brown [107] | 87.8% |

Table 3.6: The preference rate shows the percentage of comparisons in which users prefer our results.

We split the dataset randomly into a training set and a test set. We extract 500 patches from 90 training images for training and use 20 images for quantitative evaluation.

### 3.2.6 Experiments

**Comparison to prior work** We compare our model to CEILNet [30], the layer separation method by Li and Brown [107], and Pix2pix [73]. We evaluated different methods on the publicly available synthetic images from the CEILNet dataset [30] and the real images from the test set of our real-world dataset.

Our model is only trained on our generated synthetic dataset and the training set of our real-world dataset. For CEILNet, we evaluate its pre-trained model on the CEILNet synthetic images. To evaluate CEILNet on our real data, we fine-tune its model with our real training images (otherwise it performs poorly). We evaluate the approach of Li and Brown [107] with the provided default parameters. Pix2pix is a general image translation model, we train its model on our generated synthetic dataset and the training set of our collected real dataset.

| | Synthetic | | Real | |
|---|---|---|---|---|
| Method | SSIM | PSNR | SSIM | PSNR |
| Ours w/o $L_{\text{feat}}$ | 0.683 | 18.24 | 0.743 | 19.07 |
| Ours w/o $L_{\text{adv}}$ | 0.818 | 20.80 | 0.793 | 21.12 |
| Ours w/o $L_{\text{excl}}$ | 0.796 | 19.58 | 0.802 | 20.22 |
| Ours $L_{\text{adv}}$-only | 0.765 | 18.05 | 0.782 | 19.52 |
| Ours complete | **0.853** | **22.63** | **0.821** | **21.30** |

Table 3.7: Quantitative comparisons on synthetic and real images among multiple ablated models of our method. We remove each of the three losses and evaluate on the re-trained models. 'Ours $L_{\text{adv}}$-only' denotes our method trained with only an adversarial loss. Our complete model shows better performance on both synthetic and real data. We evaluate on synthetic data provided by CEILNet [30], and our real test images described in Section 3.2.5.

The quantitative results are shown in Table 3.5. We compute the PSNR and SSIM between the result transmission images of different methods and ground-truth transmission layer. We demonstrate strong quantitative performance over previous works on both synthetic and real data.

We also conduct a user study on Amazon Mechanical Turk, following the protocol by Chen and Koltun [15]. During the user study, each user is presented with a input real-world image with reflection, our predicted transmission image, and the predicted transmission image by a baseline in the same row. Then the user needs to choose an output image that is closer to the reflection-free version of the input image between the two predicted transmission images. There are 80 real-world images for comparisons from our dataset and the CEILNet dataset. The results are reported in Table 3.6. $84.2\%$ of the comparisons to CEILNet and $87.8\%$ of the comparisons to Li and Brown have our results rated to contain less reflection. The results are statistically significant with $p < 10^{-3}$ and 20 users participate in the user study.

**Qualitative results** We present qualitative results of different methods in Figure 3.11 and Figure 3.12, evaluated on real-world images from our dataset (with ground truth) and from CEILNet [30] (without ground truth), respectively.

**Controlled experiments** To analyze how each loss contributes to the final performance of our network, we remove or replace each loss in the combined objective and re-train the network. A visual comparison is shown in Figure 3.8. When we replace the feature loss $L_{\text{feat}}$ with a $L^1$ loss in color space, the output images tend to be overly-smooth; similar

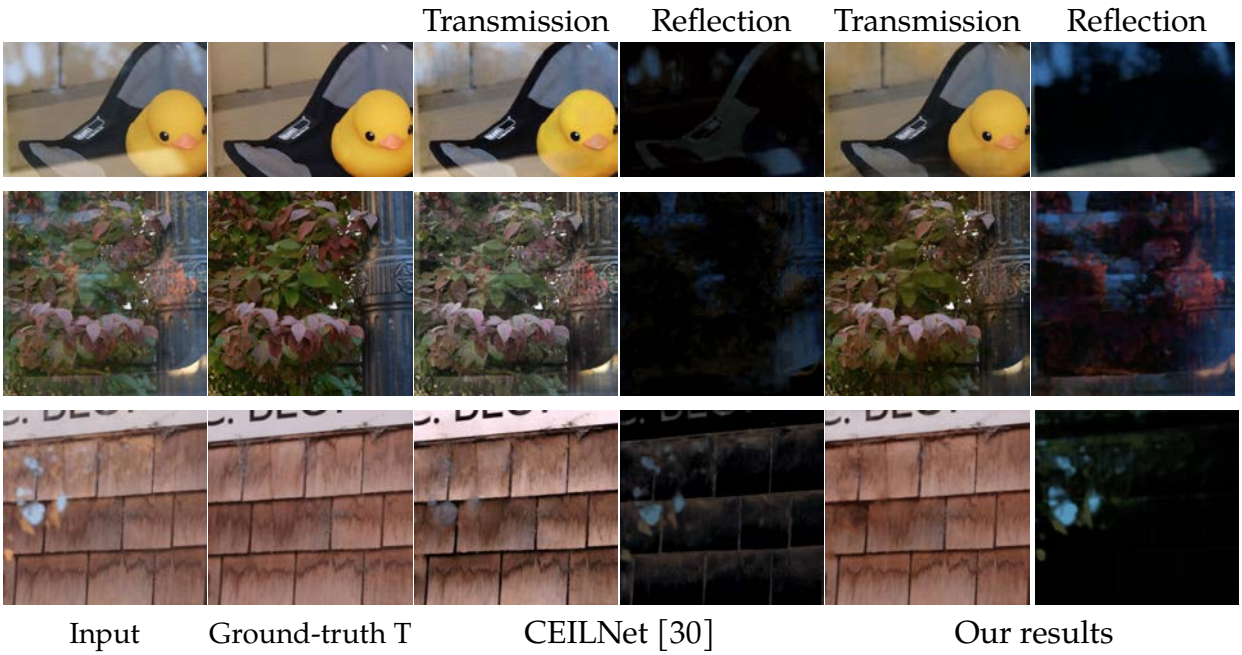



Figure 3.11: Visual results comparison between CEILNet [30] and our method, evaluated on real images from our dataset described in Section 3.2.5. From left to right: input, ground truth transmission layer, CEILNet [30] predictions and our predictions. Notice that our method produces better and cleaner predictions in both the transmission and reflection layers.


Figure 3.12: Qualitative comparisons among CEILNet [30], Li and Brown [107] and our method, evaluated on real images in the CEILNet dataset. Note that even though we have no supervision on the reflection layer for real data, our method predicts cleaner reflection layer thanks to the hybrid training scheme.


Figure 3.13: Extension applications on camera flare removal and image dehazing. For each column, from top to bottom: input, our predicted enhanced layer, our predicted removed layer.

observation is also discussed in [202, 73]. Without $L_{\text{excl}}$, we notice that visible contents of the reflection layer may appear in the transmission prediction. The adversarial refinement loss $L_{\text{adv}}$ helps recover cleaner and more natural results, as shown in (e).

The quantitative results are shown in Table 3.7. We also analyze the performance of the model with only an adversarial loss, which is similar to a conditional GAN [73].

**Extensions** We demonstrate two additional image enhancement applications, flare removal and dehazing, using our trained model to remove an undesired layer. Note that we directly apply our trained reflection removal model without training or fine-tuning on any flare removal or dehazing dataset. These two tasks can be treated as layer separation problems, similar to reflection separation. For flare removal, we aim to remove the optical artifacts of lens flare, which is caused by light reflection and scattering inside the lens. For

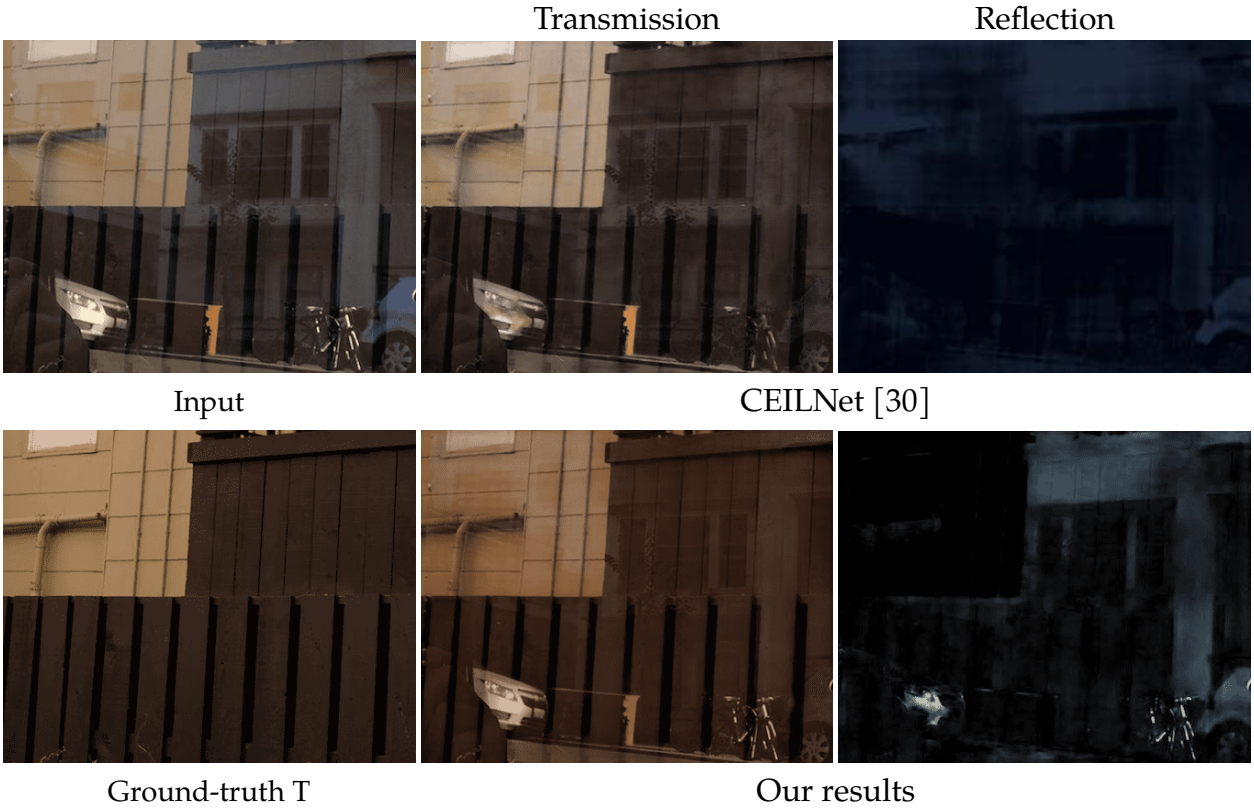



Figure 3.14: A challenging case with sharp reflection. Our method produces better reflection separation results than CEILNet, but is not able to remove reflection completely.

dehazing, we target at removing the hazy layer. The hazy images suffer from contrast loss caused by light scattering, reflection and attenuation of particles in the air. We show the extension results in Figure 3.13. Our trained model can achieve image enhancement by removing undesirable layers from the input images for flare removal and dehazing.

### 3.2.7 Discussion

We presented an end-to-end learning approach for single image reflection separation with perceptual losses and a customized exclusion loss. To decompose an image into the transmission and reflection layers, we found it effective to train a network with combined low-level and high-level image features. In order to evaluate different methods on real data, we collected a new dataset of real-world images for reflection removal that contains ground-truth transmission layers. We additionally extend our approach to two other photo enhancement applications to show generality of our approach for layer separation problems.

Although our reflection separation model outperforms state-of-the-art approaches on

both synthetic and real images, we believe the performance can be further improved. Figure 3.14 illustrates one challenging scenario where the reflection layer is almost as sharp as the transmission layer in a real-world image. Integrating additional sensor data such as depth or a second view such as a stereo pair, or collecting real-world examples with sharp reflections may address this challenge.

# Chapter 4

# Learning Better Shadow and Lighting for Casual Portraits

This chapter talks about using machine learning to enhance casual portrait photographs in the aspects of lighting and shadows, specifically foreign shadow removal, facial shadow softening, and lighting ratio balancing. Motivated by the physical tools used by photographers in studio environments, we demonstrate how Light Stage scans can be used to produce training data for facial shadow softening, and observe the value of in-the-wild images with a shadow synthesis model that accounts for the irregularity of foreign shadows in the real world. We present a mechanism for allowing convolutional neural networks to exploit the inherent bilateral symmetry of human subjects, and demonstrate that this improves the performance of facial shadow softening. Given just a single image of a human subject taken in an unknown and unconstrained environment, our complete system is able to remove unwanted foreign shadows, soften harsh facial shadows, and balance the image's lighting ratio to produce a flattering and realistic portrait image.

## 4.1 Introduction

The aesthetic qualities of a photograph are largely influenced by the interplay between light, shadow, and the subject. By controlling these scene properties, a photographer can alter the mood of an image, direct the viewer's attention, or tell a specific story. Varying the position, size, or intensity of light sources in an environment can affect the perceived texture, albedo, and even three-dimensional shape of the subject. This is especially true in portrait photography, as the human visual system is particularly sensitive to subtle changes in the appearance of human faces. For example, soft lighting (*e.g.* light from a large area light source like an overcast sky) reduces skin texture, which may cause the subject to appear younger. Conversely, harsh lighting (*e.g.* light from a small or distant source like the midday sun) may exaggerate wrinkles and facial hair, making a subject appear older. Similarly, any shadows falling on the subject's face can accentuate its three-

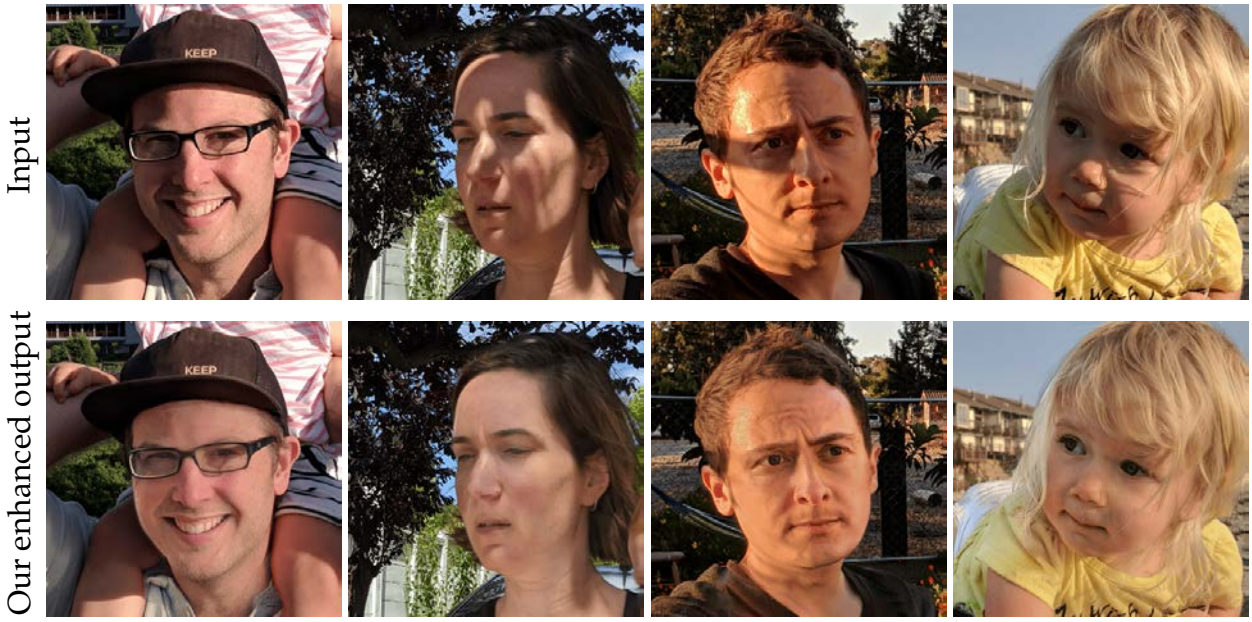



Figure 4.1: The results of our portrait enhancement method on real-world portrait photographs. Casual portrait photographs often suffer from undesirable shadows, particularly foreign shadows cast by external objects, and dark facial shadows cast by the face upon itself under harsh illumination. We propose an automated technique for enhancing these poorly-lit portrait photographs by removing unwanted foreign shadows, reducing harsh facial shadows, and adding synthetic fill lights.

dimensional structure or obfuscate it with distracting intensity edges that are uncorrelated with facial geometry. Other variables such as the lighting angle (the angle at which light strikes the subject) or the lighting ratio (the ratio of illumination intensity between the brightest and darkest portion of a subject's face) can affect the dramatic quality of the resulting photograph, or may even affect some perceived quality of the subject's personality: harsh lighting may look "serious", or lighting from below may make the subject look "sinister".

Unfortunately, though illumination is clearly critical to the appearance of a photograph, finding or creating a good lighting environment outside of a studio is challenging. Professional photographers spend significant amounts of time and effort directly modifying the illumination of existing environments through physical means, such as elaborate lighting kits consisting of scrims (cloth diffusers), reflectors, flashes, and bounce cards [51].

In this work, we attempt to provide some of the control over lighting that professional photographers have in studio environments to casual photographers in unconstrained environments. We present a framework that allows casual photographers to enhance the

quality of light and shadow in portraits from a single image after it has been captured. We target three specific lighting problems common in casual photography and uncontrolled environments:

### 4.1.1 Foreign Shadows

We will refer to any shadow cast on the subject's face by an external occluder (*e.g.* a tree, a hat brim, an adjacent subject in a group shot, the camera itself, etc.) as a *foreign shadow*. Notably, foreign shadows can result in an arbitrary two-dimensional shape in the final photograph, depending on the shape of the occluder and position of the primary, or *key*, light source. Accordingly, they frequently introduce image intensity edges that are uncorrelated with facial geometry and therefore are almost always distracting. Because most professional photographers would remove the occluder or move the subject in these scenarios, we will address this type of shadow by attempting to remove it entirely.

### 4.1.2 Facial Shadows

We will refer to any shadow cast on the face by the face itself (*e.g.* the shadow attached to the nose when lit from the side) as a *facial shadow*. Because facial shadows are generated by the geometry of the subject, these shadows (unlike foreign shadows) can only project to a small space of two-dimensional shapes in the final image. Though they may be aesthetically displeasing, the image intensity edges introduced by facial shadows are more likely than foreign shadows to be a meaningful cue for the shape of the subject. Because facial shadows are almost always present in natural lighting environments (*i.e.*, the environment is not perfectly uniform), we do not attempt to remove them entirely. We instead emulate a photographer's scrim in this scenario, which effectively increases the size of the key light and softens the edges of the shadows it casts.

### 4.1.3 Lighting Ratios

In scenes with very strong key lights (*e.g.* outdoors on a clear day), the ratio between the illumination of the brightest and darkest parts of the face may exceed the dynamic range of the camera, resulting in a portrait with dark shadows or blown out highlights. While this can be an intentional artistic choice, typical portrait compositions target less extreme lighting ratios. Professional photographers balance lighting ratios by placing a secondary, or *fill*, light in the scene opposite the key. We similarly place a virtual fill light to balance the lighting ratio and add definition to the shape of the shadowed portion of the subject's face.

Our framework consists of two machine learning models: one trained for foreign shadow removal, and another trained for handling facial shadow softening and lighting ratio adjustment. This grouping of tasks is motivated by two related observations.

Our first observation, as mentioned above, is tied to the differing relationships between shadow appearance and facial geometry. The appearance of facial shadows in the input image provides a significant cue for shape estimation, and should therefore be useful input when synthesizing an image with softer facial shadowing and a smaller lighting ratio. But foreign shadows are much less informative, and so we first identify and remove all foreign shadows before attempting to perform facial shadow manipulation. This approach provides our facial shadow model with an image in which all shadow-like image content is due to facial shadows, and also happens to be consistent with contemporary theories on how the human visual system perceives shadows [135].

Our second observation relates to training dataset requirements.

Thanks to the unconstrained nature of foreign shadow appearance, it is possible to train our first network with a synthetic dataset: 5000 "in-the-wild" images, augmented with randomly generated foreign shadows for a total of 500K training examples. This strategy is not viable for our second network, as facial shadows must be consistent with the geometry of the subject and so cannot be generated in this way. Constructing an "in-the-wild" dataset consisting entirely of images with controlled facial shadowing is also intractable. We therefore synthesize the training data for this task using one-light-at-a-time (OLAT) scans taken by a Light Stage, an acquisition setup and method proposed to capture reflectance field [19] of human faces. We use the Light Stage scans to synthesize paired harsh/soft images for use as training data. Section 4.3 will discuss our dataset generation procedure in more detail.

Though trained separately, the neural networks used for our two tasks share similar architectures: both are deep convolutional networks for which the input is a $256 \times 256$ resolution RGB image of a subject's face. The output of each network is a per-pixel and per-channel affine transformation consisting of a scaling $A$ and offset $B$, at the same resolution as the input image $I_{in}$ such that the final output $I_{out}$ can be computed as:

$$I_{out} = I_{in} \circ A + B, \tag{4.1}$$

where $\circ$ denotes per-element multiplication. This approach can be thought of as an extension of quotient images [148] and of residual skip connections [62], wherein our network is encouraged to produce output images that resemble scaled and shifted versions of the input image. The facial shadow network includes additional inputs that are concatenated onto the input RGB image: 1) two numbers encoding the desired shadow softening and fill light brightness, so that variable amounts of softening and fill lighting can be specified and 2) an additional rendition of the input image with the face mirrored about its axis of symmetry (*i.e.*, pixels corresponding to the left eye of the input are warped to the position of the right eye, and vice versa). Using a mirrored face image in this way broadens the spatial support of the first layer of our network to include the image region on the opposite side of the subject's face. This allows the network to exploit the bilateral symmetry of human faces and to easily "borrow" pixels with similar semantic meaning and texture but different lighting from the opposite side of the subject's face (see Section 4.4 for details).

In addition to the framework itself, this work presents the following technical contributions[1]:

- Techniques for generating synthetic, real-world, and Light Stage-based datasets for training and evaluating machine learning models targeting foreign shadows, facial shadows, and virtual fill lights.
- Symmetric face image generation for explicitly encoding symmetry cue into training our facial shadow model.
- Ablation studies that demonstrate our data and models achieve portrait enhancement results that outperform all baseline methods in numerical metrics and perceptual quality.

The remainder of the paper is organized as follows. Section 4.2 describes prior work in lighting manipulation, shadow removal, and portrait retouching. Section 4.3 introduces our synthetic dataset generation procedure and our real ground-truth data acquisition. Section 4.5 talks about our network architecture and training procedure. Section 4.6 shows a series of ablation studies and presents qualitative and quantitative results and comparisons. Section 4.7 discusses limitations of our approach.

## 4.2 Related Work

The detection and removal of shadows in images is a central problem within computer vision, as is the closely related problem of separating image content into reflectance and shading [65]. Many graphics-oriented shadow removal solutions rely on manually-labeled "shadowed" or "lit" regions [184, 151, 54, 3]. Once manually identified, shadows can be removed by solving a global optimization technique, such as graph cuts. Because relying on user input limits the applicability of these techniques, fully-automatic shadow detection and manipulation algorithms have also attracted substantial attention. Illumination discontinuity across shadow edges [142] can be used to detect and remove shadows [5]. Formulating shadow enhancement as local tone adjustment and using edge-preserving histogram manipulation [82] enables contrast enhancement on semantically segmented photographs. Relative differences in the material and illumination of paired image segments [55, 112] enables the training of region-based classifiers and the use of graph cuts for labeling and shadow removal. Shadow removal has also been formulated as an entropy minimization problem [35, 34], where invariant chromaticity and intensity images are used to produce a shadow mask that is then re-integrated to form a shadow-free image. These methods assume that shadow regions contain approximately constant reflectance and that image gradients are entirely due to changes in illumination, and are thereby fail when presented with complex spatially-varying textures or soft shadowing. In addition,

[1]Project website: `https://people.eecs.berkeley.edu/~cecilia77/project-pages/portrait.html`

by decomposing the shadow removal problem into two separate stages of detection and manipulation, these methods cannot recover from errors during the shadow detection step [112].

General techniques for inverse rendering [133, 147] and intrinsic image decomposition [53, 6] should, in theory, be useful for shadow removal, as they provide shading and reflectance decompositions of the image. However, in practice these techniques perform poorly when used for shadow removal (as opposed to shading removal) and usually consider cast shadows to be out of scope. For example, the canonical Retinex algorithm [65] assumes that shading variation is smooth and monochromatic and therefore fails catastrophically on simple cases such as shadows cast by the midday sun, which are usually non-smooth and chromatic (sunlit yellow outside the shadow, and sky blue within).

More recently, learning-based approaches have demonstrated a significant improvement on general-purpose shadow detection and manipulation [84, 67, 68, 22, 206, 203, 17]. However, like all learned techniques, such approaches are limited by the nature of their training data. While real-world datasets for general shadow removal are available [132, 173], they do not include human subjects and therefore are unlikely to be useful for our task, which requires the network to reason about specific visual characteristics of faces, such as the skin's subsurface scattering effect [24]. Instead, in this paper, we propose to train a model using synthetic shadows generated on images in the wild. We only use images of faces to encourage the model to learn and use priors on human faces. Earlier work has shown that training models on faces improves performance on face-specific subproblems of common tasks, such as inpainting [189, 168], super-resolution [16] and synthesis [21].

Another problem related to ours is "portrait relighting"—the task of relighting a single image of a human subject according to some desired environment map [160, 204]. These techniques could theoretically be used for our task, as manipulating the facial shadows of a subject is equivalent to re-rendering that subject under a modified environmental illumination map in which the key light has been dilated. However, as we will demonstrate (and was noted in [160]) these techniques struggle when presented with images that contain foreign shadows or high-frequency image structure due to harsh shadows in the input image, which our approach specifically addresses. Example-based portrait lighting transfer techniques [150, 152] also represent potential alternative solutions to this task, but they require a high-quality reference image that exhibits the desired lighting, and that also contains a subject with a similar identity and pose as the input image—an approach that does not scale to casual photos in the wild.

## 4.3 Data Synthesis

There is no tractable data acquisition method to collect a large-scale dataset of human faces for our task with diversity in both the subject and the shadows, as the capture process would be onerous for both the subjects (who must remain perfectly still for impractically

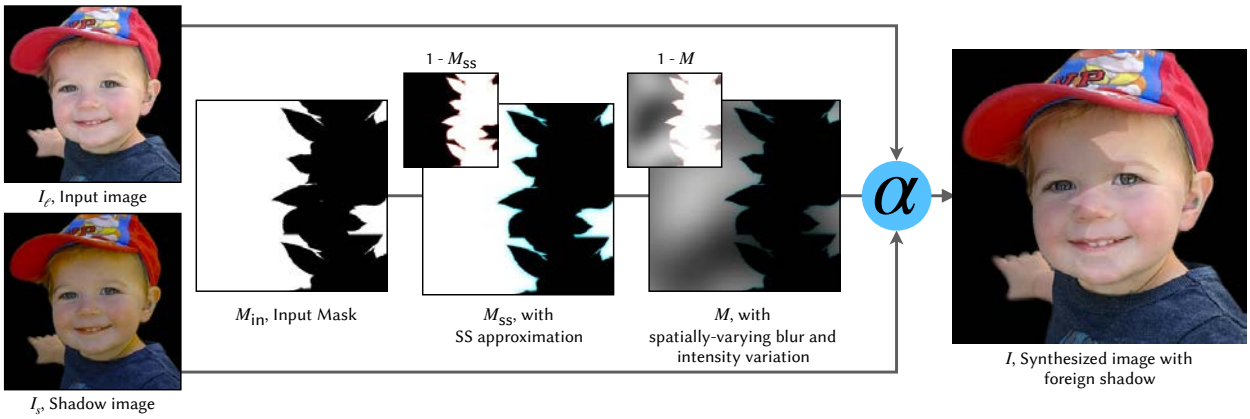



Figure 4.2: The pipeline of our foreign shadow synthesis model (Section 4.3.1). The colors of the "lit" image $I_\ell$ are randomly jittered to generate a "shadow" image $I_s$. The input mask $M_{\texttt{in}}$ shown here is generated from an object silhouette, though it may also be generated with Perlin noise. $M_{\texttt{in}}$ is subjected to a subsurface scattering (SS) approximation of human skin to generate $M_{\texttt{ss}}$, then a spatially-varying blur and per-pixel intensity variation to generate $M$. $I_\ell$ and $I_s$ are then blended according to the shadow mask $M$ to generate a training sample $I$.

long periods of time) and the photographers (who must be specially trained for the task and find thousands of willing participants in thousands of unique environments). Instead, we synthesize custom datasets for our subproblems by augmenting existing datasets—Recall that our two models require fundamentally different training data. Our foreign shadow datasets (Section 4.3.1) are based on images of faces in the wild with rendered shadows, while our facial shadow and fill light datasets (Section 4.3.2) are based on a Light Stage dataset with carefully chosen simulated environments.

### 4.3.1 Foreign Shadows

To synthesize images that appear to contain foreign shadows, we model images as a linear blend between a "lit" image $I_\ell$ and a "shadowed" image $I_s$, according to some shadow mask $M$:

$$I = I_\ell \circ (1 - M) + I_s \circ M \tag{4.2}$$

The lit image $I_\ell$ is assumed to contain the subject lit by all light sources in the scene (*e.g.* the sun and the sky), and the shadowed image $I_s$ is assumed to be the subject lit by everything other than the key (*e.g.* just the sky). The shadow mask $M$ indicates which pixels are shadowed: $M = 1$ if fully shadowed, and $M = 0$ if fully lit. To generate a training sample, we need $I_\ell$, $I_s$, and $M$. $I_\ell$ is selected from an initial dataset described below, $I_s$ is a color transform of $I_\ell$, and $M$ comes from a silhouette dataset or a pseudorandom noise function.

Because deep learning models are highly sensitive to the realism and biases of the data used during training, we take great care to synthesize as accurate a shadow mask and shadowed image as possible with a series of augmentations on $I_s$ and $M$. Figure 4.2 presents an overview of the process and below we enumerate different aspects of our synthesis model and their motivation. In Section 4.6.2, we will demonstrate their efficacy through an ablation study.

**Input Images** Our dataset is based on a set of 5,000 faces in the wild that we manually identified as not containing any foreign shadows. These images are real, in-the-wild JPEG data, and so they are varied in terms of subject, ethnicity, pose, and environment. Common accessories such as hats and scarves are included, but only if they do not cast shadows. We make one notable exception to this policy: glasses. Shadows from glasses are unavoidable and behave more like facial shadows than foreign. Accordingly, shadows from glasses are preserved in our ground truth.

**Light Color Variation** The shadowed image region $I_s$ is illuminated by a lighting environment different from that of the non-shadow region. For example, outdoor shadows are often tinted blue because when the sun is blocked, the blue sky becomes the dominant light source. To account for such illumination differences, we apply a random color jitter, formulated as a $3 \times 3$ color correction matrix, to the lit image $I_\ell$.

**Shape Variation** The shapes of natural shadows are as varied as the shapes of natural objects in the world, but those natural shapes also exhibit significant statistical regularities [70]. To capture both the variety and the regularity of real-world shadows, our distribution of input shadow masks $M_{\texttt{in}}$ is half "regular" real-world masks drawn from a dataset of 300 silhouette images of natural objects, randomly scaled and tiled; and half "irregular" masks generated using a Perlin noise function at 4 octaves with a persistence drawn uniformly at random within $[0, 0.85]$, with the initial amplitude set to 1.0.

**Subsurface Scattering** Light scatters beneath the surface of human skin before exiting, and the degree of that scattering is wavelength-dependent [58, 74, 93]: blood vessels cause red light to scatter further that other wavelengths, causing a visible color fringe at shadows. We approximate the subsurface scattering appearance by uniformly blurring $M_{\texttt{in}}$ with a different kernel per color channel, borrowing from [33]. In brief, the kernel for each channel is a sum of Gaussians $G(\sigma_{c,k})$ with weights $w_{c,k}$, such that each channel $M_c$ of the shadow mask with subsurface scattering $M_{\texttt{ss}}$ is rendered as:

$$M_c = \sum_k M_{\texttt{in}} * G(\sigma_{c,k}) w_{c,k}. \tag{4.3}$$

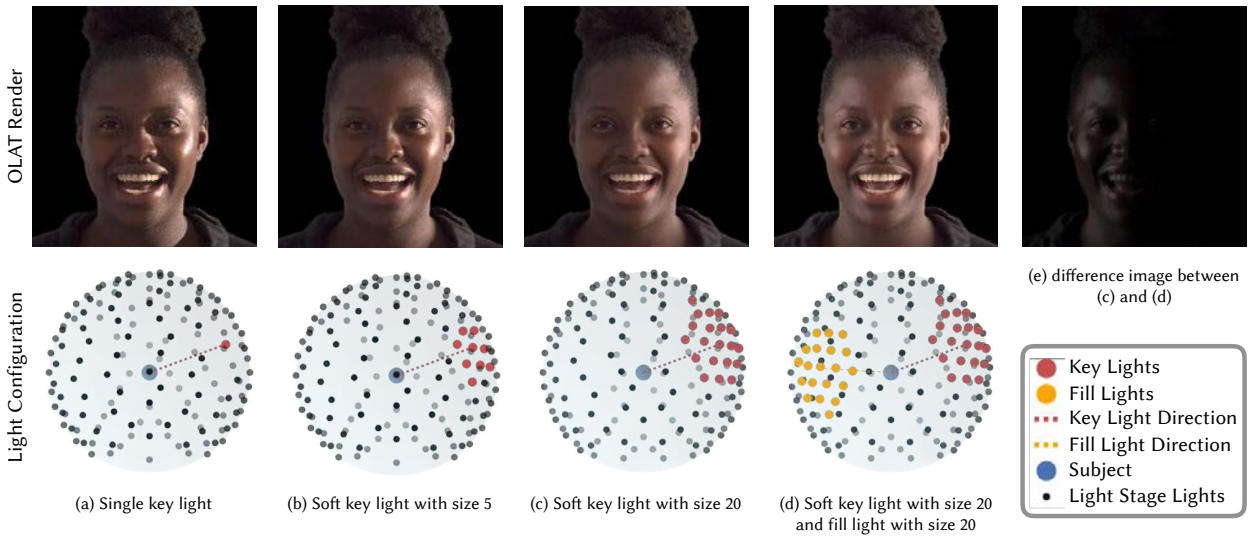



Figure 4.3: Our facial shadow synthesis model. Our input image is a OLAT render corresponding to an environment with a single key light turned on as shown in (a). To soften the shadows by some variable amount, we distribute the key's energy to a variable number of its neighbors, as shown in (b) and (c). We also add a number of fill lights on the opposite side of the Light Stage, to brighten the darker side of the face as shown in (d), with the fill light's difference image visualized in (e). For clarity, only half of the Light Stage's lights are rendered.

**Spatial Variation** The softness of the shadow being cast on a subject depends on the relative distances between the subject, the key light, and the object casting the shadow. Because this relationship varies over the image, our shadow masks incorporate a spatially-varying blur over $M_{\texttt{ss}}$. While many prior works assume that the shadow region has a constant intensity [193], we note that a partially translucent occluder or an environment violating the assumption that lights are infinitely far away will cause shadows to have different local intensities. Accordingly, we similarly apply a spatially-varying per-pixel intensity variation to $M_{\texttt{ss}}$ as well, modeled as Perlin noise at 2 octaves with a persistence drawn uniformly at random from $[0.05, 0.25]$ and an initial amplitude set to 1.0. The final mask with spatial variation incorporated is what we refer to as $M$ above.

### 4.3.2 Facial Shadows

We are not able to use "in-the-wild" images for synthesizing *facial* shadows because the highly accurate facial geometry it would require is generally not captured in such datasets. Instead, we use Light Stage data that can relight a scanned subject with perfect fidelity under any environment and select the simulated environment with care. Note that we *cannot* use light stage data to produce more accurate foreign shadows than we could using

raw, in-the-wild JPEG images, which is why we adopt different data synthesis for these two tasks.

When considering *foreign* shadows, we adopt shadow *removal* with the rationale that foreign shadows are likely undesirable from a photographic perspective and removing them does not affect the apparent authenticity of the photograph (as the occluder is rarely in frame). *Facial* shadows, in contrast, can only be *softened* if we wish to affect the mood of the photograph while remaining faithful to the scene's true lighting direction.

We construct our dataset by emulating the scrims and bounce cards employed by professional photographers. Specifically, we generate harsh/soft facial shadow pairs using OLAT scans from a Light Stage dataset. This is ideal for two reasons: 1) each individual light in the stage is designed to match the angular extent of the sun, so it is capable of generating harsh shadows, and 2) with such a dataset, we can render an image $I$ simulating an arbitrary lighting environment with a simple linear combination of OLAT images $I_i$ with weights $w_i$, *i.e.*, $I = \sum_i I_i w_i$.

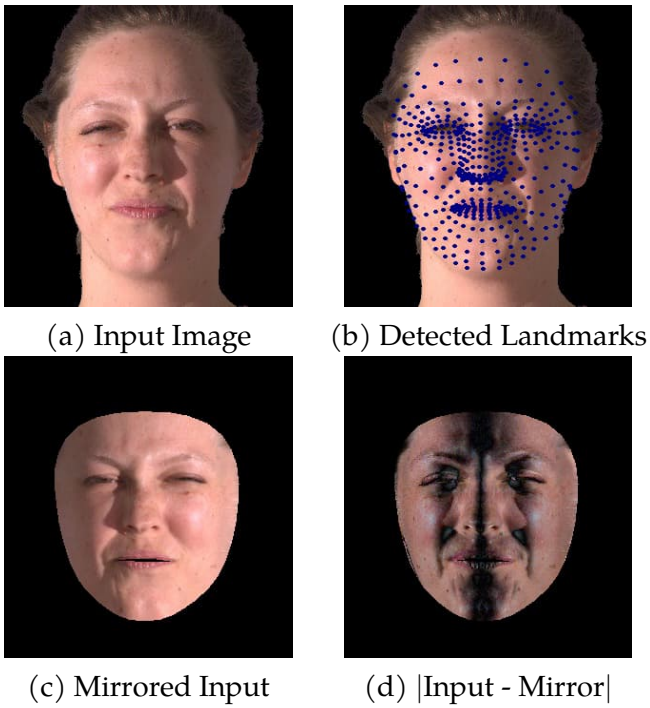
For each training instance, we select one of the $304$ lights in the stage and dub it our key light with index $i_{\texttt{key}}$, and use its location to define the key light direction $\vec{\ell}_{\texttt{key}}$. Our harsh input image is defined to be one corresponding to OLAT weights $w_i = \{P_{\text{key}}$ if $i = i_{\texttt{key}}$, $\epsilon$ otherwise$\}$, where $P_{\text{key}}$ is a randomly sampled intensity of the key light and $\epsilon$ is a small non-zero value that adds ambient light to prevent shadowed pixels from becoming fully black. The corresponding soft image is then rendered by splatting the key light energy to the set of its $m$ nearest neighboring lights $\Omega(\vec{\ell}_{\texttt{key}})$, where $m$ is drawn uniformly from a set of discrete numbers $[5, 10, 20, 30, 40]$. This can be thought of as convolving the key light source with a disc, similar in spirit to a diffuser or softbox. We then compute the location of the fill light (Figure 4.3(d)):

$$\vec{\ell}_{\texttt{fill}} = 2(\vec{\ell}_{\texttt{key}} \cdot \vec{n})\vec{n} - \vec{\ell}_{\texttt{key}}, \tag{4.4}$$

where $\vec{n}$ is the unit vector along the camera $z$-axis, pointing out of the Light Stage. For all data generation, we use a fixed fill light neighborhood size of 20, and a random fill intensity $P_{\text{fill}}$ in $[0, P_{\text{key}}/10]$. Thus, the soft output image is defined as one corresponding to OLAT weights

$$w_i = \begin{cases} P_{\text{key}}/m, & \text{if } i \in \Omega(\vec{\ell}_{\texttt{key}}) \\ P_{\text{fill}}, & \text{if } i \in \Omega(\vec{\ell}_{\texttt{fill}}) \\ \epsilon, & \text{otherwise} \end{cases} \tag{4.5}$$

To train our facial shadow model, we use OLAT images of 85 subjects, each of which was imaged under different expressions and poses, giving us in total 1795 OLAT scans to render our facial harsh shadow dataset. We remove degenerate lights that cause strong flares or at extreme angles that render too dark images, and end up using the remaining 284 lights for each view.



(a) Input Image (b) Detected Landmarks

(c) Mirrored Input (d) |Input - Mirror|

Figure 4.4: The symmetry of human faces is a useful cue for reasoning about lighting: a face's reflectance and geometry is likely symmetric, but the shadow cast upon that face is likely not symmetric. To leverage this, a landmark detection system is applied to the input image (a) and the recovered landmark (b) are used to produce a per-pixel mirrored version of the input image (c). This mirrored image is appended to the input image in our networks, which improves performance by allowing the network to directly reason about asymmetric image content (d) which is likely due to facial and foreign shadows.

## 4.4 Facial Symmetry

Human faces tend to be bilaterally symmetric: the left side of most faces closely resembles the right side in terms of geometry and reflectance, except for the occasional blemish or minor asymmetry. However, images of faces are rarely symmetric because of facial shadows. Therefore, if a neural network can easily reason about the symmetry of image content on the subject's face, it will be able to do a better job of reducing shadows cast upon that face. For this reason, we augment the image that is input to our neural networks with a "mirrored" version of that face, thereby giving the early layers of those networks the ability to straightforwardly reason about which image content is present on the opposite side of the face. Because the subject's face is rarely perfectly vertical and oriented perpendicularly to the camera's viewing direction, it is not sufficient to simply mirror the input image along the $x$-axis. We therefore estimate the geometry of the face and mirror the image using that

estimated geometry, by warping image content near each vertex of a mesh to the location of its corresponding mirrored vertex. See Figure 4.4 for a visualization.

Given an image $I$, we use the landmark detection system of [81] to produce a model of facial geometry consisting of 468 2D vertices (Figure 4.4(b)) and a mesh topology (which is fixed for all instances). For each vertex $j$ we precompute the index of its bilaterally symmetric vertex $\bar{j}$, which corresponds to a vertex $(u_{\bar{j}}, v_{\bar{j}})$ at the same position as $(u_j, v_j)$ but on the opposite side of the face. With this correspondence we could simply produce a "mirrored" version of $I$ by applying a meshwarp to $I$ where the position of each vertex $j$ is moved to the position of its mirror vertex $\bar{j}$. However, a straightforward meshwarp is prone to triangular-shaped artifacts and irregular behavior on foreshortened triangles or inaccurately-estimated keypoint locations. For this reason we instead use a "soft" warping approach based on an adaptive radial basis function (RBF) kernel: For each pixel in $I$ we compute its RBF weight with respect to the 2D locations of all vertices, express that pixel location as a convex combination of all vertex locations, and then interpolate the "mirrored" pixel location by computing the same convex combination of all *mirrored* vertex locations. Put formally, we first compute the Euclidean distance from all pixel locations to all vertex locations:

$$\mathrm{D}_{i,j} = (x_i - u_j)^2 + (y_i - v_j)^2 \tag{4.6}$$

With this we compute a weight matrix consistent of normalized Gaussian distances:

$$\mathrm{W}_{i,j} = \frac{\exp\left(-\mathrm{D}_{i,j}/\sigma_j\right)}{\sum_{j'} \exp\left(-\mathrm{D}_{i,j'}/\sigma_{j'}\right)} \tag{4.7}$$

Unlike a conventional normalized RBF kernel, $\mathrm{W}_{i,j}$ is computed using a different $\sigma$ for each of the $j$ vertices. Each vertex's $\sigma$ is selected such that each landmark's influence in the kernel is inversely proportional to how many nearby neighbors it has for this particular image:

$$\sigma_j = \underset{j'}{\mathrm{select}}\left((u_j - u_{j'})^2 + (v_j - v_{j'})^2, K_\sigma\right) \tag{4.8}$$

Where $\mathrm{select}(\cdot, K)$ returns the $K$'th smallest element of an input vector. This results in a warp where sparse keypoints have significant influence over their local neighborhood, while the influence of densely packed keypoints is diluted. This weight matrix is then used to compute the weighted average of mirrored vertex locations, and this 2D location is used to bilinearly interpolate into the input image to produce it's mirrored equivalent:

$$\bar{I} = I\left(\sum_j \mathrm{W}_{i,j} u_{\bar{j}}\,,\ \sum_j \mathrm{W}_{i,j} v_{\bar{j}}\right) \tag{4.9}$$

The only hyperparameter in this warping model is an integer value $K_\sigma$, which we set to $4$ in all experiments. This proposed warping model is robust to asymmetric expressions and poses assuming the landmarks are accurate, but is sensitive to asymmetric skin features, e.g., birthmarks.

The input to our facial shadow network is the concatenation of the input image $I$ with its mirrored version $\bar{I}$ along the channel dimension. This means that the receptive field of our CNN includes not just the local image neighborhood, but also its mirrored counterpart. Note that we do not include the mirrored image as input to our foreign shadow model, as we found it did not improve results. We suspect that this is due to the unconstrained nature of foreign shadow appearance, which weakens the assumption that corresponding face regions will have different lighting.

## 4.5 Neural Network Architecture and Training

Here we describe the neural network architectures that we use for removing foreign shadows and for softening facial shadows. As the two tasks use different datasets and there is an additional conditional component in the facial shadow softening model, we train these two tasks separately.

For both models, we employ a GridNet [38] architecture with modifications proposed in [122]. GridNet is a grid-like architecture of rows and columns, where each row is a stream that processes features with resolution kept unchanged, and columns connect the streams by downsampling or upsampling the features. By allowing computation to happen at different layers and different spatial scales instead of conflating layers and spatial scales (as U-Nets do) GridNet produces more accurate predictions as has been successfully applied to a number of image synthesis tasks [122, 123]. We use a GridNet with eight columns wherein the first three columns perform downsampling and the remaining five columns perform upsampling, and use five rows for foreign model and six rows for facial model, as we found this to work best after an architecture search.

For all training samples, we run a face detector to obtain a face bounding box, then resize and crop the face into $256 \times 256$ resolution. For the foreign shadow removal model, the input to the network is a 3-channel RGB image and the output of the model is a 3-channel scaling $A$ and a 3-channel offset $B$, which are then applied to the input to produce a 3-channel output image (Equation 4.1). For the facial shadow softening model, we additionally concatenate the input to the network with its mirrored counterpart (as per Section 4.4). As we would like our model to allow for a *variable* degree of shadow softening and of fill lighting intensity, we introduce two "knobs"—one for light size $m$ and the other for fill light intensity $P_{\text{fill}}$, which are assumed to be provided as input. To inject this information into our network, a 2-channel image containing these two values at every pixel is concatenated into both the input and the last layers of the encoders of the network.

We supervise our two models using a weighted combination of pixel-space L1 loss ($\mathcal{L}_{\text{pix}}$) and a perceptual feature space loss ($\mathcal{L}_{\text{feat}}$) which has been used successfully to train models such as image synthesis and image decomposition [15, 197, 199]. Intuitively, the perceptual loss accounts for high-level semantics in the reconstructed image but may be invariant to some non-semantic image content. By additionally minimizing a per-pixel L1 loss our model is better able to recover low-frequency image content. The perceptual

loss is computed by processing the reconstructed and ground truth images through a pre-trained VGG-19 network $\Phi(\cdot)$ and computing the L1 difference between extracted features in selected layers as specified in [197]. The final loss function is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{feat}}(\theta) &= \sum_d \lambda_d \left\| \Phi_d\left(I^*\right) - \Phi_d\left(f\left(I_{in};\theta\right)\right)\right\|_1 \\ \mathcal{L}_{\text{pix}}(\theta) &= \left\| I^* - f(I_{in};\theta)\right\|_1 \\ \mathcal{L}(\theta) &= 0.01 \times \mathcal{L}_{\text{feat}}(\theta) + \mathcal{L}_{\text{pix}}(\theta),\end{aligned} \tag{4.10}$$

where $I^*$ is the ground-truth shadow-removed or shadow-softened RGB image, $f(\cdot;\theta)$ denotes our neural network, and $\lambda_d$ denotes the selected weight for the $d$-th VGG layer. $I_{in} = I$ for foreign removal model and $I_{in} = \text{concat}(I, \bar{I}, P_{\text{fill}}, m)$ for facial shadow softening model. This same loss is used to train both models separately. We minimize $\mathcal{L}$ with respect to both of our model weights $\theta$ using Adam [88] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) for 500K iterations, with a learning rate of $10^{-4}$ that is decayed by a factor of $0.9$ every 50K iterations.

## 4.6 Experiments

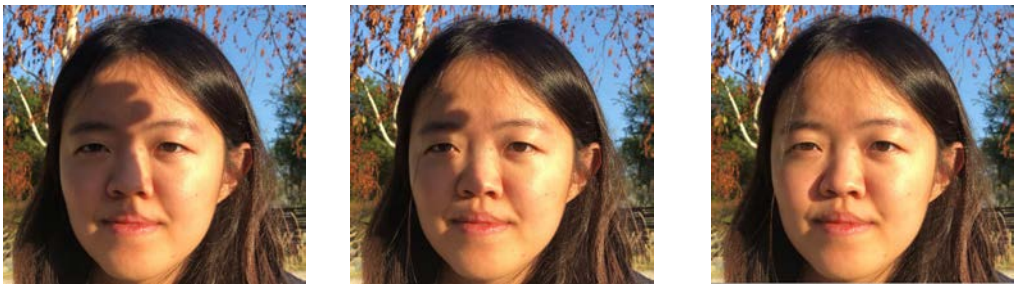
We use synthetic and real in-the-wild test sets to evaluate our foreign shadow removal model (Section 4.6.3) and our facial shadow softening model (Section 4.6.4). We also present an ablation study of the components of our foreign shadow synthesis model (Section 4.6.2) as well as of our facial symmetry modeling.

### 4.6.1 Evaluation Data

We evaluate our foreign shadow removal model with two datasets:

**(1) foreign-syn** We use a held-out set of the same synthetic data generation approach described in (Section 4.3.1), where the images (*i.e.*, subjects) and shadow masks to generate test-set images are not present in the training set.

**(2) foreign-real** We collect an additional dataset of in-the-wild images for which we can obtain ground-truth images that do not contain foreign shadows. This dataset enables the quantitative and qualitative comparison of our proposed model against prior work. This is accomplished by capturing high-framerate ($60$ fps) videos of stationary human subjects while moving a shadow-casting object in front of the subject. We collect this evaluation dataset outdoors, and use the sun as the dominant light source. For each captured video, we manually identify a set of "lit" images and a set of "shadowed" images. For each "shadowed" image, we automatically use homography to align it to each "lit" and find the one that gives the minimum mean pixel error as its counterpart. Because the foreign object is moving during capture, this collection method provides a diversity in the shape and the position of foreign shadows. In total, we capture $20$ videos of $8$ subjects during



(a) Frame 33 (Shadow 1) (b) Frame 36 (Shadow 2) (c) Frame 54 (Shadow-free)

Figure 4.5: An example of the shadow removal evaluation dataset we produce using video captured by a smartphone camera stabilized on a tripod. By filming a stationary subject under the shadow cast by a moving foreign occluder (a-c), we are able to obtain multiple input/ground-truth image pairs of the subject (a, c), (b, c). This provides us with an efficient way to collect a set of diverse foreign shadows for evaluation.

different times of day, which gives us $100$ image pairs of foreign shadow with ground truth.

We evaluate our facial shadow model with another dataset:

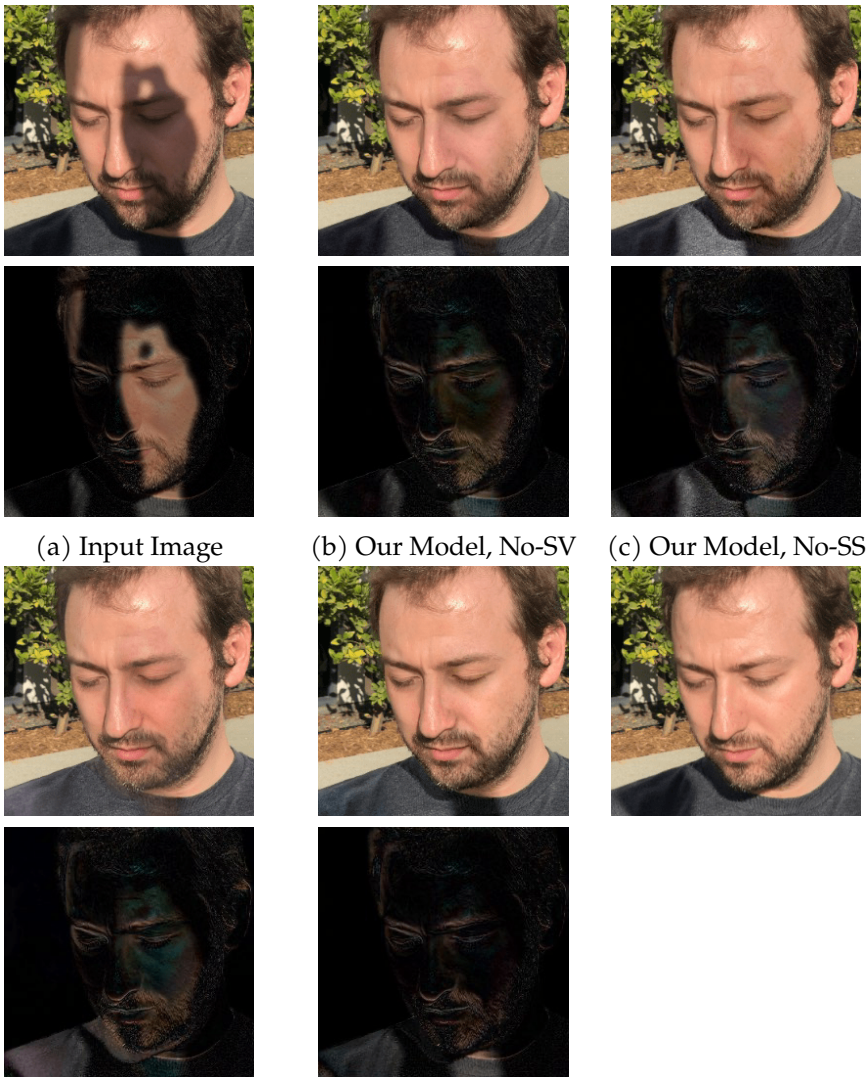
**(3) facial-syn** We use the same OLAT Light Stage data that is used to generate our facial model training data to generate a test set, by using a held-out set of $5$ subjects that are not used during training. We record each harsh input shadow image and the soft ground-truth output image along with their corresponding light size $m$ and fill light intensity $P_{\text{fill}}$ for use. Note that though this dataset is produced through algorithmic means, the ground-truth outputs are a weighted combination of real observed Light Stage images, and are therefore an accurate reflection of the true appearance of the subject up to the sampling limitations of the Light Stage hardware.

We qualitatively evaluate both our foreign shadow removal model and our facial shadow softening model using an additional dataset:

**(4) wild** We collect $100$ "in the wild" portrait images of varied human subjects that contain a mix of different foreign and facial shadows. Images are taken from the Helen dataset [97], the HDRnet dataset [45], and our own captures. These images are processed by our foreign shadow removal model, our facial shadow softening model, or both, to generate enhanced outputs that give a sense of the qualitative properties of both components of our model. See Figures 4.1, 4.8, 4.10 for results.

### 4.6.2 Ablation Study of Foreign Shadow Synthesis

Our foreign shadow synthesis technique (Section 4.3.1) simulates the complicated effect of foreign shadows on the appearance of human subjects. We evaluate this technique by



(a) Input Image (b) Our Model, No-SV (c) Our Model, No-SS

(d) Our Model, No-Color (e) Our Model (f) Ground Truth

Figure 4.6: A visualization of an ablation study of our foreign shadow removal algorithm as different aspects of our foreign shadow synthesis model (Section 4.3.1) are removed. The "No-SV", "No-SS", and "No-Color" ablations show our model trained on synthesized data *without* modeling spatial variation, approximate subsurface scattering, or color perturbation, respectively. The top row shows the images generated by each model, and the bottom row shows the difference between each output and the ground truth image (f). Our complete model (e) clearly outperforms the others. Notice the red-colored residual along the shadow edge in the model trained without approximating subsurface scattering (c), and the color inconsistency in the removed region in the model trained without color perturbation (d). A quantitative evaluation on the entire set `foreign-real` is shown in Table 4.1.

| | Rendered Test Set (`foreign-syn`) | | | Real Test Set (`foreign-real`) | | |
|---|---|---|---|---|---|---|
| Synthesis Model | PSNR↑ | SSIM↑ | LPIPS ↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| Ours, "No-Color" | 26.248 | 0.818 | 0.079 | 21.387 | 0.766 | 0.085 |
| Ours, "No-SV" | 27.546 | 0.830 | 0.058 | 22.095 | 0.782 | 0.081 |
| Ours, "No-SS" | 26.996 | 0.809 | 0.074 | 21.663 | 0.770 | 0.086 |
| Ours | 29.814 | | 0.054 | 23.816 | 0.782 | 0.074 |

Table 4.1: A quantitative ablation study of our foreign shadow removal model in terms of PSNR, SSIM, and LPIPS. Ablating any component of our removal model hurts the performance of the resulting model.

| | Rendered Test Set (`foreign-syn`) | | | Real Test Set (`foreign-real`) | | |
|---|---|---|---|---|---|---|
| Removal Model | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| Input Image | 20.657 | 0.807 | 0.206 | 19.671 | 0.766 | 0.115 |
| Guo *et al*. [55] | 19.170 | 0.699 | 0.359 | 15.939 | 0.593 | 0.269 |
| Hu *et al*. [68] | 20.895 | 0.742 | 0.238 | 18.956 | 0.699 | 0.148 |
| Cun *et al*. [17] | 22.405 | 0.845 | 0.173 | 19.386 | 0.722 | 0.133 |
| Ours | 29.814 | 0.926 | 0.054 | 23.816 | 0.782 | 0.074 |

Table 4.2: A quantitative evaluation of our foreign shadow removal model. We compare against baseline methods of [55], [68] (SRD) and [17] on both synthetic and real test sets. The input image itself is also included as point of reference. In terms of both image-quality (PSNR) and perceptual-quality (SSIM and LPIPS), our model produces better performance on all three metrics with a large margin. Visual comparisons can be seen in Figure 4.7.

removing each of the three components and measuring model performance. Our three ablations are: 1) "No-SV": synthesis without spatially varying blur or the intensity variation of the shadow, 2) "No-SS": synthesis where the approximated subsurface scattering of skin has been removed, and 3) "No-Color": synthesis where the color perturbation to generate the shadow image is not randomly changed. Quantitative results for this ablation study on our `foreign-syn` and `foreign-real` datasets can be found in Table 4.1, and qualitative results for a test image from `foreign-real` are shown in Figure 4.6.

### 4.6.3 Foreign Shadow Removal Quality

Because no prior work appears to address the task of foreign shadow removal for human faces, we compare our model against general-purpose shadow removal methods: the

| Shadow Reduction Model | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| PR-net [160] | 21.639 | 0.709 | 0.152 |
| Ours w/o Symmetry | 24.232 | 0.826 | 0.065 |
| Ours | 26.740 | 0.914 | 0.054 |

Table 4.3: A comparison of our facial shadow reduction model against the PR-net of [160] and an ablation of our model with symmetry. in terms of PSNR, SSIM, and LPIPS on the "`facial-syn`" test dataset. We see that PR-net performs poorly on images that contain harsh facial shadows, and removing the concatenated "mirrored" input during training (*i.e.*, setting $I_{\text{in}} = I$) lowers accuracy by all three metrics.

state-of-the-art learning-based method of [17][2] that uses a generative model to synthesize and then remove shadows, a customized network with attention mechanism designed by [68][3] for shadow detection and removal, and the non-learning-based method of [55] that relies on image segmentation and graph cuts. The original implementation from [55] is not available publicly, so we use a reimplementation[4] that is able to reproduce the results of the original paper. We use the default parameters settings for this code, as we find that tuning its parameters did not improve performance for our task. [68] provide two models trained on two general-purpose shadow removal benchmark datasets (SRD and ISTD), we use the SRD model as it performs better than the ISTD model on our evaluation dataset.

We evaluate these baseline methods on our `foreign-syn` and `foreign-real` datasets, as these both contain ground truth shadow-free images. We compute PSNR, SSIM [180] and a learned perceptual metric LPIPS [194] between the ground truth and the output. Results are shown in Table 4.2 and Figure 4.7. Our model outperforms these baselines by a large margin.

### 4.6.4 Facial Shadow Softening Quality

Transforming harsh facial shadows to soft in image space is roughly equivalent to relighting a face with a blurred version of the dominant light source in the original lighting environment. We compare our facial softening model against the portrait relighting method from [160], by applying a Gaussian blur to the estimated environment map from the model and then pass to the decoder for relighting. The amount of blur to apply, however, cannot map exactly to our light size parameter. We experiment with a few blur kernel values and choose the one that produces the minimum mean pixel error with the ground truth. We do this for each image, and show qualitative comparisons in Figure 4.10. In Table 4.3, we compare our model against the [160] baseline and against an ablation of our

[2]https://github.com/vinthony/ghost-free-shadow-removal
[3]https://github.com/xw-hu/DSC
[4]https://github.com/kittenish/Image-Shadow-Detection-and-Removal

(a) Input (b) Guo[55] (c) Hu[68] (d) Cun[17] (e) Our Model (f) G.T.

Figure 4.7: Foreign shadow removal results on images from our `foreign-real` test dataset. The method of [55] often incorrectly identifies dark image regions as shadows and removes them, while also failing to identify real shadows (b). The deep learning approaches of [17] and [68] (c, d) do a better job of correctly identifying shadow regions but often fail to remove shadows completely, and also change the overall brightness and tone of the image in a way that does not preserve the authenticity of the input image. Our method is able to entirely remove foreign shadows while still preserving the overall appearance of the subject (e), thereby producing output images that more closely resemble the ground truth (f).

model without symmetry, and demonstrate an improvement with respect to both. For all comparisons, we use `facial-syn`, which has ground truth soft facial shadows.

### 4.6.5 Preprocessing for Portrait Relighting

Our method can also be used as a "preprocessing" step for image modification algorithms such as portrait relighting [160, 204], which modify or replace the illumination of the input image. Though often effective, these portrait relighting techniques sometimes produce suboptimal renderings when presented with input images that contain foreign shadows or harsh facial shadows. Our technique can improve a portrait relighting solution: our model can be used to remove these unwanted shadowing effects, producing a rendering that can then be used as input to a portrait relighting solution, resulting in an improved

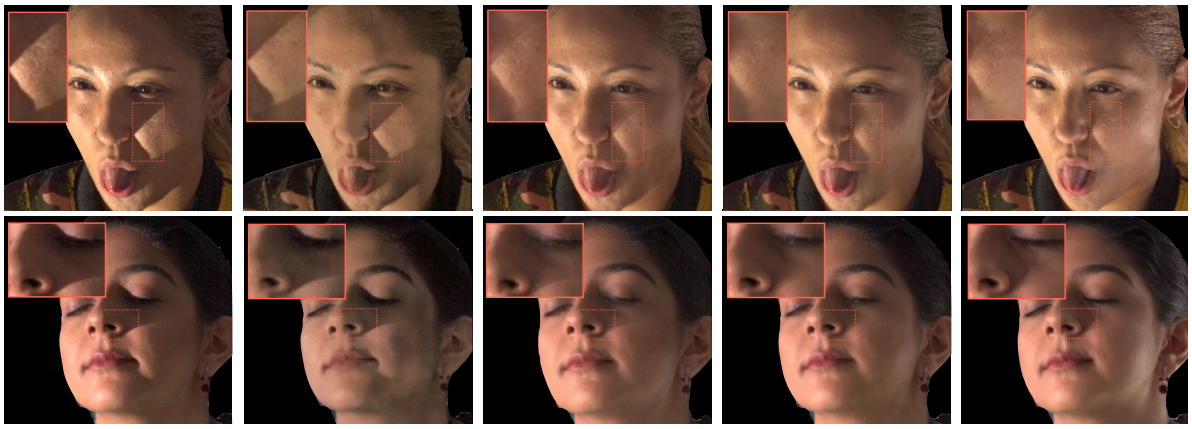
Figure 4.8: Foreign shadow removal results of our model on our `wild` test dataset. Input images that contain unwanted foreign shadows (top) are processed by our foreign shadow removal model (bottom). Though real-world foreign shadows exhibit significant variety in terms of shape, softness, and color, our foreign shadow removal model is able to successfully generalize to these challenging real-world images despite having been trained entirely on our synthetic training data (Section 4.3.1).

final rendering. See Figure 4.11 for an example.

## 4.7 Discussion

Our proposed model is not without its limitations, some of which we can identify in our `wild` dataset. When foreign shadows contain many finely-detailed structures (which are underrepresented in training), our output may retain visible residuals of those (Figure 4.12(a)). While exploiting the bilateral symmetry of the subject significantly improves our facial softening model's performance, this causes our model to sometimes fail to remove shadows that also happen to be bilaterally symmetric (Figure 4.12(b)). Because the training data of our shadow softening model is rendered by increasing the light size—a simple lighting setup that introduces bias towards generating diffused-looking images. For example, when the "light size" is set high in Figure 4.10 (c), our shadow softening model generates images with a "flat" appearance and smooths out high frequency details in the hair regions that could have been preserved if different lighting setups are used for face and hair during training data generation.

Our model assumes that shadows belong to one of two categories ("foreign" and "facial") but these two categories are not always entirely distinct and easily-separated. Because of this, sufficiently harsh facial shadows may be erroneously detected and removed



(a) Input (b) Output from PR-net (c) Output w/o symmetry (d) Output w/ symmetry (e) Ground Truth

Figure 4.9: Facial shadow softening results on `facial-syn`. We compare against the portrait relighting model (PR-net) [160] by applying a blur to its estimated environment light and relighting the input image with that blurred environment map. PR-net is able to successfully soften low frequency shadows but struggles with harsh facial shadows (b). The ablation of our model without our symmetry component (Section 4.4) also underperforms on these harsh facial shadows (c). Our complete model successfully softens these hard shadows (d), as it is able to reason about the bilateral symmetry of the subject and "borrow" pixels with similar reflectance and geometry from the opposite side of the face.

by our foreign shadow removal model (Figure 4.12(c)). This suggests that our model may benefit from a unified approach for both kinds of shadows, though this approach is somewhat at odds with the constraints provided by image formation and our datasets: a unified learning approach would require a unified source of training data, and it is not clear how existing light stage scans or in-the-wild photographs could be used to construct a large, diverse, and photorealistic dataset in which both foreign and facial shadows are present and available as ground-truth.

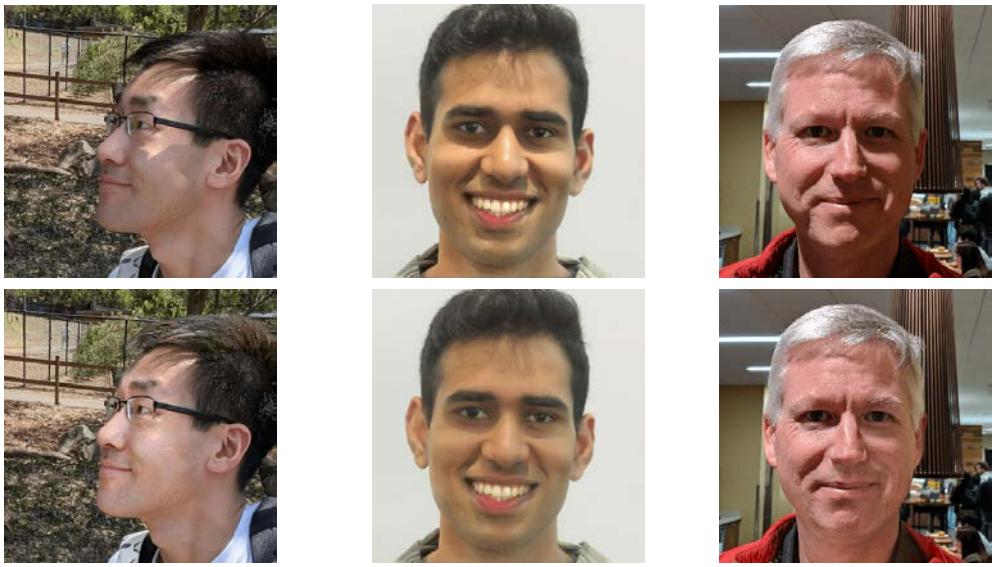
Constructing a real-world dataset for our task that contains ground-truth is challenging. Though the `foreign-real` dataset used for qualitatively evaluating our foreign shadow removal algorithm is sufficiently diverse and accurate to evaluate different algorithms, it has some shortcomings. This dataset is not large enough to be used for training, and does not provide a means for evaluating facial shadow softening. This dataset also assumes that all foreign shadows are cast by a single occluder blocking the light of a single dominant illuminant, while real-world instances of foreign shadows often involve multiple illuminants and occluders. Additionally, to satisfy our single-illuminant assumption, this dataset had to be captured in real-world environments that have one dominant light source (*e.g.* , outdoors in the midday sun). This gave us little control over the lighting environment, and resulted in images with high dynamic ranges and therefore "deep" dark

(a) Input | (b) Output ($m = 10$) | (c) Output ($m = 40$) | (d) Output ($m = 40$, $P_{\text{fill}} = P_{\text{fill}}^{\max}$) | (e) |(c) - (d)|

Figure 4.10: Facial shadow softening results on images from `wild`. Input images may contain harsh facial shadows, such as around the eyes (row 1) and by the subject's cheek (row 3). Applying our facial shadow softening model with a variable "light size" $m$ produces images with softer shadows (b, c). The specular reflection also gets suppressed, which is a desired photographic practice as specular highlights are often distracting and obscuring the surface of the subject. Additionally, the lighting ratio component of our model reduces the contrast induced by facial shadows (d) by adding a synthetic fill light with intensity $P_{\text{fill}}$, set here to the maximum value used in training (Section 4.3.2), in the direction opposite to the detected key light, as visualized in (e).

(a) Input (b) Relighting on (a) (c) Our output (d) Relighting on (c)

Figure 4.11: The portrait relighting technique of [160] provides an alternative approach for shadow manipulation. However, applying this technique to input images that contain foreign shadows and harsh facial shadows (a) often results in relit images in which these foreign and facial shadows persist as artifacts (b). If this same portrait relighting technique is instead applied to the output images of our model (c), it produces a less jarring (though still somewhat suboptimal) rendering of the subject (d).



(a) Fine-detailed shadows. (b) Symmetric facial shadows. (c) Mixed shadows.

Figure 4.12: Example failure cases from our `wild` dataset. We notice limitations of our foreign shadow removal model in handling fine-detailed structures (a), of our facial shadow softening model reducing highly facial shadows (b), and of the models not correctly separating the two types of shadows (c).

shadows, which may degrade (via noise and quantization) image content in shadowed regions. A real-world dataset that addresses these issues be invaluable for evaluating and improving portrait shadow manipulation algorithms.

# Chapter 5

# Learning to Autofocus for Casual Videography

> Always remember your focus determines your reality.
>
> — George Lucas

This chapter introduces the challenge of delivering cinema-like focus in casual videography (*i.e.*, shallow DOF with context-aware focusing). We show that a traditional approach based on physical camera auto-focus is bound to fail, because errors in focus are baked into the video and focusing correctly in real-time requires error-prone guessing about where the action will go. We embrace this insight and take a fundamentally different approach with two parts: first, committing to rendering refocusable video from deep DOF video (RVR sub-system) rather than recording shallow DOF imagery; second, looking at future video frames to make focus decisions at every point in the video (LAAF sub-system) rather than only looking at past frames.

## 5.1 Introduction

Cinematic focus is characterized by the beautiful, shallow depth of field (DOF) of large lenses, which are prized for their ability to visually isolate movie stars, control the viewer's gaze, blur out backgrounds and create gorgeous "bokeh balls" of defocused color. Focus that is tack sharp is essential, but shallow DOF makes it difficult and expensive to achieve. On a movie set, the primary camera assistant ("focus puller") must operate the camera focus controls in realtime to track moving subjects and transition focus according to the screenplay. In movies with improvisational acting, such as Coherence (2014), the cinematographer must try to anticipate what the actors will do; of course significant focus error must be accepted. It is far more common to have a movie script, and for the focus

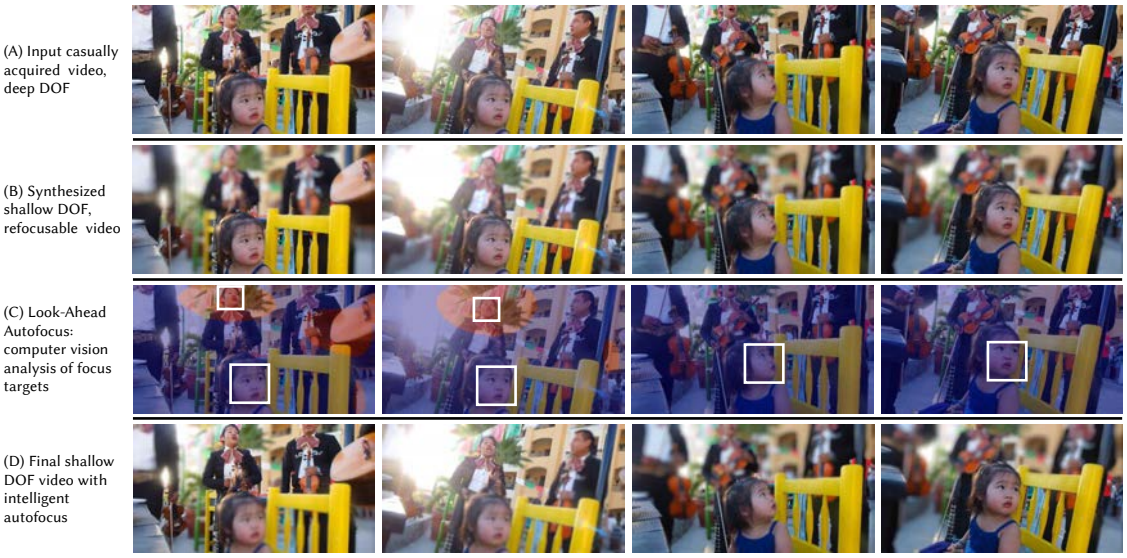



Figure 5.1: We present a new approach to pursue cinema-like focus in casual videography, with shallow depth of field (DOF) and accurate focus that isolates the subject. We start with (A) a deep DOF video shot with a small lens aperture. We use a new combination of machine learning, physically-based rendering, and temporal filtering to synthesize (B) a shallow DOF, refocusable video. We also present a novel Look-Ahead Autofocus (LAAF) framework that uses computer vision to (C) analyze upcoming video frames for focus targets. Here we see face detection (white boxes) and localization of who is speaking/singing [125] (heat map). The result is shallow DOF video (D), where LAAF tracks focus on the singer to start, and transitions focus to the child as the camera pans away from the musicians. The LAAF framework makes future-aware decisions to drive focus tracking and transitions at each frame. This presents a new framework to solve the fundamental realtime limitations of camera-based video autofocus systems.

puller to give actors markers on the ground to indicate where they should stand at specific points to facilitate highly accurate focusing.

These are the issues that make cinematic focus impossible for casual videographers, even though we would love to achieve the aesthetic. Instead, with smartphone videography we sacrifice cinematic DOF, because the small lenses cause essentially everything to be in focus at the same time. In contrast, with cameras that have larger sensors and lenses capable of cinema-like DOF, we inevitably sacrifice focus accuracy. The reason is that there is no movie script, so memorable moments and decisive actions occur unpredictably. Like the focus puller for the Coherence movie, the camera's autofocus system would need a crystal ball to perfectly track each moving subject, or decide to transition focus to a new target in anticipation of its actions taking control of the narrative.

In this paper, we argue that a new direction is necessary if we are to ever truly deliver

cinema-like focus for casual videography. An example of the kind of unprecedented results we seek is a shallow DOF video of a group conversation where the focus transitions perfectly from person to person *before* each person begins talking. Another example is a video that faithfully tracks focus on a rapidly moving soccer player, and then presciently pulls focus onto another player before she heads the ball into the goal. The key to our new approach is to commit to capturing refocusable video, and open the door to analyzing *future* video frames in order to determine whether to enable accurate tracking and anticipatory decisions about whether to transition focus to a decisive action by a new target. To enable these conventionally impossible capabilities, we contribute a framework composed of two modules[1].

- **Refocusable Video Rendering (RVR)** Rather than capturing regular video with static focus, we produce synthetic "refocusable video" where the focus depth of each frame can be computationally changed after capture. Our approach is to synthetically render a shallow DOF video, from a deep DOF video that can be recorded with a smartphone. We build on recent methods in this vein, which are limited to still photography and suffer from disturbing temporal inconsistency when applied frame by frame to videos. We extend synthetic shallow DOF to full video using a combination of machine learning, physically-based rendering and temporal filtering. For machine learning, we contribute a dataset of over 2,000 image pairs or triplets where the aperture and/or focus are varied, and explain how to use these data to train a convolutional neural network that predicts RGBD video and recovers HDR as input to Refocusable Video Rendering (RVR).

- **Look-Ahead Autofocus (LAAF) for Casual Videography** We introduce the notion of Look-Ahead Autofocus that analyzes the seconds of video frames ahead of the current frame in order to decide whether to maintain or transition the focal depth of synthetic focus rendering. We demonstrate LAAF with examples of "AI-assistance" that include: motion and face detection to focus on upcoming human actions, audio localization to focus on who is about to speak, and a machine-learning-based focus detector that shows how a large-scale video dataset can be used to help autofocus of more generic videos. We build an interactive GUI incorporating with subject tracking and automatic focus transition so that the user only makes focus choices on a few keyframes to render a video with shallow DOF and annotated focus.

## 5.2 Prior Art and Related Work

**Camera Autofocus Systems** Camera autofocus systems have generally been classified into two buckets: contrast-detection autofocus (CDAF) and phase-detection (PDAF). CDAF is

[1]Please visit our project website for accompanying videos: `https://ceciliavision.github.io/vid-auto-focus/`

slower, seeking focus by aiming to maximize image contrast as the lens focus is changed; it performs poorly for video because the "focus seeking" behavior is visible in the recorded video. Phase detection can be much faster, and is based on separately detecting and comparing light passing through different parts of the lens aperture. This was achieved in SLR cameras by reflecting light onto PDAF units that each comprised a microlens atop multiple pixels [49].

To enable PDAF in mirrorless camera designs, sensor makers began embedding microscopic PDAF units sparsely into the pixel arrays themselves [36], and advanced to the point that every imaging pixel became a PDAF unit to maximize light and autofocus-sensitive area [118, 90]. This last design is now common in smartphones [36, 104]. One might argue that such advances in physical autofocus systems are asymptotically approaching the fastest possible in many devices today. And yet, autofocus mistakes remain common and inevitable in casual videography, because the focus of each frame is "locked in" as it is being shot. A full solution is impossible because the autofocus algorithm would have to predict the future at every frame to correctly determine what to focus on, or transition focus to. This paper aims to lift this fundamental limitation by synthesizing shallow depth of field as a video postprocess, and using video autofocus algorithms that "look ahead" to make contextually meaningful predictions about what to focus on.

**Light Field Imaging** Another way to capture refocusable images is a light field camera, but the cost of light field video systems remains very high. Rather than capturing a 2D slice of the light field, as is the case with a conventional camera that samples a set of rays that converge at a single point, commercial light field cameras, *e.g.* , Lytro ILLUM[2] capture a 4D slice of the light field, trading spatial resolution for angular resolution. Light field imaging not only captures the set of rays from different viewpoints [103], but also enables physically-accurate synthetic aperture rendering and after-the-fact refocusing [121, 181]. Beyond spatial resolution trade-offs, commercial light field video cameras currently have decreased video frame rate, approximately 3 FPS rather than the desired minimum of 30 FPS. Wang *et al*. [176] propose a hybrid system using one light field camera and one DSLR camera to produce 30 FPS light field video view interpolation.

**Synthetic Defocus** To the best of our knowledge, there has not been works for rendering synthetic defocus for videos. The first step towards RVR is to accomplish single image synthetic defocus, we give a high-level review on synthetic defocus for still images.

Stereo can be used to derive this necessary depth [191, 77]. PDAF sensors described earlier can be used to provide stereo views of the scene through right and left halves of the lens aperture. This has been used to estimate depth for synthesizing defocus blur for smartphone computational photography [170, 104].

Data-driven machine learning approaches have also proven valuable in synthetic defocus tasks using single images, a lot of which are for driving scenarios using specialized

[2] `https://www.dpreview.com/products/lytro/compacts/lytro_illum`

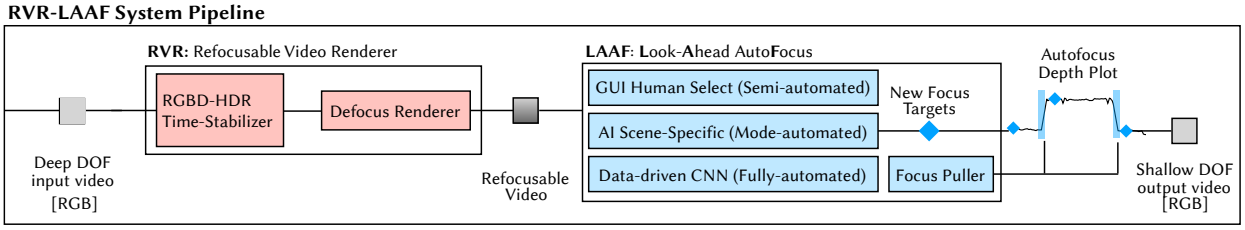



Figure 5.2: Overall system pipeline to compute shallow DOF video from a deep DOF video input. The Refocusable Video Renderer (RVR) contains a time-Stabilized RGBD-HDR (Section 5.4) that computes temporal-stabilized depth and recovers HDR. The Look-Ahead Autofocus (LAAF) (Section 5.5) pipeline consists of three approaches with different levels of automation, detecting New Focus Targets and generating autofocus depth followed by a focus puller that smooths focus transition. Output is a shallow DOF video with contextually-meaningful focus.

datasets [48, 95, 42]. MegaDepth [108] is a concurrent work that targets more generic image contents. Most related to our own work is the method of Srinivasan *et al*. [157], who use weak supervision by predicting a depth map for an input photo, passing it through a differentiable forward rendering model, and then applying a reconstruction loss to the output shallow depth-of-field photo. However, their forward model does not correctly render occlusion or salient bokeh at saturated regions. Furthermore, their method is trained on domain-specific dataset of flowers and indoor buildings and thus performance degrades when applied to other object categories. Park *et al*. [127] combines hand-crafted features with deep features to render refocusable images. However, they focus on noticeable defocus that is generated by medium-large aperture sizes. Therefore their method degrades on images taken by $\sim$f/16 and smaller aperture sizes, while this paper focuses on defocus size generated by f/20 or smaller.

A number of works also use defocus as a cue to predict depth of the scene [115]. Nayar and Nakagawa [120] propose a focus operator which compares texture variability between images to determine relative level of focus. Correspondences from light-field imaging can be used to reduce ambiguity in depth-from-defocus [165]. Suwajanakorn *et al*. [162] recently explored depth-from-defocus for smart-phone imagery and Tang *et al*. [164] generalize depth-from-defocus for unconstrained smartphone imagery in the wild using two perceptibly similar images.

**Video Analysis** Understanding video contents, such as knowing when and where activity happens or which regions are visually salient, is key to our video autofocus algorithm. Recent advances in data-driven machine learning have enabled progress in video understanding tasks such as activity classification [155, 80], activity recognition and detection [161, 175, 32, 8], which benefits LAAF in localizing action in videos. Video saliency

detects salient subjects under a more generic context, often making use of eye tracking signals [111, 177, 159]. We find saliency detection effective in proposing a coarse focus region for LAAF, and it can be combined with other detectors for finer-grained localization. A topic similar to video attention along the temporal axis is video summarization [192, 113], extracting keyframes or subshot as visual summaries for long videos. More recently, audio-visual signals have been combined jointly to learn semantically meaningful video representations, one application we use to detect and locate people who are speaking is audio source localization [125].

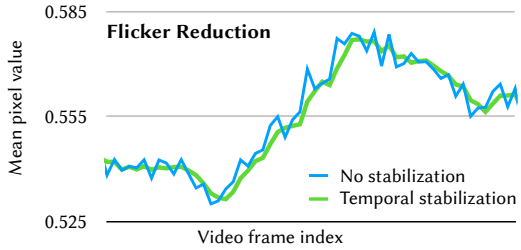
## 5.3 System Overview: RVR-LAAF

Our system (See Figure 5.2) aims to render cinematic autofocus for casual videos. It consists of two components: Refocusable Video Rendering (RVR) (Section 5.4) — rendering videos with shallow DOF focused at any depth at any time, and video Look-ahead Autofocus (LAAF) (Section 5.5) — choosing when and where to focus to make the autofocus choices contextually meaningful and visually appealing.

RVR is built upon a refocusable single frame renderer and a temporal module. We summarize our contribution in rendering refocusable video in Figure 5.3. We find it key to render RVR with temporal coherence, estimation of HDR detail, and a physically-based forward model of lens defocus. We achieve temporal stability by applying an occlusion-aware temporal filtering that is based on optical flow and robust to outliers (see Section 5.4 and Figure 5.3A). For photo-realistic rendering, we train a neural network to jointly estimate, from a single image, the defocus size and unclipped intensity value for each pixel. We find HDR recovery enables rendering of realistic bokeh (Figure 5.3B) and a correct forward model enables correct occlusion effects (Figure 5.3D). To train the network, we collect a large-scale aperture dataset that contains image pairs and triplets. We find our collected triplet dataset to improve estimation around large disparity regions (Figure 5.3C). We call our trained network a RGBD-HDR estimator (Section 5.4). RVR takes a deep DOF video that can be captured by a smartphone, and generates a shallow DOF video that can be focused on any depth at any frame.

RVR delivers video that can be focused at any depth, but the question remains: what is the correct depth to focus on at every frame? For example, retaining optical focus and simply synthesizing shallow DOF (see Figure 5.4C row 2) results in obvious focus errors and often lacks contextual meaning (e.g. focusing on one person while another person, blurred out, is speaking).

Our solution is called the LAAF framework and comprises three complementary techniques for attacking this problem of "when and where to focus". First (Section 5.5.1), LAAF contains a carefully designed user interface (RVR-LAAF GUI) that enables a user to specify only a small number of semantically meaningful "new focus targets" in a video clip – the system then tracks these subjects to maintain focus on them, and adds focus-pull transitions automatically. Second (Section 5.5.2), LAAF provides AI-based autofocus





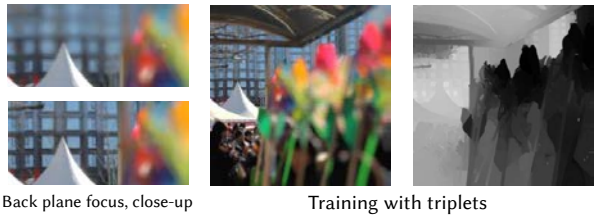
Input video frames


No stabilization, close-up

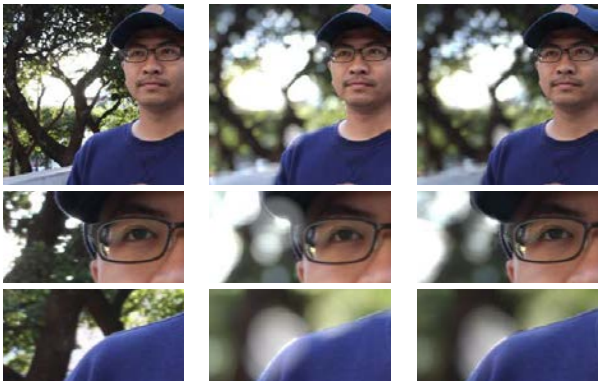

Temporal stabilization, close-up

(3A) Flicker reduction by temporal filtering. Flicker reduction is best appreciated in the accompanying video. As a proxy, we plot the mean pixel value around saturated pixels, which we find is correlated with flicker level in video. Note that temporal filtering greatly reduces the high-frequency fluctuations. The sample video frames illustrate typical levels of stabilization delivered by the temporal filtering. Note the high fluctuation in focus on the woman's face before filtering.


Input Training without triplets


Back plane focus, close-up Training with triplets

(3C) Effect of photo triplets in training. We compare the disparity map and shallow DOF rendering using networks trained w/ and w/o triplet consistency. The results with triplet consistency are geometrically more accurate in the background region of high disparity.

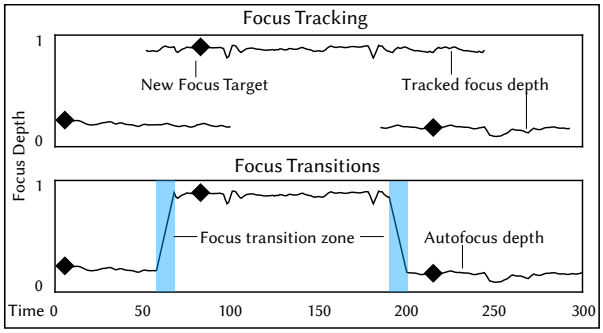

Input Srinivasan [2018] Our rendering

(3D) Comparison of synthetic shallow DOF rendering models. The in-focus regions such as the person's shoulder and hat should occlude the background defocus. Our forward model, see Section 4.1 correctly renders such occlusions while the model that uses weighted layer summation [Srinivasan et al. 2018] does not.

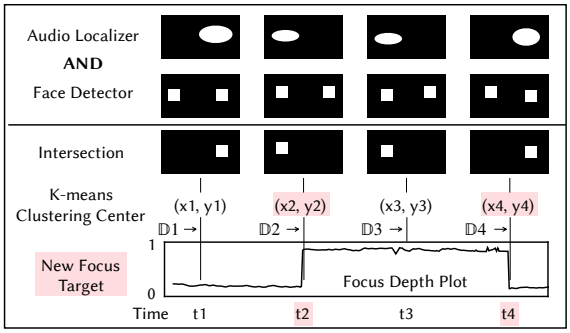

(a) Input f/22 (b) SDOF w/o HDR Recovery (c) SDOF w/ HDR Recovery

(3B) Shallow DOF Rendering with HDR Recovery. Visually salient bokeh that appears at saturated regions in (a) is suppressed in (b) without recovering a higher dynamic range. Predicted HDR recovery map enables more photo-realistic shallow DOF rendering. Rendered video results demonstrate more visually prominent differences and can be seen in the accompanying video.

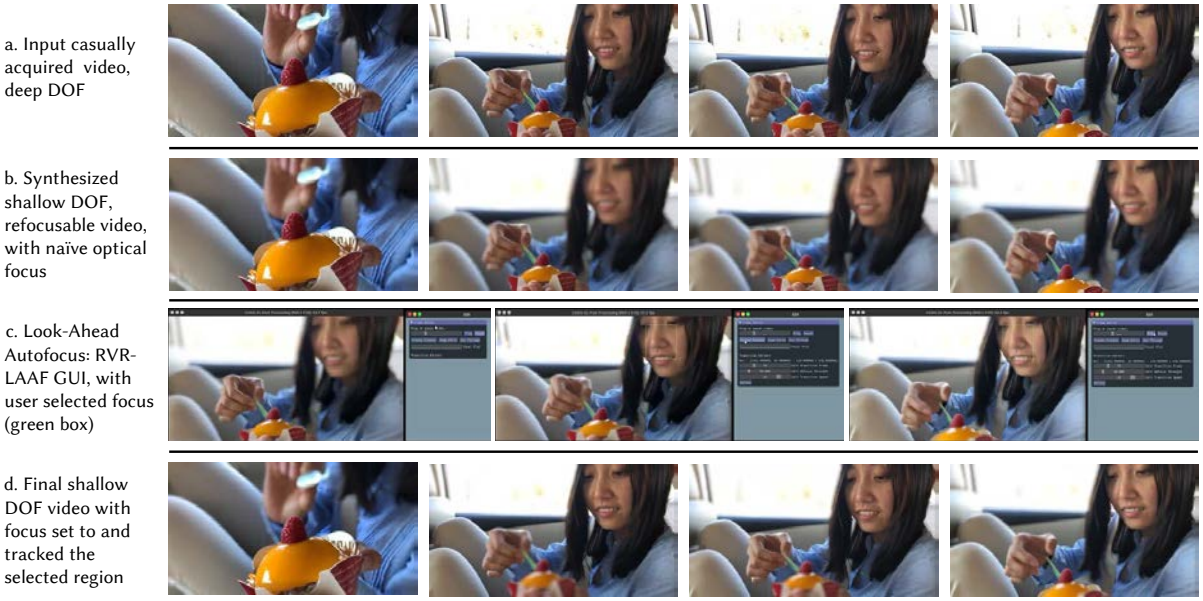
Figure 5.3: Summarized contribution of proposed Refocusable Video Renderer.




(4A) Illustration of synthetic focus pulling. Input to the module is a set of {(x, y, t)} triplets denoting new focus targets and times. Three such focus targets are shown as diamonds on the timelines above. For each of these targets, we perform focus tracking (top graph) by computing (x, y) tracking across time, and look up the focus depth from the estimated depth map. Next, we execute fast focus pull transitions (bottom graph) from one target to the next, with focus arriving at each target slightly before t to allow the viewer to visually settle before action begins.



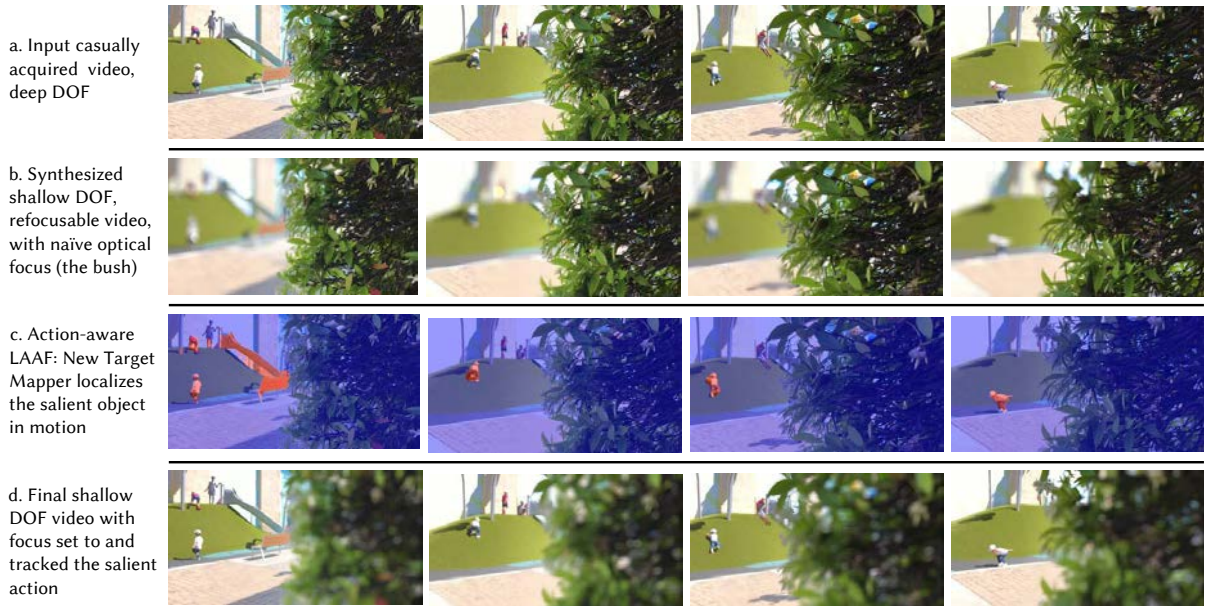

(4B) Illustration of computing New Focus Targets from scene-specific LAAF. For a given scene (eg. conversation), a set of detectors generate focus probability maps and their intersection is computed. K-means clustering finds the center of the majority cluster and generates its center (x, y), which reads from predicted disparity map to produce a focus depth plot. We apply 1-D bilateral filter to the plot and detect focus depth discontinuity to be the New Focus Targets, denoted as (x, y, t).




(4C) GUI-based video autofocus. (Row 1) Input video shows the camera shifts from the mango tart to the person. (Row 2) naïve synthetic SDOF video would not be able to resolve focus in time, blurring out the person's face. (Row 3) Our RVR-LAAF GUI is incorporated with vision-based tracking. (High-res GUI demo can be found in the accompanying video) User only needs to annotate New Focus Target (Row 3 Middle) when they identify a preferred focus subject. The GUI tracks the subject (Row 3 Right) until the next New Focus Target is selected. (Row 4) The resulting video shows autofocus that transitions correctly to the face without lagging, as the user knows camera is shift ing to a more preferred focus subject. The video result can be found in the accompanying video.

Figure 5.4: Summarized contribution of proposed Look-Ahead AutoFocus, Part I.


(4D) Audio-aware video autofocus. (Row 1) Input video shows two people discussing a bug on the ground. (Row 2) naïve synthetic SDOF video would focus on the foreground person exclusively, blurring out the speaker in the background. (Row 1 sub) Audio localization mapping, a recent advance in computer vision, identifies the active speaker at each point in the video. (Row 4) The resulting video shows autofocus that transitions smoothly from speaker to speaker, before they begin speaking. The video result can be found in the accompanying video.




(4E) Action-aware video auto-focus. (Row 1) Input video shows children playing in the background and plant in foreground, optically focused on foreground. (Row 2) Naïve synthetic SDOF video would blur out the action in the background, because the input video is optically focused on the foreground. (Row 3) Intersection of visual saliency mapping and motion detection mapping isolates action of children in the background. (Row 4) The resulting video is autofocused on the background.

Figure 5.5: Summarized contribution of proposed Look-Ahead AutoFocus, Part II.

modules that only requires the user to choose the type of scene – then, the system fully automates the task by choosing "new focus targets" intelligently. We demonstrate examples for scenes that contain conversations or actions, where the system automatically pre-focuses on each person before they speak or on actions before they occur. Third (Section 5.5.4), (Section 5.3), we demonstrate a first attempt at fully automating video autofocus using a machine learning approach – we contribute ground-truth focus annotations on a large-scale video dataset, using our RVR-LAAF GUI to create this sizable dataset efficiently.

## 5.4 Refocusable Video Rendering (RVR)

In this paper we introduce the first synthetic defocus renderer for videos. We take Monocular Depth Estimation (MDE) [157] as our launching point, and comprehensively re-work the method to make it suitable for synthesizing defocus for casual videography. This section presents the implementation details. The four critical changes are:

- Adding HDR recovery estimation, which we show is critical to achieving plausible bokeh balls that are a visual hallmark of shallow DOF videos (Figure 5.3B, Section 5.4)
- Building a superior training dataset with far greater scene diversity. This dataset contains novel "triplets" of images (Figure 5.6) that we find is important to improve depth estimation in foreground and background regions (Figure 5.3C Section 5.4).
- Correcting the forward model to make it handle occlusions correctly and so that background defocus does not incorrectly "bleed" around foreground objects (Figure 5.3D, Section 5.4)
- Adding temporal coherence to ameliorate flicker (Figure 5.3A, Section 5.4)

**RGBD-HDR Estimator** We offer an indirectly-supervised approach to estimate disparity and HDR with only aperture supervision, by training a neural network that jointly predicts a disparity map $\mathbb{D}$ and recover high dynamic range (HDR) $\mathbb{E}$ from a single image that has a deep DOF and standard dynamic range (SDR). $\mathbb{D}$ and $\mathbb{E}$ will be formally defined later. Unlike the MDE dataset that is designed to avoid saturation, our extended MDE dataset contains a diverse set of scenes that cover a wide dynamic range. We find that both disparity and HDR can be estimated purely from our extended MDE dataset, by imposing supervision on the reconstructed shallow DOF images. Alternative to our joint prediction is to estimate depth and HDR separately in sequence, for example, an HDR recovery network followed by a depth prediction network. Recent works have addressed performance on monocular depth estimation [108] and monocular HDR recovery [25]. Our sub-system on monocular shallow DOF rendering would be approximately equivalent to a composition of these state-of-the-art works. One of our advantages is that we do

not require ground truth supervision on either depth or HDR, while the aforementioned methods use direct supervision and thus require challenging data capturing and annotation to account for model generalization to different or more generic scene contents.

We formulate our joint disparity and HDR estimation network as following. Given training dataset $\mathcal{A}$ with pairs of small aperture input $I^S$ and large aperture ground truth output $I^L$, the major loss we apply to optimize the network parameters is the rendering loss:

$$L_{\text{rend}} = \|I^L - \mathcal{F}(I^S, \mathbb{D}, \mathbb{E})\|_1 \tag{5.1}$$

We use shallow DOF rendering as the objective, thereby bypassing the need to have direct supervision using disparity or HDR ground truth, which can be extremely challenging to collect and annotate.

The forward model $\mathcal{F}$ is based on an ideal thin lens model [131]. It takes a deep DOF image with predicted $\mathbb{D}$ and $\mathbb{E}$ to render a synthetic shallow DOF image. We use a disk kernel $K$ to approximate the point spread function of a defocused point through the lens. To handle occlusion, we blend layers of different disparity levels in order from back to front to prevent background blur from incorrectly bleeding around the silhouette of foreground objects, as shown in Figure 5.3D. Previous methods such as [157] simply sum up all disparity levels.

We define disparity $\mathbb{D}$, the inverse depth, in its stereo sense being created by the differing vantage points from left and right edges of the lens aperture. This amount is proportional to the defocus blur size in pixel space by a normalized scalar [63]. Assume the disparity map $\mathbb{D}$ ranges from $d_{\min}$ to $d_{\max}$ and denote disk kernel radius as $r$ ($r = 0$ at focal plane). We discretize $\mathbb{D}$ into $|d_{\max} - d_{\min}|$ levels using a soft mask $M$. We define $M$ as the matte for content at the corresponding disparity level. $M$ allows the forward model $\mathcal{F}$ to be differentiable and also stabilizes training. Formally, $M$ at disparity $d$ is defined as:

$$M(\mathbb{D}, d) = \exp(-\lambda(\mathbb{D} - d)^2) \tag{5.2}$$

$\lambda$ is empirically set to $2.8$ to model a continuous and rapid falloff across neighboring disparity levels. Because we predict signed disparity and define focal plane to have zero disparity, $r = d(d \geq 0)$ when $d$ is behind the focal plane, otherwise, $r = -d(d < 0)$. We use $I^l$ to denote the rendered shallow DOF image. Using our back-to-front forward rendering model, the shallow DOF image at disparity level $d$, denoted as $I^l_d$, is computed by blending it with its previous disparity level $I^l_{d-1}$ using the aforementioned content mask $M$:

$$I^l_d = I^l_{d-1} \cdot (1 - M(\mathbb{D}, d)) + \big(I^S \cdot M(\mathbb{D}, d)\big) \otimes K(r) \tag{5.3}$$

One of the beautiful and characteristic visual signatures of shallow DOF video are so-called "bokeh balls," which are bright disks optically created by defocusing of small, bright lights (see the background of the middle image in Figure 5.6). A numerical challenge for synthesizing such bokeh balls is that one needs HDR images to capture the very high intensity of the small lights [18]. Without HDR, the represented intensity of the small bright

lights is limited by the 8-bit fixed point precision of our input videos – when synthetically defocusing these lights, the incorrectly low intensity value spreads out over many pixels and becomes invisible rather than a bright ball of light (see middle image in Figure 5.3B). Therefore, we find HDR recovery key to render visually salient bokeh that appears at saturated regions, which has not been considered in prior synthetic defocus rendering models [92, 188, 170]. We undo gamma correction on input images to work in linear space. We predict $\mathbb{E}$ in log scale to recover a high dynamic range. Pixels that are saturated in the deep DOF image are often not saturated in its shallow DOF pair, because their energy is spread over many pixels. This provides indirect signal for the network to learn HDR recovery. We replace $I^S$ in Equation 5.3 with its HDR version $I^{S'}$, computed as:

$$I^{S'} = I^S \cdot e^{k \cdot \mathbb{E}} \tag{5.4}$$

where $k$ affects the maximum recovered saturation value and is empirically set to 50.

**Data Collection.** To train our RGBD-HDR Estimator, we build upon the Flower dataset collected in [157] and contribute the first large-scale aperture dataset that covers diverse object categories. The dataset contains 1.2K image pairs and 0.8K image triplets taken with different aperture sizes (f/2, f/8 and f/22) and focus depth. Each image pair or triplet is taken in a scripted continuous shot using Magic Lantern[3] firmware add-on for Canon EOS DSLR cameras. This minimizes misalignment among pairs/triplets during capturing time. Pixel-wise alignment is further imposed via correlation coefficient minimization [28].

**Image Pair and Triplet Supervision.** The network takes in a deep DOF image and predicts both a disparity map and a high dynamic range, which are used to render a shallow DOF image, and has ground truth to compare against. This rendering loss is back-propagated to update network parameters until convergence. We notice that the precision of large defocus values in $\mathcal{D}$ is less accurate as the gradient of the reconstruction loss decreases inversely proportional to the size of the defocus disk kernel ($\frac{\partial L_{\text{rend}}}{\partial r} \propto \frac{1}{r^2}$). This produces visual artifacts when refocusing the video to planes that are originally at large disparity. To mitigate imbalanced loss gradient back-propagated through different disparity planes, we apply a triplet consistency checking during training. Our dataset contains two types of image triplets: aperture and focal triplets. *Aperture triplets* are taken with f/2, f/8 and f/22. The estimated disparity should be able to *scale* to render both median DOF and shallow DOF images. This constraint also helps stabilize training. *Focal triplets* include a deep DOF image and two shallow DOF images focused at different depths. The estimated disparity map should be able to *shift* to render both shallow DOF images at different depths. From our thin lens model assumption, the non-linearity between change of focal plane and change of defocus blur size only depends on the object disparity and lens movement in sensor coordinate, which is relatively small to be negligible. As long as

[3] `https://magiclantern.fm/`


Figure 5.6: Example image pairs and triplets in our dataset. (Row 1) A focal triplet example, defocus map predicted from the input image is used to reconstruct the image with the same focus depth but taken with a large aperture (middle image) shifted to reconstruct the image taken with a large aperture at a different depth plane (right image). (Row 2) An example of aperture triplet, defocus map predicted from the input image is used to reconstruct the large aperture image, and scaled to reconstruct the medium aperture image. (Row 3) An example of image pairs, note that the focus is on the middle plane for this example.

the object is not too close to the camera, we can assume the shift of focal plane to be linear to the change of defocus size. For the prime lens, Canon EF 50mm f/1.8, that we use for data capturing, scene depths difference from infinity to 0.5m generate a deviation of $\sim$5% from the assumed linear model. Training with both types of triplet data helps to improve the precision of large disparity region estimation. Figure 5.3C shows a visual comparison on the estimated disparity map and back-focus shallow DOF rendering between training with and without triplet consistency data.

**Loss Functions.** We train RGBD-HDR with rendering objectives that penalize difference between rendered results and ground truth shallow DOF targets. We supervise the network with per-pixel $L^1$ loss, denoted as $L_{\text{pix}}$, as well as low-level and high-level image features denoted as $L_{\text{feat}}$, by feeding the network output and target through a pre-trained

VGG-19 network $\Phi$. We compute the $L^1$ difference between $\Phi(\mathcal{F}(I^S, \mathbb{D}, \mathbb{E}))$ and $\Phi(I^L)$ in selected feature layers.

$$\begin{aligned} L_{\text{rend}}(I^S, I^L) &= L_{\text{feat}}(I^S, I^L) + L_{\text{pix}}(I^S, I^L) \\ &= \sum_i \lambda_i \|\Phi_i(I^L) - \Phi_i(\mathcal{F}(I^S, \mathbb{D}, \mathbb{E}))\|_1 + \\ &\|I^L - \mathcal{F}(I^S, \mathbb{D}, \mathbb{E})\|_1, \end{aligned} \tag{5.5}$$

where $\Phi_i$ indicates the layer $i$ in the VGG-19 network. The weights $\{\lambda_i\}$ are used to balance different terms in the loss function. We select the layers `conv1_2`, `conv2_2`, `conv3_2`, `conv4_2`, and `conv5_2` in the VGG-19 network. These features are demonstrated to be effective for image enhancement, style transfer and many other image processing tasks [15, 75, 197].

Image triplets are applied in an adjusted rendering objective that penalizes difference from shallow DOF image $I^{L'}$ after adjusting aperture (scale) and focal plane (shift), with the adjustments linearly approximated by an affine model. Similar to $L_{\text{rend}}$, $L_{\text{adjust\_rend}}$ computes the $L^1$ difference between $\Phi(\mathcal{F}(I^S, \alpha\mathbb{D}+\beta, \mathbb{E}))$ and $\Phi(I^L)$ in the same selected feature layers. We find this image triplet training effectively improves RGBD-HDR prediction at far planes, as shown in Figure 5.3C.

We additionally incorporate an edge-aware smoothness penalty $L_{\text{smooth}}$ on the predicted disparity map by minimizing the $L^1$ norm of its gradients, weighted less on the input image edges. This ensures the predicted disparity map to be locally smooth and is formulated as:

$$L_{\text{smooth}}(\mathbb{D}) = \|\partial_x \mathbb{D}\|_1 \cdot e^{-|\partial_x I^S|} + \|\partial_y \mathbb{D}\|_1 \cdot e^{-|\partial_y I^S|}$$

Overall, we train our network by minimizing a loss function that is a weighted sum of $L_{\text{rend}}$, $L_{\text{adjust\_rend}}$, and $L_{\text{smooth}}$.

$$\begin{aligned} L_{\text{total}} = &\sum_{(I^S, I^L, I^{L'}) \in \mathcal{A}} L_{\text{rend}}(I^S, I^L) + L_{\text{adjust\_rend}}(I^S, I^{L'}) \\ &+ w_1 L_{\text{smooth}}(\mathbb{D}) \end{aligned} \tag{5.6}$$

where $w_1$ is the weight for the smoothness regularization, and is set to 10 across all experiments. When a triplet is not available for a particular example, *i.e.*, , only a pair is available, we omit $L_{\text{adjust\_rec}}(I^S, I^{L'})$ and double the weight of $L_{\text{rend}}(I^S, I^L)$.

**Training and Implementation.** We use a U-net network architecture [138] that contains an encoder-decoder structure with skip connections. All layers are followed by a leaky *ReLU* activation, except for the last prediction layer that produces $3 + N$ channels. Three of these channels are used for HDR recovery. The other N channels are used as a bilateral-space representation over luma, which are sliced with a bilateral slicing operator (where $N$ is defined by bandwidth parameters of the bilateral slicing operator) into pixel-space to

produce a 1-channel signed disparity map. In practice, we find that predicting a bilateral-space representation improves fidelity of the disparity map over predicting directly in pixel-space, particularly around edges. This is consistent with the findings of Gharbi *et al*. [45] and Barron *et al*. [7]. Positive disparity refers to planes behind the focal plane while negative disparity represents planes in front of the focal plane.

We train the network with batch size 1 on an NVIDIA Titan X GPU and weights are updated using the Adam optimizer [87] with a fixed learning rate of $10^{-4}$. A full network architecture will be made available in a code release. The network converges after $150K$ iterations. Our network is fully convolutional and can run at arbitrary image sizes. During training, we resize the images to random resolutions between $512$p and $1024$p.

**Video Temporal Consistency.** Visually, we find that the most important change when going from still images to video is to enforce temporal consistency. Independently rendering each frame with shallow DOF causes visually disturbing flickering, especially around prominent bokeh. A comparison can be found in Figure 5.3A. To impose temporal coherency, we apply a weighted temporal moving average that is occlusion-aware and robust to outliers to $\mathbb{D}$ and $\mathbb{E}$. For each target frame $I_i$, we compute $w_i, w_{i-1}, ..., w_{i-M} \in W$, where $M$ is the number of neighboring frames. We compute optical flow using a pre-trained deep neural network Flownet 2.0 [72] for consecutive pairs of frames, and align $I_{i-1}, ..., I_{i-M}$ to $I_i$ using concatenated flows. $w_i$ is computed as a weighted combination of an occlusion weight $w_i^{\text{occl}}$ and an outlier weight $w_i^{\text{med}}$.

*Occlusion Weight* Occluded pixels should be weighted little. We adopt the tactic of forward-backward consistency checking [124, 14], computing both forward $f^{i\to i+1}$ and backward optical flow $f^{i+1\to i}$ for frame pair $I_i$ and $I_{i+1}$. Consider point $p$ in $f_i$ that shifts to $p + f^{i\to i+1}(p)$. We check if we can find a point $q$ in $f_{i+1}$ such that:

$$\|(p - f^{i+1\to i}(q))\|_2 + \|(q - p + f^{i\to i+1}(p))\|_2 \leq \delta \tag{5.7}$$

where $\delta$ is a small distance threshold. If there exists such a $q$ in $f_{i+1}$, $p$ is considered as a consistent pixel in $f_i$. For each frame $I_i$ we compute such an occlusion mask and call it $w_i^{\text{occl}}$.

*Outlier Rejection* To account for optical flow inaccuracies, we classify a warped pixel as an outlier if it has very different values within its temporal moving window. We assume that outliers are sparse and thus a majority vote approach such as median filtering would be effective. For any point $p$ in frame $I_i$, its outlier weight $w(p)_i$ is set to

$$e^{-\text{median}(|p_i + f^{j\to i}(p) - p_i|_{j=i-M}^{i})}$$

We compute an outlier weight for each frame and call it $w_i^{\text{med}}$. Overall, the filtered prediction $\mathbb{D}_i$ (or $\mathbb{E}_i$) of target frame $I_i$ is computed as:

$$\mathbb{D}_i = \sum_{j=i-M,...,i-1} W_j \cdot (\mathbb{D}_i + f^{j\to i}) + W_i \cdot \mathbb{D}_i \tag{5.8}$$

where $\sum_{i-M}^{i} W_j = 1$. We set $M$ to 6 for all experiments.

**Parallelization in *Scanner*.** We have a sizable set of 100 test videos, so we use modern infrastructure to process them with RVR. We choose to use Scanner [130], which gives us the option to process on different hardware and parallelize on the cloud. With full parallelism we could in principle process all 100 videos in under a minute. In practice we use a local 4-core machine with a single Titan X GPU. On average, we process one 2 megapixel video frame in ten seconds, including RGBD-HDR inference, bi-directional optical flow computation, and temporal filtering. The current bottleneck is the flow-based temporal filtering and the fact that we do not take full advantage of Scanner's distributed processing of jobs. It remains as future work to get the total processing time down to realtime.

## 5.5 Look-Ahead Autofocus (LAAF)

Being able to synthesize refocusable videos is not a complete solution to generating a meaningful shallow DOF video – deciding on when and where to focus in the video is challenging and essential. In the conversation example shown in Figure 5.5D, the focus should shift between the active speaker and lock at the person when he/she speaks – failing to set focus correctly would cause unsynced audio and visual focus and thus confuse the viewer. On a movie set, exact focus is achieved by a movie script that exhaustively defines what should be in focus at every moment in the film, and a dedicated focus puller (the 1st Assistant Camera) who measures and marks the exact focus position. In this section, we demonstrate how LAAF uses recent computer vision advances in video understanding — analyzing semantics of current and future frames — to enable video autofocus that automates portions of focusing process in cinematography. This is, we believe for the first time, the important and previously impossible problem of delivering cinema-like focus in casual videography is shown to be tractable. We demonstrate three approaches towards semi-automated and fully-automated video autofocus, using:

- an interactive GUI with vision-based tracking and simple human focus selection (Section 5.5.1)
- scene-specific (*e.g.* conversation, action, etc.) AI-assist modules (Section 5.5.2)
- a data-driven CNN network trained from a large-scale video dataset with focus annotation, labeled using our GUI (Section 5.5.4)

Output from the above three approaches is a set of $\{(x, y, \hat{t})_i\}$ triplets denoting New Focus Targets — focus regions and times. Three such New Focus Targets are shown as diamonds on the timelines in Figure 5.4A. For each New Focus Target, we perform focus tracking by computing $(x, y)$ tracking across time, and look up the focus depth from the

estimated depth map. Next, we execute focus pull from one target to the next, with focus arriving at each target slightly before to allow the viewer to visually settle before action begins. We set a default duration of 10 frames ($\sim$ 0.67 sec) for a focus pull and linearly interpolate focal planes in between.

For real systems LAAF could happen during video capture as well. To do this, we would need to buffer a few seconds of video frames, which would be the temporal window we set to look ahead, and then pipeline the LAAF processing with video recording.

### 5.5.1 GUI-based Video Semi-Autofocus

We build an interactive RVR-LAAF GUI incorporating a vision-based tracker (*e.g.* KCF tracker [64]) such that the user only needs to specify a New Focus Target, instead of selecting a focus subject for each frame, which is extremely inefficient and impractical. The tracker tracks the selected focus region until the user pauses the video to select the next New Focus Target.

One interesting point is that RVR-LAAF GUI provides benefit even for simple scenes, such as a single person, that seem amenable to conventional autofocus. One might think that in these situations simply synthesizing shallow DOF from the recorded video would suffice. However, the issue is that synthetic defocus will amplify any misfocus error. So even in these situations, RVR-LAAF GUI allows us to increase the focus accuracy of the output video by adjusting the synthetic focal plane onto the person of interest and track the corrected subject. In other words, autofocus achieved by LAAF is essential to delivering even simple synthetic defocus video accurately.

### 5.5.2 Scene-Specific Video Autofocus

Fully-automated video autofocus without human interaction requires visual and semantic understanding of the video context. We show how LAAF incorporates recent advances in video understanding to automate, to some extent, human choices of focus selections. Faces, actions, audio sources and other salient (visually distinctive) objects, are some common subjects to set in focus in casual videography. We exploit a set of context-aware detectors $\mathcal{H}$ (*e.g.* face, action detectors) to automate the generation of New Focus Targets.

As illustrated in Figure 5.4B, we compute the intersection of a selected set of detection to identify a scene-dependent and contextually-meaningful focus region. We then apply K-means clustering (K empirically set to 4) to determine the majority clustering centroid position $(x, y)$ and read in the focus depth from the predicted disparity map $\mathbb{D}$. Frames with empty intersection use its previous disparity level. Next, we apply a bilateral filter followed by an edge detection to identify focus depth discontinuity, which marks the New Focus Target, $(x, y, t)$. Note that we use $(x, y)$ instead of $\mathbb{D}(x, y)$ to denote New Focus Target because we later use $(x, y)$ to track the subject to synthesize focus puller as shown in Figure 5.4A.

For example, we demonstrate scene-specific LAAF on two types of videos, one on action videos that use action and saliency detection ($\mathcal{H} = \{H_{\text{act}}, H_{\text{sal}}\}$) to detect salient regions that also involve action (Section 5.5.2); one on conversation videos that use audio localization and face detector ($\mathcal{H} = \{H_{\text{aud}}\, H_{\text{face}}\}$ to detect and focus on the person who is speaking (Section 5.5.2). These two types of videos are commonly seen in both casual videography and professional filmmaking. According to a video essay that educates focus rack and summarizes the best 30 rack focus examples throughout the film history (years 1963-2016) [4], half of the focus racks are triggered by human action or audio source change.

**Action-aware LAAF.** We collect a set of videos that involve unexpected actions, which trigger focus depth change towards the action subject. An example scenario is a background subject of interest enter the frame unexpectedly while focus is at the foreground, and the result we seek is to shift focus at, or a few frames before, the subject entering the view. We use an action localizer ($H_{\text{act}}$) and a salient object detector ($H_{\text{sal}}$) to generate probable focus regions that are both salient and involve actions. $H_{\text{act}}$ is based on optical flow and deepmatch [136] for action localization and tracking, and is used as video pre-processing step for several computer vision tasks such as unsupervised video feature learning [129]. $H_{\text{sal}}$ is based on a still image saliency detection [66] work, which trains a deep network to compute a saliency heat map that identifies visually distinctive objects and regions in an image. As illustrated in Figure 5.5E, LAAF analyzes future frames (Row 1) and detects the child's action that is about to happen. Instead of always keeping focus on the foreground bush, LAAF is able to shift focus to the background before the child slides down the hill (Row 4).

**Audio-aware LAAF.** We demonstrate LAAF applied to another video collection of conversational scenes. The goal is to place focus right before the person who is about to speak. We use an audio localizer ($H_{\text{aud}}$) and face detector ($H_{\text{face}}$) to compute probable focus regions of the person who is speaking. $H_{\text{aud}}$ employs a recent work on audio localization [125], where the authors train a deep network to learn multisensory representation using the fact that visual and audio signals often align temporally. $H_{\text{face}}$ is a machine-learning-based face detector from dlib[5]. As illustrated in Figure 5.5D, without LAAF the focus is always on the front person, blurring out the person in the back even when she starts talking (Row 2). LAAF analyzes future frames to understand who's speaking, and is able to correctly shift focus a few frames before the person starts to talk (Row 4).

### 5.5.3 New Focus Targets Evaluation.

To evaluate the set of New Focus Targets (x,y,t) generated from scene-specific LAAF, we compute the difference on focus plane in disparity units $\Delta d = \mathbb{D}(x, y) - \mathbb{D}(\hat{x}, \hat{y})$, and the

[4] https://www.youtube.com/watch?v=tT_qv9ptauU&t=75s

[5] http://dlib.net/

temporal position offset in number of frames $\Delta t = t - \hat{t}$, where $(\hat{x}, \hat{y}, \hat{t})$ is the ground truth annotated New Focus Targets, using RVR-LAAF GUI.

### 5.5.4 Data-driven Video Autofocus

To make autofocus fully automated on videos with any scene content, we present a first attempt at training a CNN (AF-Net) to predict New Focus Targets, replacing the scene-dependent detectors used in scene-specific LAAF. This is enabled using our RVR-LAAF GUI to annotate New Focus Targets on a large-scale video dataset. AF-Net takes in a sequence of frames centered at the query frame to predict the focus region $(x, y)$ and the probability of the query frame being a New Focus Target.

The key to LAAF is to analyze past and, particularly, future frames. A major challenge that arises is to have the network cover a wide temporal span of frames in a manner that is efficient in memory and computation. We introduce a temporal aggregation architecture consisting of two CNNs with different temporal receptive field sizes, as illustrated in Figure 5.7. The temporal receptive field is determined by the number of temporal units $f_s$, and the temporal coverage $f_T$ inside each temporal unit. CNN-1 has a temporal receptive field of $f_T$ to predict the focus region of the middle query frames ($i_1$, ..., $i_s$). CNN-2 takes in feature maps generated from CNN-1, and predicts the probability of the global center frame ($i_c$), efficiently seeing a wider temporal receptive field of $f_s \cdot f_T$ frames. Loss of the network is a weighted sum of the evaluation metrics $\Delta d$ and $\Delta t$ we described in Section 5.5.2.

However, evaluating $\Delta d$ requires disparity maps for all videos. We find that large-scale public video datasets that contains deep DOF are heavily-compressed [1, 177] and cannot be processed by our RGBD-HDR or have tracking be applied with high fidelity. We thus made 2

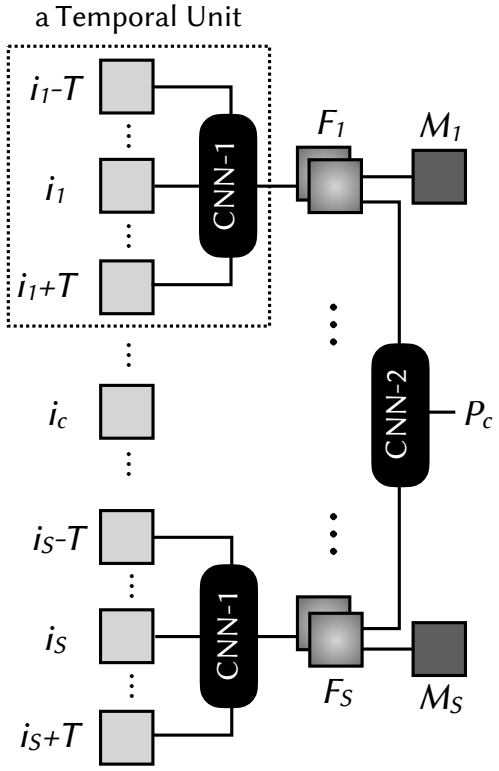



Figure 5.7: AF-Net architecture predicts the focus region $(M)$ for $i_1, ..., i_s$ and the probability $(P)$ of the center frame $i_c$ being a New Focus Target. The key of AF-Net is to cover a wide temporal range such that it sees past and future frames to make the prediction. In this illustration, CNN-1 has a temporal receptive field $(f_T)$ of $2T$ and CNN-2 has a wide temporal receptive field of $2T \cdot s$, where $s$ is the number of temporal units $(f_s)$.

adjustments to training our AF-Net. First, instead of using $\Delta d = \mathbb{D}(x, y)$, we use its proxy $(x, y)$ as supervision. Second, we choose to use a eye-tracking dataset DHF1K [177] and a filtered version of its ground truth eye fixation map to supervise $(x, y)$, which we find to be highly correlated with focus region we annotated at New Focus Targets; supervision for $t$ comes from our annotation. We call our annotated focus video dataset VAF.

We remove unsuitable videos (*e.g.* with jump cuts) and end up with 419 videos ($640 \times 360$) of 20-60 seconds each in full frame rate (30 FPS). We reduce temporal sampling rate by a factor of 5 and set temporal stride of AF-Net to be 5. Both CNN-1 and CNN-2 architecture resembles ResNet-10, followed by a global average pooling for CNN-2. Training schedules are set to be the same as training our RGBD-HDR net.

## 5.6 Results

We evaluate the key components of RVR (Section 5.6.1), and follow the metric described in Section 5.5.2 to quantitatively evaluate LAAF (Section 5.6.2). We also show a comparison against the autofocus system inside a high-end consumer camera Olympus EM1.2 in Section 5.6.5.

### 5.6.1 RVR Evaluation

We render shallow DOF video using forward model in Equation 5.3 frame by frame, and compare rendering results with and without temporal coherency in Figure 5.3A. The focal planes are set to be the same for each comparison. We also show a forward model with and without occlusion-awareness in Figure 5.3D, and with and without HDR recovery in Figure 5.3B.

Evaluating intermediate network predictions on disparity and HDR maps against ground truth provides additional insight on rendering performance. For depth sensing, advanced range sensors such as LiDAR captures high-quality dense depth maps. Lightweight RGB-D cameras such as Intel RealSense [83] have achieved adequate resolution for some consumer-grade applications, but still suffer from noisy output and limited precision around object edges. It still remains a challenge to obtain accurate depth for casual videos using portable devices. A survey on RGBD camera is written by Zollhöfer [207]. Current high-end smartphones such as iPhone X supports depth measurement using dual-pixels and dedicated post-processing to generate smooth, edge-preserving depth maps. A recent paper [174] on monocular shallow DOF synthesis uses iPhone to construct the iPhone Depth Dataset for model training and testing. In a similar manner, we capture 50 test images using an iPhone X and extract disparity map as a proxy for ground truth to evaluate our predicted disparity map. We also apply a state-of-the-art monocular depth estimator, MegaDepth [108], to these test images for comparison. We follow the quality metric proposed in [143] to compute the RMS (root-mean-squared) error measured in disparity units between the predicted disparity map and its ground truth. For HDR evaluation, it is even more chal-

lenging to capture ground truth HDR using existing hardware sensors. We choose to use the public HDR test dataset constructed from exposure stacks from HDRCNN [25], and use their metric by computing the mean square error in the log space of the predicted linear image and its ground truth.

Figure 5.8A shows a histogram on disparity evaluation between our prediction and that from MegaDepth. Our indirect method without depth supervision produces comparable performance with MegaDepth that requires ground truth depth for training. In Figure 5.8B, we plot the histogram on HDR evaluation of the test images from [25]. It is expected that HDRCNN generates better quantitative performance as the model is trained with ground truth supervision, while our model is trained without direction supervision on HDR and on a different dataset. We provide an alternative way to recover HDR without ground truth required, and are able to produce HDR maps with adequate quality to render shallow DOF images (see Figure 5.3B). Importantly, per frame estimation of depth and HDR is not sufficient for rendering refocusable video, as we have shown that temporal coherence is also critical in Figure 5.3A.

We design RVR as a flexible system to incorporate future works that improve upon disparity and HDR estimation for still images and even for videos, or use camera sensors that support depth and HDR video streaming. A recent work [170] uses dual-pixel imagery to estimate scene disparity for smartphone photography. When dual-pixel imagery is available, RVR could use its depth estimation as $\mathbb{D}$ in the pipeline.

### 5.6.2 LAAF Evaluation

We evaluate the predicted New Focus Target $(x, y, t)$ against ground truth $(\hat{x}, \hat{y}, \hat{t})$ (from GUI annotation) using $\Delta d$ and $\Delta t$ (See Section 5.5.2). For each test video, we compute the average focus depth difference $\overline{|\Delta d|}$ across all frames, and the average temporal position difference $\overline{|\Delta t|}$ across all New Focus Targets.

**GUI-based Semi-Autofocus** We show an example in Figure 5.4C using RVR-LAAF GUI to annotate a New Focus Target (on the person's face). Figure 5.4C Row 3 presents that the user only needs to annotate the New Focus Target — selecting a focus region and creating a tracker, and the GUI will then track the selected region. The GUI also features fine tuning on the defocus strength and the focus puller duration to account for different story tone and visual sensitivity. High-res version of the GUI is shown in the accompanying video.

**Evaluation on Action-aware Autofocus** We test LAAF on 11 casually collected videos with unexpected action that triggers focus depth change. We use action-aware LAAF with $\mathcal{H} = \{H_{\text{sal}}, H_{\text{act}}\}$, as described in Section 5.5.2 to compute New Focus Targets. Among all 11 test videos, 8 (72%) videos achieve $|\Delta t| < 15$, generating New Focus Targets that on average offset by less than half a second. Most test videos use LAAF to locate focus regions

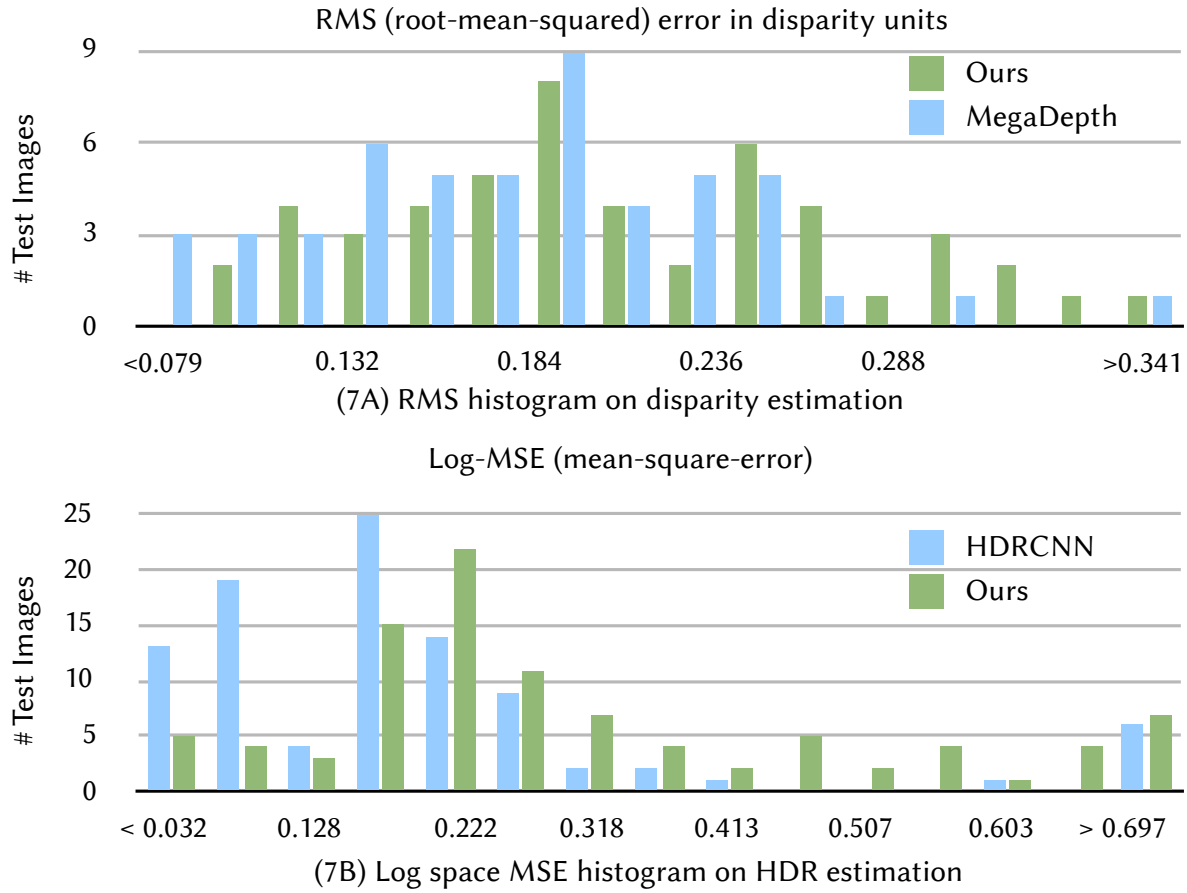



Figure 5.8: Evaluation on predicted disparity and HDR against ground truth. Depth ground truth is obtained using dual-pixel on an iPhone X. HDR ground truth is from the public dataset described in [25].

within 2 depth planes difference from ground truth. Quantitative results are shown in Table 5.1. A qualitative result is shown in Figure 5.5E.

**Evaluation on Data-driven Autofocus Detector** Among the 419 VAF videos, 40 videos are held out as test set for evaluation. To evaluate New Focus Target, we run inference on a 100-frame clip to get the probability of each frame being the New Focus Target and select the frame with the highest probability. Note that for a test clip that does not have New Focus Targets (no focus depth changes required), we only evaluate $\Delta d$ but not $\Delta t$. We also test on 30 collected smartphone videos, among which 11 are the action videos used to evaluate action-aware LAAF in Table 5.1. $\overline{|\Delta t|}$ on the 11 action videos increase from $15.1$ (using our specialized action-aware LAAF) to $21$ (using AF-Net). However, AF-Net has


Figure 5.9: Data-driven AF-Net to predict New Focus Targets for video autofocus. (Row 1) Input video shows a dumpling is picked up from the plate. (Row 2) naïve synthetic SDOF video would easily suffer from mis-focus when applied shallow DOF. (Row 3) Focus region prediction from AF-Net; green marks focus with higher probability. Second frame (bounded by red) is the predicted temporal position of the New Focus Target, when the person picks up the dumpling. (Row 4) The resulting video shows autofocus that tracks accurately on the dumpling. Video result can be found in the accompanying video.

| Video ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\overline{\lvert\Delta d\rvert}$ | 1.3 | 1.4 | 1.2 | 2.2 | 2.0 | 1.4 | 1.6 | 1.2 | 1.8 | 4.7 | 3.1 |
| $\overline{\lvert\Delta t\rvert}$ | 6 | 8 | 3 | 15 | 7 | 10 | 14 | 20 | 7 | $> 30$ | 16 |

Table 5.1: New Focus Target evaluation of action-aware LAAF on 11 casually collected videos using metrics introduced in Section 5.5.2. $\overline{\lvert\Delta d\rvert}$ (in disparity units) evaluates the performance of computed depth plane over all frames for each video. $\overline{\lvert\Delta t\rvert}$ (in number of frames) evaluates the average performance of temporal position of all New Focus Targets. Ground truth is obtained from hand-annotation using our RVR-LAAF GUI.

| Dataset | **VAF**-test | **Ours**-test | **Ours**-action-test | |
|---|---|---|---|---|
| Method | AF-Net | AF-Net | Action-LAAF | AF-Net |
| $\overline{\lvert\Delta d\rvert}$ | — | 2.64 | 1.99 | 2.21 |
| $\overline{\lvert\Delta t\rvert}$ | 19.8 | 24.5 | 15.1 | 21.0 |

Table 5.2: Results using AF-Net to predict New Focus Targets on VAF test and our collected videos. We compute the average $\overline{|\Delta t|}$ across all test videos. $\overline{|\Delta d|}$ is evaluated only on our test videos as we do not have disparity for VAF. We find AF-Net to perform slightly worse than action-aware LAAF on the action videos, but achieves reasonable performance on generic video contents, *e.g.* 1 second difference in $t$ on all test videos. Ground truth for our collected test videos are from hand-annotation using RVR-LAAF GUI.

the advantage of handling generic video contents and is able to achieve $\overline{|\Delta t|}$ less than 1 second on all test videos (see Table 5.2). This indicates the potential of having a large-scale annotated video dataset with temporal aggregation (AF-Net) to tackle the challenging problem of video autofocus. One example result is shown in Figure 5.9. Video results are at 05:58 in the accompanying video.

### 5.6.3 Limitation of AF-Net

While AF-Net explores the potential of using machine learning for video autofocus, its requirement of large-scale video focus annotation is expensive and the dataset we accumulated was of modest size. Recently, unsupervised visual feature learned from large-scale unlabeled videos has shown to be effective for video understanding, such as object tracking via video colorization [169] and audio localization by learning to temporally align audio and video [125]. We believe video autofocus can be likewise addressed by self-supervision to learn from unlabeled internet-scale videos such as public movie clip dataset [137].

### 5.6.4 Analysis of Artifacts in Output Video

The accompanied video discussed in this section can be found here: `https://youtu.be/FqQQw3DGI9I`. The performance of our RVR-LAAF system is commensurate with a first prototype according to this new approach to the video auto-focus problem. For example, visible artifacts present in rendered videos due to imperfect disparity and HDR estimation. In this section, we classify the video artifacts visible in the results and discuss their causes by inspection and analysis of the real video results presented in previous sections, as well as experiments on a rendered video clip (see Supplementary Video B ) that provides "ground truth" for comparison.

For the synthetic scene, we used a clip from the Blender Open Movie titled "Sintel" (2010), rendering this scene with a simulation of a small aperture similar to the deep DOF video captured by a modern smartphone camera. In Supplementary Video B , we show ablation results that compare ground truth disparity or HDR with estimations using our method and Eilertson *et al*. [26] for HDR and MegaDepth [108] for disparity. It is important to note that there is a large domain shift between this synthetic rendering and the real training data used in all the estimation methods described, which disadvantages the estimated results. Nevertheless, the comparisons against ground truth provide empirical clues to support technical dissection of which errors in the system are associated with which classes of video artifacts.

The most visually prominent set of artifacts is due to errors in estimated disparity. There are several classes of visual artifacts that may be seen. First, depth estimation across boundaries is imperfect, which is residual error in spite of our bilateral space processing. These edge errors cause prominent visual artifacts when in-focus regions are incorrectly blurred. Examples can be seen in the synthetic video result at 00:25 (ear) and in real video results at 05:13 (front person's right shoulder). The second class of visual artifact is splotchiness of synthesized defocus blur, due to incorrect spatial variation of depth estimates on flat regions. This error tends to appear in regions at large disparity, and is residual error in spite of our triplet training procedure (Section 5.4) that helps to effectively reduce this problem. Examples can be seen in the synthetic video result at 00:31 (in the highlight region of the background arm), and in the real video result at 06:20 (background segments are incorrectly rendered sharper). The third class of visual artifact related to disparity estimation error is temporal fluctuation of the defocus blur, typically in regions of background. Examples of this error can be seen in the background of synthetic video results around 0:14, and in the real video results at 04:27 (behind conversation) and 05:20 (behind dog). This is residual temporal error in spite of compensation by our temporal stabilization module.

A second set of artifacts is related to errors in HDR estimation. The most common examples of this error manifests as missing bokeh balls (false negative) in output video, while hallucinated bokeh balls (false positive) are generally rare. For example, real video results at 02:20 fail to recover HDR specular highlights on the glistening sea surface and are therefore missing salient bokeh balls expected in that region. Synthetic video results at 00:43 underestimate the HDR value of the background figure's arm and renders a darker defocused highlight.

### 5.6.5 Compare Against Market Camera

High-end consumer cameras with state-of-the-art autofocus technology still suffers from mis-focus, especially during a rapid subject change that requires focus to resolve accordingly. We capture a pair of videos of the same scene with a Olympus EM1.2 under f/2.8 and a smartphone. RVR-LAAF is then applied to the video from the smartphone to render shallow DOF with autofocus using our GUI. We compare the two videos and show

Input Synthetic Anamorphic Bokeh

Figure 5.10: Cinematic bokeh rendering. Our system takes in the deep DOF image (on the left), and renders a cinematic bokeh using the predicted disparity and HDR map, and a lens shape that approximates the cinematography lens ARRI Master Anamorphic.

RVR-LAAF tracks focus accurately while the DSLR fails to transition focus when making large subject change.

Another interesting application is that we can simulate effect produced by expensive camera lenses. We apply our system to demonstrate a cinematic bokeh rendering in Figure 5.10 that approximates the ARRI Master Anamorphic[6] lens for professional cinematography. Anamorphic lenses are prized by certain cinematographers, in part because of their ellipsoidal, decidedly non-circular, defocus blur, as seen in Figure 5.10.

## 5.7 Discussion

In this paper, we built our RVR-LAAF prototype as a proof-of-concept for two main reasons. First, to show that we can achieve fundamentally better video auto-focus decisions by re-structuring the problem this way. And second, to show that it is tractable to attack the two sub-problems posed by this approach: synthesizing refocusable video and computing meaningful look-ahead autofocus decisions today. Regarding synthetic refocusable video, we summarized a broad array of technical approaches that the imaging and computational photography communities are actively advancing, from light field imaging to novel sensor designs to machine learning for depth inference. We are confident that performance and quality will improve rapidly. Regarding the problem of computing meaningful look-ahead auto-focus decisions, we hope to have clearly conveyed the idea that this problem is also tractable and ripe for research. We believe that this area can also advance rapidly, given the broad range of current research in computer vision that can be brought to bear.

[6] http://www.arri.com/camera/cine_lenses/prime_lenses/anamorphic/

# Chapter 6

# Conclusion

As a photography enthusiast, I always like bringing a camera everywhere I go, and have been using a Fujifilm mirrorless in the past few years. I take photos casually of any interesting object or event, but occasionally am disappointed by the photos later displayed on the screen, – cluttered, distracting, missing the best moment of an expression or action – which are far less impressive than the image recorded in my mind. Partly, this is because I am not a great photographer, but I also believe that our brain plays an important role cleaning, refining and fixing the imperfect image projected onto our retina. Our brain understands what it sees and processes as the way we prefer, either it is our friend burst into laughter, or looking through a window to something we have yearned for, or enjoying a mid-day picnic with families on a sunny day, we can easily focus our attention on our friend, ignore the window reflection and will likely not realize the distracting shadows on others' faces until we take out our camera and snap a picture. Our brain, in these cases, helps us naturally attend to people we know, filter out distraction (the reflection) and extract information of a face about its semantics rather than its apparent look.

I believe it is necessary for a camera to gain similar understanding as our brain does to achieve the similar functionality of cleaning, refining and fixing imagery. If there is a task human visual system can easily do while a camera struggles with, the knowledge the human brain uses to accomplish the task will likely benefit the camera too. One simple example is getting the correct white balance – we have no problem identifying the skin tone of a person lit by colorful neon lights, or perceiving the colors on snowy days – two challenging scenarios of auto-white-balance even for the best consumer cameras. By working on this thesis, I learned, and also showed that the solution to bring such understanding into a camera system is to learn from data. We can use machine learning to enhance image resolution, to remove unwanted contents (*e.g.* reflection and shadow), and to identify semantically meaningful regions of a video for autofocus.

The efforts of developing machine learning algorithms need to go along with using the right data and finding the appropriate evaluation metrics. A slight shift between training and testing dataset distribution can degrade model performance. Many computational photography algorithms target mobile devices, whose input images will have character-

istics seen in mobile photography – deep depth of field, lens flare artifact, shot noise in low-light regions, obstructions or distractors that clutter the scene, to name a few. Either the model uses smartphone images for training, or it incorporates relevant data augmentations to align the train-test data distributions. While there are well-established perceptual metrics such as PSNR, many computational photography tasks rely on human eyeballing (*e.g.* a user study) for quality evaluation. It is challenging to quantitatively measure reflection removal quality with no ground truth, or the quality of video autofocus where the ground truth itself may be a perspective of the viewer. An evaluation framework that incorporates semantic quality, viewing condition and even user input will likely benefit the development of computational photography algorithms.

Future cameras are blessed with stronger computing power and breakthroughs in sensors and lenses, opening up wider opportunities for machine learning and intelligence. Mobile devices are equipped with more than one cameras that by nature acquires stereo, becoming more similar to human perception mechanism and even surpassing it because the cameras also have different focal lengths. SPAD sensors that were once only for scientific imaging are demonstrating practicality towards consumer cameras, leveraging burst photography techniques such as robust temporal alignment, distortion-free image warping and handling of dynamics and motion. Video are now shot in 4K resolution on most mobile devices, which will benefit from more efficient data encoding and processing to make the best use out of the resolution. LiDAR depth sensing has just launched on mobile devices with accurate depth data to better understand the 3D space. Though LiDAR raw data is still sparse, it will motivate depth densification and refinement that have been already researched in the field of autonomous driving.

One of the interesting and controversial topics in computational photography is authenticity versus manipulation. Here I refer manipulation as the editing algorithms to improve image quality (similar to the ones shown in this thesis), as opposed to adding creativity or artistic aspects. Manipulation methods are sometimes seen as the culprit for altering the reality and defeating the purpose of an immersive photography experience. I see in a different way. I think the key to preserving authenticity is via regulations of how the images are presented and distributed. One recent effort is the Content Authenticity Initiative [1], which is building a system to provide provenance and history for digital media. I always respect every aspect of photography and I am willing to spend efforts and time on this subject. However, I think it is even better to share the beauty of photography to a wider audience who may not be able to devote the same commitment. This is the goal of what I hope computational photography can achieve. After all, the moment our brain processes what we see, the reality has been altered by us in a unique way. What we see is already a projected reality, so is every photograph.

I expect to see cameras to incorporate with machine intelligence at a flashing speed in the next few years. Personally, I will still learn to be a serious photographer; I will wait for the golden hour to get the best lighting; I will bring a tripod to capture the night sky; I will

[1] https://contentauthenticity.org

exchange physical lenses to suit different types of scenes. But for my friends, families who do not have time for photography, I will encourage them to bring a smartphone, because I hope they capture the world with a similar high quality as I do, and I am confident that casual imaging can get there. I look forward to the day when people are not limited by their devices or capturing constraints, but only by where they go and what they see.

# Bibliography


[1] Sami Abu-El-Haija et al. "Youtube-8m: A large-scale video classification benchmark". In: *arXiv preprint arXiv:1609.08675* (2016).

[2] Amit Agrawal et al. "Removing photography artifacts using gradient projection and flash-exposure sampling". In: *TOG* (2005).

[3] Eli Arbel and Hagit Hel-Or. "Shadow removal using intensity surfaces and texture anchor points". In: *IEEE TPAMI* (2010).

[4] Nikolaos Arvanitopoulos, Radhakrishna Achanta, and Sabine Süsstrunk. "Single Image Reflection Suppression". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[5] Masashi Baba and Naoki Asada. "Shadow Removal from a Real Picture". In: *SIGGRAPH*. 2003.

[6] Jonathan T. Barron and Jitendra Malik. "Shape, Illumination, and Reflectance from Shading". In: *TPAMI* (2015).

[7] Jonathan T Barron et al. "Fast bilateral-space stereo for synthetic defocus". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[8] Hakan Bilen et al. "Dynamic image networks for action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[9] Yochai Blau et al. "The 2018 PIRM Challenge on Perceptual Image Super-Resolution". In: *Proceedings of the European Conference on Computer Vision Workshops*. 2018.

[10] Sean Borman and Robert L Stevenson. "Super-resolution from image sequences-a review". In: *Midwest symposium on circuits and systems*. IEEE. 1998, pp. 374–378.

[11] M.C. Brown. *Libyan Sugar*. Colección de Fotolibros. Twin Palms, 2016. ISBN: 9781936611096. URL: https://books.google.com/books?id=ZfG-ngEACAAJ.

[12] Joan Bruna, Pablo Sprechmann, and Yann LeCun. "Super-resolution with deep convolutional sufficient statistics". In: *ICLR*. 2015.

[13] Chen Chen et al. "Learning to See in the Dark". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[14] Qifeng Chen and Vladlen Koltun. "Full flow: Optical flow estimation by global optimization over regular grids". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[15] Qifeng Chen and Vladlen Koltun. "Photographic image synthesis with cascaded refinement networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[16] Yu Chen et al. "Fsrnet: End-to-end learning face super-resolution with facial priors". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[17] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. "Towards Ghost-free Shadow Removal via Dual Hierarchical Aggregation Network and Shadow Matting GAN". In: 2020.

[18] Paul E Debevec and Jitendra Malik. "Recovering high dynamic range radiance maps from photographs". In: (2008).

[19] Paul Debevec et al. "Acquiring the reflectance field of a human face". In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 2000, pp. 145–156.

[20] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[21] Emily L Denton, Soumith Chintala, Rob Fergus, et al. "Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks". In: *NIPS*. 2015.

[22] Bin Ding et al. "ARGAN: Attentive Recurrent Generative Adversarial Network for Shadow Detection and Removal". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

[23] Chao Dong et al. "Image Super-Resolution Using Deep Convolutional Networks". In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2016).

[24] Craig Donner and Henrik Wann Jensen. "A Spectral BSSRDF for Shading Human Skin". In: (2006).

[25] Gabriel Eilertsen et al. "HDR image reconstruction from a single exposure using deep CNNs". In: *ACM Transactions on Graphics (TOG)* (2017).

[26] Gabriel Eilertsen et al. "HDR image reconstruction from a single exposure using deep CNNs". In: *ACM Transactions on Graphics (TOG)* (2017).

[27] Abbas El Gamal and Helmy Eltoukhy. "CMOS image sensors". In: *IEEE Circuits and Devices Magazine* 21.3 (2005), pp. 6–20.

[28] G. D. Evangelidis and E. Z. Psarakis. "Parametric Image Alignment Using Enhanced Correlation Coefficient Maximization". In: *PAMI* (2008).

[29] Georgios D Evangelidis and Emmanouil Z Psarakis. "Parametric image alignment using enhanced correlation coefficient maximization". In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2008).

[30] Qingnan Fan et al. "A Generic Deep Architecture for Single Image Reflection Removal and Image Smoothing". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[31] Sina Farsiu, Michael Elad, and Peyman Milanfar. "Multiframe demosaicing and super-resolution of color images". In: *IEEE Trans. Image Processing* (2006).

[32] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[33] Randima Fernando. *GPU Gems: Programming Techniques, Tips and Tricks for Real-Time Graphics*. Pearson Higher Education, 2004.

[34] Graham D Finlayson, Mark S Drew, and Cheng Lu. "Entropy minimization for shadow removal". In: *IJCV* (2009).

[35] Graham D Finlayson, Steven D Hordley, and Mark S Drew. "Removing Shadows from Images". In: *Proceedings of the European Conference on Computer Vision*. 2002.

[36] R Fontaine. "A survey of enabling technologies in successful consumer digital imaging products". In: *Proceedings of the international image sensors workshop*. 2017.

[37] Eric R Fossum et al. "The quanta image sensor: Every photon counts". In: *Sensors* 16.8 (2016), p. 1260.

[38] Damien Fourure et al. "Residual Conv-Deconv Grid Network for Semantic Segmentation". In: *Proceedings of the British Machine Vision Conference*. 2017.

[39] William T. Freeman, Thouis R. Jones, and Egon C. Pasztor. "Example-Based Super-Resolution". In: *IEEE Computer Graphics and Applications* (2002).

[40] Kun Gai, Zhenwei Shi, and Changshui Zhang. "Blind separation of superimposed moving images using image statistics". In: *IEEE PAMI* 34 (2012).

[41] Yaroslav Ganin and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation". In: *International conference on machine learning*. PMLR. 2015, pp. 1180–1189.

[42] Ravi Garg et al. "Unsupervised cnn for single view depth estimation: Geometry to the rescue". In: *Proceedings of the European Conference on Computer Vision*. 2016.

[43] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. "Image Style Transfer Using Convolutional Neural Networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[44] Karl R Gegenfurtner, Marina Bloj, and Matteo Toscani. "The many colours of 'the dress'". In: *Current Biology* 25.13 (2015), R543–R544.

[45] Michaël Gharbi et al. "Deep bilateral learning for real-time image enhancement". In: *SIGGRAPH*. 2017.

[46] Michaël Gharbi et al. "Deep joint demosaicking and denoising". In: *ACM Trans. on Graphics (TOG)* (2016).

[47] Daniel Glasner, Shai Bagon, and Michal Irani. "Super-resolution from a single image". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2009.

[48] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. "Unsupervised monocular depth estimation with left-right consistency". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[49] Norman Goldberg. *Camera technology: the dark side of the lens*. 1992.

[50] Ian Goodfellow et al. "Generative adversarial nets". In: *NIPS*. 2014.

[51] Christopher Grey. *Master lighting guide for portrait photographers*. Amherst Media, 2014.

[52] Herbert Gross, Fritz Blechinger, and Bertram Achtner. *Handbook of optical systems*. Vol. 1. Wiley Online Library, 2005.

[53] Roger Grosse et al. "Ground Truth Dataset and Baseline Evaluations for Intrinsic Image Algorithms". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2009.

[54] Maciej Gryka, Michael Terry, and Gabriel J. Brostow. "Learning to Remove Soft Shadows". In: *ACM TOG* (2015).

[55] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. "Paired regions for shadow detection and removal". In: *TPAMI* (2012).

[56] Xiaojie Guo, Xiaochun Cao, and Yi Ma. "Robust separation of reflection from multiple images". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[57] Byeong-Ju Han and Jae-Young Sim. "Reflection Removal Using Low-Rank Matrix Completion". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[58] Pat Hanrahan and Wolfgang Krueger. "Reflection from Layered Surfaces Due to Subsurface Scattering". In: *SIGGRAPH*. 1993.

[59] Bharath Hariharan et al. "Hypercolumns for object segmentation and fine-grained localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[60] Samuel W Hasinoff et al. "Burst photography for high dynamic range and low-light imaging on mobile cameras". In: *ACM Transactions on Graphics (TOG)* 35.6 (2016), pp. 1–12.

[61] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[62] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[63] Robert T Held et al. "Using blur to affect perceived distance and size". In: *ACM Transactions on Graphics (TOG)* (2010).

[64] João F Henriques et al. "High-speed tracking with kernelized correlation filters". In: *PAMI* (2015).

[65] Berthold K. P. Horn. "Determining lightness from an image". In: *Computer Graphics and Image Processing* (1974).

[66] Qibin Hou et al. "Deeply supervised salient object detection with short connections". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[67] Xiaowei Hu et al. "Direction-Aware Spatial Context Features for Shadow Detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[68] X Hu et al. "Direction-aware Spatial Context Features for Shadow Detection and Removal." In: *IEEE transactions on pattern analysis and machine intelligence* (2019).

[69] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. "Single image super-resolution from transformed self-exemplars". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[70] Jinggang Huang, Ann B Lee, and David Mumford. "Statistics of range images". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2000.

[71] Matthias Hullin et al. "Physically-based real-time lens flare rendering". In: (2011), pp. 1–10.

[72] Eddy Ilg et al. "Flownet 2.0: Evolution of optical flow estimation with deep networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[73] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[74] Henrik Wann Jensen et al. "A Practical Model for Subsurface Light Transport". In: *SIGGRAPH*. 2001.

[75] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *Proceedings of the European Conference on Computer Vision*. 2016.

[76] Alexia Jolicoeur-Martineau. "The relativistic discriminator: A key element missing from standard GAN". In: *ICLR*. 2019.

[77] Neel Joshi and Larry Zitnick. *Micro-Baseline Stereo*. Tech. rep. 2014.

[78] John Kahrs et al. "Pixel cinematography: a lighting approach for computer graphics". In: *Notes for Course* 30 (1996).

[79] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. "The fusiform face area: a module in human extrastriate cortex specialized for face perception". In: *Journal of neuroscience* 17.11 (1997), pp. 4302–4311.

[80] Andrej Karpathy et al. "Large-scale video classification with convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[81] Yury Kartynnik et al. "Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs". In: (2019).

[82] Liad Kaufman, Dani Lischinski, and Michael Werman. "Content-Aware Automatic Photo Enhancement". In: *Computer Graphics Forum*. Vol. 31. 8. 2012, pp. 2528–2540.

[83] Leonid Keselman et al. "Intel realsense stereoscopic depth cameras". In: *CVPR Workshops*. 2017.

[84] Salman H Khan et al. "Automatic Shadow Detection and Removal from a Single Image". In: *IEEE TPAMI* (2015).

[85] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. "Deeply-recursive convolutional network for image super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[86] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate Image Super-Resolution Using Very Deep Convolutional Networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[87] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *ICLR*. 2015.

[88] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *ICLR*. 2015.

[89] Rudolf Kingslake. "The development of the zoom lens". In: *Journal of the SMPTE* (1960).

[90] Masahiro Kobayashi et al. "A low noise and high sensitivity image sensor with imaging and phase-difference detection AF in all pixels". In: *ITE Trans. on Media Technology and Applications* (2016).

[91] Naejin Kong, Yu-Wing Tai, and Joseph S Shin. "A physically-based approach to reflection separation: from physical modeling to constrained optimization". In: *PAMI* (2014).

[92] Martin Kraus and Magnus Strengert. "Depth-of-field rendering by pyramidal image processing". In: *CGF* (2007).

[93] Aravind Krishnaswamy and Gladimir VG Baranoski. "A biophysically-based spectral model of light interaction with human skin". In: *Computer Graphics Forum*. 2004.

[94] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[95] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. "Semi-supervised deep learning for monocular depth map prediction". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[96] Wei-Sheng Lai et al. "Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[97] Vuong Le et al. "Interactive facial feature localization". In: *Proceedings of the European Conference on Computer Vision*. 2012.

[98] Christian Ledig et al. "Photo-realistic single image super-resolution using a generative adversarial network". In: *arXiv preprint arXiv:1609.04802* (2016).

[99] Christian Ledig et al. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[100] Anat Levin and Yair Weiss. "User assisted separation of reflections from a single image using a sparsity prior". In: *IEEE PAMI* (2007).

[101] Anat Levin, Assaf Zomet, and Yair Weiss. "Learning to perceive transparency from the statistics of natural scenes". In: *NIPS*. 2003.

[102] Anat Levin, Assaf Zomet, and Yair Weiss. "Separating reflections from a single image using local features". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2004.

[103] Marc Levoy and Pat Hanrahan. "Light Field Rendering". In: (1996).

[104] Marc Levoy and Yael Pritch. *Portrait mode on the Pixel 2 and Pixel 2 XL smartphones*. Blog. 2017.

[105] Xin Li and Michael T. Orchard. "New edge-directed interpolation". In: *IEEE Trans. Image Processing* (2001).

[106] Yu Li and Michael S Brown. "Exploiting reflection change for automatic reflection removal". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013.

[107] Yu Li and Michael S Brown. "Single image layer separation using relative smoothness". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[108] Zhengqi Li and Noah Snavely. "MegaDepth: Learning Single-View Depth Prediction from Internet Photos". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[109] Orly Liba et al. "Handheld mobile photography in very low light". In: *ACM Transactions on Graphics (TOG)* 38.6 (2019), pp. 1–16.

[110] Bee Lim et al. "Enhanced deep residual networks for single image super-resolution". In: *CVPR Workshops*. 2017.

[111] Nian Liu et al. "Learning to predict eye fixations via multiresolution convolutional neural networks". In: *IEEE trans. on neural networks and learning systems* (2018).

[112] Li-Qian Ma et al. "Appearance harmonization for single image shadow removal". In: *Computer Graphics Forum*. 2016.

[113] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. "Unsupervised video summarization with adversarial lstm networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[114] Pascal Mamassian, David C Knill, and Daniel Kersten. "The perception of cast shadows". In: *Trends in cognitive sciences* 2.8 (1998), pp. 288–295.

[115] George Mather. "Image Blur as a Pictorial Depth Cue". In: *Proc. Biological Sciences* (1996).

[116] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. "The contextual loss for image transformation with non-aligned data". In: *Proceedings of the European Conference on Computer Vision*. 2018.

[117] Ben Mildenhall et al. "Burst Denoising with Kernel Prediction Networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[118] Atsushi Morimitsu et al. *A 4M pixel full-PDAF CMOS image sensor with 1.58 $\mu$m 2$\times$ 1 On-Chip Micro-Split-Lens technology*. Tech. rep. 2015.

[119] Junichi Nakamura. *Image sensors and signal processing for digital still cameras*. CRC press, 2017.

[120] S. K. Nayar and Y. Nakagawa. "Shape from focus". In: *PAMI* (1994).

[121] Ren Ng et al. *Light Field Photography with a Hand-held Plenoptic Camera*. Tech. rep. 2005.

[122] Simon Niklaus and Feng Liu. "Context-aware synthesis for video frame interpolation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[123] Simon Niklaus et al. "3D Ken Burns effect from a single image". In: *ACM TOG* (2019).

[124] Abhijit S Ogale, Cornelia Fermuller, and Yiannis Aloimonos. "Motion segmentation using occlusions". In: *PAMI* (2005).

[125] Andrew Owens and Alexei A Efros. "Audio-Visual Scene Analysis with Self-Supervised Multisensory Features". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[126] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[127] Jinsun Park et al. "A unified approach of multi-scale deep and hand-crafted features for defocus estimation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[128] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. "Super-resolution image reconstruction: a technical overview". In: *IEEE signal processing magazine* 20.3 (2003), pp. 21–36.

[129] Deepak Pathak et al. "Learning Features by Watching Objects Move". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[130] Alex Poms et al. "Scanner: Efficient Video Analysis at Scale". In: (2018).

[131] Michael Potmesil and Indranil Chakravarty. "Synthetic image generation with a lens and aperture camera model". In: *ACM Transactions on Graphics (TOG)* (1982).

[132] Liangqiong Qu et al. "DeshadowNet: A Multi-Context Embedding Deep Network for Shadow Removal". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[133] Ravi Ramamoorthi and Pat Hanrahan. "A Signal-Processing Framework for Inverse Rendering". In: *SIGGRAPH*. 2001.

[134] Rolf Reber, Norbert Schwarz, and Piotr Winkielman. "Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?" In: *Personality and social psychology review* 8.4 (2004), pp. 364–382.

[135] Ronald A Rensink and Patrick Cavanagh. "The Influence of Cast Shadows on Visual Search". In: *Perception* (2004).

[136] Jerome Revaud et al. "Deepmatching: Hierarchical deformable dense matching". In: *IJCV* (2016).

[137] Anna Rohrbach et al. "A Dataset for Movie Description". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[138] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *MICCAI*. 2015.

[139] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *IJCV* (2015).

[140] Mehdi SM Sajjadi, Bernhard Schölkopf, and Michael Hirsch. "EnhanceNet: Single image super-resolution through automated texture synthesis". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[141] Bernard Sarel and Michal Irani. "Separating transparent layers through layer information exchange". In: *Proceedings of the European Conference on Computer Vision* (2004).

[142] Imari Sato, Yoichi Sato, and Katsushi Ikeuchi. "Illumination from Shadows". In: *IEEE TPAMI* (2003).

[143] Daniel Scharstein and Richard Szeliski. "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms". In: *IJCV* (2002).

[144] Yoav Y Schechner, Nahum Kiryati, and Ronen Basri. "Separation of transparent layers using focus". In: *IJCV* (2000).

[145] Eli Schwartz, Raja Giryes, and Alexander M. Bronstein. "DeepISP: Toward Learning an End-to-End Image Processing Pipeline". In: *IEEE Trans. Image Processing* (2019).

[146] Sidney A Self. "Focusing of spherical Gaussian beams". In: *Applied optics* 22.5 (1983), pp. 658–661.

[147] Soumyadip Sengupta et al. "SfSNet: Learning Shape, Refectance and Illuminance of Faces in the Wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[148] Amnon Shashua and Tammy Riklin-Raviv. "The Quotient Image: Class-Based Re-Rendering and Recognition with Varying Illuminations". In: *IEEE TPAMI* (2001).

[149] YiChang Shih et al. "Reflection removal using ghosting cues". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[150] YiChang Shih et al. "Style Transfer for Headshot Portraits". In: *SIGGRAPH* (2014).

[151] Yael Shor and Dani Lischinski. "The shadow meets the mask: Pyramid-based shadow removal". In: *Computer Graphics Forum*. 2008.

[152] Zhixin Shu et al. "Portrait lighting transfer using a mass transport approach". In: *SIGGRAPH*. 2017.

[153] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *ICLR*. 2015.

[154] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *ICLR* (2015).

[155] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild". In: *arXiv preprint arXiv:1212.0402* (2012).

[156] Ofer Springer and Yair Weiss. "Reflection separation using guided annotation". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 1192–1196.

[157] Pratul P Srinivasan et al. "Aperture Supervision for Monocular Depth Estimation". In: (2018).

[158] Chao Sun et al. "Automatic reflection removal using gradient intensity and motion cues". In: *Proceedings of the 2016 ACM on Multimedia Conference*. 2016.

[159] Meijun Sun et al. "SG-FCN: A Motion and Memory-Based Deep Learning Model for Video Saliency Detection". In: *IEEE Trans. on Cybernetics* (2018).

[160] Tiancheng Sun et al. "Single Image Portrait Relighting". In: *SIGGRAPH* (2019).

[161] Jaeyong Sung et al. "Unstructured human activity detection from rgbd images". In: *ICRA*. 2012.

[162] S. Suwajanakorn, C. Hernandez, and S. M. Seitz. "Depth from focus with your mobile phone". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[163] Richard Szeliski, Shai Avidan, and P Anandan. "Layer extraction from multiple images containing reflections and transparency". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2000.

[164] Huixuan Tang et al. "Depth from Defocus in the Wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[165] Michael W. Tao et al. "Depth from Combining Defocus and Correspondence Using light-Field Cameras". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013.

[166] Hui Tian. "Noise analysis in CMOS image sensors". PhD thesis. Stanford University, 2000.

[167] Carlo Tomasi and Roberto Manduchi. "Bilateral filtering for gray and color images". In: *Proceedings of the IEEE International Conference on Computer Vision*. 1998.

[168] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. "Deep image prior". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[169] Carl Vondrick et al. "Tracking emerges by colorizing videos". In: *Proceedings of the European Conference on Computer Vision*. 2018.

[170] Neal Wadhwa et al. "Synthetic Depth-of-Field With A Single-Camera Mobile Phone". In: *ACM Transactions on Graphics (TOG)* (2018).

[171] Renjie Wan et al. "Benchmarking Single-Image Reflection Removal Algorithms". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[172] Renjie Wan et al. "Depth of field guided reflection removal". In: *ICIP*. 2016.

[173] Jifeng Wang, Xiang Li, and Jian Yang. "Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[174] Lijun Wang et al. "DeepLens: shallow depth of field from a single image". In: *ACM Transactions on Graphics (TOG)" (2018).*

[175] Limin Wang, Yu Qiao, and Xiaoou Tang. "Action recognition with trajectory-pooled deep-convolutional descriptors". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[176] Ting-Chun Wang et al. "Light field video capture using a learning-based hybrid imaging system". In: *ACM Transactions on Graphics (TOG)* (2017).

[177] Wenguan Wang et al. "Revisiting Video Saliency: A Large-scale Benchmark and a New Model". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[178] Xintao Wang et al. "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks". In: *Proceedings of the European Conference on Computer Vision*. 2018.

[179] Xintao Wang et al. "Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[180] Zhou Wang et al. "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE TIP* (2004).

[181] Bennett Wilburn et al. "High Performance Imaging Using Large Camera Arrays". In: (2005).

[182] David Wong. *Opening Scene Godfather*. `https://youtu.be/OIBpHO1gZgQ`. Feb. 2014.

[183] Bartlomiej Wronski et al. "Handheld multi-frame super-resolution". In: *ACM Transactions on Graphics (TOG)* 38.4 (2019), pp. 1–18.

[184] Tai-Pang Wu et al. "Natural shadow matting". In: *ACM TOG* (2007).

[185] Tianfan Xue et al. "A computational approach for obstruction-free photography". In: *ACM Trans. Graph.* 34.4 (2015).

[186] Yuichiro Yamashita and Shigetoshi Sugawa. "Intercolor-Filter Crosstalk Model for Image Sensors With Color Filter Array". In: *IEEE Trans. on Electron Devices* (2018).

[187] Wenming Yang et al. "Deep learning for single image super-resolution: A brief review". In: *IEEE Transactions on Multimedia* 21.12 (2019), pp. 3106–3121.

[188] Yang Yang et al. "Virtual DSLR: High Quality Dynamic Depth-of-Field Synthesis on Mobile Platforms". In: *Digital Photography and Mobile Imaging*. 2016.

[189] Raymond A Yeh et al. "Semantic image inpainting with deep generative models". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[190] Fisher Yu and Vladlen Koltun. "Multi-Scale Context Aggregation by Dilated Convolutions". In: *ICLR*. 2016.

[191] Zhan Yu et al. "Dynamic Depth of Field on Live Video Streams: A Stereo Solution". In: *CGI*. 2011.

[192] Ke Zhang et al. "Video summarization with long short-term memory". In: *Proceedings of the European Conference on Computer Vision*. 2016.

[193] Ling Zhang et al. "Effective Shadow Removal Via Multi-Scale Image Decomposition". In: *The Visual Computer* (2019).

[194] Richard Zhang et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[195] Richard Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[196] Xuaner Cecilia Zhang et al. "Portrait Shadow Manipulation". In: *ACM Transactions on Graphics (TOG)* 39.4 (2020).

[197] Xuaner Zhang, Ren Ng, and Qifeng Chen. "Single Image Reflection Removal with Perceptual Losses". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[198] Xuaner Zhang, Ren Ng, and Qifeng Chen. "Single image reflection separation with perceptual losses". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[199] Xuaner Zhang et al. "Synthetic defocus and look-ahead autofocus for casual videography". In: *ACM Transactions on Graphics (TOG)* 38.4 (2019), pp. 1–16.

[200] Xuaner Zhang et al. "Zoom to learn, learn to zoom". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3762–3770.

[201] Yulun Zhang et al. "Residual dense network for image super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[202] Hang Zhao et al. "Loss Functions for Neural Networks for Image Processing". In: *IEEE Trans. Computational Imaging* (2017).

[203] Quanlong Zheng et al. "Distraction-aware shadow detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019, pp. 5167–5176.

[204] Hao Zhou et al. "Deep Single-Image Portrait Relighting". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

[205] Ruofan Zhou, Radhakrishna Achanta, and Sabine Süsstrunk. "Deep residual network for joint demosaicing and super-resolution". In: *Color and Imaging Conference*. 2018.

[206] Lei Zhu et al. "Bidirectional Feature Pyramid Network with Recurrent Attention Residual Modules for Shadow Detection". In: *Proceedings of the European Conference on Computer Vision*. 2018.

[207] Michael Zollhöfer et al. “State of the Art on 3D Reconstruction with RGB-D Cameras”. In: *Computer Graphics Forum*. 2018.