UNIVERSITY OF CALIFORNIA SAN DIEGO

Fluorescence, Scattering and Refraction in Computer Vision, with a Taste of Deep Learning

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Zachary Murez

Committee in charge:

David Kriegman, Chair
Ravi Ramamoorthi, Co-Chair
Jules Jaffe
Jurgen Schulze
Zhuowen Tu

2018

The dissertation of Zachary Murez is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2018

DEDICATION

I dedicate this to my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work would not have been possible without all the help and support I have received throughout the years. Thanks go first and foremost to my advisor David Kriegman for taking me under his wing and training me to be an independent researcher. His guidance and encouragement throughout the process (both the ups and the downs) was invaluable. I greatly appreciate his going above and beyond what was required of an adviser while on leave of absence (all those late night skype meetings). I also want to thank my co-adviser Ravi Ramamoorthi for taking me in during David's physical absence. He helped focus my research, allowing me to be more productive and get more work published. His dedication and insight helped bring me to the finish line.

Next I wish to acknowledge my co-authors and collaborators. I am grateful to Tali Treibitz for the hands-on help getting my research career started and showing me the ropes around the lab. Without her help we never would have manged to go from the initial conception of the idea of shape from fluorescence to ECCV submission in under a month. I wish to thank Zhengqin Li for collaborating with me on the water project and Manmohan Chandracker for his mentorship. I thank the entire viscomp lab for their friendship and the great work environment (particularly Sam Kwak for always keeping the lab interesting). And thanks to my friends for always being there for me.

Last, but not necessarily least, thanks to my family, without whose support I could not have done it. Thanks to my sister Andi for being the best sister a brother could ask for. Thanks to my dad Jim for getting me into technology at a young age, and continually giving me crazy ideas that sometimes turned into good ones. Thanks also for all the suggestions and help constructing many of the experimental setups necessary to perform the experiments in this thesis. And thanks to my mom Melanie for always loving and encouraging me, and for proof-reading much of my work.

The work was supported by NSF grant ATM-0941760, ONR grant N00014-15-1-2013, W.M. Keck Foundation, and by the UC San Diego Center for Visual Computing. We gratefully

acknowledge the support of NVIDIA Corporation with the donation of a Titan X Pascal GPU used for this research.

This dissertation is based on the following published papers, which were co-authored with others:

- Chapter 2 is a reformatted version of "Shape from Fluorescence," T. Treibitz, Z. Murez, B. G. Mitchell, D. Kriegman, *European Conference for Computer Vision (ECCV)* 2012 [TMMK12]. The dissertation author was the primary investigator and author of this paper.

- Chapter 3 is a reformatted version of "Photometric Stereo in a Scattering Medium," Z. Murez, T. Treibitz, D. Kriegman, R. Ramamoorthi, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2016 [MTRK17]. The dissertation author was the primary investigator and author of this paper.

- Chapter 4 is a reformatted version of "Learning to See through Turbulent Water", Z. Li, Z. Murez, D Kriegman, R. Ramamoorthi, M. Chandraker, *IEEE Winter Conf. on Applications of Computer Vision (WACV)* 2018 [LMK$^+$18]. The dissertation author was the primary investigator and author of this paper.

- Chapter 5 is a reformatted version of "Image to Image Translation for Domain Adaptation", Z. Murez, S. Kolouri, D Kriegman , R. Ramamoorthi, K. Kim, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2018. [MKK$^+$17] The dissertation author was the primary investigator and author of this paper.

VITA

2011	B. S. in Mathematics , Yale University
2011	B. S. in Computer Science , Yale University
2015	M. S. in Computer Science , University of California, San Diego
2018	Ph. D. in Computer Science, University of California, San Diego

PUBLICATIONS

Z. Murez, S. Kolouri, D Kriegman , R. Ramamoorthi, K. Kim "Image to Image Translation for Domain Adaptation", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2018.

Z. Li, Z. Murez, D Kriegman, R. Ramamoorthi, M. Chandraker "Learning to See through Turbulent Water", *IEEE Winter Conf. on Applications of Computer Vision (WACV)* 2018.

J. Tian, Z. Murez, T. Cui, Z Zhang, D Kriegman , R. Ramamoorthi "Depth and Image Restoration from Light Field In a Scattering Medium", *IEEE International Conference on Computer Vision (ICCV)* 2017.

Z. Murez, T. Treibitz, D. Kriegman, R. Ramamoorthi, "Photometric Stereo in a Scattering Medium," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2016.

Z. Murez, T. Treibitz, D. Kriegman, R. Ramamoorthi, "Photometric Stereo in a Scattering Medium," *IEEE International Conference on Computer Vision (ICCV)* 2015.

Veesler, David, et al. "Maximizing the potential of electron cryomicroscopy data collected using direct detectors." *Journal of structural biology 184.2* (2013): 193-202.

T. Treibitz, Z. Murez, B. G. Mitchell, D. Kriegman, "Shape from Fluorescence," *European Conference for Computer Vision (ECCV)* 2012.

ABSTRACT OF THE DISSERTATION

Fluorescence, Scattering and Refraction in Computer Vision, with a Taste of Deep Learning

by

Zachary Murez

Doctor of Philosophy in Computer Science

University of California San Diego, 2018

David Kriegman, Chair
Ravi Ramamoorthi, Co-Chair

Physics based vision attempts to model and invert light transport in order to extract information (such as 3D shape and reflectance properties) about a scene from one or more images. In order for the inversion of the model to be tractable, many simplifying assumptions about the physics are made that may or may not hold in practice.

On the other-hand, learning based vision ignores the underlying physics and instead models observations of the world statistically. A prime example of this is deep learning, which has recently revolutionized computer vision tasks such as classification, detection, and segmentation.

These two approaches to vision have traditionally been relatively disjoint, but are beginning to see some overlap. This thesis extends the state-of-the-art on both sides as well as brings them closer together.

First the novel use of imaging fluorescence for 3D reconstruction from shape from shading and photometric stereo is proposed. This is achieved by leveraging the previously unexploited fact that fluorescence emission is isotropic making it an ideal input for algorithms that assume Lambertian reflectance. In addition, fluorescence can be combined with reflectance to resolve the Generalized Bas relief ambiguity in uncalibrated photometric stereo. Furthermore, it is observed that when a material fluoresces a different color than it reflects, inter-reflections do not exist, which typically causes problems for photometric stereo.

Second, photometric stereo is extended to work in participating media by accounting for how scattering affects image formation. The first insight is that in this situation fluorescence can be used to optically remove backscatter which significantly improves the signal-to-noise ratio compared to image subtraction methods. Second, it is justified, through extensive simulations, that forward scatter from the light to the object can be calibrated out and effectively ignored. Finally, using deconvolution to handle forward scatter blur from the object to the camera, a phenomenon which is often ignored in computer vision, is proposed.

Next the problem of single image dynamic refractive distortion correction is tackled. Previous work has attacked this problem using physics based approaches and as such requires additional information, such as high frame rate video or templates, to handle its under-constrained nature. Instead, using deep learning to learn image and distortion priors which can be used to undistort a single image is proposed. The initial attempt to train the model using synthetically generated data failed to generalize to real data, so instead a special new large scale dataset for this problem was collected.

Finally, the failure to train the model using synthetic data prompted the investigation of domain adaptation. A novel framework for unsupervised domain adaptation building off the

ideas of adversarial discriminative feature matching and image-to-image translation is proposed. Many previous works can be seen as special cases of this general framework. The method is validated by achieving state-of-the-art results on common domain adaptation benchmarks, but may be particularly useful for traditionally physics based problems where synthetic data is easy to generate but real data is hard to annotate.

Chapter 1

Introduction

Physics based vision attempts to model and invert light transport in order to extract information (such as 3D shape and reflectance properties), about a scene from one or more images. In order for the inversion of the model to be tractable, many simplifying assumptions about the physics are made, that may or may not hold in practice.

On the other-hand, learning based vision ignores the underlying physics and instead models observations of the world statistically. A prime example of this is deep learning, which has recently revolutionized computer vision tasks such as classification, detection, and segmentation.

These two approaches to vision have traditionally been relatively disjoint, but are beginning to see some overlap. The work of this thesis advances the field from both sides as well as brings them a bit closer together. Chapters 2 through 4 examine a variety of physical effects that are often ignored in the image formation models employed in computer vision. Figure 1.1 shows a schematic diagram of these effects. Chapters 2 and 3 relax many of the assumptions made by traditional shape from shading and photometric stereo algorithms by analyzing the physics of fluorescence and scattering. In chapter 4 we examine the problem of single image dynamic refractive distortion correction. Here, since the problem is severely under-determined, we deviate from traditional physics based approaches, and propose a deep learning solution. In the process,

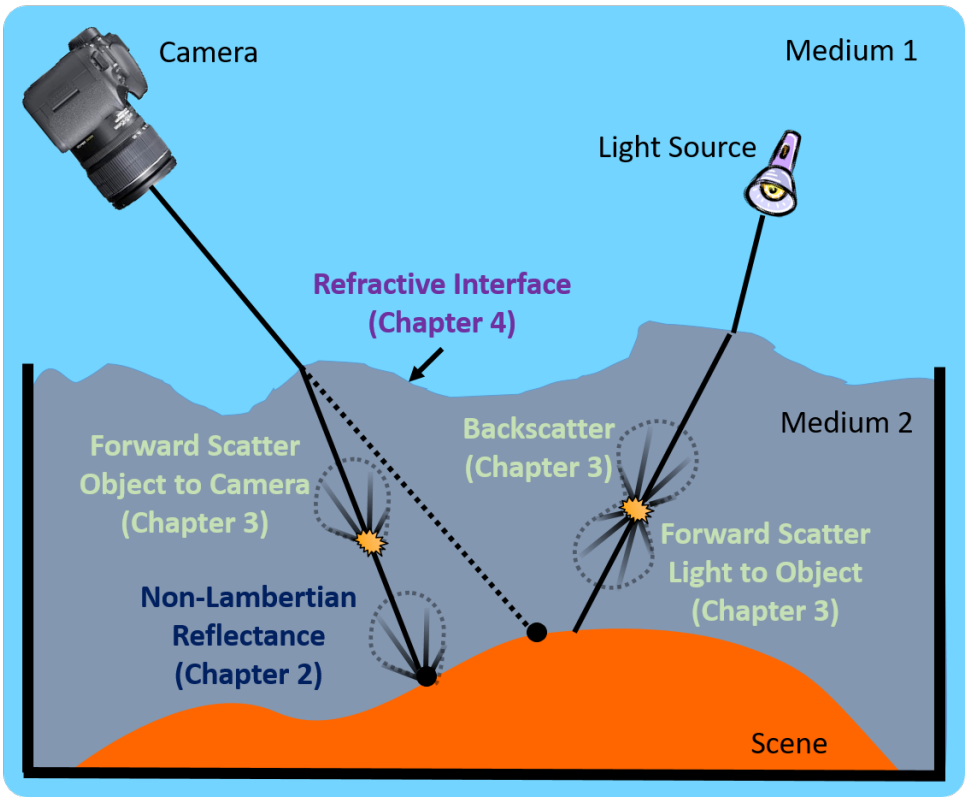

Figure 1.1: General image formation diagram. Often in computer vision it is assumed that there is no medium and that the surface reflects diffusely. In this thesis we examine cases of non-Lambertian reflectance, volumetric scattering within a medium and refraction between two media.

we realized that getting a deep network trained on synthetic data to generalize to real test data is difficult. Motivated by this, in Chapter 5 we propose a novel method for unsupervised domain adaptation. While this is a general tool for any domain shift, we believe it will be particular useful for physics based problems where synthetic data is easy to generate, but real data is hard for humans to annotate (for example people are bad at judging absolute reflectance).

Complementary to triangulation based methods for 3D reconstruction (such as binocular stereo and structure from motion), radiometric based methods (such as shape from shading and photometric stereo) recover 3D shape by analyzing how light is reflected by the surface. In particular, the brightness of a pixel can be related to the orientation of the surface normal with respect to the observation and illumination directions. While this relationship can be quite

complex in general, it is often assumed that the surface reflects light isotropically (also known as Lambertian reflectance) in which case the relationship simplifies to

$$L(x) = \rho(x) \max\left(0, \mathbf{D} \cdot \hat{\mathbf{N}}(x)\right) \tag{1.1}$$

where L is the measured pixel irradiance, ρ is the spatially varying albedo which is the fraction of light which is reflected by the material, $\mathbf{D}$ is the light direction vector scaled by its intensity, $\hat{\mathbf{N}}$ is the unit surface normal vector, and x is the pixel location. This relationship can be rewritten as

$$L(x) = \mathbf{D} \cdot \mathbf{N}(x) \tag{1.2}$$

where the albedo has been absorbed into the scaled surface normal $\mathbf{N}$ and we have assumed that the point is not in shadow allowing us to drop the $\max(0, \cdot)$. Now given three or more images under different but known light directions, $\mathbf{N}$ can found by solving the system of linear equations given by eq. 1.2. Once the normals are known, they can be integrated to recover the surface heights. This is known as photometric stereo and was origionally proposed by Woodham [Woo80]. Furthermore, when fewer than three images are given, the under-determined system can still be solved by imposing additional constraints, such as the surface being continuous and integrable, leading to the method of shape from shading.

These algorithms rely on many assumptions (which often do not hold in practice) to reduce the complex relationship between pixel intensity and surface normal direction to the simple one given by eq.1.2. In Chapter 2 we propose using fluorescent imaging to help relax some of these commonly made assumptions. Beyond day glow highlighters and psychedelic black light posters, it has been estimated that fluorescence is a property exhibited by 20% of objects. When a fluorescent material is illuminated with a short wavelength light, it re-emits light at a longer wavelength isotropically in a similar manner as a Lambertian surface reflects light. This hitherto neglected property opens the doors to using fluorescence to reconstruct 3D shape with some of

the same techniques as for Lambertian surfaces – even when the surface's reflectance is highly non-Lambertian.

Single image shape-from-shading and calibrated Lambertian photometric stereo can be applied to fluorescence images to reveal 3D shape. When performing uncalibrated photometric stereo, both fluorescence and reflectance can be used to recover Euclidean shape and resolve the generalized bas relief ambiguity. Finally for objects that fluoresce in wavelengths distinct from their reflectance (such as plants and vegetables), reconstructions do not suffer from problems due to inter-reflections. We validate these claims through experiments.

In Chapter 3 we extend the use of photometric stereo to scattering media such as turbid water, biological tissue and fog. Its use here has been limited until now, because of forward scattered light from both the source and object, as well as light scattered back from the medium (backscatter). Here we make three contributions to address the key modes of light propagation, under the common single scattering assumption for dilute media.

First, we show through extensive simulations that single-scattered light from a source can be approximated by a point light source with a single direction. This alleviates the need to handle light source blur explicitly. Next, we model the blur due to scattering of light from the object. We measure the object point-spread function and introduce a simple deconvolution method. Finally, we show how imaging fluorescence emission where available, eliminates the backscatter component and increases the signal-to-noise ratio. Experimental results in a water tank, with different concentrations of scattering media added, show that deconvolution produces higher-quality 3D reconstructions than previous techniques, and that when combined with fluorescence, can produce results similar to that in clear water even for highly turbid media.

In Chapters 2 and 3, we used physical models to extend radiometric based 3D reconstruction algorithms. In Chapter 4 we consider the problem of imaging through dynamic refractive media, such as looking into turbulent water, or through hot air. This is challenging since light rays are bent by unknown amounts leading to complex geometric distortions. Inverting these

distortions and recovering high quality images is an inherently ill-posed problem, leading previous works to require extra information such as high frame-rate video or a template image, which limits their applicability in practice.

Instead, we propose training a deep convolution neural network to undistort dynamic refractive effects using only a single image. The neural network is able to solve this ill-posed problem by learning image priors as well as distortion priors. Our network consists of two parts, a warping net to remove geometric distortion and a color predictor net to further refine the restoration. Adversarial loss is used to achieve better visual quality and help the network hallucinate missing and blurred information. Unlike prior works on water undistortion, our method is trained end-to-end, only requires a single image and does not use a ground truth template at test time.

Our first attempt to train the network with synthetically distorted data failed to generalize to the real test data, and is what motivated the work of Chapter 5. Instead, here we trained the network by generating our own semi-real dataset. Imagenet images were displayed on a monitor placed under a tank of water and re-imaged from above. Although images of a monitor under a tank are not exactly the same as images of a real scene, this produced real enough looking data to allow the network to generalize to images of real objects captured in the wild.

As observed above, training deep models on synthetic data often fails to generalize to real data. This is a specific case of the more general problem of domain shift, where a network trained on a training set fails to generalize to the test set due to differences in the data distributions between the two sets. In Chapter 5 we propose a general framework for unsupervised domain adaptation, which allows deep neural networks trained on a source domain to be tested on a different target domain without requiring any training annotations in the target domain. This is particularly useful for training networks to solve traditionally physics based problems, where synthetic data is easy to generate, but real data is hard, if not impossible, for a human to annotate.

This domain adaptation is achieved by adding extra networks and losses that help regularize the features extracted by the backbone encoder network. To this end we propose the novel use of the recently proposed unpaired image-to-image translation framework to constrain the features extracted by the encoder network. Specifically, we require that the features extracted are able to reconstruct the images in both domains. In addition we require that the distribution of features extracted from images in the two domains are indistinguishable. Many recent works can be seen as specific cases of our general framework. We apply our method for domain adaptation between MNIST, USPS, and SVHN datasets, and Amazon, Webcam and DSLR Office datasets in classification tasks, and also between GTA5 and Cityscapes datasets for a segmentation task. We demonstrate state of the art performance on each of these datasets.

Finally, in Chapter 6 we conclude with some known limitations of our methods and directions for future work.

Chapter 2

Shape From Fluorescence

When a material fluoresces, it absorbs light at a shorter wavelength and emits it at a longer wavelength. For example, when illuminated by blue light, Chlorophyll fluoresces red (Fig. 2.1). Minerals emit a wide variety of visible colors when illuminated by ultraviolet (UV) light. While it has been reported that 20% of materials fluoresce [Bar99], most models of color and reflectance in computer vision neglect fluorescence.

Radiometric methods for reconstructing shape such as single image shape-from-shading, photometric stereo, and passive photometric stereo from motion develop explicit models relating illumination, surface reflectance, and image irradiance. Central to these methods is the bidirectional reflectance distribution function (BRDF) which can be considered as being wavelength dependent. For example, in classic photometric stereo of surfaces with arbitrary reflectance, reflectance maps are either derived from the BRDF for each light source direction [Woo80] or measured from a reference object of the same material [Sil80]. Under Lambertian reflectance, the measured image irradiance is independent of the viewing direction. Then, surface normals are readily estimated from three or more images as the solution of a linear system of equations [Woo80], and a spatially varying albedo can also be estimated. In addition, many computer vision algorithms that rely on the constant brightness assumption (e.g., optical flow, dense stereo matching, space carving)

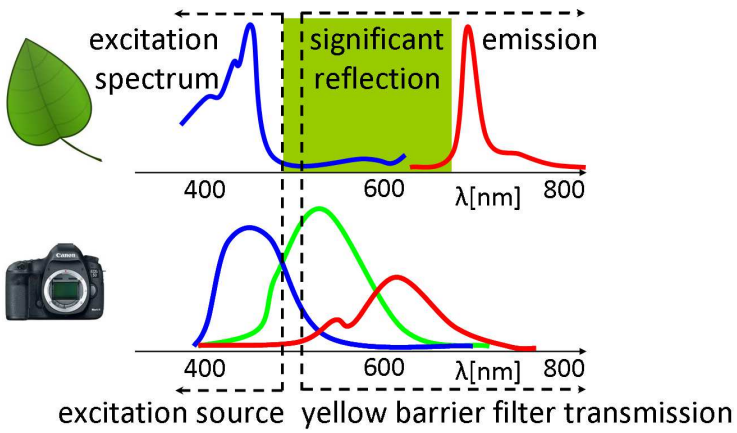

Figure 2.1: [Top] Chlorophyll-a excitation and emission spectra. The excitation peak is at 430nm and the emission peak is at 685nm. Wavelengths at the green range are not absorbed and thus reflected, giving the plants their prominent green color. [Bottom] Spectral sensitivities of the RGB channels in a Canon 5DII camera. The excitation spectrum has almost no overlap with the red camera channel. The yellow barrier filter over the lens enables imaging of green fluorescence (such as Green Fluorescent Proteins) without contamination of the excitation light. Although theoretically it seems that the yellow filter is not necessary for red fluorescence imaging, in practice, the excitation intensity is much stronger than the emission intensity, and so any leakage of longer wavelength of light from the blue source would reflect and appear in the red channel. The barrier filter reduces leakage and leads to correct exposure of the red channel.

implicitly or explicitly assume that surfaces are Lambertian. Unfortunately few materials are truly Lambertian except manufactured materials like Spectralon, and so one must often contend with non-Lambertian surfaces and this significantly complicates reconstruction compared to the simplicity and robustness of methods for Lambertian surfaces.

Until now, fluorescent materials were mainly used in computer vision and graphics to enhance visibility and contrast when imaging in scattering media [Gui90]. The property that the fluorescence color is distinct from the illumination color is used to optically filter out at the camera the wavelengths corresponding to the illuminant while allowing the wavelengths of the fluorescence emission from the object to pass. Fluorescence was also used to reconstruct transparent objects. Ihrke et al. [IGM05] added a fluorescent dye to water to enable reconstruction of flow from a video sequence. Hullin et al. [HFI$^+$08] immerse transparent objects in a fluorescent solution to enable range scanning techniques.

Interestingly, fluorescent emissions are almost always isotropic [Gla95], radiating energy equally in all directions, albeit at a different wavelength than the incident illumination. Consequently, it has the same behavior as an ideal Lambertian surface. In this chapter, we exploit this phenomenon to estimate shape from fluorescence in images. Contemporary with the origional publication of the paper related to this chapter, Sato et al. [SOS12a] presented similar contributions.

2.1 Image Formation

2.1.1 Reflectance

For simplicity of development, we consider an object surface point to be illuminated by a single light source from the incident direction in polar coordinates $(\theta_\mathrm{i}, \phi_\mathrm{i})$ with an incident radiance $L_\mathrm{i}(\lambda)$. The object is viewed by a camera from the direction $(\theta_\mathrm{r}, \phi_\mathrm{r})$, also in polar coordinates. The amount of light reflected from the object point towards the camera [Sze10] is expressed by

$$L_r(\lambda) = L_i(\lambda) F(\lambda, \boldsymbol{\theta}) \cos\theta_\mathrm{i} \ , \tag{2.1}$$

where $F(\lambda, \boldsymbol{\theta})$, is the BRDF at this object point, and $\boldsymbol{\theta} = (\theta_\mathrm{i}, \phi_\mathrm{i}, \theta_\mathrm{r}, \phi_\mathrm{r})$, are incident and viewing directions in spherical angles relative to a local coordinate system defined by the surface normal. The $\cos\theta_\mathrm{i}$ term is a foreshortening factor, as the exposed surface area decreases as the angle between the surface normal and illumination direction increases.

The intensity of a pixel in color channel c with sensitivity $z_c(\lambda)$ is

$$I_c = k(\gamma) \cos\theta_\mathrm{i} \int_\Lambda z_c(\lambda) L_\mathrm{i}(\lambda) F(\lambda, \boldsymbol{\theta}) d\lambda \ , \tag{2.2}$$

where Λ is the range of visible wavelengths. Here k is the ratio between image irradiance and scene radiance [Hor86], which depends on the effective f-number, lens transfer function, etc. In addition, k depends on γ, the angle between the projected ray to the optical axis [Hor86]. Most algorithms reconstructing shape from illumination assume the camera is viewing all object points from the same direction (orthographic projection), so that $k(\gamma)$ becomes a constant, or estimate $k(\gamma)$ through calibration. Either way, from here on we assume the dependency on γ was corrected and treat k as a constant.

Typical BRDFs can be divided into two types of reflections. A *specular reflection* is mirror-like where most of the incident light at the surface is reflected and concentrated about the reflection angle $\theta_r = \theta_i$. On the other hand, a *diffuse* surface reflects light towards a wide range of angles. An ideal *Lambertian* surface reflects light uniformly towards all directions, and the BRDF is equal to the surface albedo

$$F(\lambda, \boldsymbol{\theta}) = F(\lambda) \ . \tag{2.3}$$

Plugging Eqs. (2.1,2.3) into Eq. (2.2) yields the image intensity for a Lambertian surface

$$I_c = a_c \vec{L} \cdot \vec{N} \ , \tag{2.4}$$

where $\vec{L}$ is a illumination direction scaled by the light source strength and $\vec{N}$ is the unit surface normal. The term a_c is the sensed color of the object in the channel c, as a function of the albedo, light spectrum and the channel sensitivity

$$a_c = k(\gamma) \int_\Lambda z_c(\lambda) L_i(\lambda) F(\lambda) d\lambda \ . \tag{2.5}$$

Ideal Lambertian reflectance hardly exists in nature. Thus, there is a wide literature on models of non-Lambertian reflectance that aim to characterize either very specific material

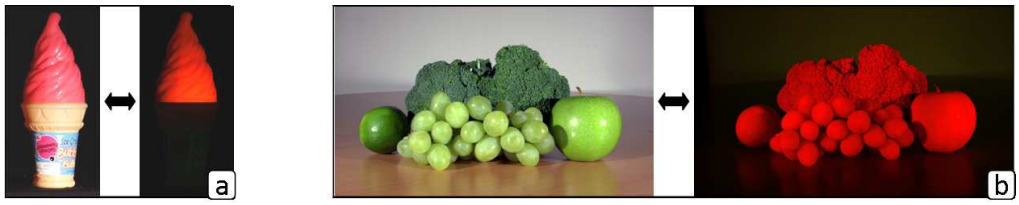

Figure 2.2: Common fluorescent objects. In each example, the color image on the left was acquired under white light whereas the image on the right was acquired under blue light with a yellow filter over the lens. (a) Plastic objects (as well as papers) often fluoresce under a wide excitation spectrum. They are made fluorescent at the same color they reflect to appear especially bright. The color of the "ice cream" on the right is due to fluorescence; since the cone does not fluoresce and only reflects, it appears black in the fluorescence image. (b) Green plants and vegetables contain Chlorophyll that fluoresces red. The image on the right was taken with a camera with a removed IR filter, for increased sensitivity for Chlorophyll fluorescence [TNR+12].

classes with a few parameters or a wide range of materials with many parameters. Some of the prominent ones are Oren-Nayar [ON95], Phong [Pho75], Cook-Torrance [CT82] and Ashikmin-Shirley [AS00].

2.1.2 Fluorescence

Stokes fluorescence, the common observed type of fluorescence, is the re-emission of photons having longer wavelengths than the absorbed photons [Gui90]. Any fluorescent molecule has two characteristic spectra. The excitation spectrum is the relative efficiency of different wavelengths of the exciting radiation to cause fluorescence. The emission spectrum is the relative intensity of radiation emitted at various wavelengths (example in Fig. 2.1). To image just the fluorescence, an excitation filter is mounted on a light source, or a narrow band light source such as a laser or LED is used. In addition, an emission filter is mounted on the camera. These filters are designed to have sharp boundaries, and to have negligible overlap in their transmittance spectra. Recently, Zhang and Sato [ZS11] proposed a method for separating reflectance and fluorescence by using two images illuminated by distinct colored light sources, that can have

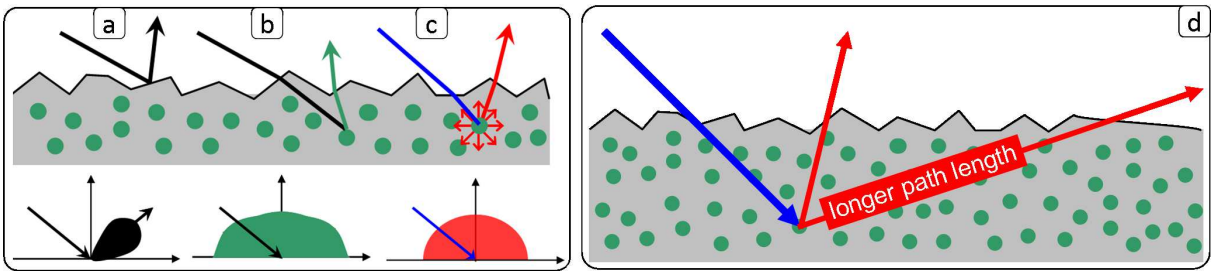

Figure 2.3: Scattering effects in a fluorescent object: (a) Specular reflections from the surface are concentrated in a specular lobe around the specular angle $\theta_r = \theta_i$ and often resemble the illumination color; (b) Light that penetrates the surface undergoes scattering and eventually part of it is reflected back to the viewer taking on the body color. The shape of the diffuse BRDF, which includes ideal Lambertian reflectance, varies among surfaces and depends on the surface roughness [ON95], among other factors; (c) Fluorescence emission of particles is often isotropic, and thus the surface emission is close to be ideally Lambertian; (d) Differences in optical path length from the fluorescence emission to the surface might harm isotropy due to absorption.

some overlap. In microscopy, narrow band filters are used to image various fluorescent molecules, that are later unmixed [ASW10] to yield the separate emitted fluorescence signals.

Fluorescence is more abundant in everyday life than is usually acknowledged, and examples of fluorescent objects are shown in Fig. 2.2. Many papers and plastics are made fluorescent to have a more striking color. Bleaching detergent is fluorescent and as a result, white papers and many cloths fluoresce as well. In nature, green fruits, vegetables and plants fluoresce due to Chlorophyll. This photosynthetic pigment strongly absorbs blue and red irradiance and reflects green. Part of the blue irradiance is converted to fluorescence in the red spectrum. Excitation and emission spectra of Chlorophyll are shown in Fig. 2.1. Corals fluoresce in red and green as they contain both Chlorophyll and Green Fluorescent Proteins. While less significant in every day life because of atmospheric attenuation, many objects fluoresce when excited by UV light. Many minerals fluoresce under UV in colors of red, orange, yellow, green, blue, violet, etc. Porphyrins in the human skin fluoresce under UV and are used for diagnosing dermatological conditions, while ”black light posters” are more entertaining. In addition, some fluorescent materials emit light in the infrared.

Within the graphics community, Glassner [Gla95] followed by Wilkie et al. [WWLP06] rendered fluorescent objects as diffuse, and Hullin et al. [HHA+10] claim that they are "weakly directional". These works explain this observation by the fact that the fluorescence emission does not originate from the surface, but from subsurface processes. In addition, it is important to realize that even before the subsurface scattering takes place, the angular distribution of fluorescence from particles itself was measured to be isotropic in most cases [KLK78, GVK93]. This isotropic emission from the particles manifests as isotropic emission from the surface if the fluorescent particles are uniformly spread in the object and are relatively close to the surface. The reflectance and scattering processes are demonstrated in Fig. 2.3. Incident light may specularly reflect off of the surface and the observed color is often that of the illuminant (a). Diffuse reflectance may not be ideally Lambertian (b), and the fluorescence emission (c) is observed to be closer to ideal Lambertian. For subsurface particles, differences in optical path length from the fluorescence emission to the surface may manifest as a non-isotropic fluorescence emission, due to absorption [CKSSM85] of the object material (d), especially in wider viewing angles. This means that not all fluorescence emissions are ideally Lambertian throughout the entire viewing range. Characterizing the nature of these cases requires future work. Here, following the above analysis, in the next section we provide an empirical demonstration of fluorescent emission that follows Lamberts law, and in the rest of the chapter we demonstrate how to use this insight for shape estimation.

2.2 Angular Dependency of Fluorescent Emission

In this section we verify the claim that the fluorescence emission is well approximated by an ideal Lambertian model through a simple experiment. We image a cylinder, illuminated with a distant collimated light source, such that the light source direction and strength is uniform across

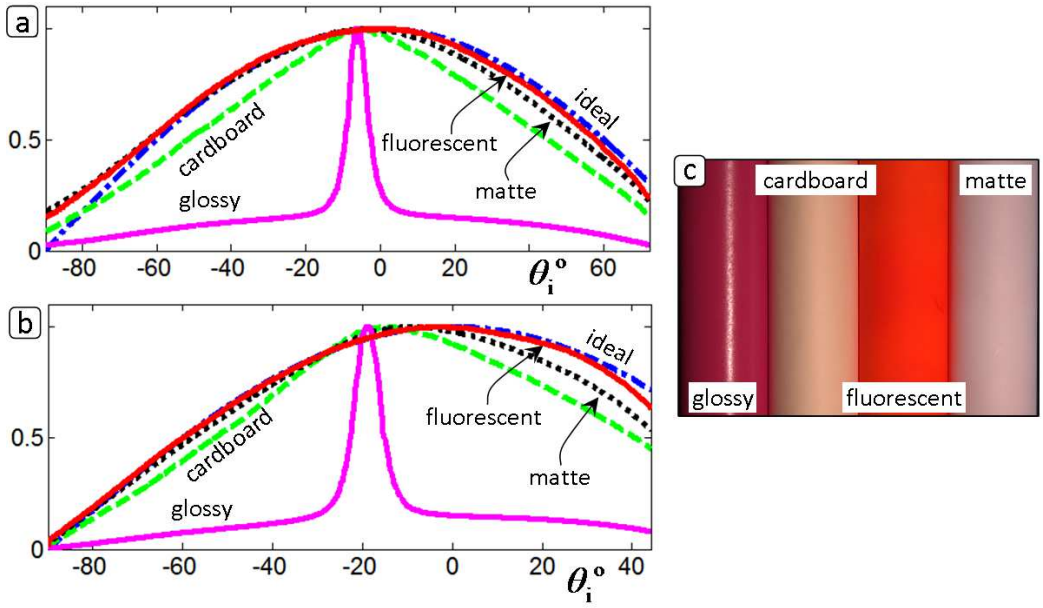

Figure 2.4: Examination of the angular dependency of the fluorescent emission: (a,b) The normalized intensity of several spray-painted paper sheets wrapped around a cylinder is plotted as a function of incident angle θ_i, for two light source positions. The incident angle was estimated using a spherical light probe. The specular reflection from the glossy paint indicates the location of the specular peak. The non-ideal diffuse peaks are shifted from the ideal Lambertian peak towards the specular direction. From all the measured surfaces, the fluorescent surface is the closest to the ideal reflection. Interestingly, the reflection from the cardboard is significantly less diffuse than that of the matte paint; (c) Images of the four materials. The cylinders are adjacent for display purposes, but the measurements were done separately to avoid inter-reflections.

the cylinder. The cylinder provides imaging of all incident directions in a single image. The light source direction is determined using a light probe (mirrored sphere) in the scene.

To compare reflectance properties, we spray-painted paper sheets with several types of paint: matte, glossy, and fluorescent. These were imaged under a few light source directions. Fig. 2.4 depicts the intensity plots as a function of θ_i for two distinct light source positions in addition to the expected ideal Lambertian. The normal direction at each pixel is obtained from its position on the cylinder and the known radius of the cylinder. Measurements along lines with constant surface were averaged to remove noise, and the brightness was normalized to have a consistent maximum intensity. At all measured light source directions, the fluorescent surface is closer to the ideal reflection than the other surfaces.

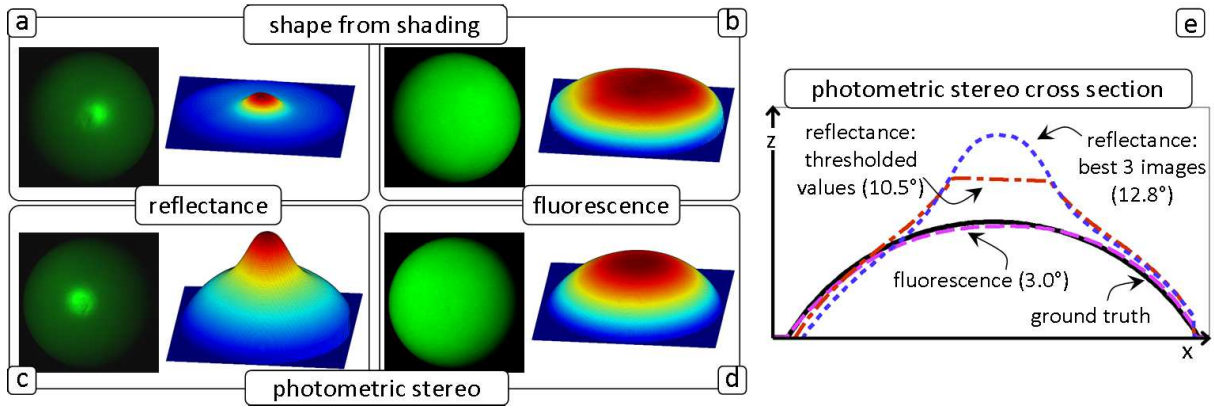

Figure 2.5: Shape reconstructions of a sphere spray painted with green fluorescent paint, using reflectance and fluorescence channels. In both setups, the green channel is used. (a,b) Reconstruction using shape from shading. The reconstruction from reflectance has a clear bump from specularities. In the reconstruction from fluorescence, the reconstruction is not ideal but closely resembles the shape of a sphere. (c,d) Reconstruction by photometric stereo. In the reconstruction from reflectance, for each pixel, we use the 3 frames that are the least bright in this pixel, to reduce the effect of the specular reflection. Still, the reconstruction from fluorescence is almost ideal, whereas the reconstruction from reflectance suffers from its non-Lambertian nature. Note that both images in (b,d) appear very diffuse. (e) Cross sections of the different reconstructions. Values in parenthesis depict the mean error in degrees of the normal angles in each reconstruction.

In all experiments reported in this chapter we use raw images acquired either with a Canon 1D mark IV or with a Canon 5DII illuminated with a Lowel pro-light tungsten halogen light source. The reflectance images were imaged with this setup as-is and all the fluorescence images were imaged using a blue filter on the light source and a yellow filter mounted on the lens acquired from NightSea LLC. For Chlorophyll fluorescence images, we use a Canon 5DII with a removed IR filter, for increased sensitivity for Chlorophyll fluorescence [TNR+12].

2.3 Shape from Shading

Eq. (2.4) is used in various methods for reconstructing 3D shape as it provides a simple relation between image intensity and object normals, provided the surface is indeed ideally Lambertian and the light source is distant. In shape from shading methods, a single image is

used to estimate the shape under the assumption of constant albedo [ZTCS99]. A unit normal at each pixel is defined by two unknowns, and Eq. (2.4) provides a single constraint. Then, various additional constraints, such as integrability and smoothness are applied to obtain a solution. There is also a bulk of literature on numerical methods for integrating the normals into a smooth surface [ZTCS99, DFS08]. The majority of the existing methods assume Lambertian reflectance and even then it is commonly acknowledged that results are usually unsatisfactory. A few methods consider non-Lambertian reflectance models, that are known a-priori, e.g. [AF06].

Here we show that the Lambertian-like fluorescence emission is an ideal input to shape from shading. We imaged a sphere spray-painted with green fluorescent paint. In each case (reflectance and fluorescence) the green image channel is the input to shape from shading. The reconstruction was performed using code based on [TS94], and the recovered shapes are shown in Fig. 2.5. The reflectance suffers from the specular reflection, whereas in the reconstruction from fluorescence, the reconstruction is not ideal but very closely resembles the shape of a sphere.

2.4 Calibrated Photometric Stereo

As stated in the previous section, shape from a single image usually produces unsatisfiable results due to its under-determined nature and the required constant albedo assumption. Therefore, in photometric stereo methods at least three images are taken [Woo80], in order to have a fully determined set of equations from the form of Eq. (2.4). The problem can be cast as a matrix equation, where for each pixel

$$\vec{I}_{k\times 1} = \vec{S}_{k\times 3} N_{3\times 1} \ . \tag{2.6}$$

Here S is a $k \times 3$ matrix, representing k distinct light source directions normalized by their intensity. The pixel values under each light source are stacked in $\vec{I}_{k\times 1}$. In the calibrated case, the light source directions $\vec{S}$ are known and thus $\vec{N}$ can be solved for directly from Eq. (2.6) given at least three linearly independent light source locations.

The majority of photometric stereo methods assumes Lambertian surfaces, while some works assume a different known reflectance function, explicit or parametric, e.g. [TD91]. There are also attempts to estimate spatially varying BRDFs [GCHS10]. Considerable effort is given to remove specularities in the images, usually assuming dichromatic BRDFs [SI94, ZMKB08a].

Here, we show how the fluorescence signal provides an ideal input to Eq. (2.6) and alleviates the need for more complex methods. In Fig. 2.5 we show the result of photometric stereo on a sphere spray painted with green fluorescent spray. In the reconstruction from reflectance, we tried two methods to reduce the effect of the non-Lambertian reflectance. In one, for each object point we used only the three frames where the corresponding pixel was the least bright. In the second, we manually thresholded the brightest areas, and only use the non-thresholded frames for each object point. A normal pointing in the z direction was assigned to object points that did not have at least three input frames after thresholding. Still, the reconstruction from fluorescence is much closer to the ground truth. The cross sections of the different reconstructions and the mean error of the reconstructed normal angles are depicted in Fig 2.5(e).

Reconstruction of more complex objects is shown in Fig. 2.6. We show reconstruction for a fluorescent plastic bottle (top) and for a (real) green bell pepper (bottom). In addition, Fig. 2.7 shows another reconstruction of a fluorescent toy squirt gun. All objects fluoresce as-is and were not painted (see Sec. 2.2). In all cases the reflectance image (left) has clear specularities that harm the reconstruction. The fluorescence images (right) appear very diffuse and indeed the reconstruction based on them does not suffer from problems common to non-ideal diffuse surfaces. For all objects the fluorescence image is taken from the red camera channel, and the reflectance is from the blue channel. The integration part in the reconstruction uses the method in [FC88].

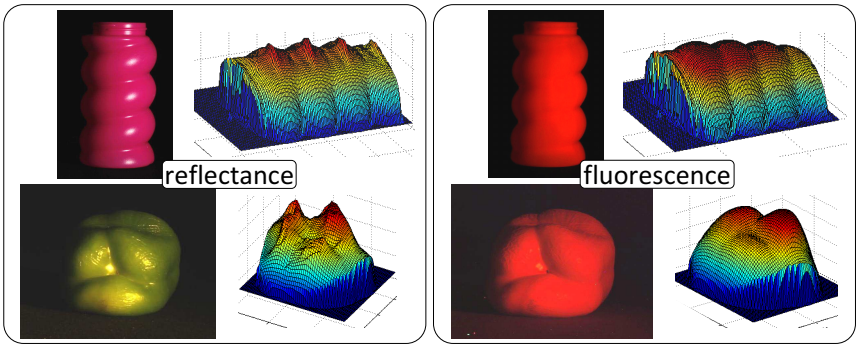

Figure 2.6: 3D reconstruction using photometric stereo applied to a fluorescent pink plastic bottle (top) and to a (real) green bell pepper (bottom). These objects fluoresce as-is and were not painted. [Left] The reflectance images have clear specularities that harm the reconstruction. [Right] The fluorescence images are very diffuse and thus the reconstruction based on them does not suffer from these problems. For all objects the fluorescence image was taken from the red camera channel, and the reflectance image was from the blue channel. Seven images were used to reconstruct the bottle while nine were used for the pepper.

2.5 Uncalibrated Photometric Stereo

The input to uncalibrated photometric stereo is a set of images of an object in fixed pose under unknown lighting conditions. i.e., no information about the light source strength or direction is needed [Hay94, YS97]. The basic equation describing the problem is then

$$\vec{I}_{k\times j} = \vec{S}_{k\times 3} N_{3\times j} \ , \tag{2.7}$$

where j is the number of pixels in the image. The unknown light source directions $\vec{S}$ are considered uniform for all j pixels in the image. Image intensity of all pixels under all light sources $\vec{I}_{k\times j}$ is used as an input, to simultaneously estimate $\vec{S}$ and the normal directions at every pixel, $N_{3\times j}$. When the object is Lambertian and assuming only local reflectance, Eq. (2.7) can be solved using SVD and applying integrability constraints up to a generalized bas relief (GBR)

tranformation [YS97, BKY99] in the form

$$G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \mu & \nu & \tau \end{pmatrix} , \tag{2.8}$$

where μ, ν and τ are the three degrees of freedom of the tranformation.

For strictly Lambertian surfaces, inter-reflections in concave regions can resolve the GBR [CKK05] as can heuristics like minimization of the entropy of albedo [AMK07]. However, both constraints are not always effective. Alternatively, the GBR can be resolved using isotropy together with Helmholtz reciprocity [TMQ$^+$07]. This method can only be accomplished when the Lambertian component of reflectance can be separated from the specular component (e.g., using polarization [NFB97] or the SUV color space under the dichromatic reflectance model [ZMKB08b]).

Here, we offer a method to resolve the GBR in color images when the fluorescence is in one color channel and other color channels contain reflectance images of a specular surface. The strategy is simple. Because the emitted radiance due to fluorescence behaves like a Lambertian surface, the surface can be reconstructed up to a GBR tranformation from the fluorescence channel of multiple images under unknown lighting using the method of Yuille and Snow [YS97]. The GBR can then be resolved from the specularities detected in the reflectance channel using the method of Drbohlav and Chaniler [DC05]. This method uses normals from at least two specular points to impose constraints on the reconstruction of the form

$$\vec{v} = 2(\vec{n} \cdot \vec{l})\vec{n} - \vec{l} , \tag{2.9}$$

where $\vec{v}$ is the viewing direction, $\vec{n}$ is the unknown normal at the specular point and $\vec{l}$ is the unknown light source direction at the specific image containing the specular point. The constraint

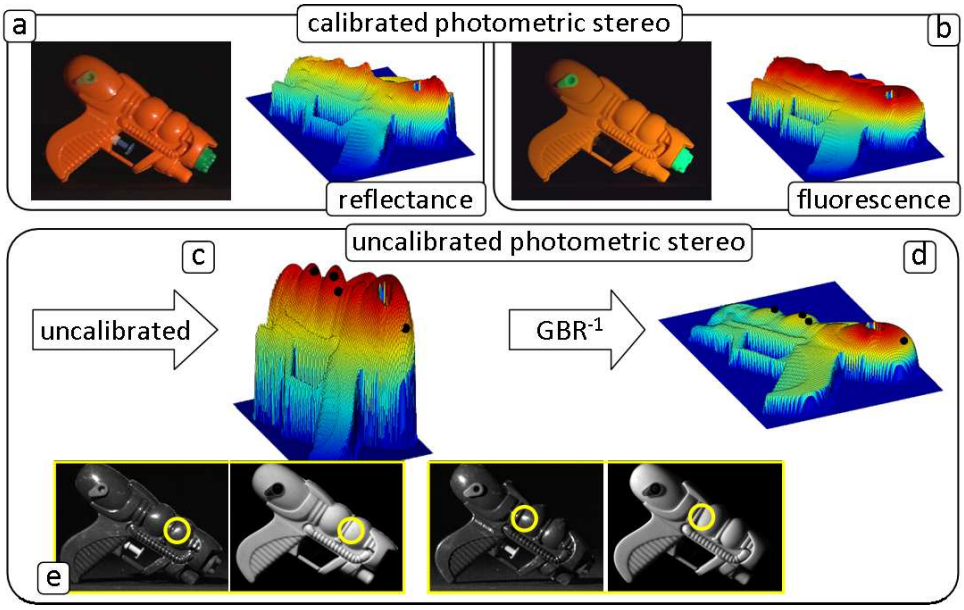

Figure 2.7: (a,b) Calibrated photometric stereo from fluorescence, demonstrated on a fluorescent orange toy squirt gun, using 12 images. The fluorescence signal is diffuse, providing a 3D reconstruction that does not suffer from bumps due to specularities. (c-e) Here we show how specular pixels from the *reflectance* image (originally not used for reconstruction) can be used to resolve GBR ambiguity in uncalibrated photometric stereo from the *fluorescence* channel. (c) The solution for the uncalibrated case, based on [YS97], using the input images used in (b). The shape of the squirt gun is apparent, but skewed with a GBR. Then, specular pixels are chosen (marked by black spheres). These pixels imply constraints that allow recovery of the correct GBR tranformation, following Eq. (2.9). The successfully transformed reconstruction is shown in (d). In (e) two examples of the actual input images (both reflectance and fluorescence) are shown and the specularity location is marked. Note that the reflection channel is used solely for locating the specularity while surface is densely reconstructed from the more stable fluorescence channel which does not exhibit specularities at these (or any) pixels.

in Eq. (2.9) is imposed on the solution of the uncalibrated step [YS97] to estimate the correct GBR tranformation.

In [DC05], the normals reconstructed from the specular pixels in the reflectance image are actually used in Eq. (2.9) to impose the constraint for surface reconstruction. Thus, they are prone to errors and to an unstable reconstruction. As opposed to that, in our case we have a clear advantage as the normals in the uncalibrated solution are obtained from the Lambertian *fluorescence* channel, and the *reflectance* channel is used solely to locate the specular pixels. We

show an example for this method in Fig. 2.7(c-e). The uncalibrated case is solved up to a GBR ambiguity, and then the tranformation is estimated based on specularities. In our experience, this method sometimes actually produces results that are better than the calibrated results due to better accuracies in estimating the light source directions.

2.6 Mutual Illumination in Reconstruction

Nearly all radiometric methods for shape reconstruction (shape-from-shading, photometric stereo) are based on local illumination models wherein the image irradiance is a function of the direct illumination from the light source onto an imaged surface patch and the patch properties. Unless methods to separate local and global illumination [NKGR06] are applied, the inter-reflections of light from other surfaces in the scene onto the patch are neglected and left unmodeled. However, when rendering scenes in computer graphics, global illumination methods achieve their realism precisely because they account for inter-reflections. Forsyth and Zisserman showed that the image intensity of a dihedral (where two planar faces meet and form a concavity) becomes brighter closer to the corner due to mutual illumination, and the increase of brightness is related to the albedos of the two faces [FZ89]. When performing reconstruction using shape-from-shading or photometric stereo, this increased brightness results in shallower reconstructed concavities than the true depth. Nayar et. al. [NIK90] called this the pseudo surface, and introduced an iterative method for estimating the true surface from the pseudo surface by estimating an inter-reflection kernel at each step.

Consider a simple concavity illuminated solely by wavelengths within its fluorescence excitation spectrum. Some of the incident light will reflect from the surface according to its BRDF, and some amount will excite fluorescence that will be emitted according to the material's emission spectrum. In general, the reflected light from the fluorescence emission can interreflect within the concavity. However, when the material *does not* reflect light at the wavelengths of

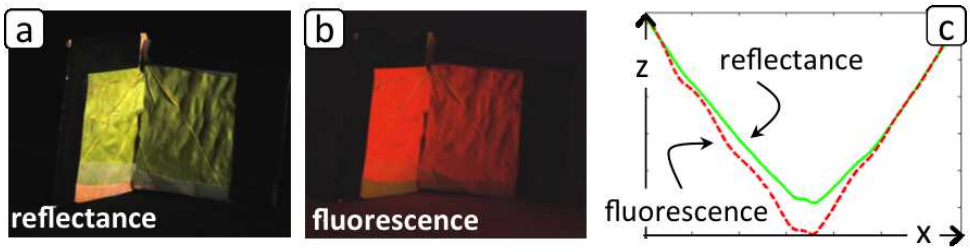

Figure 2.8: To demonstrate how fluorescence helps avoid mutual illumination, we covered a V-shaped cardboard with leaves. (a,b) The reflectance and the fluorescence images of the covered cardboard; (c) The shape was reconstructed using calibrated photometric stereo for both the reflectance and fluorescence channels. The reconstructed depth (at a certain cross section) is plotted for the reflectance and fluorescence images. While the reconstruction from the reflectance channel shows classical bending of the corner (pseudo surface) as described in [NIK90], the reconstruction from the fluorescence channel maintains sharp edges clearly demonstrating the reduction of the effect of inter-reflections.

the fluorescence emission, there will not be any subsequent interreflections of the fluorescence emissions. Consequently, reconstruction methods that only assume local Lambertian reflectance will be effective for images of such materials.

Contrary to [SOS12a], we found that most artificial fluorescent objects and materials reflect the color they fluoresce, as they are made fluorescent to appear brighter than they are, in the same color. However, many natural objects such as leaves, vegetables and fruits, fluoresce at different wavelengths than they reflect. For example, green leaves reflect green light but absorb red and blue light. When lit by blue light, the Chlorophyll in the leaves will fluoresce red (Fig. 2.1). Because the leaf absorbs red light, we do not expect appreciable brightening due to inter-reflections. Consequently, when performing reconstruction, local illumination models will be adequate to correctly estimate the depth, and methods such as [NKGR06, NIK90] should be unnecessary.

To demonstrate this claim on a simple case, we created a concave dihedral (V-shaped) out of cardboard and glued leaves to it, such that it is covered by the leaves. Using the reflectance and fluorescence channels separately, the shape was reconstructed using calibrated photometric stereo (example images shown in Fig. 2.8a,b, total images was 10). To clearly demonstrate the effect,

we plot a cross section of the reconstructed height for the two channels, (Fig. 2.8c). Whereas the reconstruction from the fluorescence channel shows a sharp corner, the reconstruction from the reflectance channel is shallower and the smooth rounding of the corner (pseudo surface) is similar to what is described in [NIK90].

2.7 Summary

Imaging fluorescence through inexpensive off-the-shelf filters opens an enchanting world of glowing objects and objects that alter their color appearance. It is surprising to discover that many every day objects fluoresce in the visible spectrum under UV or blue light and therefore can be imaged with conventional cameras. In this chapter we showed the benefit of using fluorescence for 3D reconstruction methods since fluorescence emissions are often isotropic much like ideal Lambertian reflectance. This insight can be used, for example, for 3D reconstruction of fluorescent corals underwater, where texture correspondences are sometimes problematic, given the water optical properties are taken into account. We showed that there is an advantage of imaging a fluorescent object over a matte object, as specularities from the reflectance channel can provide additional information regarding light source directions relative to the object. In addition, we showed how in some cases fluorescence can avoid mutual illumination problems.

Chapter 2 is a reformatted version of "Shape from Fluorescence," T. Treibitz, Z. Murez, B. G. Mitchell, D. Kriegman, *European Conference for Computer Vision (ECCV)* 2012 [TMMK12]. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Photometric Stereo in a Scattering Medium

In the previous chapter we examined how properties of fluorescence make it an ideal input for shape from shading and photometric stereo. In this chapter, we focus on photometric stereo, and extend its use to participating media, such as in fog, haze, water, or biological tissue. Obtaining 3D information in this case is difficult because of scattering [GRG+13, KDCS08, PPV08].

Unlike in air, in a scattering medium, light propagation is affected by scattering which degrades the performance of photometric algorithms unless accounted for. Distance dependent attenuation caused by the medium has been dealt with in the past [KFB92]. Here, our contributions lie in handling three scattering effects (Fig. 3.1), based on a single scatter model [SRNN05]: 1) light traveling *from the source to the object* is blurred due to forward scattering; 2) light traveling *from the object to the camera* is blurred due to forward scattering; 3) light traveling *from the source is scattered back towards the camera* without hitting the object. This is known as backscatter and is an additive component that veils the object. All these effects are distance dependent and thus depend on the object 3D surface: the property we aim to reconstruct. To

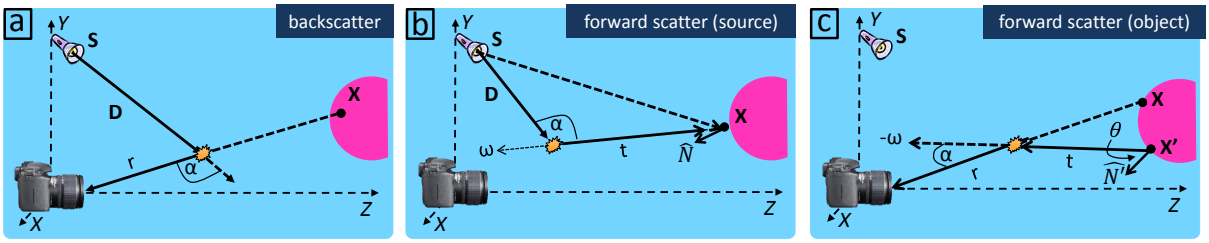

Figure 3.1: A perspective camera is imaging an object point at **X**, with a normal **N**, illuminated by a point light source at **S**. The object is in a scattering medium, and thus light may be scattered in the three ways shown, detailed in Sec. 3.2.

handle this we introduce the *small surface variations* approximation for the object (Sec. 3.2), that assumes surface changes are small relative to the distance from the object (that is assumed to be known). This assumption removes the dependence on the unknown surface heights Z, but unlike the common distant light/camera approximations, it still allows for dependencies on spatial locations X and Y. One important consequence of this is the ability to model anisotropic light sources, which is not possible for distant lights.

Forward scatter was previously compensated for iteratively for both pathways (light to object and object to camera) simultaneously [NZH02]. We analyze the paths separately. The resulting algorithm is simpler, requires fewer images and yields better results. First, consider the blurring of light traveling *to* the object from the source (Fig. 3.1b). The photometric stereo formulation assumes a point light source, illuminating from a single direction. However, if the source is scattered by the medium, this no longer holds: the point light source is spread, and the direction of light rays incident on the object changes. Nonetheless, for a Lambertian surface, illuminated from a variety of directions, we still get a linear equation between the image intensities and the surface normals [Sha92]. Here, we show through simulations in a large variety of single scattering media, that a forward-scattered light source illuminating a Lambertian surface can be well approximated by a non-blurred light source in an effective purely absorbing medium (Sec. 3.5). This allows for much easier calibration in practice.

Next, we observe that the blur caused by scattering *from* the object to the camera (Fig. 3.1c) significantly affects the shape of the surface reconstructed by photometric stereo. This important effect has been neglected in many previous works. In general, the point-spread function (PSF) for an object is spatially varying and dependent on the unknown scene depths. However, we demonstrate that a spatially invariant approximation can still achieve good results, when calibrated for the desired medium and approximate object distance. Although this means we must capture an additional calibration image for each medium and working distance, we do not believe this will be too cumbersome based on previous experience in the field. We estimate the PSF and use it to deconvolve the images after backscatter has been removed (Sec. 3.6). These corrected images are used as input to a linear photometric stereo algorithm to recover the surface normals, which are then integrated. This results in much higher quality 3D surfaces across varying turbidity levels.

Finally, consider the backscatter component (Fig. 3.1a). In a previous work (Tsiotsios et al. [TAKD14]) backscatter was calibrated and subtracted from the input images. This is similar in spirit to, and can in fact be used in conjunction with, ambient light subtraction, however underwater light is rapidly attenuated and thus ambient light is often minimal. However, when backscatter is strong relative to the object signal, subtracting it after image formation leads to lack of dynamic range and lower signal-to-noise ratios (SNR) [TS12] that significantly degrades deblurring and reconstruction. Here we show that if the object fluoresces, this can be leveraged to optically remove the backscatter prior to image formation (Sec. 3.7). Fluorescence is the re-emission of photons in wavelengths longer than the excitation light [Gui90], and therefore the backscatter can be eliminated by optically blocking the excitation wavelengths and imaging only the fluorescence emission. This improves SNR, especially in high turbidity. This approach is feasible as many natural underwater objects such as corals and algae fluoresce naturally.

We demonstrate our method experimentally in a water tank (Sec. 3.9) with varying turbidity levels. Deblurring can be used separately or combined with fluorescence imaging to significantly improve the quality of photometric stereo reconstructions.

3.1 Previous Work

The traditional setup for photometric stereo assumes a Lambertian surface, orthographic projection, distant light sources, and a non-participating medium [Woo80]. However, underwater light is exponentially attenuated with distance, and thus the camera and lights must be placed close to the scene for proper illumination. This means that the orthographic camera model, and the distant light assumptions, are no longer valid. In addition, attenuation and scattering by the medium need to be accounted for. These effects were partially considered in previous works.

Near-Field Effects and Exponential Attenuation: Photometric stereo in air was solved with perspective cameras [LK96, TK05], nearby light sources [KB91], or both [GT96, ISI90]. Kolagani et al. [KFB92] uses a perspective camera, nearby light sources and includes exponential attenuation of light in a medium. Their formulation leads to nonlinear solutions for the normals and heights. We handle these near field effects but linearize the problem (Sec. 3.2).

Photometric Stereo with Backscatter: Narasimhan et al. [NNSK05] handles backscatter and attenuation, with the assumption of distant light sources and an orthographic camera. Tsiotsios et al. [TAKD14] extends this to nearby point sources and assumes the backscatter saturates close to the camera and thus does not depend on the unknown surface height. Then, it can be calibrated and subtracted from the images. We use the method in [TAKD14] in one of our variants.

Backscatter Removal: Backscatter was previously removed for visibility enhancement, by structured light [GNS08], range-gating [KDCS08], or using polarizers [TS09]. Nevertheless, these methods do not necessarily preserve photometric information. It is sometimes possible to reduce backscatter by increasing the camera light source separation [GNS08, Jaf90], but this often leads to more shadowed regions, creating problems for photometric stereo.

Fluorescence Imaging: Removing scatter using fluorescence is used in microscopy [TT07], where many objects of interest are artificially dyed to fluoresce. Hullin et al. [HFI$^+$08] imaged objects immersed in a fluorescent liquid to reconstruct their 3D structure. It was recently shown

that the fluorescence emission yields photometric stereo reconstructions [SOS12b, TMMK12] in air that are superior to reflectance images as the fluorescence emission behaves like a Lambertian surface due to its isotropic emission.

Deblurring Forward Scatter: Zhang et al. [ZN02] and Negahdaripour et al. [NZH02] handle blur caused by forward scatter using the PSF derived in [Jaf90, McG80]. Their PSF depends on the unknown distances, as well as three empirical parameters, and affects both the path from the light source to the object and from the object to the camera. They iteratively deconvolve and update the depths until a good result is achieved. Trucco et al. [TOA06] simplify the PSF of [Jaf90, McG80] to only depend on two parameters while assuming the depth is known. Our PSF is nonparametric, independent of the unknown depths and only affects the path to the camera, which allows for a direct solution without iteration. While we look at a Lambertian surface in a scattering medium, Inoshita et al. [IMMY14] and Dong et al. [DMZP14] consider the problem of photometric stereo in air on a surface that exhibits subsurface scattering, which blurs the radiance across the surface. They deconvolve the images to improve the quality of the normals recovered using linear photometric stereo. Tanaka et al. [TMK$^+$15] also model forward scatter blur as a depth dependent PSF and combine it with multi (spatial) frequency illumination to recover the appearance of a small number of inner slices of a translucent material.

3.2 Overview and Assumptions

In this section we introduce the image formation model, considering each of the modes of light propagation in a single scattering medium, as shown in Fig. 3.1. We derive expressions for each component in the following sections.

Consider a perspective camera placed at the origin, with the image (x,y) coordinates parallel to the world's (X,Y) axes, and the Z-axis aligned with the camera's optical axis. Let the point $\mathbf{X} = (X,Y,Z)$ be the point on the object's surface along the line of sight of pixel $\mathbf{x} = (x,y)$.

Our Algorithm

Input:

3 or more images: L^i
source positions: $\mathbf{S}^i$
mean depth: $\bar{Z}$
backscatter images: L_b^i
scattering parameters: PSF and $\tilde{\sigma}$ (Sec. 3.8.3)

Output:

normals: N
surface heights: Z

1: **if** reflectance images **then**
2: subtract backscatter: $L^i - L_b^i$ (Sec. 3.4)
3: **end if**
4: deblur images: $L_o^i \leftarrow h^{-1} * (L^i - L_b^i)$ (Eq. 3.19)
5: solve linear PS: $N_j \leftarrow [\tilde{L}_j^{eq}]^{-1}[L_{o_j}]$ (Eq. 3.10,3.12)
6: integrate normals: $Z \leftarrow \int N$

Figure 3.2: Our Algorithm. Steps 2 and 4 are applied to each image i independently. Step 5 is applied to each pixel j independently with the data from each image i stacked into a matrix.

Let $\mathbf{S}$ be the world coordinates of a point light source, and define $\mathbf{D}(\mathbf{X}) = \mathbf{S} - \mathbf{X}$ as the vector from the object to the source.

We assume a single scattering medium which allows us to express the radiance L_o reflected by a surface point as the sum of two terms:

$$L_o(\mathbf{x}) = L_d(\mathbf{x}) + L_s(\mathbf{x}) \tag{3.1}$$

where L_d is the direct radiance from the source (Sec. 3.3), and L_s is the radiance from the source which is scattered from other directions onto $\mathbf{X}$ (Sec. 3.5 and Fig. 3.1b).

Next, we express the radiance arriving at the camera as the sum of three terms:

$$L(\mathbf{x}) = L_o(\mathbf{x})e^{-\sigma\|\mathbf{X}\|} + L_b(\mathbf{x}) + L_c(\mathbf{x}) \tag{3.2}$$

where L_o is the light reflected by the surface point $\mathbf{X}$ which arrives at the camera without undergoing scattering. Note that it is attenuated by $e^{-\sigma\|\mathbf{X}\|}$ where σ is the extinction coefficient. L_b is composed of rays of light emitted by the source that are scattered into $\mathbf{x}$'s line of sight before hitting the surface (Sec. 3.4 and Fig. 3.1a). This term is known as backscatter. Finally, L_c, is composed of rays of light reflected by other points on the surface that are scattered into pixel $\mathbf{x}$'s line of sight (Sec. 3.6 and Fig. 3.1c).

In order to write analytic expressions for these terms and derive a simple solution we make **two assumptions**. First, the surface is Lambertian with a spatially varying albedo $\rho(\mathbf{X})$. Second, we assume that surface variations in height are small compared to object distance from the camera. We call this the *small surface variations approximation* and note that it is weaker than the common distant light sources and orthographic projection approximations. Let $\bar{Z}$ be the average Z coordinate of the surface (assumed to be known). Then, the approximation claims that for every point on the surface: $|Z(\mathbf{X}) - \bar{Z}| \ll \bar{Z}, \forall \mathbf{X}$.

The approximation results in a weak perspective such that the projection $\mathbf{x}$ of $\mathbf{X}$ in the image plane is given by

$$\mathbf{x} = \left(f\frac{X}{\bar{Z}}, f\frac{Y}{\bar{Z}} \right)^t \quad ; \quad \mathbf{X} = \left(\frac{\bar{Z}}{f}x, \frac{\bar{Z}}{f}y, \bar{Z} \right)^t, \tag{3.3}$$

where f is the known focal length. Note that for a given pixel, since we know its (x,y) coordinates and the average object distance $\bar{Z}$, the world coordinates $\mathbf{X}$ are known. Specifically, $\mathbf{D}(\mathbf{X})$ is independent of the unknown object height Z but still depends on X and Y, whereas in the distant light sources approximation $\mathbf{D}(\mathbf{X})$ is a constant.

Outline of Our Method

Given an input image L we eliminate the backscatter L_b by one of two methods. The first follows [TAKD14]: backscatter from each light source is measured by imaging it with no objects

in the scene, and then the measured backscatter is subtracted from the input images. In the second, backscatter is optically eliminated using fluorescence as we explain in Sec. 3.7. Once backscatter is removed, the resulting images are deblurred, using a calibrated PSF, to recover L_o (Sec. 3.6, Eq. 3.19). Next we write L_o as a linear equation between the unknown surface normals, albedo and an equivalent light source (Sec. 3.5, Eq. 3.10), which we approximate as an effective point source in a purely absorbing medium with effective extinction coefficient (Sec. 3.5, Eq. 3.12). With a minimum of 3 images under distinct light locations the normals can be solved for, as in conventional photometric stereo. The normals are then integrated to recover a smooth surface. This algorithm is summarized in Fig. 3.2.

3.3 Direct Radiance

First, consider the direct reflected radiance from a Lambertian surface[Woo80]:

$$L_d(\mathbf{x}) = I(\mathbf{X})\frac{\rho(\mathbf{X})}{\pi}\hat{\mathbf{D}}(\mathbf{X})\cdot\hat{\mathbf{N}} \ , \tag{3.4}$$

where $\hat{\mathbf{N}}$ is the unit surface normal and $\hat{\mathbf{D}}$ is the normalized source-to-object vector. The radiance on the object surface $I(\mathbf{X})$ depends on the radiant intensity I_0 of the source in direction[1] $(-\hat{\mathbf{D}})$:

$$I(\mathbf{X}) = \Big(I_0\big(-\hat{\mathbf{D}}(\mathbf{X})\big)e^{-\sigma\|\mathbf{D}(\mathbf{X})\|}\Big)/\|\mathbf{D}(\mathbf{X})\|^2 \ . \tag{3.5}$$

Eq. 3.5 accounts for nearby angularly-varying sources, exponential attenuation along the optical path length with extinction coefficient σ, and inverse-square distance falloff.

[1] the direction is negative as we consider outgoing rays from the source.

Can Distance-Dependent Falloff Be Neglected?

We have introduced a near-field source, but often, in photometric stereo, the distant light source assumption is used, as it simplifies the mathematical development, computation of shape, and calibration of implemented systems because the light source direction can be treated as a constant, and the incident irradiance does not depend upon depth allowing it to also be treated as a constant. In this section, we explore whether similar simplifications are possible for photometric stereo in a medium.

In a medium, the incident irradiance falls off as function of distance due to the product of two factors; free space falloff and medium attenuation. Consider the situation depicted in Fig. 3.3a for which two points P and P_Δ are illuminated by a light source that is respectively at distances d and $d+\Delta$ from the points. The ratio Ψ of the irradiance at P and P_Δ can be expressed as $\Psi = E_{P_\Delta}/E_P = \Psi_{\text{freespace}}\Psi_{\text{medium}}$. The irradiance of a point light source propagating in a medium at distance d, falls off by $1/d^2$, and is attenuated by the medium by $e^{-\beta d}$. Thus,

$$\Psi_{\text{freespace}} = d^2/(d+\Delta)^2 \quad , \quad \Psi_{\text{medium}} = e^{-\beta\Delta} \quad . \tag{3.6}$$

Interestingly, $\Psi_{\text{freespace}}$ depends on both the absolute distance d to the light source as well as Δ, while Ψ_{medium} is independent of d, and depends only on the path difference Δ and on the attenuation coefficient β. To get an idea of object dimensions where the variation in incident irradiance is small and might be treated as constant, let us consider an example where the path difference Δ yields $\Psi = 0.9$ (i.e., a 10% difference in the incident irradiance at P and P_Δ). Figure 3.3b shows a plot of Δ vs d where $\Psi_{\text{freespace}} = 0.9$, and we see that the object's size can increase linearly with distance. Figure 3.3b shows a plot of Δ vs β the attenuation parameter where $\Psi_{\text{medium}} = 0.9$, and we see that as the medium becomes murkier the size decreases and is independent of distance. In other words, the opportunity to neglect distance-dependent falloff of lighting in a medium depends upon the clarity of the medium, even when the distance is large

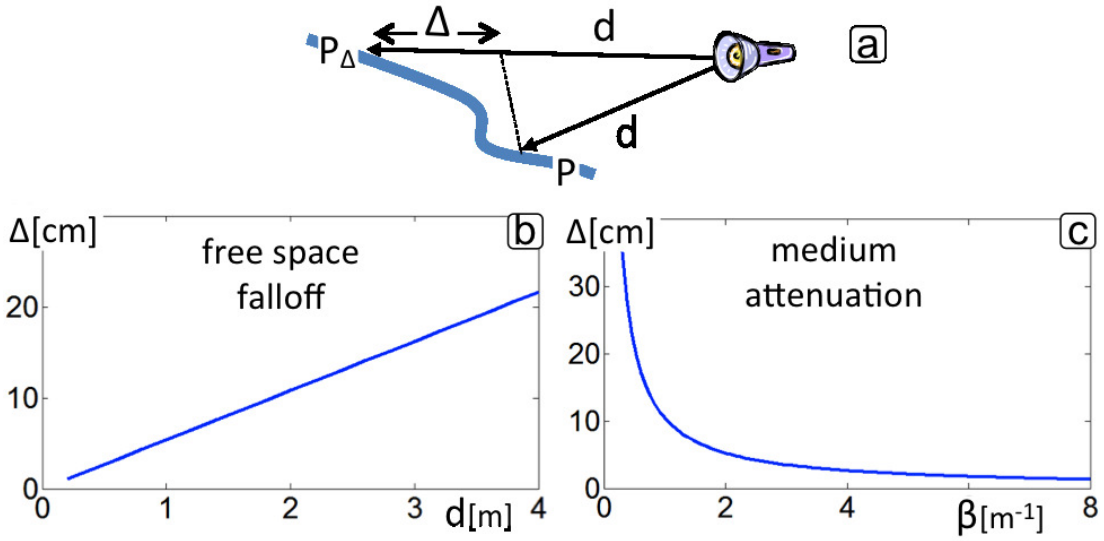

Figure 3.3: (a) Light source can be considered distant if the irradiant light intensity across it is uniform. (b) Path length differences yield 10% intensity difference of point light source intensity due to free space falloff, as a function of d. (c) Path length differences along an object that yield 10% difference in medium attenuation across it, as a function of β. For $\beta > 1m^{-1}$, which represents fairly clear water, path lengths greater than $10cm$ already result in noticeable intensity changes, ruling out the distant light source assumption.

enough to allow the freespace falloff to be ignored. Due to this we chose to use the *small surface variations approximation* instead of the distant source approximation.

3.4 Backscatter

Light is scattered as it travels through a medium. The fraction of light scattered to each direction is determined by the phase function $P(\alpha)$, where $\alpha \in [0, 2\pi]$ is the angle between the original ray direction and scattered ray, and β is the scattering coefficient.

Light which is scattered directly into the camera by the medium without reaching the object is termed *backscatter* and is given by [NGD+06, SRNN05] (Fig. 3.1a):

$$L_b(\mathbf{x}) = \beta \int_0^{\|\mathbf{X}\|} I(r\hat{\mathbf{X}}) P(\alpha) e^{-\sigma r} dr \tag{3.7}$$

The integration variable r is the distance from the camera to the imaged object point $\mathbf{X}$ along the line of sight (LOS), that is a unit direction $\hat{\mathbf{X}}$. The scattering angle α is given by $\cos(\alpha) = \hat{\mathbf{D}} \cdot \hat{\mathbf{X}}$ (recall that $\mathbf{D}$ is the direction to the light) and $I(r\hat{\mathbf{X}})$ is the direct radiance of the source at point $r\hat{\mathbf{X}}$ as defined in Eq. 3.5.

Note that for the *small surface variations approximation*, $\mathbf{X}$ and hence the limits of the integral are known for a given pixel. Therefore, the backscatter does not depend on the unknown height of the object and is a (different) constant for each pixel, similar to Tsiotsios et al. [TAKD14].

Instead of analytically computing L_b, we found that it was easier and more accurate to directly measure it using the calibration method of [TAKD14]: for each light an image is captured with no object in the field-of-view.

3.5 Single Scattered Source Radiance

Because of the medium, light rays that are not originally pointed at an object point may be scattered and reach it from the entire hemisphere of directions Ω (Fig. 3.1b), termed forward scattered radiance

$$L_s(\mathbf{x}) = \frac{\rho(\mathbf{X})}{\pi} \int_{\boldsymbol{\omega} \in \Omega} L_i(\boldsymbol{\omega})(\boldsymbol{\omega} \cdot \hat{\mathbf{N}})\, d\boldsymbol{\omega} \ . \tag{3.8}$$

where $L_i(\boldsymbol{\omega})$ is the total radiance scattered into the direction $\boldsymbol{\omega}$ and is given by

$$L_i(\boldsymbol{\omega}) = \beta \int_{t=0}^{\infty} I(\mathbf{X} + t\boldsymbol{\omega}) P(\alpha) e^{-\sigma t} dt \ , \tag{3.9}$$

where t is the distance from the object, and the angle α is given by $\cos(\alpha) = \hat{\mathbf{D}}(\mathbf{X} + t\boldsymbol{\omega}) \cdot \boldsymbol{\omega}$. Note that $\mathbf{D}$ is the direction of the integration point to the light.

Substituting Eqs. 3.4,3.8 into Eq. 3.1 and rearranging yields:

$$L_o(\mathbf{X}) = L_d(\mathbf{X}) + L_s(\mathbf{X}) = \frac{\rho(\mathbf{X})}{\pi} \boldsymbol{L}^{\text{eq}}(\mathbf{X}) \cdot \hat{\mathbf{N}} \ , \tag{3.10}$$

where

$$\boldsymbol{L}^{\text{eq}}(\mathbf{X}) = I(\mathbf{X})\hat{\mathbf{D}}(\mathbf{X}) + \int_{\boldsymbol{\omega} \in \Omega} L_i(\boldsymbol{\omega})\boldsymbol{\omega} d\boldsymbol{\omega} \ . \tag{3.11}$$

Here, the direct light as well as the integrated scattered contributions can be thought of as an equivalent distant source. However, this equivalent source may be different (in direction and magnitude) for each surface point, and thus is not a true distant source. Furthermore, the integration domain Ω in Eq. 3.11 depends on $\hat{\mathbf{N}}$, preventing Eq. 3.10 from giving us a simple linear equation for the unknown normals and albedo.

We next show through simulations, that for a wide variety of media, $L^{\text{eq}}(\mathbf{X})$ can be approximated as an effective point source in a purely absorbing medium with effective extinction coefficient. This eliminates the non-linearity in Eq. 3.10 allowing for a linear solution for the normals given 3 or more images.

Effective Point Source Simulations

We approximate $L^{\text{eq}}(\mathbf{X})$ as:

$$\tilde{\boldsymbol{L}}^{\text{eq}}(\mathbf{X}) \approx \frac{\kappa I_0\big(-\hat{\mathbf{D}}(\mathbf{X})\big) e^{-\tilde{\sigma}\|\mathbf{D}(\mathbf{X})\|}}{\|\mathbf{D}(\mathbf{X})\|^2} \hat{\mathbf{D}}(\mathbf{X}) \ , \tag{3.12}$$

which has the same form as Eq. 3.5, but the extinction coefficient is replaced by the effective extinction coefficient $\tilde{\sigma}$, and the intensity is scaled by κ. The effective source has the same position $\mathbf{S}$ and intensity distribution I_0 as the real source. Note that κ is a global brightness scale,

which is the same for all the lights, and thus does not need to be explicitly calibrated, as it cancels out in the normal estimation.

The intuition for why the source direction is unchanged is visualized in Fig. 3.4a. Although light is arriving from the entire hemisphere of directions, the vector sum of most of these directions lies in the original direction due to symmetry. Only the area of asymmetric scattering does not have symmetrical rays since the symmetrical rays lie below the visible hemisphere (in attached shadow) for the surface point. Although these asymmetric rays could potentially shift the equivalent direction, their contribution is often small for two reasons. First, these rays correspond to larger scattering angles, which are often much weaker than for rays with smaller scattering angles. Second, these paths are on average longer than for paths with smaller scattering angles and thus are more attenuated. Guided by this intuition we formulated the effective source approximation, and verified it's accuracy through extensive simulations in a wide variety of media.

For our simulations we used an isotropic point source at a distance d from a Lambertian surface patch with an angle ϕ between the surface normal and light direction. Note that the parametrization of a surface patch by d and ϕ fully parametrize the space of possible surface patches. For the scattering function, we used the common Henyey-Greenstein phase function [HG41], which can represent a large space of scattering functions by tuning a single parameter $g \in [-1,1]$. In water, g is usually between $0.7-0.9$ [NGD$^+$06].

We compute $L_o(d,\phi)$ for $d \in [200,600]$mm, $\phi \in [0,\pi]$, for a variety of media given by $\beta \in [0,0.005]\text{mm}^{-1}$ and $g \in [0,0.9]$. Note that we choose I_0 such that $L_o(200,0)$ is normalized to 1. To reduce the number of parameters we set $\sigma = \beta$, which does not influence the analysis[2].

[2]In general $\beta \leq \sigma$, but since β purely scales L_s, a smaller value of beta would make L_o closer to L_d and thus $\tilde{L}_0$ would be an even better fit than we calculated.

For each parameter pair g, β we compute the approximation parameters κ and $\tilde{\sigma}$ by minimizing:

$$\min_{\kappa,\tilde{\sigma}} \sum_{d,\phi} |L_o(d,\phi) - \tilde{L}_o(d,\phi)|^2 \ . \tag{3.13}$$

The error in the approximation is then given by the residuals $\mathrm{RE}(d,\phi) = |L_o(d,\phi) - \tilde{L}_o(d,\phi)|$.

The residuals for $g = 0.8$, $\beta = 0.0026$ (common in our setup) are plotted in Fig. 3.4b. We can see that the difference between the approximation and true values are small for all values of d and ϕ. Note that the error is largest near $\phi = 90^\circ$, where the area of asymmetric scattering is largest.

For each $\beta \in [0, .005]\mathrm{mm}^{-1}$ and $g \in [0, .9]$ we compute the mean residual (MRE) over d, ϕ and plot it in Figure 3.4c. We see that the MRE is less than 2% across all medium conditions tested justifying the approximation. Further, we see that the MRE increases slowly with scattering coefficient β, and more rapidly with phase parameter g. As such, in water, which is mostly forward scattering (g between $0.7 - 0.9$), our approximation is very accurate, even for highly turbid media. On the other hand, if scattering is more isotropic (g close to zero), then the approximation might not be valid for large β.

3.6 Single Scatter Object Blur

Similar to the light source blur, radiance from the object is also blurred while it propagates to the camera (Fig. 3.1c). As we demonstrate, this effect deteriorates the performance of photometric stereo, although it has been neglected in previous works [SRNN05].

The contribution of object blur to the pixel intensity is computed by integrating light scattered into the LOS of $\mathbf{X}$ from all other points on the surface:

$$L_c(\mathbf{x}) = \beta \int_{r=0}^{\|\mathbf{X}\|} \int_{\boldsymbol{\omega}\in\Omega} L_o(\mathbf{X}') P(\alpha) e^{-\sigma(t+r)} \, d\boldsymbol{\omega} dr. \tag{3.14}$$

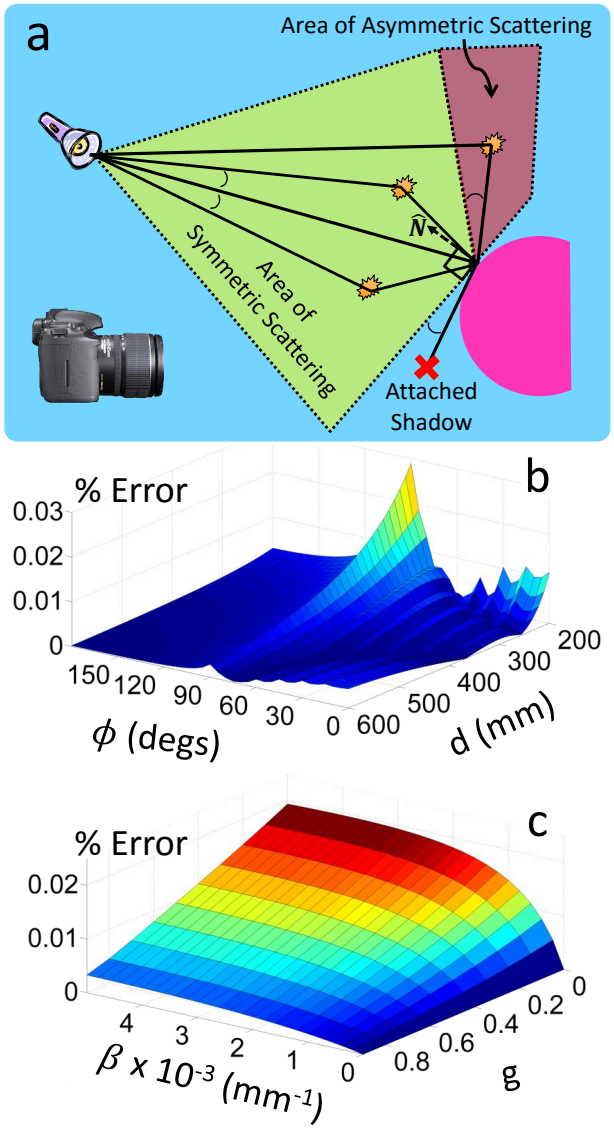

Figure 3.4: a) Diagram of the intuition for the effective source approximation. Although light is arriving from the entire hemisphere of directions, the vector sum of most of these directions lies in the original direction due to symmetry. Only the area of asymmetric scattering does not have symmetrical rays since the symmetrical rays lie below the visible hemisphere (in attached shadow) for the surface point. The contribution from these rays are often small. b) The relative error between $L_o(d,\phi)$ and $\tilde{L}_o(d,\phi)$ for $g = 0.8$ and $\beta = 0.0026$. Note that the spike only reaches 3% and is located at $\phi = 90^\circ$ where $\tilde{L}_o = 0$ due to shadowing. ϕ near 90° and above is not usually relevant for photometric stereo. c) The mean relative error between L_o and $\tilde{L}_o$ for $\beta \in [0, 0.005]\text{mm}^{-1}$ and $g \in [0, 0.9]$. The approximation errors are small over a wide variety of media.

Here r is the distance along the LOS, $\mathbf{X}'$ is the object surface point intersected by the ray starting at point $r\mathbf{X}$ in direction $\boldsymbol{\omega}$. Its radiance is $L_o(\mathbf{X}')$ and its distance to the scatter point in the LOS is given by $t = \|r\hat{\mathbf{X}} - \mathbf{X}'\|$ with scattering angle $\cos\alpha = \boldsymbol{\omega}\cdot(-\hat{\mathbf{X}})$.

We now show that L_o can be recovered from $L_o e^{-\sigma\|\mathbf{X}\|} + L_c$ by deconvolution with a constant PSF.

Deblurring Object Scatter

First we rewrite Eq. 3.14 to integrate over the area of the object surface $dA = d\boldsymbol{\omega}\cdot t^2/\cos\theta$ instead of solid angle $d\boldsymbol{\omega}$, where t is the distance from $\mathbf{X}'$ to the scattering event, and θ is the angle between the normal at $\mathbf{X}'$ and the ray of light before scattering. Eq. 3.14 now becomes

$$L_c(\mathbf{x}) = \beta \int_{\mathbf{X}'} L_o(\mathbf{X}') \int_0^{\|\mathbf{X}\|} P(\alpha) e^{-\sigma(t+r)} \frac{\cos\theta}{t^2}\, dr\, dA. \tag{3.15}$$

Now we define the scattering kernel

$$\begin{aligned} K(\mathbf{X},\mathbf{X}') = {} & \delta(\mathbf{X}-\mathbf{X}')e^{-\sigma\|\mathbf{X}\|} + \\ & \beta \int_0^{\|\mathbf{X}\|} P(\alpha) e^{-\sigma(t+r)} \frac{\cos\theta}{t^2} dr, \end{aligned} \tag{3.16}$$

where $\delta(\mathbf{X}-\mathbf{X}')$ is the Dirac delta function. Now,

$$L_o(\mathbf{x})e^{-\sigma\|\mathbf{X}\|} + L_c(\mathbf{x}) = \int_{\mathbf{X}'} K(\mathbf{X},\mathbf{X}') L_o(\mathbf{X}')\, dA(\mathbf{X}') \ . \tag{3.17}$$

In general, the kernel K, depends on $\mathbf{X}$, $\mathbf{X}'$ and the unknown normals $\hat{\mathbf{N}}'$. For an orthographic camera viewing a plane at constant depth, K is shift invariant and Eq. 3.17 can be written as a convolution with a PSF. Motivated by this, we found empirically that for a given $\bar{Z}$ it is approximately shift invariant (and rotationally symmetric).

Denoting the PSF as h, we get

$$L_o(\mathbf{x})e^{-\sigma\|\mathbf{X}\|} + L_c(\mathbf{x}) \approx h * L_o \ . \tag{3.18}$$

We emphasize here that we have shown that under a single scattering model, the forward scatter from the object can be written as an integral transform with kernel K. This justifies approximating the forward scattering as a PSF which is not obvious in the form of Eq. 3.14.

We solve Eq. 3.18 for L_o by writing the image as a column vector and representing the convolution as a matrix operation

$$(L - L_b) = HL_o \tag{3.19}$$

where we have substituted the known backscatter compensated image $(L - L_b)$ for $L_o e^{-\sigma\|\mathbf{X}\|} + L_c$, and H is the matrix representation of h. Here H is a large nonsparse matrix and thus storing it in memory and directly inverting it is infeasible. Instead we solve the linear system of Eq. 3.19 using conjugate gradient descent. This requires only the matrix vector operation which can be computed as a convolution and implemented using a Fast Fourier Transform (FFT).

3.7 Backscatter Removal Using Fluorescence

While we are able to subtract the backscatter component, it is an additive component that effectively reduces the dynamic range of the signal from the object, degrades the image quality and reduces SNR [TS12]. As such it is beneficial to optically remove it when imaging. Here, we use the observation that for fluorescence images taken with non-overlapping excitation and emission filters, there is no backscatter in the image (Fig. 3.5a). In fluorescence imaging, the signal of interest is composed of wavelengths that are longer than that of the illumination, and a barrier filter on the camera is used to block the reflected light. The backscatter is composed of light scattered by the medium *before* it reaches the object. Thus, the backscatter has the same

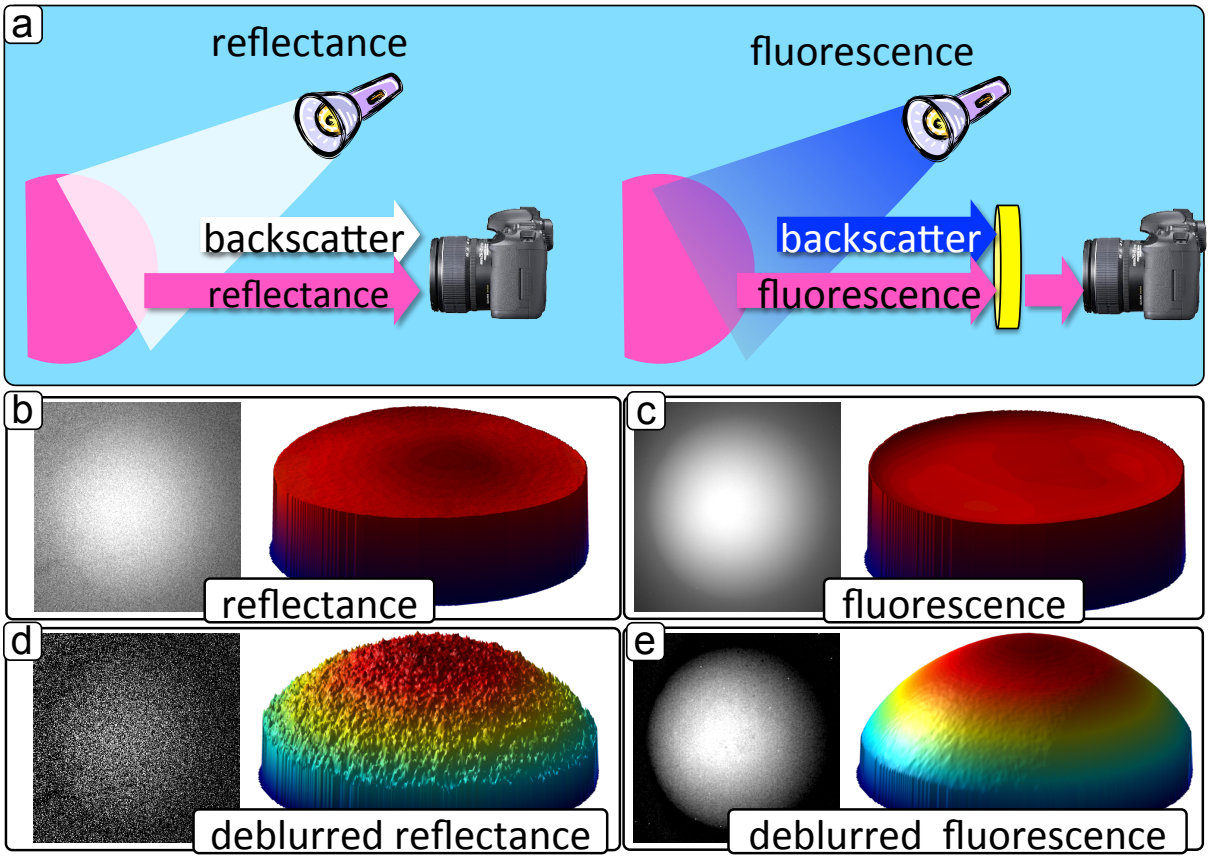

Figure 3.5: a) Backscatter is caused by light that is scattered into the camera by a medium, *before* it reaches the object and has the same color as the illumination. Thus, the barrier filter used to block fluorescence excitation also blocks backscatter, while imaging the signal from the object. We use this property to remove backscatter in input images. b) Photometric stereo reconstruction of a fluorescent sphere using backscatter subtracted reflectance images. One of the input images is shown on the left, with visible noise and blur. Blur in the input images flattens the reconstruction. c) Looking at fluorescence images as an input, the backscatter is eliminated while maintaining a higher SNR. However the blur still flattens the reconstruction. d) Deblurring the backscatter subtracted images recovers the general shape but suffers from noise as seen by the spiky surface. e) Deblurring the fluorescence results in the correct shape with much less noise.

spectral distribution as the light source, which is blocked by the barrier filter on the camera. This insight enables imaging without loss of dynamic range even in highly turbid media. Compared to a backscatter subtracted reflectance image, a fluorescence image has less noise (Fig. 3.5b,c). This difference becomes even more apparent after deconvolution (Fig. 3.5d,e).

In addition, in [SOS12b, TMMK12] it was shown that the fluorescence emission acts as a Lambertian surface in photometric reconstructions. Thus, imaging fluorescence has an additional advantage as it relaxes the need for a Lambertian surface.

In the development of our algorithm we assumed a single set of medium parameters β, σ and $P(\alpha)$. However these quantities are in general wavelength dependent. In reflectance imaging, the wavelength of the light is the same on both pathways: light to object, and object to camera. However, in fluorescence imaging they are different. Nevertheless, the only parameters that require calibration in our solution are the effective extinction coefficient $\tilde{\sigma}$ and the PSF. The parameter $\tilde{\sigma}$ is estimated for the excitation wavelength and the PSF is estimated for the emission wavelength, and as such we do not need to calibrate any extra parameters in the case of fluorescence imaging.

3.8 Implementation

3.8.1 Experimental Setup

Our setup is shown in Fig. 3.6. We used a Canon 1D camera with a 28mm lens placed 2cm away from a 10 gallon glass aquarium. All sides except the front (where the camera looks in) were painted black to reduce reflection. In addition, a black panel was suspended just below the surface of the water to remove reflection from the air-water interface. The objects were placed at an average distance of 40cm from the front of the tank. For point illumination we used Cree XML - RGBW Star LEDs. The LEDs were water proofed by coating the electrical terminals with epoxy. Reflection images were taken under white illumination while fluorescent images were

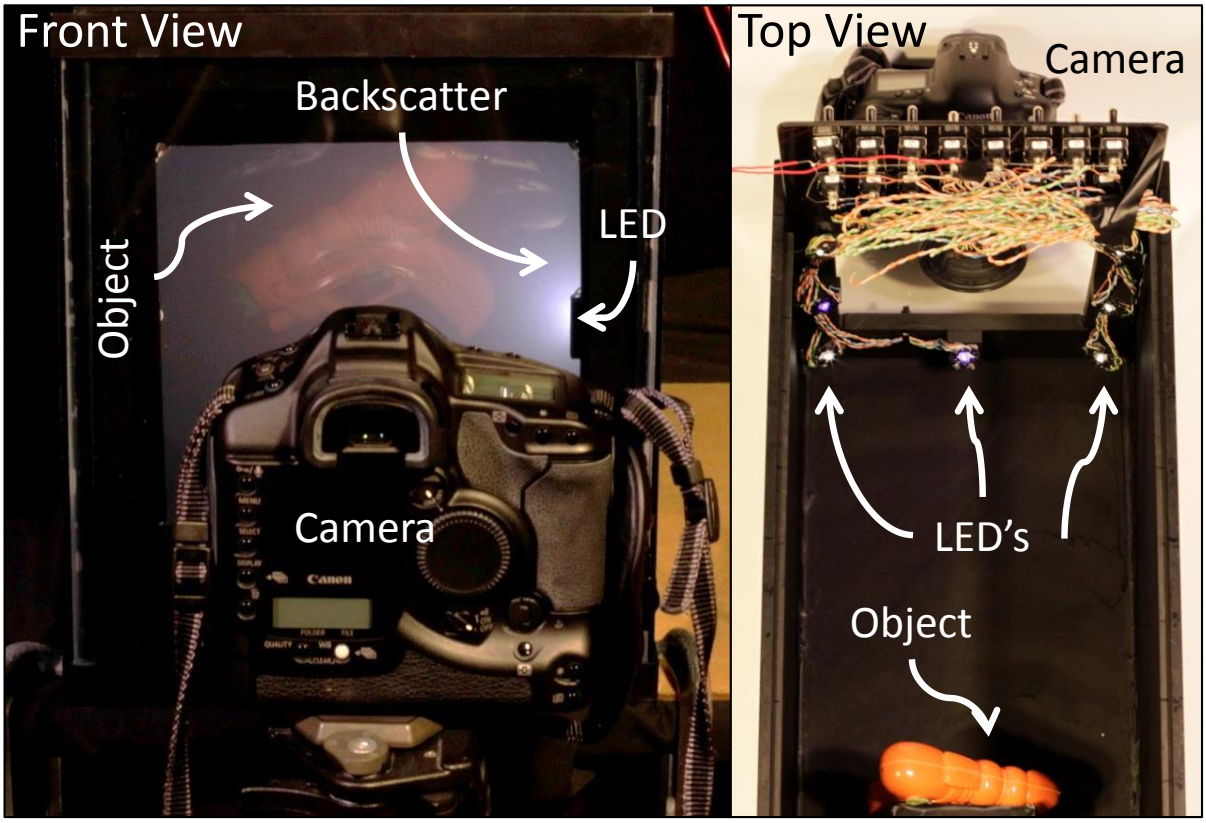

Figure 3.6: [Left] Our experimental setup consists of a camera looking through a glass port into a tank. [Right] 8 LEDs are mounted inside the tank around the camera port illuminating the object placed at the back of the tank.

	Level 1	Level 2	Level 3	Level 4
Objects	Spherical Cap & Lobster			
Milk (ml)	1.25	2.50	3.75	5.00
juice (ml)	15.0	30.0	45.0	60.0
$\beta(\times 10^{-3} mm^{-1})$	.602	1.20	1.81	2.41
$\sigma(\times 10^{-3} mm^{-1})$	.642	1.28	1.93	2.57
Objects	Toy Gun & Mask			
Milk (ml)	1.25	2.50	3.75	5.00
juice (ml)	0	0	0	0
$\beta(\times 10^{-3} mm^{-1})$	.602	1.20	1.81	2.41
$\sigma(\times 10^{-3} mm^{-1})$	.602	1.20	1.81	2.41

Figure 3.7: Tabulated values for the amount of milk and grape juice added in our experiments, and the associated scattering and extinction coefficients. The coefficients were computed using the data provided in [NGD+06].

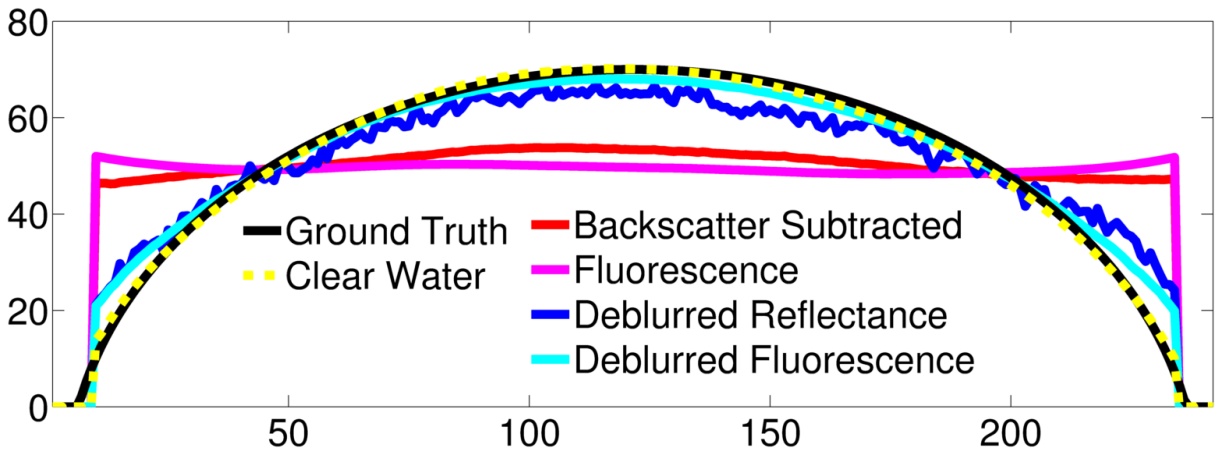

Figure 3.8: Cross-sections of the spherical cap reconstruction in turbid medium using various methods compared to ground truth. The clear water reconstruction resembles the ground truth. Only correcting for the backscatter (by subtraction or fluorescence) yields flattened results. Deblurring the backscatter subtracted images recovers the shape but is degraded by noise (the surface is jagged). Deblurring the fluorescence images produces the best results.

taken under blue illumination with a Tiffen #12 emission filter on the camera. We used tap water, and the turbidity was increased using a mixture of whole milk and grape juice (milk is nearly purely scattering, while grape juice is nearly purely absorbing and thus by mixing them we can achieve a variety of scattering conditions [NGD$^+$06]). The LEDs were mounted inside the tank on a square around the camera, four on the corners and four on the edges. Their positions were measured. Images were acquired in Raw mode which is linear and the normal integration was done using the method of [ARC06].

3.8.2 Geometric and Radiometric Calibration

Images of a checkerboard (in clear water) were used to calibrate the intrinsic camera parameters (implicitly accounting for refraction) [Bou]. The location of each light was measured using a ruler, and transformed to the camera reference frame. To calibrate each light's angular intensity distribution we imaged a matte painted (assumed to be Lambertian) plane at a known

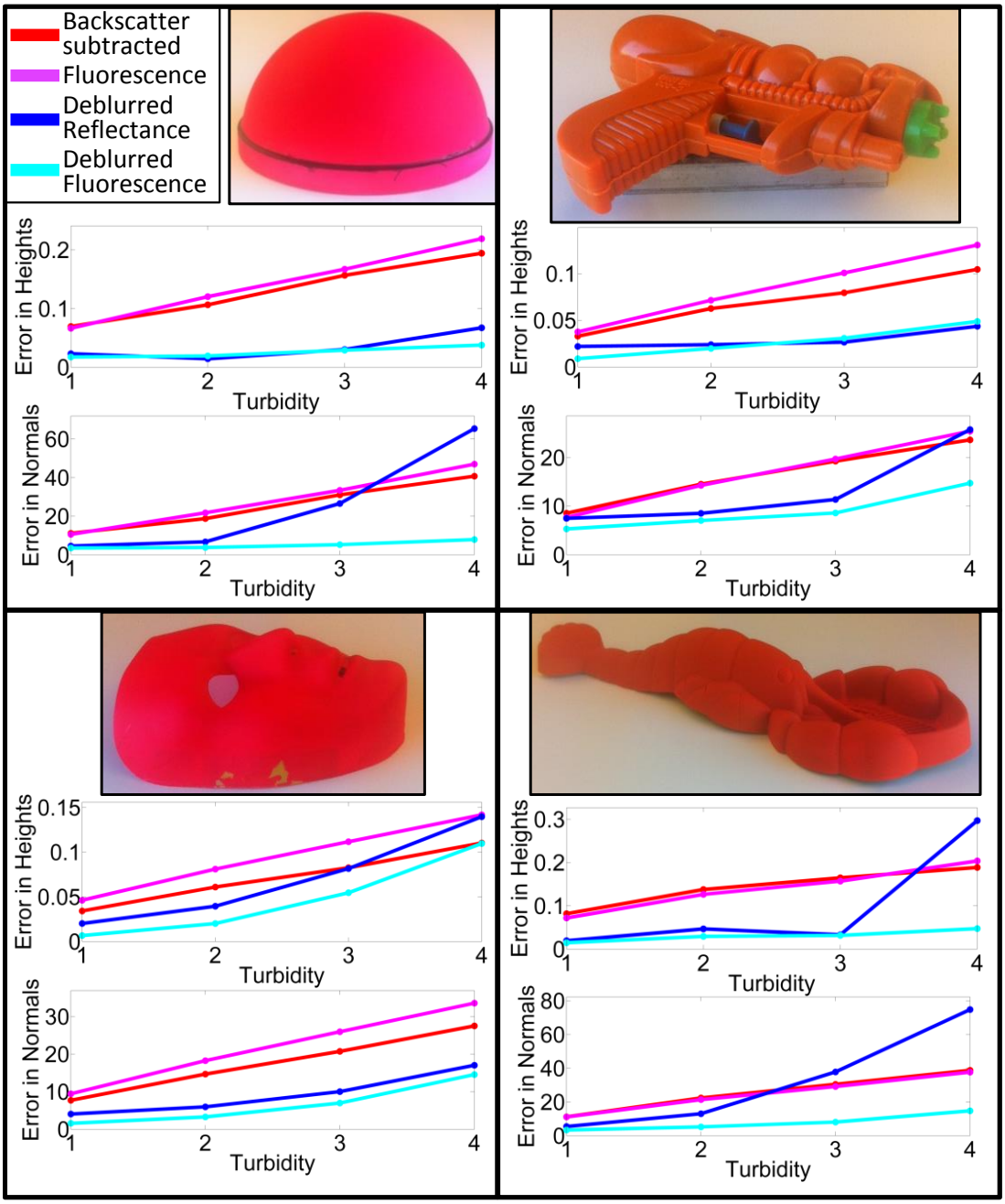

Figure 3.9: Errors in the reconstructions of four objects as a function of turbidity, compared to clear water reconstruction. Top rows are average percent errors in heights and bottom rows are average angular errors in normals. Removing backscatter by either subtraction or using fluorescence performs similarly. Deblurring the backscatter compensated images significantly improves the reconstructions. In high turbidity where the backscatter is strong compared to the object signal deblurring the backscatter subtracted images degrades due to noise, while deblurring the fluorescence suffers less, as the fluorescence images have a higher SNR.

position under illumination from each light in clear water. Using Eq. 3.4, the known geometry, and $\sigma = 0$, we compute $I_0(-\hat{\mathbf{D}})$, the angular dependence of the light source.

3.8.3 Calibration of Medium Parameters

The backscatter component is measured using the calibration method of [TAKD14]. For each light an image is captured with no object in the field-of-view and subsequently subtracted from future reflectance images. This is not used when imaging fluorescence.

Our method works independent of how the PSF is calibrated and thus a variety of methods could be used including that of Narasimhan et. al. [NN03]. Here we chose a procedure using a calibration target similar to [JSK08] due to its ease of implementation. We use a matte painted checkerboard which is imaged with its axis aligned to the image plane at the approximate depth of the objects we plan to reconstruct. As the PSF is rotationally symmetric its parameters are the values along a radius $[h_0, ..., h_s]$, where h_0 is the center value and h_s is the value on the support radius s. The PSF and the effective extinction coefficient $\tilde{\sigma}$ are estimated by optimizing

$$\min_{\tilde{\sigma}} \min_{h_0 \ldots h_s} \sum_{\mathbf{x}} \| h * L_o(\mathbf{x}, \tilde{\sigma}) - (L - L_b) \| \tag{3.20}$$

where L is the image of the checkerboard in the medium and L_b is the image of the backscatter. L_o is computed from Eq. 3.10 using the calibrated lights, known geometry, and registering the checkerboard albedo, measured in clear water to the image in turbid water. The inner optimization is an overdetermined linear system holding $\tilde{\sigma}$ fixed. We sweep over the values of $\tilde{\sigma}$ and choose the one with the minimum error. Note that the PSF is not normalized due to loss of energy (attenuation) from the object to the camera.

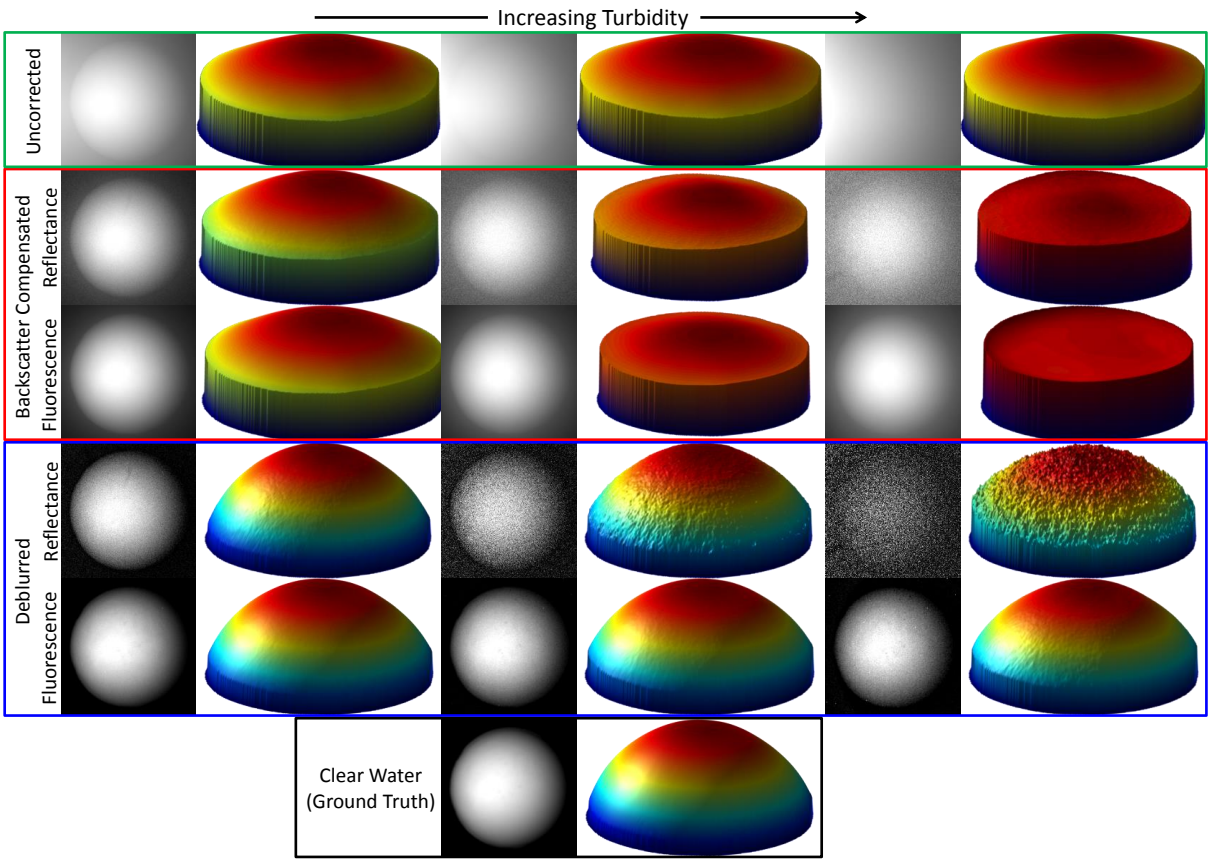

Figure 3.10: Input images and resulting surface reconstructions of the spherical cap. The columns depict three levels of increasing turbidity from left to right. [1st row] result of standard photometric stereo (scattering is ignored). The shape is not reconstructed correctly. [2nd row] Result of removing the backscatter as in [TAKD14]. The reconstruction is improved but still unsatisfactory. [3rd row] Using fluorescence to remove backscatter. The result is basically the same as backscatter subtraction. [4th row] result of deblurring the backscatter subtracted images. This recovers the shape quite well when the SNR is not too low. However this is not the case in high turbidity. [5th row] result of deblurring the fluorescence images. Here the SNR remains high even in high turbidity and thus we continue to get excellent quality reconstructions. Note the roughness on the fourth row, second column due to noise. [Bottom row] Clear water reconstruction (ground truth).

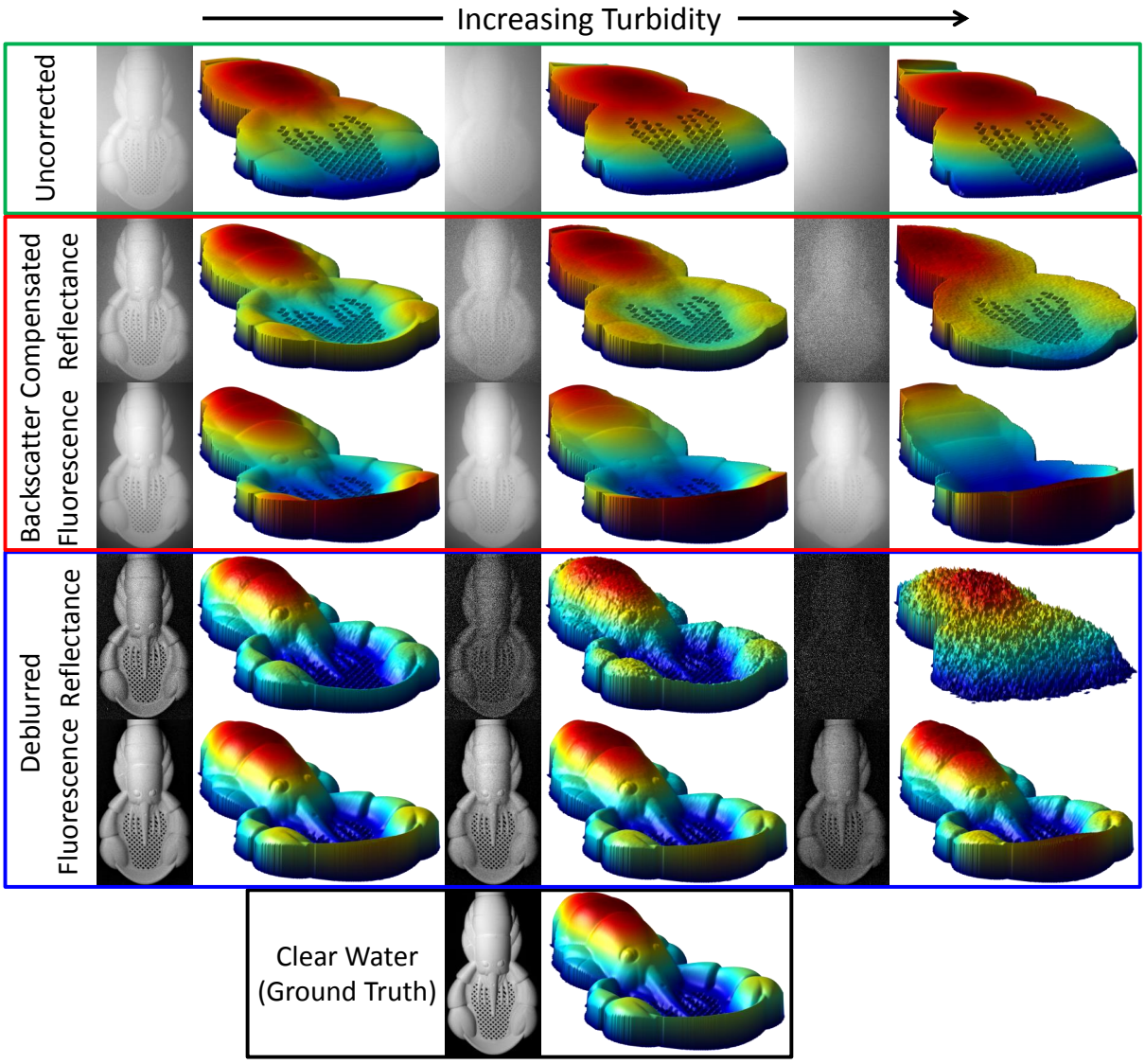

Figure 3.11: Input images and resulting surface reconstructions of the toy lobster. The columns depict three levels of increasing turbidity from left to right. [1st row] result of standard photometric stereo (scattering is ignored). The shape is not reconstructed correctly. [2nd row] Result of removing the backscatter as in [TAKD14]. The reconstruction is improved but still unsatisfactory. [3rd row] Using fluorescence to remove backscatter. The result is basically the same as backscatter subtraction. [4th row] result of deblurring the backscatter subtracted images. This recovers the shape quite well when the SNR is not too low. However this is not the case in high turbidity. [5th row] result of deblurring the fluorescence images. Here the SNR remains high even in high turbidity and thus we continue to get excellent quality reconstructions. Note the roughness on the fourth row, second column due to noise. [Bottom row] Clear water reconstruction (ground truth).

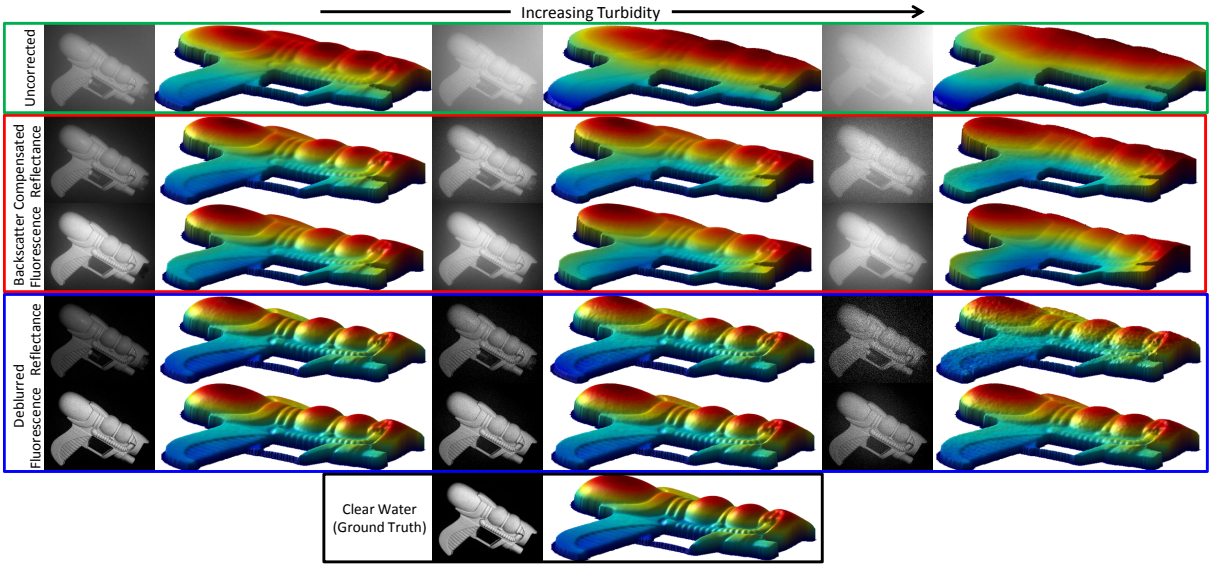

Figure 3.12: Input images and resulting surface reconstructions of the toy squirt gun. The columns depict three levels of increasing turbidity from left to right. [1st row] result of standard photometric stereo (scattering is ignored). The shape is not reconstructed correctly. [2nd row] Result of removing the backscatter as in [TAKD14]. The reconstruction is improved but still unsatisfactory. [3rd row] Using fluorescence to remove backscatter. The result is basically the same as backscatter subtraction. [4th row] result of deblurring the backscatter subtracted images. This recovers the shape quite well when the SNR is not too low. However this is not the case in high turbidity. [5th row] result of deblurring the fluorescence images. Here the SNR remains high even in high turbidity and thus we continue to get excellent quality reconstructions. Note the roughness on the fourth row, third column due to noise. The object signal is stronger in this case than the lobster and sphere since the medium doesn't contain juice which increases attenuation. [Bottom row] Clear water reconstruction (ground truth).

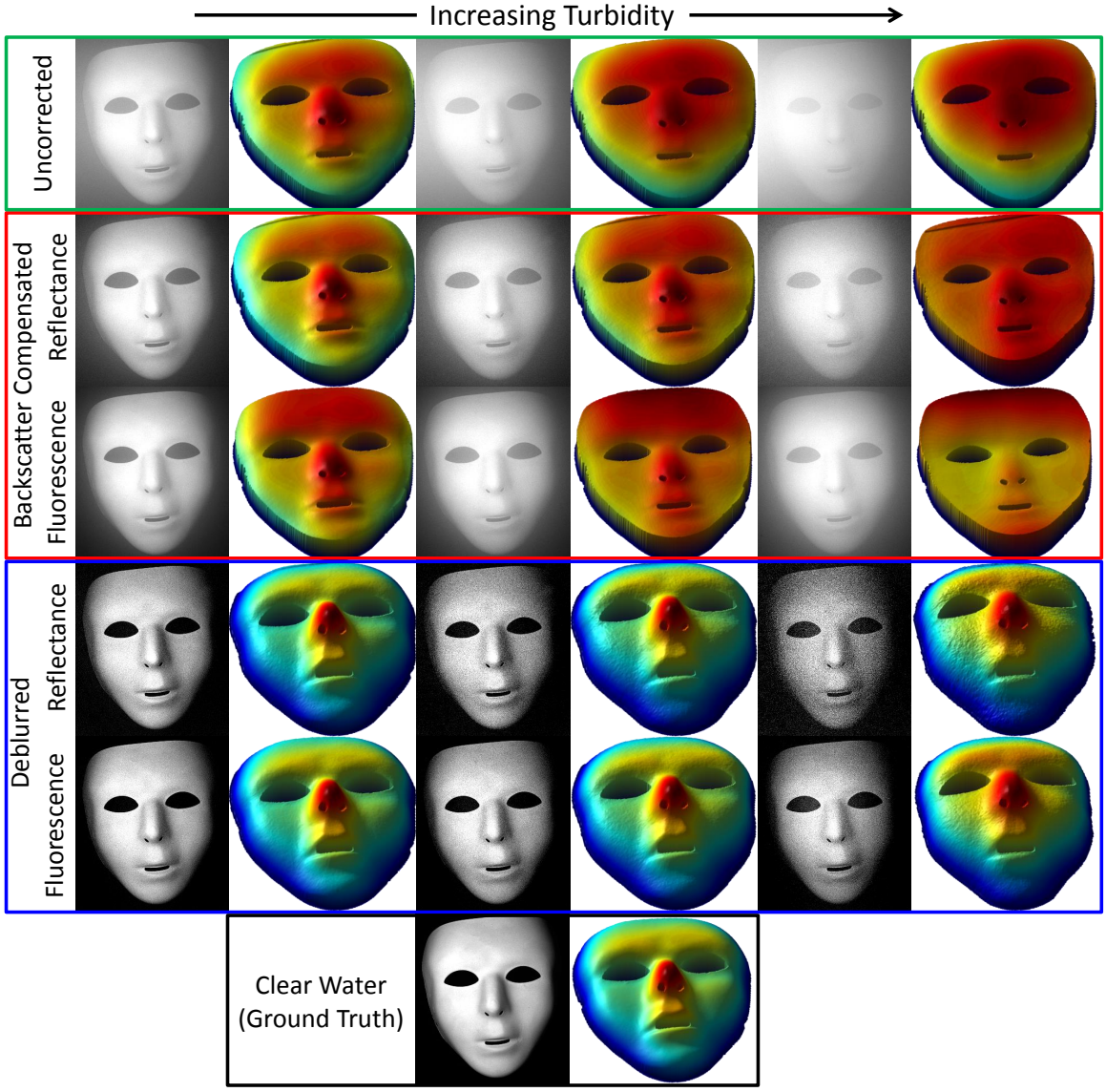

Figure 3.13: Input images and resulting surface reconstructions of the mask. The columns depict three levels of increasing turbidity from left to right. [1st row] result of standard photometric stereo (scattering is ignored). The shape is not reconstructed correctly. [2nd row] Result of removing the backscatter as in [TAKD14]. The reconstruction is improved but still unsatisfactory. [3rd row] Using fluorescence to remove backscatter. The result is basically the same as backscatter subtraction. [4th row] result of deblurring the backscatter subtracted images. This recovers the shape quite well when the SNR is not too low. However this is not the case in high turbidity. [5th row] result of deblurring the fluorescence images. Here the SNR remains high even in high turbidity and thus we continue to get excellent quality reconstructions. Note the roughness on the fourth row, third column due to noise. Similar to the toy gun, the object signal is stronger in this case than the lobster and sphere since the medium doesn't contain juice which increases attenuation. [Bottom row] Clear water reconstruction (ground truth).

3.9 Results

We imaged four objects: a spherical cap (Fig. 3.10), a plastic toy lobster (Fig. 3.11), a plastic toy squirt gun (Fig. 3.12), and a fluorescent painted mask (Fig. 3.13) in clear water as well as four increasing turbidities. Each turbidity level corresponded to adding 1.25ml of milk to the 10 gallon tank. For the spherical cap and the lobster, we also added 15ml of grape juice per turbidity level to increase absorption. In this case, since attenuation is exponential with distance, the signal from the object, which travels further, is relatively weaker than the backscatter which comes mostly from shorter paths. This exacerbates the loss of signal-to-noise ratio in backscatter subtracted images. To get an idea of the true scattering parameters of our various media we use the data provided in [NGD+06]. Tabulated values are shown in Fig. 3.7.

We employ two error metrics to evaluate the quality of our reconstructions: The mean absolute difference in heights (Err $Z = \text{mean}(Z - Z_{gt})$) and the mean angular error in the normals (Err $N = \text{mean}(\text{acos}(N \cdot N_{gt}))$, where Z_{gt} and N_{gt} are the ground truth heights and normals. Note that during integration, random noise in the normals cancels out locally, resulting in reconstructions with the correct overall shape, but with rough surfaces. As such Err Z captures systematic errors that affect the overall shape, but is less sensitive to noise in the normals.

We see that the reconstructed spherical cap in clear water nearly perfectly matches the ground truth (Fig. 3.8) with an Err Z of 1.4% and Err N of 3°. This justifies our use of clear water reconstructions as ground truth for the other objects where true ground truth is not available.

The quality of results as a function of turbidity level is demonstrated in Fig. 3.9. The plots show how Err Z and Err N increase for each method as the turbidity increases, where the lowest error is achieved using the deblurred fluorescence images. In the highest turbidity level, the deblurred reflectance image often performs worse than all other methods, as the deblurring degrades with noise.

Figures 3.10,3.11,3.12,3.13 depict an input image and the resulting reconstruction for the spherical cap, the toy lobster, the toy squirt gun, and the mask respectively. In each figure, the rows show various reconstruction methods and the columns show the results for turbidity levels 2-4. The bottom row shows an input image and reconstruction in clear water which are treated as ground truth.

In all results, reconstructions from uncorrected images are flattened. Removing backscatter, either by backscatter subtraction (current state-of-the-art [TAKD14]), or using fluorescence, but without handling blur, also produces flattened results. For lower turbidities deblurring backscatter subtracted images produces excellent results, but in the highest turbidity, where the backscatter dominates the signal, using fluorescence reduces the noise and results in a smoother surface.

3.10 Summary

In this chapter, we have developed a comprehensive and novel solution for photometric stereo in a scattering medium. We address each of the three key modes of single scattering, showing how a scattered light source can be modeled as an unscattered point light source, accounting for blur due to scattering from the object through a novel deconvolution framework, and demonstrating how fluorescence imaging can optically eliminate backscatter, increasing SNR in high turbidity. With the simple *small surface variations approximation*, we reduce the problem to a linear system for the surface normals, almost identical to conventional photometric stereo. Our practical methods for deconvolution and fluorescence can be combined to produce reconstructions almost as accurate as those obtained in air, and significantly better than previous methods.

Chapter 3 is a reformatted version of "Photometric Stereo in a Scattering Medium," Z. Murez, T. Treibitz, D. Kriegman, R. Ramamoorthi, *IEEE Transactions on Pattern Analysis and*

Machine Intelligence 2016 [MTRK17]. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Learning to See through Turbulent Water

In the previous chapter we considered photometric stereo when the camera and scene are submerged in the same medium and accounted for volumetric scattering. In this chapter we consider the imaging scenario in which the camera views a scene through a refractive medium, in which the interface is constantly changing. Two common examples of this occur when looking from air into water with a turbulent surface and imaging through a medium with temperature variations that gives rise to atmospheric refraction or mirages. In all such cases, the scene appears distorted due to the bending of light as it passes through the refractive interface.

Removing such distortions from a single image is challenging since the shape of the interface is not known a priori and must be estimated simultaneously with the latent image. The problem is similar to blind deconvolution, but the kernel is spatially varying and can be much larger than what is typically considered in image deblurring. As such, most previous works [DDR06, DR07, EISV04, TN09] assume an input video instead of a single frame.

In contrast, we attempt to solve the single image undistortion problem by building upon the recent success of deep convolutional neural networks at solving image-to-image translations [IZZE16a]. Our hypothesis is that the space of natural images as well as the space of natural refractive distortions is structured enough that a neural network can learn a reasonable mapping

Figure 4.1: Top: Input and our result on a scene captured in the wild. Note the distortions to the ladder on the top of the fire truck and the landing skids of the helicopter. Bottom: Our laboratory setup for generating large amounts of training data.

between distorted input images and undistorted output images. We demonstrate that it is in fact the case by training a network end-to-end for our task.

Although, in principle, a purely convolutional and deconvolutional network could learn the complex mapping between distorted images and undistorted images directly, we find training such a network to be difficult. Instead we propose a two-step framework to address the nature of images observed through dynamic refraction. The first step outputs a warping field and applies it to the input image to undistort it. Note that we can apply the warping in a differentiable manner by using bilinear sampling. While such a warping network is able to remove many of the geometric distortions, there is often information lost during image formation due to blurring and holes induced by the complex shape of the interface. To correct for this, we train another color network that takes the output of the warp net and hallucinates plausible details. Both the networks are trained together in an end-to-end manner. As has been observed in prior work [LTH+16], when the network is trained solely with the L1 or L2 loss, the output images are blurry. To combat this, our network is also trained with adversarial [GPAM+14a] and perceptual losses [LTH+16].

To train the network we need a large number of input distorted and ground truth image pairs. Unlike previous works that use computer graphics simulations to generate voluminous

data, we find that our application demands a narrower domain gap between training and testing. Since no such dataset with real images currently exists, we construct a new large scale dataset by displaying ImageNet images on a monitor placed under a glass tank full of water and capturing images from above. We demonstrate that by training our network on this dataset we are able to generalize to images of real objects, even in completely different environments (see Fig 4.1). Our method consistently produces high quality undistorted images from a single distorted input, in contrast to the recent end-to-end learning framework of [IZZE16b]. Our dataset and code will be publicly released to stimulate further research towards this challenging problem.

In summary, we make the following contributions: 1) propose using deep learning to solve the as yet unattempted problem of single image distortion removal, 2) design a new special network architecture that takes advantage of the physical image distortion model, 3) construct a large scale image dataset that can be used to train our network, 4) show high quality results on real objects imaged through diverse distortions in various settings.

4.1 Related Work

Imaging Through Refractive Distortions: Water distortion removal is an extremely challenging problem due to its inherent ill-posed nature. To the best of our knowledge, all previous methods assume additional information beyond a single input image.

One common approach is to use a video sequence of a still scene under varying distortions. Murase et al. [Mur92] proposes the common assumption that the water surface slant is Gaussian with mean zero over time. This means that the temporal average of frames will give a reasonable undistorted image. This suggests the method known as lucky imaging in which the image with the least distortion is chosen as the restoration. Going beyond this, Efros et al. [EISV04] divide the images into patches and choose the best patch for each location across the video sequence and stitch the results into the final result. Donate et al. [DDR06, DR07] improve this method by

further removing the motion blur and by using k-means clustering to reduce the number of patches being considered for the patch selection process. Wen et al. [WLFL10] combines lucky imaging with Fourier domain spectral analysis for better reconstruction. Tian et al. [TN09] propose a compact spatial distortion model based on the wave equation and use it to design an image restoration technique specifically for water distortion. Periodicity and smoothness constraints for water surfaces are used as regularization to help avoid poor local minima. Oreifej et al. [OSPS11] propose an iterative two stage restoration in which the first stage robustly aligns the frames to the temporal mean image and the second stage removes sparse noise using a low rank assumption.

Another branch of works focuses on recovering the shape of the water surface from a distorted and non-distorted image pair. Note that this is a slightly different problem than ours as the desired non-distorted image is assumed known. In this case the problem can be posed as an image alignment problem seeking the warping field that warps the distorted image to the undistorted one. Tian et al. [TN12] develop a data-driven gradient descent algorithm that iteratively recovers the warping field. They first generate a large set of training samples with known distortions. Then in each iteration, they find the nearest neighbor of the current distorted image in the training set and use its distortion parameters to warp the distorted image back to the template. Tian et al. [TN15] further develop a hierarchical structure which needs much less training samples and can consider global and local distortion simultaneously. Zhang et al. [ZLG$^+$14] uses defocus and distortion cues from a video along with a non-distorted template to solve for both the water surface and object depth.

Alterman et al. [ASS16] considers the problem of multi-view stereo through a dynamic refractive interface. They use multiple cameras along a wide baseline to observe a scene under uncorrelated distortions and recover sparse point clouds. Xue et al. [XRW$^+$14] estimate flow velocity in a dynamic refractive medium using optical flow.

CNNs for Estimating Transformations: Siamese networks have been used for estimating rigid or non-rigid transformations between two images for tasks such as motion estimation or matching [ACM15, KJC16]. In contrast, we use a single image for undistortion, since the ground truth target image is not known at test time. The spatial transformer has been proposed as a trainable module in classification networks by Jaderberg et al. [JSZK15] to estimate parametric transformations, with a convolutional variant used for correspondence learning in [CGSC16]. Non-parametric transformations in the form of a shape basis representation are estimated in [YZC16] to handle articulations. In contrast to those works, we address the problem of distortions induced by waves on the surface of water, which is not a parametric transformation and often too complex to be representable by a small number of bases.

Image-to-Image Deep Learning: Although deep learning first saw great success on the problem of image classification [KSH12], it has also proven very successful on image to image problems such as semantic segmentation [LSD15]. Recently many works have trained convolutional/ deconvolutional networks to perform a variety of image to image problems, such as image super-resolution [DLHT16, LTH+16], image colorization [DRF15, DLYF16, ZIE16, ISSI16], image inpainting [PKD+16], image style transfer [LW16], image manipulation guided by user constraints [ZKSE16] and image de-raining [ZSP17].

Many of these works rely on generative adversarial networks (GANs), which have recently shown promise at the task of natural image generation [GPAM+14a]. A GAN consists of two networks: a generator, whose task is to generate realistic looking images, and a discriminator, whose job is to label images from the generator as fake and real images as real. These two networks are trained together forcing the generator to learn to produce realistic images. Despite recent work on improving the training of GANs [RMC15], the resulting images are not yet of high quality. However, when the generator is conditioned on an input image and can can be trained with a traditional loss, such as L1 or L2, in addition to the adversarial loss, the results are

much more impressive [LTH+16, IZZE16b]. The adversarial loss drives the results away from the mean/median image that is learned from solely the L2/L1 loss, which allows the network to learn to predict more detailed, less blurry, realistic looking images.

Isola et al. [IZZE16b] build upon these to propose a general framework for image-to-image translation problems that involves training a convolutional/ deconvolutional network on input and output image pairs using a combination of L1 pixel loss and an adversarial loss. Although this can be used to solve our problem in principle, our experiments indicate that their general purpose net has difficulty in learning to correct geometric distortions in practice.

4.2 Model

We train a deep neural network to take in images distorted by a dynamic refractive interface and output the undistorted image that would have been observed without an interface. Although, in theory, a purely convolutional/ deconvolutional architecture such as [IZZE16b] could learn this complex mapping, we find it does not perform well in practice (see Figure 4.4). Unlike most previous image-to-image networks [IZZE16b, LTH+16, ZSP17], we draw inspiration from the physical image formation model to help simplify the problem for the network.

Let $\mathbf{I}(\mathbf{x})$ be the image that would have been observed without any refractive distortion and $\tilde{\mathbf{W}}(\mathbf{x})$ be a 2D warping field that corresponds to the distortion induced by the refractive interface. When the height of the variations of the water are small compared to the depth of the scene and the height of the camera, $\tilde{\mathbf{W}}$ is linearly related to the gradient of the surface height $\nabla Z(\mathbf{x})$. Then the observed, distorted image $\mathbf{J}(\mathbf{x})$ is given by

$$\mathbf{J}(\mathbf{x}) = \mathbf{I}(\mathbf{x} + \tilde{\mathbf{W}}(\mathbf{x})) \tag{4.1}$$

Unfortunately, inverting 4.1 is difficult not only since both $\mathbf{I}(\mathbf{x})$ and $\tilde{\mathbf{W}}$ are unknown, but also because the mapping need not be one-to-one.

Inspired by this, we train our network to predict the inverse warping field $\mathbf{W}(\mathbf{x})$ such that

$$\mathbf{I}(\mathbf{x}) = \mathbf{J}(\mathbf{x} + \mathbf{W}(\mathbf{x})) \tag{4.2}$$

Thus, given a predicted warping field $\mathbf{W}(\mathbf{x})$ from our network, we can easily compute the desired undistorted image by interpolation of the input image. We use bilinear interpolation since it is differentiable, which allows end-to-end training. Here we have taken advantage of the fact that we know the mapping between input and output images to be a warp. By performing the warping explicitly through interpolation, we do not require the network to learn to do it through convolutions.

However as stated above, the forward warping need not be one-to-one and thus information may be lost in the distorted image $\mathbf{J}(\mathbf{x})$. This is often observed as blurring, double images and singularities. To handle this, we train a second image-to-image network, which we call the color network, that takes the unwarped image $\mathbf{J}(\mathbf{x} + \mathbf{W}(\mathbf{x}))$ and outputs our final image. The goal of this second network is to add back details lost during the warping and correct other artifacts that the warping network could not handle (partly due to its limited modeling).

Let our warping network be denoted as $\mathbf{W}_\theta$ and our color network as $\mathbf{C}_\phi$, where θ and ϕ are the learnable parameters of each network, respectively. Then our full generator network is given by

$$\mathbf{G}_{\theta\phi}(\mathbf{J}(\mathbf{x}), \mathbf{x}) = \mathbf{C}_\phi(\mathbf{J}(\mathbf{x} + \mathbf{W}_\theta(\mathbf{J}(\mathbf{x})), \mathbf{x}), \tag{4.3}$$

which we train end-to-end.

4.2.1 Network Architecture

The architectures of our warping network $\mathbf{W}_\theta$ and color network $\mathbf{C}_\phi$ are inspired by Ledig et al. [LTH+16] and Isola et al. [IZZE16b], but we make a few important changes to better suit

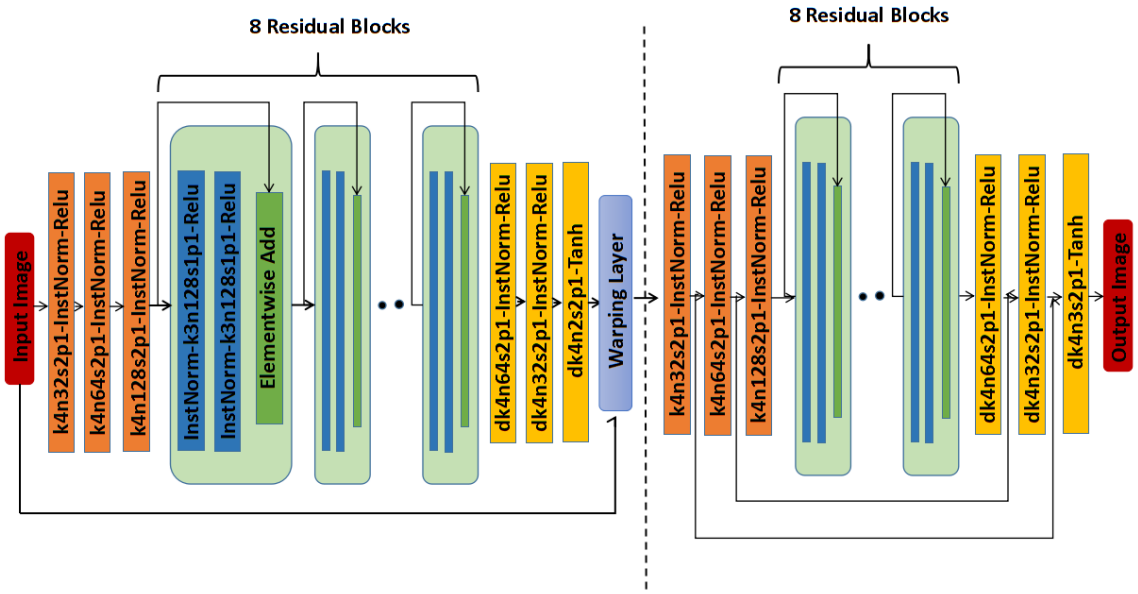

Figure 4.2: The network structure of our generator. For each convolutional layer, k represents the kernel size, n represents the number of feature maps, s represents the stride and p represents the padding size. Here, d represents the transpose convolutional layer.

our problem. Both our networks have the same general structure with only a few differences, which we now discuss in detail.

Both nets consist of three stride 2, size 4 convolution layers, followed by eight residual blocks, followed by three stride 2, size 4 deconvolution layers (see Figure 4.2). The output feature dimensions are 32, 64, 128, 64, 32, x, where x is a two channel warping field for the warp net and a 3 channel RGB image for the color net. Each residual block consists of two stride 1, size 3, dim 128 convolution layers followed by an additive skip connection, following the design of [HZRS16a]. We also add concatenation skip connections between corresponding convolution and deconvolution layers of the color net to help maintain fine details in the output image. This is not necessary for the warp net. Note that a similar, but much shallower, two stage network structure was proposed in [KWR16] for the problem of lightfield interpolation where the warps are small.

We find that normalization plays an important role in generalizing from our training set to real objects that have somewhat different color statistics (see Section 4.3). With standard batch normalization, we achieve the best results (in terms of L1 loss) on the training set, but observe bright blob artifacts when testing on real objects. This is due to the network over fitting to the color statistics of our training images. The problem is not alleviated solely by using instance normalization as suggested by [UVL16] because unlike them, we expect the network to preserve the brightness and contrast of the input image. To address this, we use instance normalization throughout our network, but save the mean and variance extracted from the input layer and use it to scale and shift the output.

4.2.2 Training Objective

We train our network by minimizing the L1 loss in pixel space

$$L_{con} = \sum_{\mathbf{x}} |\mathbf{I}(\mathbf{x}) - \mathbf{G}_{\theta\phi}(\mathbf{J}(\mathbf{x}), \mathbf{x})| \tag{4.4}$$

which we call the content loss. However, the L1 loss alone trains the network to predict the median image, which is often blurry and lacking in high frequency details. As in [IZZE16b, LTH+16] we also train our network with an adversarial loss to help encourage the predicted images to reside on the natural image manifold. This forces the network to produce sharp images with more fine details, and even hallucinate missing information from large distortions.

We train an additional discriminator network D_γ to distinguish between undistorted images from the generator and the natural non-distorted images, while the generator is trained to fool the discriminator. During the training process, the discriminator and generator are trained in an alternating manner to solve the min-max problem

$$\min_{\theta,\phi} \max_{\gamma} \; \mathbf{E}[\log D_\gamma(\mathbf{I})] + \mathbf{E}[\log(1 - D_\gamma(G_{\theta\phi}(\mathbf{J}))\,)] \tag{4.5}$$

For more stable training we use the Least Squares GAN objective [MLX+16]

$$L_{adv} = -(D_{\gamma}(G_{\theta\phi}(\mathbf{J})) - 1)^2 \tag{4.6}$$

The discriminator architecture follows the guidelines proposed in [RMC15]. We use 7 convolutional layers with kernel size 4 and stride 2 and increasing feature dimension (32,64,128,256,512,512,1). Each convolution except the last is followed by batch normalization and LeakyReLU activations. The last output is followed by a sigmoid activation. We also try the *PatchGAN* [IZZE16a] by decreasing the receptive field of discriminator to 70×70 and apply it through the image convolutionally. However, in our case, using *PatchGAN* does not improve the image quality.

Although the adversarial loss encourages more details, it also introduces some artifacts due to the unstable nature of GAN training. To combat this we follow [LTH+16] and add a perceptual loss defined by

$$L_{per} = \sum_{\mathbf{x}} |\psi(\mathbf{I}(\mathbf{x})) - \psi(\mathbf{G}_{\theta\phi}(\mathbf{J}(\mathbf{x}), \mathbf{x}))|, \tag{4.7}$$

where ψ is the output of an intermediate feature layer of a pretrained convolutional neural net. In our implementation we use the output of the conv4_3 layer of VGG.

Our final loss function is a weighted combination of the 3 losses

$$L = L_{con} + \lambda_{adv} L_{adv} + \lambda_{per} L_{per} \tag{4.8}$$

Training Detail: We largely follow the training scheme in [LTH+16] and [RMC15]. We first train the network with L1 loss alone from scratch and then fine tune the network adding adversarial loss and perceptual loss. The weight for adversarial loss and perceptual loss are 0.0005 and 0.3 respectively. Compared with [LTH+16], our weight for adversarial loss and perceptual loss is much lower and we do not remove L1 loss when fine tuning the network. This is because we

observe that L1 loss is important for our problem and if we remove L1 loss the network will not generate reasonable results. When training with L1 loss, we set the learning rate to be 0.001 and divide it by 10 after 15000 iterations. We train the network for 30,000 iterations with a batch size of 32. Then we fine tune the network adding perceptual loss and adversarial loss for 2000 iterations with learning rate 0.0002 and batch size 16.

4.3 Training Data

To train our deep network, we need a large training set. However, collecting a large number of images distorted by a water surface along with the corresponding non-distorted ground truth is challenging. There are no such existing large scale datasets. Tian et al. [TN09] provide a small dataset but that is not nearly enough to train a deep network.

Synthetic data is a natural option, but we found generating diverse enough water surfaces to be challenging. We tried using Gaussian Processes and the wave equation as in [TN09], as well as perturbing the surface with random Gaussian shaped drops. In each case, the network quickly over fit to the particular distribution of water surfaces generated and failed to generalize to real images. Creating diverse synthetic water surfaces is an interesting direction for future work.

Instead, we choose to construct a large dataset of distorted and non-distorted image pairs by capturing images of ImageNet images displayed under a water surface (see Figure 4.1). We place a computer monitor under a glass tank, which is filled with approximately 13cm water. The water is kept in motion using a small agitating pump. A Cannon 5D Mark IV is placed approximately 1.5m above the tank. Images are resampled using bilinear interpolation to fill the available screen space in the tank, after which the captured image is tightly cropped to its original shape and downsampled to its original size. The camera is set to f/1.2, ISO100, with exposure time of 1/320s. The camera is manually focused just beyond the monitor as this slight defocus removes the Moiré pattern observed in properly focused images.

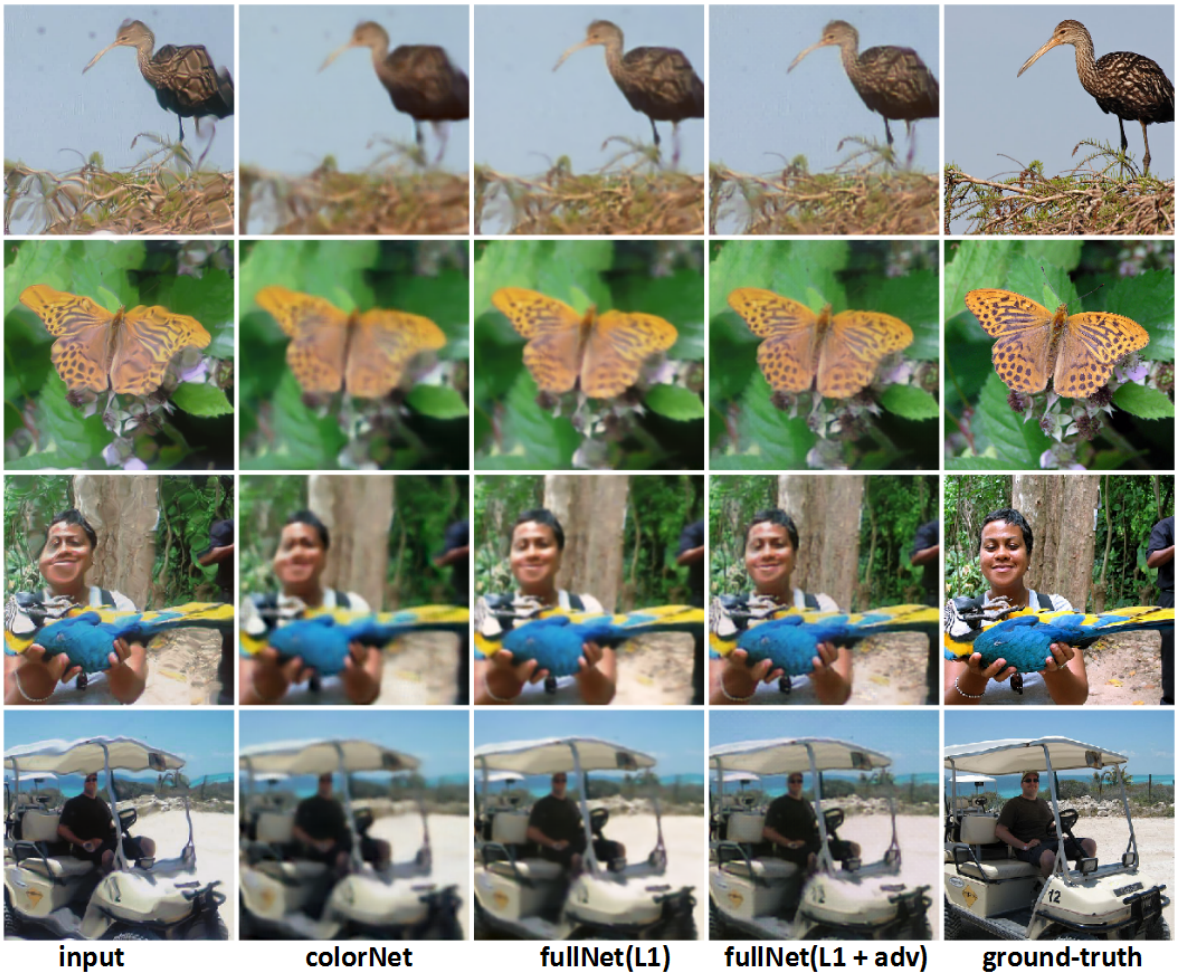

Figure 4.3: Qualitative results for ablative study on ImageNet validation test set. From left to right: input image, color net with L1 loss, warp+color net with L1 loss, warp+color net with L1+Adv+Per losses, ground truth. We observe that estimating the undistortion with the warp net significantly improves the geometry, while the adversarial loss allows better perceptual alignment to ground truth.

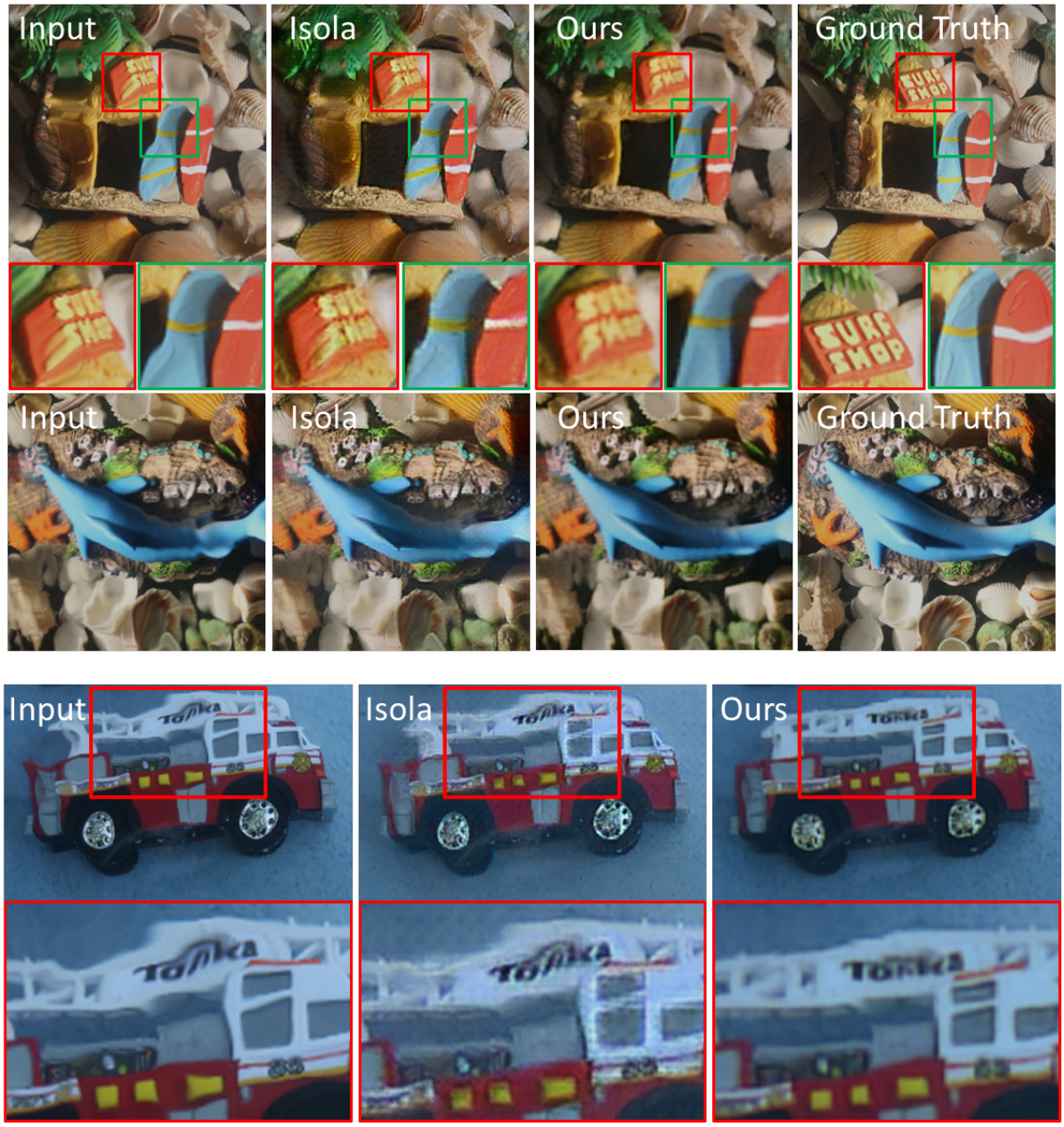

Figure 4.4: Results on real objects demonstrating generalization. (Rows 1 and 2) From left to right: input image in a larger tank, result of Isola et al. [IZZE16b], our result and ground truth. We observe that our framework that uses problem structure and careful normalizations produces better geometric undistortion and color outputs. (Bottom row) We show another example of further generalization by acquiring an image in a fountain pool. We see more significant contrasts relative to [IZZE16b], with clearly better undistortion performance for our method.

Table 4.1: Quantitative results for the ImageNet validation set. The network is trained with L1 loss, whereby we observe that L1 error reduces for the full network compared to warp or color net alone. As expected, the adversarial loss increases the L1 error, but allows for better appearance. For completion, we also show other metrics not directly related to the training, such as MSE, PSNR and SSIM.

Method	L1	MSE	PSNR	SSIM
colorNet	20.318	998.631	18.841	0.470
warpNet	20.140	961.978	19.035	0.490
fullNet(L1)	19.091	902.032	19.306	0.502
fullNet(adv+L1+per)	19.109	894.178	19.348	0.499

The process of displaying and recapturing the images changes the color space slightly due to nonlinear gamma curves and pixel sensitivity. To handle this, we pretrain a small color correction net that consists of 6 convolutional layers with receptive fields of size 1 to minimize the L1 distance between the captured image and the original ImageNet image. This mostly solves the problem, however we find that proper normalization in the net (as described in Section 4.2.1) is important for generalization to real objects. We collect 324,452 images from all 1000 ImageNet categories. We withhold 5 images from each category to form a validation set of 5000 images.

We note that creating a large real image dataset for 3D underwater scenes in the wild is extremely difficult. The intent of our training data collection is to easily generate sufficient volume in conditional similar but not identical to the application scenario. The choice of using flat images is a deliberate one, sacrificing realism for quantity. This is in line with several studies that use simulations for generating training data. Our laboratory setup similarly allows collecting large-scale data, but with reduced domain gap. Our experiments demonstrate generalization from the laboratory tank setup with flat images to real images of non-flat objects and in wild settings.

4.4 Results

We show results on our validation set captured by displaying ImageNet images on a monitor under a water surface, as well as images of real objects underwater. To demonstrate

generalization ability, the real objects are imaged in a different larger tank, as well as outdoors in a fountain pool.

In Figure 4.3, we show our results and an ablation study on the ImageNet validation set. In addition to the input image (column 1), our result (column 4) and ground truth (column 5), we also show two ablation results. The first is our color net alone trained with only the L1 loss (column 2). The second is our full generator architecture but trained with only the L1 loss (column 3). The five rows show the outputs for different input images.

We observe that the color net alone struggles to remove the large geometric distortions. Adding the warp net that accounts for the structure of the problem results in significantly better geometric undistortion, while also producing good colors. Next, adding the adversarial and perceptual loss has the effect of recovering sharp detailed images that are perceptually closer to the ground truth.

Figure 4.4 shows our results on real objects as well as a comparison to a state-of-the-art method for image to image translation [IZZE16b]. This method is similar to our color net alone but does not generalize well to real data due to the normalization issues discussed above. It also does not take advantage of domain knowledge that the transformation is a warp. Due to these factors, we observe that our method produces results that are closer to ground truth as compared to [IZZE16b]. This is emphasized by the insets, showing better geometric warp estimation in regions with long edges and also better color estimates than [IZZE16b] which produces subtle checkerboard artifacts.

Although no previous work in the water undistortion literature attempts the problem of single image blind undistortion, we note that methods such as [TN12] estimate a warping field by assuming the ground truth nondistorted image is known. In comparison, we do not require the assumption of a template, which might not exist in wild settings. Even in lab settings, acquiring a template requires careful alignment of images before and after the water surface is agitated. Other

works such as [TN09] additionally assume high frame rate video inputs, whereas we require only a single image.

Finally, in Figures 4.1 and 4.4, we show example outputs on an underwater sequence captured in a wild setting. We use the same network trained on ImageNet images observed through distortions in a tank, but the test images in this experiment are acquired outdoors at a water fountain. While there is no available ground truth, it is observed that the network generalizes quite well to this unseen condition, as reflected by the undistortion output that preserves edge shapes and displays plausible colors.

4.5 Summary

We have proposed a novel approach that uses deep learning to solve the previously unattempted problem of using a single image to remove distortions due to a refractive interface such as water surface. Since a turbulent water surface induces distortions that are too complex to be modeled as parametric or basis transformations, we use domain knowledge to model the distortion as a warp. This is different from general purpose image to image translation networks, which does not utilize problem structure. We demonstrate in experiments that our formulation as an end-to-end trainable two-stage network that estimates geometry and color, along with careful consideration of normalizations, leads to better results and generalization ability. To train our network, we collected a large scale dataset in lab settings with displayed images and show that it generalizes to images of real scenes imaged in different settings including unconstrained ones.

Chapter 4 is a reformatted version of "Learning to See through Turbulent Water", Z. Li, Z. Murez, D Kriegman, R. Ramamoorthi, M. Chandraker, *IEEE Winter Conf. on Applications of Computer Vision (WACV)* 2018 [LMK+18]. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Image to Image Translation for Domain Adaptation

In the previous chapter we trained a network using a dataset collected by imaging a computer monitor under a tank of turbulent water. We resorted to collecting this new dataset because our attempts to train using purely synthetic data failed to generalize to real data. Motivated by this, in this chapter we propose a new domain adaptation method.

The recent unprecedented advances in computer vision and machine learning are mainly due to: 1) deep (convolutional) neural architectures, and 2) existence of abundant labeled data. Deep convolutional neural networks (CNNs) [KSH12, HZRS16b, HLWvdM16] trained on large numbers of labeled images (tens of thousands to millions) provide powerful image representations that can be used for a wide variety of tasks including recognition, detection, and segmentation. On the other hand, obtaining abundant annotated data remains a cumbersome and expensive process in the majority of applications. Hence, there is a need for transferring the learned knowledge from a source domain with abundant labeled data to a target domain where data is unlabeled or sparsely labeled. The major challenge for such knowledge transfer is a phenomenon known as

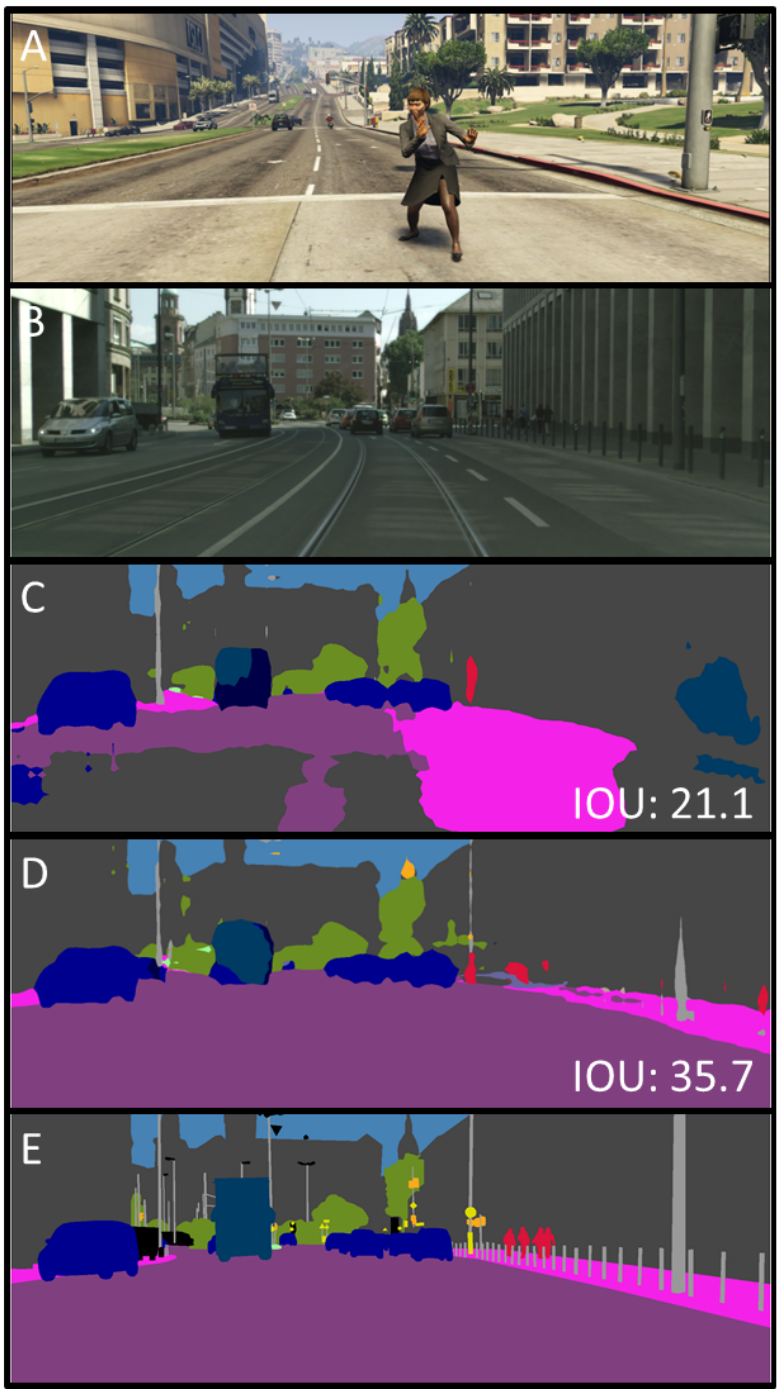

Figure 5.1: A) Sample image from the synthetic GTA5 dataset. B) Input image from the real Cityscapes dataset. C) Segmentation result trained on GTA5 dataset without any domain adaptation. D) Ours. E) Ground truth. We can see that our adaptation fixes large areas of simple mistakes on the road and sidewalk and building on the right. We also partially detect the thin pole on the right. The mean Intersection Over Union (IOU) values are reported.

domain shift [GSH+09], which refers to the different distribution of data in the target domain compared to the source domain.

To further motivate the problem, consider the emerging application of autonomous driving where a semantic segmentation network is required to be trained to detect roads, cars, pedestrians, etc. Training such segmentation networks requires semantic, instance-wise, dense pixel annotations for each scene, which is excruciatingly expensive and time consuming to acquire. To avoid human annotations, a large body of work focuses on designing photo-realistic simulated scenarios in which the ground truth annotations are readily available. Synthia [RSM+16], Virtual KITTI [GWCV16], and GTA5 [RVRK16] datasets are examples of such simulations, which include a large number of synthetically generated driving scenes together with ground truth pixel-level semantic annotations. Training a CNN based on such synthetic data and applying it to real-world images (i.e. from a dashboard mounted camera), such as the Cityscapes dataset [COR+16], will give very poor performance due to the large differences in image characteristics which gives rise to the domain shift problem. Figure 5.1 demonstrates this scenario where a network is trained on the GTA5 dataset [RVRK16], which is a synthetic dataset, for semantic segmentation and is tested on the Cityscapes dataset [COR+16]. It can be seen that with no adaptation the network struggles with segmentation (Figure 5.1, C), while our proposed framework ameliorates the domain shift problem and provides a more accurate semantic segmentation.

Domain adaptation techniques aim to address the domain shift problem, by finding a mapping from the source data distribution to the target distribution. Alternatively, both domains could be mapped into a shared domain where the distributions are aligned. Generally, such mappings are not unique and there exist many mappings that align the source and target distributions. Therefore various constraints are needed to narrow down the space of feasible mappings. Recent domain adaptation techniques parameterize and learn these mappings via deep neural networks [THDS15, LCWJ15, THSD17, LZHFF17, SBCC17, SPT+]. In this paper, we propose a unifying, generic, and systematic framework for unsupervised domain adaptation, which is

broadly applicable to many image understanding and sensing tasks where training labels are not available in the target domain. We further demonstrate that many existing methods for domain adaptation arise as special cases of our framework.

While there are significant differences between the recently developed domain adaptation methods, a common and unifying theme among these methods can be observed. We identify three main attributes needed to achieve successful unsupervised domain adaptation: 1) domain agnostic feature extraction, 2) domain specific reconstruction, and 3) cycle consistency. The first requires that the distributions of features extracted from both domains are indistinguishable (as judged by an adversarial discriminator network). This idea was utilized in many prior methods [HWYD16, GL15, GUA$^+$16], but alone does not give a strong enough constraint for domain adaptation knowledge transfer, as there exist many mappings that could match the source and target distributions in the shared space. The second is requiring that the features are able to be decoded back to the source and target domains. This idea was used in Ghifary et al. [GKZ$^+$16] for unsupervised domain adaptation. Finally, the cycle consistency is needed for unpaired source and target domains to ensure that the mappings are learned correctly and they are well-behaved, in the sense that they do not collapse the distributions into single modes [ZPIE17]. Figure 5.2 provides a high-level overview of our framework.

The interplay between the 'domain agnostic feature extraction', 'domain specific reconstruction with cycle consistency', and 'label prediction from agnostic features' enables our framework to simultaneously learn from the source domain and adapt to the target domain. By combining all these different components into a single unified framework we build a systematic framework for domain knowledge transfer that provides an elegant theoretical explanation as well as improved experimental results. We demonstrate the superior performance of our proposed framework for segmentation adaptation from synthetic images to real world images (See Figure 5.1 as an example), as well as for classifier adaptation on three digit datasets. Furthermore, we

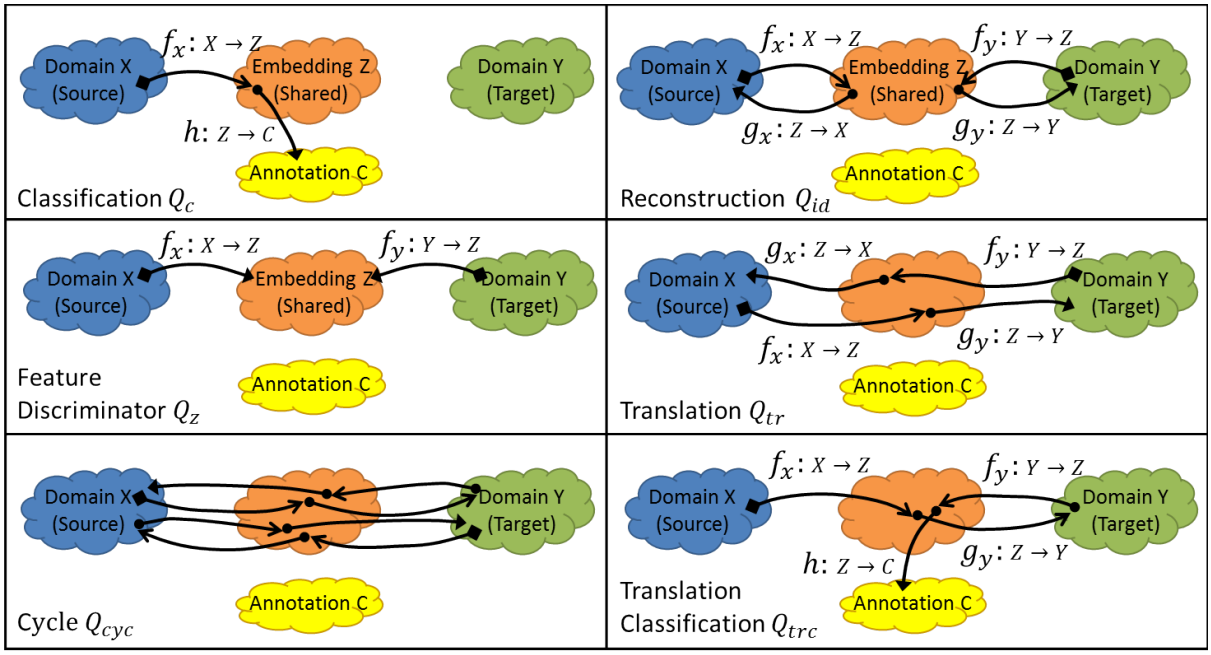

Figure 5.2: The detailed system architecture of our I2I (image to image) Adapt framework. The pathways to the loss modules denote the inputs to these modules, which are used for training. Best viewed in color.

show that many of the State Of the Art (SOA) methods can be viewed as special cases of our proposed framework. Code is available at `https://github.com/zmurez/I2IAdapt`.

5.1 Related Work

There has been a plethora of recent work in the field of visual *domain adaptation* addressing the domain shift problem [GSH+09], otherwise known as the *dataset bias problem*. The majority of recent work use deep convolutional architectures to map the source and target domains into a shared space where the domains are aligned [THZ+14, THDS15, THSD17, HWYD16, SBCC17]. These methods widely differ on the architectures as well as the choices of loss functions used for training them. Some have used Maximum Mean Discrepancy (MMD) between the distributions of the source and target domains in the shared space [LCWJ15], while others have used correlation maximization to align the second-order statistics of the domains. Another popular

and effective choice is maximizing the confusion rate of an adversarial network, that is required to distinguish the source and target domains in the shared space [THDS15, HWYD16, GL15, GUA+16, GKZ+16]. Other approaches include the work by Sener et al. [SSSS16], where the domain transfer is formulated in a transductive setting, and the Residual Transfer Learning (RTL) approach [LZWJ16] where the authors assume that the source and target classifiers only differ by a residual function and learn these residual functions.

Our work is primarily motivated by the work of Hoffman et al. [HWYD16], Isola et al. [IZZE16b], Zhu et al. [ZPIE17], and Ghifary et al. [GKZ+16]. Hoffman et al. [HWYD16] utilized fully convolutional networks with domain adversarial training to obtain domain agnostic features (i.e. shared space) for the source and target domains, while constraining the shared space to be discriminative for the source domain. Hence, by learning the mappings from source and target domains to the shared space (i.e. f_x and f_y in Figure 5.2), and learning the mapping from the shared space to annotations (i.e. h in Figure 5.2), their approach effectively enables the learned classifier to be applicable to both domains. The Deep Reconstruction Classification Network (DRCN) of Ghifary et al. [GKZ+16], utilizes a similar approach but with a constraint that the embedding must be decodable, and learns a mapping from the embedding space to the target domain (i.e. g_y in Figure 5.2). The image-to-image translation work by Isola et al. [IZZE16b] maps the source domain to the target domain by an adversarial learning of f_x and g_y and composing them $g_y \circ f_x : X \to Y$. In their framework the target and source images were assumed to be paired, in the sense that for each source image there exists a known corresponding target image. This assumption was lifted in the follow-up work of Zhu et al. [ZPIE17] and Royer et al.[RBG+17], where cycle consistency was used to learn the mappings based on unpaired source and target images. While the approaches of Isola et al. [IZZE16b] and Zhu et al. [ZPIE17] do not address the domain adaptation problem, yet they provide a baseline for learning high quality mappings from a visual domain into another.

The patterns that collectively emerge from the mentioned papers [THZ+14, HWYD16, IZZE16b, GKZ+16, ZPIE17], are: a) the shared space must be a discriminative embedding for the source domain, b) the embedding must be domain agnostic, hence maximizing the similarity between the distributions of embedded source and target images, c) the information preserved in the embedding must be sufficient for reconstructing domain specific images, d) adversarial learning as opposed to the classic losses can significantly enhance the quality of learned mappings, e) cycle-consistency is required to reduce the space of possible mappings and ensure their quality, when learning the mappings from unpaired images in the source and target domains. Our proposed method for unsupervised domain adaptation unifies the above-mentioned pieces into a generic framework that simultaneously solves the domain adaptation and image-to-image translation problems.

There have been other recent efforts toward a unifying and general framework for deep domain adaptation. The Adversarial Discriminative Domain Adaptation (ADDA) work by Tzeng et al. [THSD17] is an instance of such frameworks. Tzeng et al. [THSD17] identify three design choices for a deep domain adaptation system, namely a) whether to use a generative or discriminative base, whether to share mapping parameters between f_x and f_y, and the choice of adversarial training. They observed that modeling image distributions might not be strictly necessary if the embedding is domain agnostic (i.e. domain invariant).

Very similar ideas to ours have been published concurrently by Hoffman et al. [HTP+17] and Liu et al. [LBK17].

5.2 Method

Consider training images $x_i \in X$ and their corresponding annotations/labels $c_i \in C$ from the source domain (i.e. domain X). Note that c_i may be image level such as in classification or pixel level in the case of semantic segmentation. Also consider training images $y_j \in Y$ in the

target domain (i.e. domain Y), where we do not have corresponding annotations for these images. Our goal is then to learn a classifier that maps the target images, y_js, to labels C. We note that the framework is readily extensible to a semi-supervised learning or few-shot learning scenario where we have annotations for a few images in the target domain. Given that the target domain lacks labels, the general approach is to learn a classifier on the source domain and adapt it in a way that its domain distribution matches that of the target domain.

The overarching idea here is to find a joint latent space, Z, for the source and target domains, X and Y, where the representations are domain agnostic. To clarify this point, consider the scenario in which X is the domain of driving scenes/images on a sunny day and Y is the domain of driving scenes on a rainy day. While 'sunny' and 'rainy' are characteristics of the source and target domains, they are truly nuisance variations with respect to the annotation/classification task (e.g. semantic segmentation of the road), as they should not affect the annotations. Treating such characteristics as structured noise, we would like to find a latent space, Z, that is invariant to such variations. In other words, domain Z should not contain domain specific characteristics, hence it should be domain agnostic. In what follows we describe the process that leads to finding such a domain agnostic latent space.

Let the mappings from source and target domains to the latent space be defined as $f_x : X \to Z$ and $f_y : Y \to Z$, respectively (See Figure 5.2). In our framework these mappings are parameterized by deep convolutional neural networks (CNNs). Note that the members of the latent space $z \in Z$ are high dimensional vectors in the case of image level tasks, or feature maps in the case of pixel level tasks. Also, let $h : Z \to C$ be the classifier that maps the latent space to labels/annotations (i.e. the classifier module in Figure 5.2). Given that the annotations for the source class X are known, one can define a supervised loss function to enforce $h(f_x(x_i)) = c_i$:

$$Q_c = \sum_i l_c \left(h(f_x(x_i)), c_i\right) \tag{5.1}$$

where l_c is an appropriate loss (e.g. cross entropy for classification and segmentation). Minimizing the above loss function leads to the standard approach of supervised learning, which does not concern domain adaptation. While this approach would lead to a method that performs well on the images in the source domain, $x_i \in X$, it will more often than not perform poorly on images from the target domain $y_j \in Y$. The reason is that, domain Z is biased to the distribution of the structured noise ('sunny') in domain X and the structured noise in domain Y ('rainy') confuses the classifier $h(\cdot)$. To avoid such confusion we require the latent space, Z, to be domain agnostic, so it is not sensitive to the domain specific structured noise. To achieve such a latent space we systematically introduce a variety of auxiliary networks and losses to help regularize the latent space and consequently achieve a robust $h(\cdot)$. The auxiliary networks and loss pathways are depicted in Figure 5.2. In what follows we describe the individual components of the regularization losses.

1. First of all Z is required to preserve the core information of the target and source images and only discard the structured noise. To impose this constraint on the latent space, we first define decoders $g_x : Z \to X$ and $g_y : Z \to Y$ that take the features in the latent space to the source and target domains, respectively. We assume that if Z retains the crucial/core information of the domains and only discards the structured noise, then the decoders should be able to add the structured noise back and reconstruct each image from their representation in the latent feature space, Z. In other words, we require $g_x(f_x(\cdot))$ and $g_y(f_y(\cdot))$ to be close to identity functions/maps. This constraint leads to the following loss function:

$$Q_{id} = \sum_i l_{id}\left(g_x(f_x(x_i)), x_i\right) + \sum_j l_{id}\left(g_y(f_y(y_j)), y_j\right) \tag{5.2}$$

   where $l_{id}(\cdot,\cdot)$ is a pixel-wise image loss such as the L_1 norm.

2. We would like the latent space Z to be domain agnostic. This means that the feature representations of the source and target domain should not contain domain specific information. To achieve this, we use an adversarial setting in which a discriminator $d_z : Z \to \{c_x, c_y\}$

tries to classify if a feature in the latent space $z \in Z$ was generated from domain X or Y, where c_x and c_y are binary domain labels (i.e. from domain X or domain Y). The loss function then can be defined as the certainty of the discriminator (i.e. domain agnosticism is equivalent to fooling the discriminator), and therefore we can formulate this as:

$$Q_z = \sum_i l_a\left(d_z(f_x(x_i)), c_x\right) + \sum_j l_a\left(d_z(f_y(y_j)), c_y\right) \tag{5.3}$$

where $l_a(\cdot,\cdot)$ is an appropriate loss (the cross entropy loss in traditional GANs [GPAM$^+$14b] and mean square error in least squares GAN [MLX$^+$16]). The discriminator is trained to maximize this loss while the discriminator is trained to minimize it.

3. To further ensure that the mappings f_x, f_y, g_x, and g_y are consistent we define translation adversarial losses. An image from target (source) domain is first encoded to the latent space and then decoded to the source (target) domain to generate a 'fake' (translated) image. Next, we define discriminators $d_x : X \to \{c_x, c_y\}$ and $d_y : Y \to \{c_x, c_y\}$, to identify if an image is 'fake' (generated from the other domain) or 'real' (belonged to the actual domain). To formulate this translation loss function we can write:

$$Q_{tr} = \sum_i l_a\left(d_y(g_y(f_x(x_i))), c_x\right) + \sum_j l_a\left(d_x(g_x(f_y(y_j)), c_y\right) \tag{5.4}$$

4. Given that there are no correspondences between the images in the source and target domains, we need to ensure that the semantically similar images in both domains are projected into close vicinity of one another in the latent space. To ensure this, we define the cycle consistency losses where the 'fake' images generated in the translation loss, $g_x(f_y(y_j))$ or $g_y(f_x(x_i))$, are encoded back to the latent space and then decoded back to their original space. The entire cycle should be equivalent to an identity mapping. We can

Table 5.1: Showing the relationship between the existing methods and our proposed method.

Method	λ_c	λ_z	λ_{tr}	λ_{idX}	λ_{idY}	λ_{cyc}	λ_{trc}
[HWYD16]	✓	✓					
[THSD17]	✓	✓					
[GKZ+16]	✓				✓		
[SBCC17]	✓		✓				
[ZPIE17]			✓			✓	
Ours	✓	✓	✓	✓	✓	✓	✓

formulate this loss as follows:

$$Q_{cyc} = \sum_i l_{id}\left(g_x(f_y(g_y(f_x(x_i)))), x_i\right) + \sum_j l_{id}\left(g_y(f_x(g_x(f_y(y_j)))), y_j\right) \tag{5.5}$$

5. To further constrain the translations to maintain the same semantics, and allow the target encoder to be trained with supervision on target domain 'like' images we also define a classification loss between the source to target translations and the original source labels:

$$Q_{trc} = \sum_i l_c\left(h(f_y(g_y(f_x(x_i)))), c_i\right) \tag{5.6}$$

Finally, by combining these individual losses we define the general loss to be,

$$Q = \lambda_c Q_c + \lambda_z Q_z + \lambda_{tr} Q_{tr} + \lambda_{id} Q_{id} + \lambda_{cyc} Q_{cyc} + \lambda_{trc} Q_{trc} \tag{5.7}$$

A variety of prior methods for domain adaptation are special cases of our framework. Table 5.1 summarizes which hyperparameters to include and which are set to zero to recover these prior methods.

Table 5.2: Performance of various prior methods as well as ours and ablations on digits datasets domain adaptation. MNIST → USPS indicates MNIST is the source domain (labels available) and USPS is the target domain (no labels available). Results reported are classification error rate (lower is better). Blue is best prior method, bold is best overall. Our results are considerably better than the prior state of the art.

Method						MNIST → USPS	USPS → MNIST	SVHN → MNIST
Source only						24.8	42.9	39.7
Gradient reversal [GUA+16]						28.9	27.0	26.1
Domain confusion [THDS15]						20.9	33.5	31.9
CoGAN [LT16]						8.8	10.9	-
ADDA [THSD17]						10.6	9.9	24.0
DTN [TPW16]						-	-	15.6
WDAN [YDL+17]						27.4	34.6	32.6
PixelDA [BSD+17]						4.1	-	-
DRCN [GKZ+16]						8.2	26.3	18.0
Gen to Adapt [SBCC17]						7.5	9.2	15.3
λ_z	λ_{tr}	λ_{id_X}	λ_{id_Y}	λ_{cyc}	λ_{trc}	Ours		
						8.9	33.0	28.5
✓						2.1	2.8	19.9
			✓			2.1	28.5	23.7
			✓		✓	1.4	22.6	29.5
✓			✓			1.3	3.0	12.1
✓			✓		✓	**1.2**	**2.4**	**9.9**
✓	✓	✓	✓		✓	2.5	3.6	10.0
✓	✓	✓	✓	✓	✓	1.5	2.6	10.4

Table 5.3: Accuracy (larger is better) of various methods on the Office dataset consisting of three domains: Amazon (A), Webcam (W) and DSLR (D). A $\rightarrow$ W indicates Amazon is the source domain (labels available) and Webcam is the target domain (no labels available). Bold is best. Our method performs best on 4 out of 6 of the tasks.

Method					A $\rightarrow$ W	W $\rightarrow$ A	A $\rightarrow$ D	D $\rightarrow$ A	W $\rightarrow$ D	D $\rightarrow$ W
Domain confusion [THZ$^+$14]					61.8	52.2	64.4	21.1	98.5	95.0
Transferable Features [LCWJ15]					68.5	53.1	67.0	54.0	99.0	96.0
Gradient reversal [GUA$^+$16]					72.6	52.7	67.1	54.5	99.2	96.4
DHN [VECP17]					68.3	53.0	66.5	55.5	98.8	96.1
WDAN [YDL$^+$17]					66.8	52.7	64.5	53.8	98.7	95.9
DRCN [GKZ$^+$16]					68.7	**54.9**	66.8	**56.0**	99.0	96.4
λ_z	λ_{tr}	λ_{id_X}	λ_{id_Y}	λ_{trc}	Ours					
					59.1	46.4	61.0	45.3	98.0	92.8
✓					70.8	49.0	67.1	43.4	98.2	90.8
			✓		61.1	49.6	67.3	49.8	99.0	94.7
✓			✓	✓	71.2	49.1	70.9	45.5	97.8	94.3
✓	✓	✓	✓	✓	**75.3**	52.1	**71.1**	50.1	**99.6**	**96.5**

Table 5.4: Performance (Intersection over Union) of various methods on driving datasets domain adaptation. Above the line uses the standard dilated ResNet as the encoder. Our method performs the best overall and on all sub categories except two. Switching to a DenseNet encoder beats the previous method even without domain adaptation. DenseNet plus our method significantly out performs the previous method. Blue is best with ResNet, Bold is best overall.

Method	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Source only	31.9	18.9	47.7	7.4	3.1	16.0	10.4	1.0	76.5	13.0	58.9	36.0	1.0	67.1	9.5	3.7	0.0	0.0	0.0	21.1
FCNs in the Wild [HWYD16]	67.4	29.2	64.9	15.6	8.4	12.4	9.8	2.7	74.1	12.8	66.8	38.1	2.3	63.0	9.4	5.1	0.0	3.5	0.0	27.1
Ours	85.3	**38.0**	71.3	18.6	16.0	18.7	12.0	4.5	72.0	43.4	63.7	43.1	3.3	76.7	14.4	12.8	0.3	9.8	0.6	31.8
Source only - DenseNet	67.3	23.1	69.4	13.9	14.4	21.6	**19.2**	**12.4**	78.7	24.5	74.8	49.3	3.7	54.1	8.7	5.3	2.6	6.2	**1.9**	29.0
Ours - DenseNet	**85.8**	37.5	**80.2**	**23.3**	**16.1**	**23.0**	14.5	9.8	**79.2**	**36.5**	**76.4**	**53.4**	**7.4**	**82.8**	**19.1**	**15.7**	**2.8**	**13.4**	1.7	**35.7**

5.3 Experiments

The loss function is optimized via the ADAM method with learning rate 0.0002 and betas 0.5 and 0.999, in an end-to-end manner. The discriminative networks, d_x, d_y, and d_z are trained in an alternating optimization alongside with the encoders and decoders.

To further constrain the features that are learned we share the weights of the encoders. We also share the weights of the first few layers of the decoders. To stabilize the discriminators we train them using the Improved Wasserstein method [GAA+17]. The loss of the feature discriminator (Q_z) is only backpropagated to the generator for target images (we want the encoder to learn to map the target images to the same distribution as the source images, not vice versa). Likewise, the translation classification loss (Q_{trc}) is only backpropagated to the second encoder and classifier (f_y and h). This prevents the translator (g_y) from cheating by hiding class information in the translated images.

5.3.1 MNIST, USPS, and SVHN digits datasets

First, we demonstrate our method on domain adaptation between three digit classification datasets, namely MNIST [LBBH98], USPS [Hul94], and the Street View House Numbers (SVHN) [NWC+11] datasets. We followed the experimental protocol of [GUA+16, THDS15, LT16, THSD17, SBCC17] where we treated one of the digit datasets as a labeled source domain and another dataset as unlabeled target domain. We trained our framework for adaptation from MNIST→ USPS, USPS → MNIST, and SVHN → MNIST. Figure 5.3 shows examples of MNIST to SVHN input and translated images.

For a fair comparison with previous methods, our feature extractor network (encoder, f_x and f_y) is a modified version of LeNet [LBBH98]. Our decoders (i.e. g_x and g_y) consist of three transposed convolutional layers. Our image discriminators consist of three convolutional layers and our feature discriminator consists of three fully connected layers. All images from MNIST

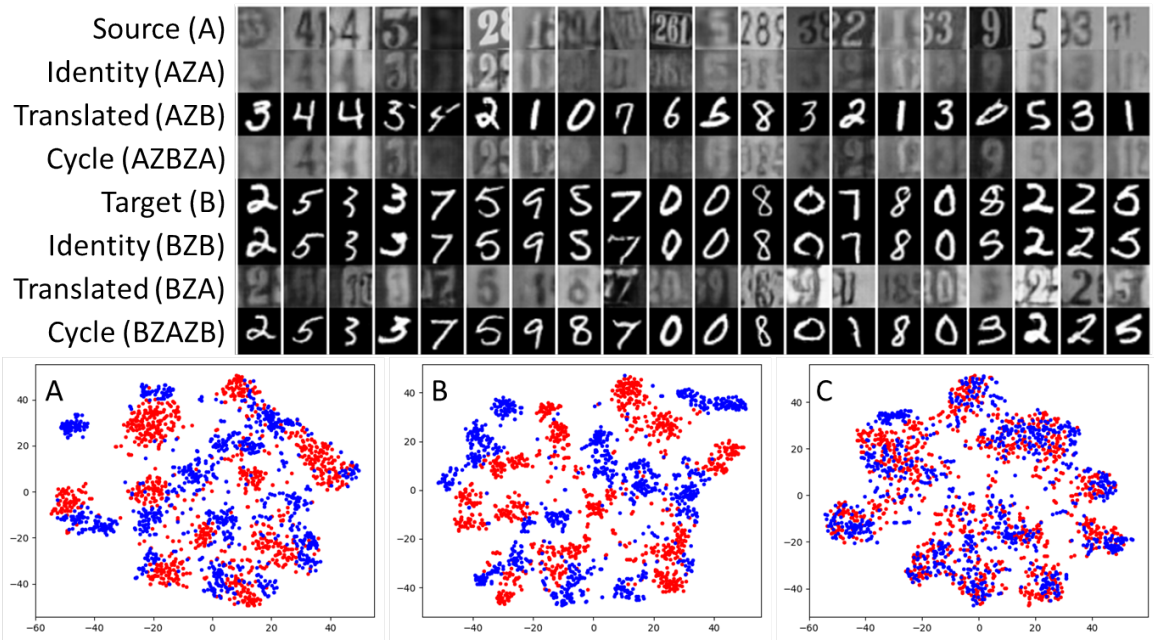

Figure 5.3: Top) Image to image translation examples for MNIST to SVHN. Bottom) TSNE embedding visualization of the latent space. Red are source images, Blue are target images. A) No adaptation. B) Image to image adaptation without latent space discriminator. C) Full adaptation.

and USPS were bilinearly upsampled to 32x32. Images from SVHN were converted to gray scale. We also included very simple data augmentation in the form of random translations and rotations as it helped a lot. For our hyperparameters we used: $\lambda_c = 1.0$, $\lambda_z = 0.05$, $\lambda_{id} = 0.1$, $\lambda_{tr} = 0.02$, $\lambda_{cyc} = 0.1$, $\lambda_{trc} = 0.1$.

We compare our method to nine prior works (see Table. 5.2). Our method consistently out performs the prior state of the art by a significant margin. We also show ablations to analyze how much each of the loss terms contributes to the overall performance. First note that even our results without domain adaption (top row of Table. 5.2 below the line) are better than many prior methods. This is purely due to the simple data augmentation. Next we see that λ_z and λ_{id_Y} both improve results a lot. However the combination of them with λ_{trc} produces the best results. This is especially apparent in SVHN to MNIST, which is the most challenging. Finally, in this case, the remaining losses don't help any further.

Figure 5.3 A,B,C show TSNE embeddings of the features extracted from the source and target domain when trained without adaptation, with image to image loss only, and our full model. It can be seen that without adaptation, the source and target images get clustered in the feature space but the distributions do not overlap which is why classification fails on the target domain. Just image to image translation is not enough to force the distributions to overlap as the networks learn to map source and target distributions to different areas of the feature space. Our full model includes a feature distribution adversarial loss, forcing the source and target distributions to overlap, while image translation makes the features richer yielding the best adaptation results.

5.3.2 Office dataset

The Office dataset [SKFD10] consists of images from 31 classes of objects in three domains: Amazon (A), Webcam (W) and DSLR (D) with 2817, 795 and 498 images respectively. Our method performs the best in four out of six of the tasks (see Table 5.3). The two tasks that ours is not best at consist of bridging a large domain shift with very little training data in the source domain (795 and 498 respectively). Here the ablations show that the translation loss (Q_{tr}) helps.

For our encoder we use a ResNet34 pretrained on ImageNet. The encoder is trained with a smaller learning rate (2×10^{-5}), to keep the weights closer to their good initialization. Images are down sampled to 256x256 and then a random crop of size 224x244 is extracted. The final classification layer is applied after global average pooling. Our decoders consist of 5 stride 2 transposed convolutional layers. The image discriminators consist of 4 stride 2 convolutional layers. The feature discriminator consists of 3 1x1 convolutions followed by global average pooling.

Our hyperparameters were: $\lambda_c = 1.0$, $\lambda_z = 0.1$, $\lambda_{tr} = 0.005$, $\lambda_{id} = 0.2$, $\lambda_{cyc} = 0.0$, $\lambda_{trc} = 0.1$.

5.3.3 GTA5 to Cityscapes

We also demonstrate our method for domain adaptation between the synthetic (photorealistic) driving dataset GTA5 [RVRK16] and the real dataset Cityscapes [COR+16]. The GTA5 dataset consists of 24,966 densely labeled RGB images of size 1914×1052, containing 19 classes that are compatible with the Cityscapes dataset (See Table 5.4). The Cityscapes dataset contains 5,000 densely labeled RGB images of size 2040×1016 from 27 different cities. Here the task is pixel level semantic segmentation. Following the experiment in [HWYD16], we use the GTA5 images as the labeled source dataset and the Cityscapes images as the unlabeled target domain.

We point out that the convolutional networks in our model are interchangeable. We include results using a dilated ResNet34 encoder for fair comparison with previous work, but we found from our experiments that the best performance was achieved by using our new Dilated Densely-Connected Networks (i.e. Dilated DenseNets) for the encoders which are derived by replacing strided convolutions with dilated convolutions [YKF17] in the DenseNet architecture [HLWvdM16]. DenseNets have previously been used for image segmentation [JDV+17] but their encoder/decoder structure is more cumbersome than what we proposed. We use a series of transposed convolutional layers for the decoders.

Our decoders consist of a stride 1 convolutional layer followed by 3 stride 2 transposed convolutional layers. The image discriminators consist of 4 stride 2 convolutional layers. We did not include the cycle consistency constraint due to memory issues. Due to computational and memory constraints, we down sample all images by a factor of two prior to feeding them into the networks. Output segmentations are bilinearly up sampled to the original resolution. We train our network on 256x256 patches of the down sampled images, but test on the full (downsampled) images convolutionally. Our hyperparameters were: $\lambda_c = 1.0$, $\lambda_z = 0.01$, $\lambda_{tr} = 0.04$, $\lambda_{id} = 0.2$, $\lambda_{cyc} = 0.0$, $\lambda_{trc} = 0.1$.

Our encoder architecture (dilated ResNet/DenseNet) is optimized for segmentation and thus it is not surprising that our translations (see Figure. 5.4) are not quite as good as those

reported in [ZPIE17]. Qualitatively, it can be seen from Figure 5.4 that our segmentations are much cleaner compared to no adaptation. Quantitatively (see Table 5.4), our method outperforms the previous method [HWYD16] on all categories except 3, and is 5% better overall. Furthermore, we show that using Dilated DenseNets in our framework, increases the SOA by 8.6%.

5.4 Summary

We have proposed a general framework for unsupervised domain adaptation which encompasses many recent works as special cases. Our proposed method simultaneously achieves image to image translation, source discrimination, and domain adaptation.

Our implementation outperforms state of the art on adaptation for digit classification and semantic segmentation of driving scenes. When combined with the DenseNet architecture our method significantly outperforms the current state of the art.

Chapter 5 is a reformatted version of "Image to Image Translation for Domain Adaptation", Z. Murez, S. Kolouri, D Kriegman , R. Ramamoorthi, K. Kim, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2018. [MKK+17] The dissertation author was the primary investigator and author of this paper.

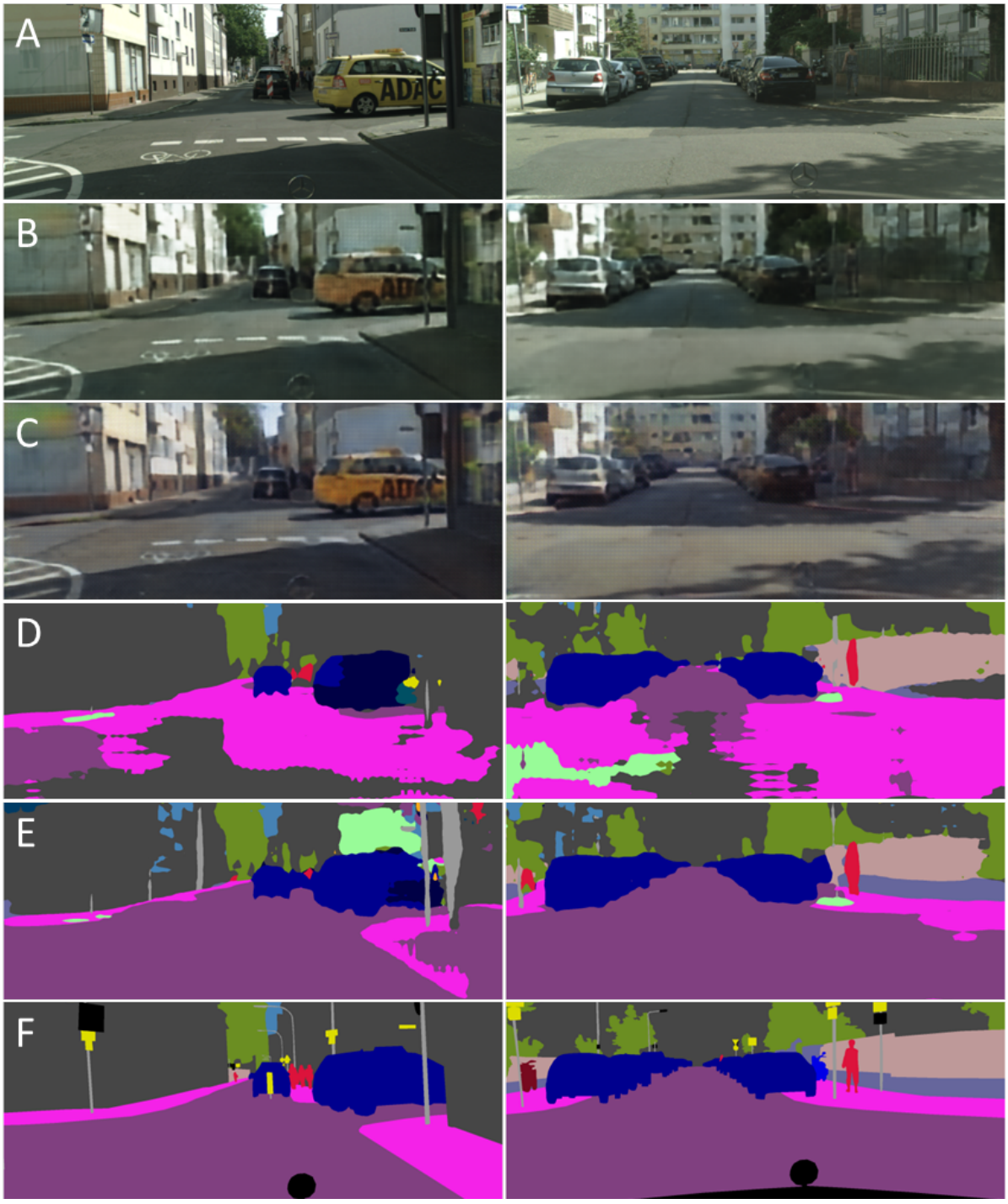

Figure 5.4: A) Input image from real Cityscapes dataset. B) Identity mapped image. C) Translated image. D) Segmentation without domain adaptation. E) Our Segmentation. F) Ground truth. Although our image translations might not be as visually pleasing as those in [ZPIE17] (our architecture is not optimized for translation), they succeed in their goal of domain adaptation.

Chapter 6

Conclusion

We have presented work which relaxes some of the assumptions of shape from shading and photometric stereo within a physics based framework. We have also shown that deep learning can be applied to problems traditionally approached by physics based methods. This is especially successful when inspiration from the physical models can be incorporated, either into the architecture, or training data. Finally we proposed a novel method for unsupervised domain adaptation within a deep learning framework. This has many potential applications, but one of the most exciting is training with synthetic data, especially for physics based problems where human annotations are often difficult to collect.

In chapter 2 we saw that due to its isotropic emission, fluorescence is an ideal input to shape from shading and photometric stereo algorithms. We also showed that in some cases, fluorescence can remove inter-reflections, and that when combined with reflectance can resolve the Generalized Bas Relief ambiguity of uncalibrated photometric stereo. However, in some cases, the fluorescence emission might not be ideally Lambertian due to differences in subsurface absorption. The characterization of such cases requires more measurements of different fluorescent materials. When the emitted radiance due to fluorescence is isotropic, other computer vision methods that rely on the constant brightness assumption such as binocular stereopsis, multi view reconstruction,

and optical flow estimation can be applied to fluorescence images. However, fluorescence typically does not exhibit the same type of spatial/texture variation as would be found with reflectance texture, and this might provide an alternate set of challenges.

In chapter 3 we modeled the affects of scattering to improve the quality of photometric stereo in a scattering medium. We showed that fluorescence can be used to optically remove backscatter with much higher signal-to-noise ratios than subtraction methods. We also showed that forward scatter from the source can be calibrated out, and that forward scatter blur from the object can be removed with deconvolution. Although our theory only applies to a single scattering medium, in practice, our calibrated PSF may be taking multiple scattering effects into account. Extending our theory to multiple scattering would provide further insight. Future work includes removing the need to know the average object distance, removing the *small surface variations approximation*, and an automated PSF calibration procedure for varying turbidities and depths.

In chapter 4 we showed that a deep convolutional neural network can be trained to solve the traditionally physics based problem of dynamic refractive distortion correction. We collected a large scale dataset by imaging a monitor under a tank of turbulent water. Future work includes training the network on purely synthetic data, with the aid of domain adaptation. Another direction is to train the network with real data but without paired examples. By this we mean that we have access to distorted images and non-distorted images, but not to the non-distorted image corresponding to a particular distorted image. The advantage is that this data is easy to capture, as it only becomes difficult to capture when you require ground truth pairs. The best solution is probably a mixture of unpaired examples combined with synthetic data. Other future work includes dealing with other distortions including volumetric scattering and surface reflection.

In chapter 5 we proposed a novel unsupervised domain adaptation method based upon adversarial discriminative feature matching and image-to-image translation. We evaluated our method by achieving state-of-the-art results on benchmark domain adaptation datasets, but believe this will be particularly useful for traditionally physics based problems where synthetic data is

easy to generate but real data is hard to annotate (for example the problem addressed in chapter 4). Other directions for future work include multi-domain adaptation as well as combining unsupervised domain adaptation with a few examples in the target domain. Furthermore, in these cases each domain need not have annotations for all the classes, yielding a zero-shot-learning hybrid.

Computer vision has recently gone from a purely academic venture to something that is being widely used throughout industry to solve real world problems. However most of this success has been limited to solving classification, detection and segmentation problems using deep learning. This thesis has pushed the state-of-the-art in physics based shape reconstruction as well as deep learning, and has brought the two closer together. Hopefully this will provide a stepping stone for further unifying deep learning with physics based vision and lead to similar levels of success and adoption in those problems.

Bibliography

[ACM15] P. Agrawal, J. Carreira, and J. Malik. Learning to See by Moving. In *ICCV*, 2015.

[AF06] A.H. Ahmed and A.A. Farag. A new formulation for shape from shading for non-lambertian surfaces. In *Proc. IEEE CVPR*, 2006.

[AMK07] Neil G. Alldrin, Satya Mallick, and David Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. In *Proc. IEEE CVPR*, 2007.

[ARC06] Amit Agrawal, Ramesh Raskar, and Rama Chellappa. What is the range of surface reconstructions from a gradient field? In *Proc. ECCV*, pages 578–591. Springer, 2006.

[AS00] M. Ashikhmin and P. Shirley. An anisotropic phong BRDF model. *J. of graphics tools*, 5(2):25–32, 2000.

[ASS16] Marina Alterman, Yoav Schechner, and Yohay Swirski. Triangulation in random refractive distortions. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[ASW10] M. Alterman, Y.Y. Schechner, and A. Weiss. Multiplexed fluorescence unmixing. In *Proc. IEEE ICCP*, 2010.

[Bar99] K. Barnard. Color constancy with fluorescent surfaces. In *Proc. IS&T/SID Seventh Color Imaging Conference: Color Science, Systems and Applications*, 1999.

[BKY99] P.N. Belhumeur, D.J. Kriegman, and A.L. Yuille. The bas-relief ambiguity. *IJCV*, 35(1), November 1999.

[Bou] J. Y. Bouguet. Camera calibration toolbox for matlab. `vision.caltech.edu/bouguetj/calib\_doc`.

[BSD+17] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3722–3731, 2017.

[CGSC16] Christopher B. Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. WarpNet: Weakly supervised matching for single-view reconstruction. In *NIPS*, 2016.

[CKK05] Manmohan Chandraker, Fredrik Kahl, and David Kriegman. Reflections on the generalized bas-relief ambiguity. In *Proc. IEEE CVPR*, 2005.

[CKSSM85] D.J. Collins, D.A. Kiefer, J.B. Soohoo, and I. Stuart McDermid. The role of reabsorption in the spectral distribution of phytoplankton fluorescence emission. *Deep Sea Research Part A. Oceanographic Research Papers*, 32:983–1003, 1985.

[COR+16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[CT82] R.L. Cook and K.E. Torrance. A reflectance model for computer graphics. *ACM TOG*, 1(1):7–24, 1982.

[DC05] O. Drbohlav and M. Chaniler. Can two specular pixels calibrate photometric stereo? In *Proc. IEEE ICCV*, 2005.

[DDR06] Arturo Donate, Gary Dahme, and Eraldo Ribeiro. Classification of textures distorted by waterwaves. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 421–424. IEEE, 2006.

[DFS08] J.D. Durou, M. Falcone, and M. Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding*, 109(1):22–43, 2008.

[DLHT16] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[DLYF16] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, and David Forsyth. Learning diverse image colorization. *arXiv preprint arXiv:1612.01958*, 2016.

[DMZP14] Bo Dong, Kathleen D Moore, Weiyi Zhang, and Pieter Peers. Scattering parameters and surface normals from homogeneous translucent materials using photometric stereo. In *Proc. IEEE CVPR*, pages 2299–2306, 2014.

[DR07] Arturo Donate and Eraldo Ribeiro. Improved reconstruction of images distorted by water waves. In *Advances in Computer Graphics and Computer Vision*, pages 264–277. Springer, 2007.

[DRF15] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 567–575, 2015.

[EISV04] Alexei A Efros, Volkan Isler, Jianbo Shi, and Mirkó Visontai. Seeing through water. In *NIPS*, volume 17, pages 393–400, 2004.

[FC88] R.T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Trans. PAMI*, 10(4):439–451, 1988.

[FZ89] D. Forsyth and A. Zisserman. Mutual illumination. In *Proc. IEEE CVPR*, 1989.

[GAA+17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[GCHS10] D.B. Goldman, B. Curless, A. Hertzmann, and S.M. Seitz. Shape and spatially-varying BRDFs from photometric stereo. *IEEE Trans. PAMI*, 32:1060–1071, 2010.

[GKZ+16] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[GL15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by back-propagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[Gla95] A.S. Glassner. *Principles of Digital Image Synthesis, Ch. 17*. Morgan Kaufmann Publishers, 1995.

[GNS08] M. Gupta, S.G. Narasimhan, and Y.Y. Schechner. On controlling light transport in poor visibility environments. In *Proc. IEEE CVPR*, 2008.

[GPAM+14a] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[GPAM+14b] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[GRG+13] Nuno Gracias, Pere Ridao, Rafael Garcia, J Escartin, Michel L'Hour, Franca Cibecchini, Ricard Campos, Marc Carreras, David Ribas, Narcıs Palomeras, Lluis Magi, Albert Palomer, Tudor Nicosevici, Ricard Prados, Ramon Hegedus, Laszlo Neumannk, Francesco de Filippo, and Angelos and Mallios. Mapping the moon: Using a lightweight auv to survey the site of the 17th century ship la lune. In *Proc. MTS/IEEE OCEANS*. IEEE, 2013.

[GSH+09] Arthur Gretton, Alexander J Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. 2009.

[GT96] Mauricio Galo and Cl6sio L Tozzi. Surface reconstruction using multiple light sources and perspective projection. In *Proc. IEEE ICIP*, volume 1, pages 309–312, 1996.

[GUA+16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[Gui90] G.G. Guilbault. *Practical fluorescence*, volume 3. CRC, 1990.

[GVK93] HR Gordon, KJ Voss, and KA Kilpatrick. Angular distribution of fluorescence from phytoplankton. *Limnology and oceanography*, pages 1582–1586, 1993.

[GWCV16] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2016.

[Hay94] H. Hayakawa. Photometric stereo under a light-source with arbitrary motion. *JOSA-A*, 11(11):3079–3089, November 1994.

[HFI+08] M.B. Hullin, M. Fuchs, I. Ihrke, H.P. Seidel, and H. Lensch. Fluorescent immersion range scanning. *ACM TOG*, 27(3), 2008.

[HG41] Louis G Henyey and Jesse L Greenstein. Diffuse radiation in the galaxy. *The Astrophysical Journal*, 93:70–83, 1941.

[HHA+10] M.B. Hullin, J. Hanika, B. Ajdin, H.P. Seidel, J. Kautz, and H. Lensch. Acquisition and analysis of bispectral bidirectional reflectance and reradiation distribution functions. *ACM TOG*, 29(4), 2010.

[HLWvdM16] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[Hor86] B. Horn. *Robot vision, Ch. 10*. The MIT Press, 1986.

[HTP+17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[Hul94] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

[HWYD16] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[HZRS16a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[HZRS16b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[IGM05] I. Ihrke, B. Goidluecke, and M. Magnor. Reconstructing the geometry of flowing water. In *Proc. IEEE ICCV*, 2005.

[IMMY14] Chika Inoshita, Yasuhiro Mukaigawa, Yasuyuki Matsushita, and Yasushi Yagi. Surface normal deconvolution: Photometric stereo for optically thick translucent objects. In *Proc. ECCV*, pages 346–359. Springer, 2014.

[ISI90] Yuji Iwahori, Hidezumi Sugie, and Naohiro Ishii. Reconstructing shape from shading images under point light source illumination. In *Proc. IEEE ICPR*, volume 1, pages 83–87, 1990.

[ISSI16] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)*, 35(4):110, 2016.

[IZZE16a] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[IZZE16b] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[Jaf90] Jules S Jaffe. Computer modeling and the design of optimal underwater imaging systems. *IEEE J. Oceanic Engineering*, 15(2):101–111, 1990.

[JDV+17] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1175–1183. IEEE, 2017.

[JSK08] Neel Joshi, Richard Szeliski, and David Kriegman. PSF estimation using sharp edge prediction. In *Proc. IEEE CVPR*, pages 1–8. IEEE, 2008.

[JSZK15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[KB91] Byungil Kim and Peter Burger. Depth and shape from shading using the photometric stereo method. *CVGIP: Image Understanding*, 54(3):416–427, 1991.

[KDCS08] D.M. Kocak, F.R. Dalgleish, F.M. Caimi, and Y.Y. Schechner. A focus on recent developments and trends in underwater imaging. *Marine Technology Society Journal*, 42(1):52–67, 2008.

[KFB92] N. Kolagani, J.S. Fox, and D.R. Blidberg. Photometric stereo using point light sources. In *Proc. IEEE Int. Conf. Robotics and Automation*, pages 1759–1764, 1992.

[KJC16] A. Kanazawa, D. W. Jacobs, and M. Chandraker. WarpNet: Weakly supervised matching for single-view reconstruction. In *CVPR*, 2016.

[KLK78] JP Kratohvil, M.P. Lee, and M. Kerker. Angular distribution of fluorescence from small particles. *Applied Optics*, 17(13), 1978.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[KWR16] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, 35(6):193, 2016.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LBK17] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[LCWJ15] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

[LK96] Kyoung Mu Lee and C-C Jay Kuo. Shape from photometric ratio and stereo. *J. Vis. Comm. & Image Rep.*, 7(2):155–162, 1996.

[LMK+18] Zhengqin Li, Zak Murez, David Kriegman, Ravi Ramamoorthi, and Manmohan Chandraker. Learning to see through turbulent water. In *Applications of Computer Vision (WACV), 2018 IEEE Winter Conference on*, pages 512–520. IEEE, 2018.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[LT16] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.

[LTH+16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[LW16] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[LZHFF17] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li Fei-Fei. Label efficient learning of transferable representations across domains and tasks. In *Conference on Neural Information Processing Systems (NIPS)*, 2017.

[LZWJ16] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[McG80] BL McGlamery. A computer model for underwater camera systems. In *Ocean Optics VI*, pages 221–231. International Society for Optics and Photonics, 1980.

[MKK+17] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. *arXiv preprint arXiv:1712.00479*, 2017.

[MLX+16] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, and Zhen Wang. Multi-class generative adversarial networks with the l2 loss function. *arXiv preprint arXiv:1611.04076*, 2016.

[MTRK17] Zak Murez, Tali Treibitz, Ravi Ramamoorthi, and David J Kriegman. Photometric stereo in a scattering medium. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1880–1891, 2017.

[Mur92] Hiroshi Murase. Surface shape reconstruction of a nonrigid transparent object using refraction and motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(10):1045–1052, 1992.

[NFB97] S.K. Nayar, X.S. Fang, and T. Boult. Separation of reflection components using color and polarization. *IJCV*, 21(3):163–186, 1997.

[NGD+06] Srinivasa G Narasimhan, Mohit Gupta, Craig Donner, Ravi Ramamoorthi, Shree K Nayar, and Henrik Wann Jensen. Acquiring scattering properties of participating media by dilution. *ACM Transactions on Graphics (TOG)*, 25(3):1003–1012, 2006.

[NIK90] S.K. Nayar, K. Ikeuchi, and T. Kanade. Shape from interreflections. In *Proc. IEEE ICCV*, 1990.

[NKGR06] S.K. Nayar, G. Krishnan, M.D. Grossberg, and R. Raskar. Fast separation of direct and global components of a scene using high frequency illumination. *ACM TOG*, 25(3):935–944, 2006.

[NN03] Srinivasa G Narasimhan and Shree K Nayar. Shedding light on the weather. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–665. IEEE, 2003.

[NNSK05] Srinivasa G Narasimhan, Shree K Nayar, Bo Sun, and Sanjeev J Koppal. Structured light in scattering media. In *Proc. IEEE ICCV*, pages 420–427, 2005.

[NWC+11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5, 2011.

[NZH02] Sharriar Negahdaripour, H Zhang, and X Han. Investigation of photometric stereo method for 3-D shape recovery from underwater imagery. In *Proc. MTS/IEEE OCEANS*, volume 2, pages 1010–1017, 2002.

[ON95] M. Oren and S.K. Nayar. Generalization of the lambertian model and implications for machine vision. *IJCV*, 14(3):227–251, 1995.

[OSPS11] Omar Oreifej, Guang Shu, Teresa Pace, and Mubarak Shah. A two-stage reconstruction approach for seeing through water. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1153–1160. IEEE, 2011.

[Pho75] B.T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.

[PKD+16] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[PPV08] Ruggero Pintus, Simona Podda, and Massimo Vanzi. An automatic alignment procedure for a four-source photometric stereo technique applied to scanning electron microscopy. *IEEE Trans, Instrumentation and Measurement*, 57(5):989–996, 2008.

[RBG+17] Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Moressi, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. *arXiv preprint arXiv:1711.05139*, 2017.

[RMC15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[RSM+16] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016.

[RVRK16] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016.

[SBCC17] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. *arXiv preprint arXiv:1704.01705*, 2017.

[Sha92] Amnon Shashua. *Geometry and photometry in 3D visual recognition*. PhD thesis, Massachusetts Institute of Technology, 1992.

[SI94] Y. Sato and K. Ikeuchi. Temporal-color space analysis of reflection. *JOSA A*, 11(11):2990–3002, 1994.

[Sil80] William M Silver. Determining shape and reflectance using multiple images. Master's thesis, Massachusetts Institute of Technology, 1980.

[SKFD10] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. *Computer Vision–ECCV 2010*, pages 213–226, 2010.

[SOS12a] I. Sato, T. Okabe, and Y. Sato. Bispectral photometric stereo based on fluorescence. In *Trans. IEEE CVPR*, 2012.

[SOS12b] Imari Sato, Takahiro Okabe, and Yoichi Sato. Bispectral photometric stereo based on fluorescence. In *Proc. IEEE CVPR*, 2012.

[SPT+] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training.

[SRNN05] Bo Sun, Ravi Ramamoorthi, Srinivasa G Narasimhan, and Shree K Nayar. A practical analytic single scattering model for real time rendering. *ACM Transactions on Graphics (TOG)*, 24:1040–1049, 2005.

[SSSS16] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 2110–2118, 2016.

[Sze10] R. Szeliski. *Computer vision: Algorithms and applications*. Springer, 2010.

[TAKD14] Chourmouzios Tsiotsios, Maria E Angelopoulou, Tae-Kyun Kim, and Andrew J Davison. Backscatter compensated photometric stereo with 3 sources. In *Proc. IEEE CVPR*, pages 2259–2266, 2014.

[TD91] H.D. Tagare and R.J.P. Defigueiredo. A theory of photometric stereo for a class of diffuse non-lambertian surfaces. *IEEE Trans. PAMI*, 13(2):133–152, 1991.

[THDS15] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[THSD17] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *arXiv preprint arXiv:1702.05464*, 2017.

[THZ+14] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[TK05] Ariel Tankus and Nahum Kiryati. Photometric stereo under perspective projection. In *Proc. IEEE ICCV*, 2005.

[TMK+15] Kenichiro Tanaka, Yasuhiro Mukaigawa, Hiroyuki Kubo, Yasuyuki Matsushita, and Yasushi Yagi. Recovering inner slices of translucent objects by multi-frequency illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5464–5472, 2015.

[TMMK12] Tali Treibitz, Zak Murez, B Greg Mitchell, and David Kriegman. Shape from fluorescence. In *Proc. ECCV*, 2012.

[TMQ+07] Ping Tan, Satya Mallick, Long Quan, David Kriegman, and Todd Zickler. Isotropy, reciprocity and the generalized bas-relief ambiguity. In *Proc. IEEE CVPR*, 2007.

[TN09] Yuandong Tian and Srinivasa G Narasimhan. Seeing through water: Image restoration using model-based tracking. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2303–2310. IEEE, 2009.

[TN12] Yuandong Tian and Srinivasa G Narasimhan. Globally optimal estimation of nonrigid image distortion. *International journal of computer vision*, 98(3):279–302, 2012.

[TN15] Yuandong Tian and Srinivasa G Narasimhan. Theory and practice of hierarchical data-driven descent for optimal deformation estimation. *International Journal of Computer Vision*, 115(1):44–67, 2015.

[TNR+12] T. Treibitz, B. P. Neal, P. Roberts, D. I. Kline, O. Beijbom, S. Belongie, B. G. Mitchell, J. Jaffe, and D. Kriegman. Wide field of view full spectrum fluorescence imaging for coral ecology. In *International Coral Reef Symposium*, 2012.

[TOA06] Emanuele Trucco and Adriana T Olmos-Antillon. Self-tuning underwater image restoration. *IEEE J. Oceanic Engineering*, 31(2):511–519, 2006.

[TPW16] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[TS94] Ping-Sing. Tsai and M. Shah. Shape from shading using linear approximation. *Image and Vision Computing*, 12(8):487–498, 1994.

[TS09] T. Treibitz and Y. Y. Schechner. Active polarization descattering. *IEEE Trans. PAMI*, 31:385–399, 2009.

[TS12] Tali Treibitz and Yoav Y Schechner. Resolution loss without imaging blur. *JOSA A*, 29(8):1516–1528, 2012.

[TT07] V Viktorovich Tuchin and V Tuchin. *Tissue optics: light scattering methods and instruments for medical diagnosis*, volume 13. SPIE press Bellingham, 2007.

[UVL16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[VECP17] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5385–5394. IEEE, 2017.

[WLFL10] Zhiying Wen, Andrew Lambert, Donald Fraser, and Hongdong Li. Bispectral analysis and recovery of images distorted by a moving water surface. *Applied optics*, 49(33):6376–6384, 2010.

[Woo80] R.J. Woodham. Photometric method for determining surface orientation from multiple images. *Opt. Eng.*, 19(1):139–144, January 1980.

[WWLP06] A. Wilkie, A. Weidlich, C. Larboulette, and W. Purgathofer. A reflectance model for diffuse fluorescent surfaces. In *Proc. int. conf. Comp. graphics & interactive techniques in Australasia and Southeast Asia*, pages 321–331, 2006.

[XRW+14] Tianfan Xue, Michael Rubinstein, Neal Wadhwa, Anat Levin, Fredo Durand, and William T Freeman. Refraction wiggles for measuring fluid depth and velocity from video. In *European Conference on Computer Vision*, pages 767–782. Springer, 2014.

[YDL+17] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2281, 2017.

[YKF17] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. *arXiv preprint arXiv:1705.09914*, 2017.

[YS97] A. Yuille and D. Snow. Shape and albedo from multiple images using integrability. In *Proc. IEEE CVPR*, 1997.

[YZC16] Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation networks for object landmark localization. In *ECCV*, 2016.

[ZIE16] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.

[ZKSE16] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[ZLG+14] Mingjie Zhang, Xing Lin, Mohit Gupta, Jinli Suo, and Qionghai Dai. Recovering scene geometry under wavy fluid via distortion and defocus analysis. In *European Conference on Computer Vision*, pages 234–250. Springer, 2014.

[ZMKB08a] T. Zickler, S.P. Mallick, D.J. Kriegman, and P.N. Belhumeur. Color subspaces as photometric invariants. *IJCV*, 79:13–30, 2008.

[ZMKB08b] Todd Zickler, Satya Mallick, David Kriegman, and Peter N. Belhumeur. Color subspaces as photometric invariants. *IJCV*, 79:13–30, 2008.

[ZN02] Shaomin Zhang and Shahriar Negahdaripour. 3-D shape recovery of planar and curved surfaces from shading cues in underwater images. *IEEE J. Oceanic Engineering*, 27(1):100–116, 2002.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.

[ZS11] C. Zhang and I. Sato. Separating reflective and fluorescent components of an image. In *Proc. IEEE CVPR*, 2011.

[ZSP17] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *arXiv preprint arXiv:1701.05957*, 2017.

[ZTCS99] R. Zhang, P.S. Tsai, J.E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE Trans. PAMI*, 21(8):690–706, 1999.