CroMo: Cross-Modal Learning for Monocular Depth Estimation

Yannick Verdié[1], Jifei Song[1], Barnabé Mas[1,2], Benjamin Busam[1,3], Aleš Leonardis[1], Steven McDonagh[1]

[1] Huawei Noah's Ark Lab   [2] École Polytechnique   [3] Technical University of Munich

{yannick.verdie,jifei.song,ales.leonardis,steven.mcdonagh}@huawei.com

barnabe.mas@polytechnique.edu   b.busam@tum.de

Abstract

Learning-based depth estimation has witnessed recent progress in multiple directions; from self-supervision using monocular video to supervised methods offering highest accuracy. Complementary to supervision, further boosts to performance and robustness are gained by combining information from multiple signals. In this paper we systematically investigate key trade-offs associated with sensor and modality design choices as well as related model training strategies. Our study leads us to a new method, capable of connecting modality-specific advantages from polarisation, Time-of-Flight and structured-light inputs. We propose a novel pipeline capable of estimating depth from monocular polarisation for which we evaluate various training signals. The inversion of differentiable analytic models thereby connects scene geometry with polarisation and ToF signals and enables self-supervised and cross-modal learning.

In the absence of existing multimodal datasets, we examine our approach with a custom-made multi-modal camera rig and collect CroMo; the first dataset to consist of synchronized stereo polarisation, indirect ToF and structured-light depth, captured at video rates. Extensive experiments on challenging video scenes confirm both qualitative and quantitative pipeline advantages where we are able to outperform competitive monocular depth estimation methods.

1. Introduction

Modern vision sensors are able to leverage a variety of light properties for optical sensing. Common RGB sensors, for instance, use colour filter arrays (CFA) over a pixel sensor grid to separate incoming radiation into specified wavebands. This allows a photosensor to detect wavelength-separated light intensity and enables the acquisition of familiar visible spectrum images. Wavelength is however only one property of light capable of providing information.

Light polarisation defines another property and describes the oscillation direction of an electromagnetic wave. While the majority of natural light sources (*e.g.* the sun) emit unpolarised light, consisting of a mixture of oriented oscillations, surface reflection from non-metallic objects can
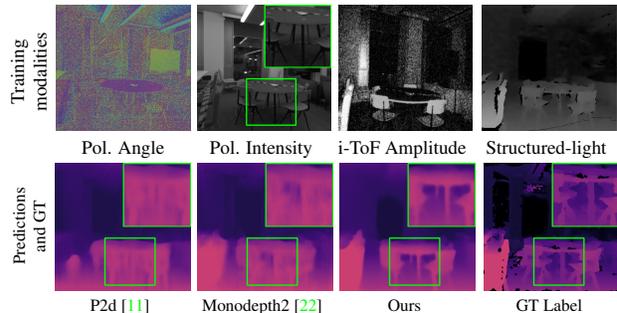

Figure 1. Top row: polarisation input signal (Pol.) visualised as Angle and Intensity. Additionally; Time-of-Flight Amplitude (i-ToF) and structured-light sensor co-modalities, exploitable during training. Bottom row: monocular depth estimation, using the Pol. input. Uni-modal model training of p2d [11] and the monodepth2 architecture [22], compared with cross-modal training (Ours).

linearly polarise the light. Such polarised light then contains surface structure information, retrievable using analytic physical models [8]. This information can be used to harness the depth cues offered by this light property. Polarimetric imagery is a *passive* example for depth estimation. Passive sensors have acceptable resolution and dense depth however there exist well understood capture situations that prove challenging (*e.g.* textureless surface regions).

However further known properties of light (*i.e.* speed) provide yet more information. Indirect Time-of-Flight (i-ToF) cameras are *active* light sensors and use a pulsed, near infrared light source to measure object and surface distance. Further active sensors use structured-light and these emit known infrared patterns and use stereoscopic imaging to measure the distance to the surface. While i-ToF and structured-light cameras have clear advantages, such as the ability to function in low-light scenarios and good short range precision, they are susceptible to specular reflections, ambient light and range remains limited.

We argue that novel combination of *active* and *passive* light sensors offers new possibilities. We can exploit such a combination to take advantage of the discussed, modality-specific strengths and weaknesses. We observe that (1) differing visual modalities offer information cues about com-

plimentary aspects of the world and (2) there exist clear trade-offs between the complexity of capture sensor setups and the resulting data diversity and quality, accessible for supervision signals. This motivates us to systematically investigate these considerations and provide insight into training data capture design decisions and the related pay-offs. Our study results in the proposal of a framework capable of exploiting available supervision signals and is tailored to benefit from the particular strengths of unique modalities.

We instantiate our ideas by bringing together the physical understanding of Polarisation and i-ToF in a data driven fashion. In practice this affords an inference pipeline that estimates depth from a single polarisation image. We train a convolutional neural network (CNN), with cross-modal fusion using differentiable physical models. We establish a dataset comprising Ground Truth depth obtained via Multi-View Stereo (MVS) reconstruction [52] that enjoys access to information rich, full video sequences. We carry out extensive experimental work to establish the efficacy of our proposed monocular depth estimation strategies.

Our **contributions** can be summarised as:

1. **Novel multi-modal method**. We propose a multi-modal training approach that allows for monocular depth estimation from polarisation images. We propose (i) differentiable analytic formulae that define modal-specific loss terms, (ii) cross-modal consistency joint-training towards improved real-world depth estimation from a single polarisation image, (iii) architectural components that increase predicted depth sharpness (see Fig. 1).

2. **CroMo dataset and training modalities study**. We provide a systematic analysis of the benefits afforded when multiple image modalities are available at training time, for monocular depth estimation. Investigation and exposure of improvements are enabled by the unique Cross-Modality video dataset[1]. Our multiple hardware-synchronized cameras capture, for the first time, stereo polarisation (Pol), indirect Time-of-Flight (i-ToF) and structured-light images from active sensing.

The remaining sections of the paper are thus organised: Sec. 2 provides brief review of depth estimation with respect to relevant modalities and previous work considering multiple information signals. Sec. 3 presents our model capable of monocular depth estimation from polarisation imagery and our cross-modal training procedure. In Sec. 4 we introduce CroMo, our novel multi-modal dataset, Sec. 5 reports experimental work validating our contributions and Sec. 6 provides discussion and future research issues.

2. Related Work

To the best of our knowledge this is the first work to study end-to-end monocular depth inference, utilising cross-modal information from Time-of-Flight (i-ToF), active stereo and polarisation modalities during training. We briefly review the literature most closely related to the main components of our investigation and proposed framework.

2.1. Monocular depth estimation

Estimating depth from a single image constitutes a hard, ill-posed problem. Pioneering work on supervised monocular depth estimation [42] used synthetic samples during training. Synthetic data was also previously used in conjunction with stereo network distillation [25] for this task. To improve accuracy and convergence speed, [17] introduce a spatially-increasing discretisation. However, acquiring ground truth depth data remains a difficult task [20].

To overcome the difficulty of collecting accurate ground truth signal, multiple works [19, 61] investigate a consistency loss by leveraging stereo imagery during training, towards self-supervision. While being undoubtedly path-breaking, the initial methods suffered from a non-differentiable sampling step. Godard *et al*. [21] formulated a fully-differentiable pipeline with left-right consistency checks during training and have also explored the temporal components [22], even in challenging setups such as night scenes [55]. These methods predict depth with RGB input, while we utilise polarisation images.

Monocular Polarisation Previous work use monocular polarisation imagery to recover depth. One route to overcome Shape from Polarisation (SfP) ambiguities is to use orthographic camera models to express polarisation intensity in terms of depth [64]. Atkinson *et al*. [8] compute depth without knowing the light direction through a non-linear optimization framework and yet assume fully diffuse surfaces. Linear systems have also been constructed for the task [53] by adding shape from shading equations. While theoretically interesting, the orthographic assumption has restricted their application to synthetic lab environments.

Learning based Polarisation Due to lack of reasonably-sized datasets, only a limited number of works focus on learning with polarisation. Ba *et al*. [9] provide polarisation images together with a set of plausible inputs from a physical model to estimate surface normals. The work of [34] apply polarisation for instance segmentation of transparent objects and [37] learn de-glaring of images with semi-transparent objects. Recently, Blanchon *et al*. [11] extended the work of [22] with complementary polarimetric cues. In contrast to them, we invert a physical model to enable self-supervision through consistency cycles and additionally study the benefit of co-modal i-ToF information.

Learning based i-ToF i-ToF sensors acquire distance information by estimating the time required for an emitted light pulse to be reflected [65]. Sensors measure either the time (direct) or the phase (indirect) difference between emitted and received light. The modality enjoys high precision for short range distances [27], yet suffers from limited spatial resolution and noise [15], which constitute challenging fac-

tors for any learning-based approach. Obtaining reliable signals from specular surfaces is difficult and inherent Multi Path Interference (MPI), often manifests as noisy measurements and artifacts. Synthetic training is also explored for raw i-ToF input data in end-to-end fashion [5,24,57]. However, the ability to account for real world domain shifts is limited. In [4] a GAN is employed towards addressing such domain adaption issues on a limited dataset.

i-ToF depth improvement MPI can be considered a critical issue and error mitigation has been the focus of a body of work [60]. Two-path approximations [23] have been used within optimization schemes [14,35] and multiple frequencies are used to constrain the problem [16]. Kadambi *et al*. [33] propose a hardware solution to address scenes with translucent objects and a number of scholars incorporate light transport information to correct for MPI [2,26,44,46].

2.2. Depth with multiple sensors

Depth completion has been carried out via combining multiple input modalities, for example, a sparse but accurate LiDAR signal in combination with RGB [58]. It is difficult to address sparse signals with CNNs [41] and LiDAR sensors can produce problematic artifacts resulting in unreliable Ground Truth depth estimates [39]. One strategy towards removing dependence on this form of supervision are self-supervision cues however these fall behind supervised pipelines in terms of accuracy [40].

i-ToF and *x* Confidence-based combination of i-ToF depth and classical RGB stereo is explored with the network architecture of [3] and a semi-supervised approach for this combination is explored by [47] in a generic framework. While these approaches improve upon the individual depth estimates, they rely on a late fusion paradigm. Son *et al*. [54] use a robotic arm to collect 540 real training images of short range scenes with structured light ground truth.

By inserting micro linear polarizers [45] in front of a photo-receiver chip, Yoshida *et al*. [62] build an i-ToF sensor capable of acquiring both i-ToF depth and polarisation scene cues. Combination of both the absolute depth (i-ToF) and relative shape (polarisation cues) allowed reconstruction of depth for specular surfaces. While this pipeline requires i-ToF and polarisation input to solve an optimization problem, we alternatively explore cross-modal self-supervised training and single image inference.

Depth from multi-view Polarisation Another route to predict depth is the use of more than one polarisation image [7] which enables methods based on physical models. An RGB+Polarisation pair can provide sharp depth maps with stereo vision [66]. Other methods [12] use more than two polarisation images. Despite the sharpness of the results, the difficulty to acquire multi-view polarisation images is still a major hurdle. Atkinson *et al*. [6] combine polarisation methods with photometric stereo. Two images of a scene, from an identical view point yet with dif-

ferent light exposures, are leveraged. An extension dealing with mixed reflectivity is established via a combined photometric-polarisation linear system in [38] and Garcia *et al*. [18] solve for polarisation normals using circularly polarised light. Traditional multi-view methods also benefit from polarisation. Miyazaki *et al*. [43] recover surfaces of black objects using polarisation physics and space carving.

Depth refinement with Polarisation Consumer depth estimation tools progress significantly in recent years however their predictions are noisy and lack details. Using polarisation cues, [32] enhance sharper depth maps from RGBD cameras by differentiating their depth maps to resolve polarisation ambiguities and perform mutual optimization.

Despite clear improvements in monocular depth estimation methods, their performance remains bounded by the chosen modality hence calling for multi-modal depth estimation. Our method alleviates this problem with a learning based approach. During training we leverage complementary modalities such that our model can compensate the drawbacks of the single modality used at inference time.

3. Method

Our multi-modal monocular depth investigation leads to a new model architecture that accounts explicitly for prediction blur and introduces two novel analytic losses. We discuss these components in the following sections.

3.1. Architecture

Our architecture employs multiple encoder-decoder networks illustrated in Fig. 3a. We observe that monocular depth estimation methods often incur blurry image predictions and we address this problem by introducing architectural components that account for prediction blur. Firstly convolutions in our encoders are coupled with gated convolution. Our network then composes a traditional U-Net [51] with skip connections and the gated convolutions [63]. The encoder utilises a ResNet [28] style block, while the decoder is a cascade of convolutions with layer resizing.

Secondly, drawing on the fact that Displacement Fields (DF) can be utilised to aid sharpness [49], we estimate a DF using a self-supervised sharpening decoder. Depth pixels with strong local disparity have values redefined to mirror a nearest neighbour that does not exhibit strong local disparity. Groundtruth (GT) displacement fields can thus be defined for each predicted depth during training ("on-the-fly"), guiding our dedicated displacement field prediction. We inspect predicted depth with and without our DF strategy and observe significant improved sharpness, most evident when employing 3D visualisations (Fig. 4).

3.2. Loss Formulation

Our study considers multiple modalities and various sensor configurations at training time. We explore several loss terms to exploit our unique setting (see Tab. 3). Loss terms
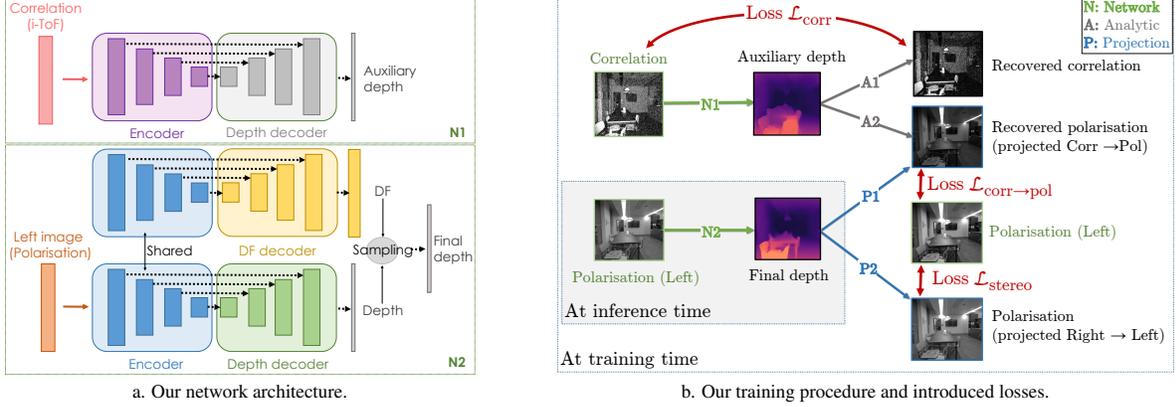
a. Our network architecture.

b. Our training procedure and introduced losses.

Figure 3. Our full model with modality specific losses $\mathcal{L}_{\text{corr}}$, $\mathcal{L}_{\text{corr}\rightarrow\text{pol}}$ and $\mathcal{L}_{\text{stereo}}$ (see Sec. 3.1 for further details).
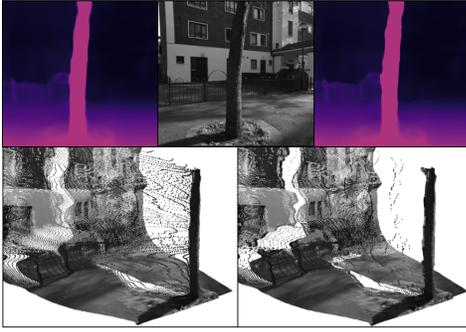

Figure 4. Effect of the *Displacement Fields* (DF). Left top (bottom): predicted depth (point cloud) without DF. Center top: polarisation intensity. Right top (bottom): predicted depth (point cloud) with DF. Flying pixels, visible in 3D, are clearly reduced.

in our training procedure are enabled through both coordinate frame projections (P1, P2) and analytic transforms (A1, A2) of individual network (N1, N2) outputs (see Figs. 3a, 3b). We firstly process input modalities individually using distinct networks. These ingest i-ToF correlation and left polarisation images respectively and output initial depth maps. We propose two analytic losses, derived from properties of i-ToF and polarisation, to train and link the networks. We train the i-ToF module without ground truth and also leverage the available multi-modal information through *image recovery* via related analytical formulae (A1, A2). Strategically similar to previous work [13], at inference time we require only a single modality (in our case polarisation), and can discard network N1 completely.

Terms $\mathcal{L}_{\text{corr}\rightarrow\text{pol}}$ and $\mathcal{L}_{\text{corr}}$ evaluate discrepancies between each input image and respective *recovered* images, obtained using auxiliary and final depth maps (Fig. 3b). Our individual branches share information through the loss term $\mathcal{L}_{\text{corr}\rightarrow\text{pol}}$. Explicitly, we recover a polarisation image from an *auxiliary* depth map and then project this, using projection P1, to the polarisation sensor frame of reference via the final depth map D_{pol}. Finally our third loss term $\mathcal{L}_{\text{stereo}}$ is used to train the polarisation network (N2) by

comparing the right polarisation image, projected using the predicted depth D_{pol}, with the left polarisation image. We next provide details of our analytical formulae for *image recovery* and the loss terms that enable our training procedure.

Depth to polarisation (A2) Polarisation cameras capture polarised intensity along directions φ_{pol}. The measured intensity is given by [66]

$$i_{\varphi_{pol}} = i_{un} \cdot (1 + \rho \, \cos(2\varphi_{pol} - 2\phi))$$
$$\text{with } \varphi_{pol} \in \left\{ 0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4} \right\} \quad (1)$$

where φ_{pol} is the polariser angle, i_{un} is the intensity of unpolarised light, ρ is the degree of linear polarisation and ϕ is the Angle of Polarisation (AoP). The polarisation parameters $\rho \in \{\rho_s, \rho_d\}$ and $\phi \in \{\phi_s, \phi_d\}$ depend on local reflection type, either *diffuse* (d) or *specular* (s) as follows:

$$\begin{cases} \rho_s = \dfrac{2\sin^2(\theta)\cos(\theta)\sqrt{\eta^2 - \sin^2(\theta)}}{\eta^2 - \sin^2(\theta) - \eta^2\sin^2(\theta) + 2\sin^4(\theta)} \\[4mm] \rho_d = \dfrac{(\eta - 1/\eta)^2 \sin^2(\theta)}{2 + 2\eta^2 - (\eta + 1/\eta)^2 \sin^2(\theta) + 4\cos(\theta)\sqrt{\eta^2 - \sin^2(\theta)}} \end{cases} \quad (2)$$

with $\theta \in [0, \pi/2]$ the viewing angle and η the object refractive index, typically 1.5, and

$$\begin{cases} \phi_d & = \alpha \, [\pi] & \text{if the pixel is diffuse} \\ \phi_s & = \alpha + \frac{\pi}{2} \, [\pi] & \text{if the pixel is specular} \end{cases} \quad (3)$$

The π-ambiguity is denoted as $[\pi]$ in (3), and α denotes the azimuth angle of the surface normal **n**. Azimuth angle α and viewing angle θ are obtained as

$$\cos(\theta) = \frac{\mathbf{n} \cdot \mathbf{v}}{\|\mathbf{n}\|\|\mathbf{v}\|} \quad \text{and} \quad \tan(\alpha) = \frac{n_y}{n_x}, \quad (4)$$

with **v** the viewing vector defined as the vector pointing toward the camera center from the 3D point $P(x, y)$ corresponding to pixel (x, y) with depth $d(x, y)$ and **n** the

outward pointing normal vector, defined as the cross product of the partial derivatives with respect to x and y [66]:

$$\mathbf{n} = \begin{bmatrix} -f_y \partial_x d(x,y) \\ -f_x \partial_y d(x,y) \\ (x-c_x)\partial_x d(x,y) + (y-c_y)\partial_y d(x,y) + d(x,y) \end{bmatrix} \quad (5)$$

with f_x, f_y, c_x, c_y the camera intrinsics.

Hence, from a given depth map d, one can compute the azimuth angle α and the viewing angle θ using (4) and (5), followed by the polarisation parameters ρ and ϕ with (2) and (3). The polarisation images for diffuse and specular surfaces $\widehat{I}_{\text{pol}}^{\text{diffuse}}$ and $\widehat{I}_{\text{pol}}^{\text{specular}}$ are finally recovered using the calculated polarisation parameters with (1).

Depth to correlation (A1) Indirect ToF measures the correlation between a known emitted signal and the received signal. The emitted signal at frequency f_{M} is a sinusoid:

$$g(t) = 2\cos(2\pi f_{\text{M}} t) + 1 \quad (6)$$

and the signal, reflected by the scene, is of the form [30]

$$f(t) = \alpha \cos(2\pi f_{\text{M}} t + 2\pi f_{\text{M}} \tau) + \beta \quad (7)$$

where the τ is the time delay between the emitted signal $g(t)$ and the reflected signal $f(t)$. The i-ToF measurement $c(x)$ is the correlation between the two signals:

$$c(x) = \lim_{T \to \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) g(t-x)\, dt \\ = \alpha \cos(2\pi f_{\text{M}} x + 2\pi f_{\text{M}} \tau) + \beta \quad (8)$$

where we only consider the direct reflection signal and ignore the multipath interference (MPI) and sensor imperfections. We are interested in the phase φ, proportional to the depth d between the objects in the scene and the sensor:

$$\varphi = (2\pi f_{\text{M}} \tau)\,[2\pi] = \left(d \cdot \frac{4\pi f_{\text{M}}}{c}\right)[2\pi] \quad (9)$$

where c is the speed of light and $[2\pi]$ represents the 2π-ambiguity. Using the four bucket strategy [36] to sample $c(x)$ at four positions, where $2\pi f_{\text{M}} x \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$, four measurements $\{c(x_0), c(x_1), c(x_2), c(x_3)\}$ can be obtained to recover the phase φ, the amplitude α and the intensity β.

$$\tan(\varphi) = \frac{c(x_3) - c(x_1)}{c(x_0) - c(x_2)} \quad (10)$$

$$\alpha = \frac{1}{2}\sqrt{(c(x_3) - c(x_1))^2 + (c(x_1) - c(x_0))^2} \quad (11)$$

$$\beta = \frac{1}{4}\sum_{i=0}^{3} c(x_i) \quad (12)$$

Hence, from a given depth d, one can compute the phase φ using (9) and then reformulate the recovered i-ToF correlation using (10), (11) and (12) in turn to form $\widehat{I}_{\text{corr}}$.

Stereo loss $\mathcal{L}_{\text{stereo}}$ This loss requires that left and right image pairs are accessible during training. While only the left image I_l is fed to the network, the right image I_r can guide the model towards generating valid depth, and vice versa. More formally, let K_l and K_r be camera matrices with intrinsic parameters for left and right images respectively, and D a depth map on the left reference frame. Let $T_{\text{left} \to \text{right}}$ denote the transformation that moves 3D points from the left coordinate system to the right. An image coordinate transformed from left coordinate p_l to the right image is

$$p_{\text{left} \to \text{right}} = K_r \cdot T_{\text{left} \to \text{right}} \cdot D(p_l) \cdot K_l^{-1} \cdot p_l \quad (13)$$

A backward differentiable warping [31] is used to reproject an image onto the left view as $I_{\text{right} \to \text{left}}$.

We form a stereo loss $\mathcal{L}_{\text{stereo}}$, and related mask loss $\mathcal{L}_{\text{mask}}$ similarly to [22], which aid network training and deal with occluded pixels respectively as

$$\mathcal{L}_{\text{stereo}} = E_{\text{pe}}\left(I_l, I_{\text{right} \xrightarrow{D_{pol}} \text{left}}\right) \quad (14)$$

$$\mathcal{L}_{\text{mask}} = E_{\text{pe}}\left(I_l, I_{\text{right} \xrightarrow{D\infty} \text{left}}\right) \quad (15)$$

where the photometric error is similar to [22]:

$$E_{\text{pe}}(I_x, I_y) = \alpha \frac{1 - \text{SSIM}(I_x, I_y)}{2} + (1 - \alpha)\|I_x - I_y\|_1 \quad (16)$$

Analytical losses $\mathcal{L}_{\text{corr}}$ and $\mathcal{L}_{\text{corr} \to \text{pol}}$ Depth D_{corr} is firstly inferred directly from i-ToF correlation input, and then two recovered images $\widehat{I}_{\text{corr}}$ and $\widehat{I}_{\text{pol}}$ are formed. Recovered images represent the 'ideal' input for each modality, i-ToF and polarisation respectively, conditioned on the inferred depth. Since $\widehat{I}_{\text{pol}}$ is generated from D_{corr}, we reproject it using D_{pol} to form a recovered final polarisation image $\widehat{I}^i_{\text{corr} \xrightarrow{D_{pol}} \text{pol}}, i \in \{\text{diffuse, specular}\}$. In each case, discrepancies between the recovered image and the true input image provide a strong indication of the quality of the generated depth. We use this signal to guide the network. Formally

$$\mathcal{L}_{\text{corr}} = E_{\text{pe}}\left(I_{\text{corr}}, \widehat{I}_{\text{corr}}\right) \quad (17)$$

$$\mathcal{L}_{\text{corr} \to \text{pol}} = \min_{i \in \{\text{diffuse, specular}\}} \left\{ E_{\text{pe}}\left(I_l, \widehat{I}^i_{\text{corr} \xrightarrow{D_{pol}} \text{pol}}\right) \right\} \quad (18)$$

where I_l is the left polarisation image, $\widehat{I}_{\text{corr} \xrightarrow{D_{pol}} \text{pol}}$ the recovered polarisation image aligned to I_l, I_{corr} the i-ToF correlation input, and $\widehat{I}_{\text{corr}}$ the recovered correlation image. We use a $\min$ operator for $\mathcal{L}_{\text{corr} \to \text{pol}}$ to lift the problem of classifying a pixel as *diffuse* or *specular* by computing both possibilities and letting the network select the best solution.

Finally, following [59], we use an additional loss $\mathcal{L}_{\text{struct}}$ in the objective function, derived from structured-light information (see appendix for further detail).

5

In summary, depending on the input modalities available at training time, we can add or remove the introduced losses $\mathcal{L}_{\text{corr}}$, $\mathcal{L}_{\text{corr}\to\text{pol}}$, $\mathcal{L}_{\text{stereo}}$ and $\mathcal{L}_{\text{struct}}$ as appropriate. We explicitly note that hyper parameter tuning for balancing of these loss terms is *not* required, due to our formulation.

Our total loss $\mathcal{L}$ can then be formulated as:

$$\mathcal{L} = \min_{i \in \{\text{mask}, \text{stereo}, \text{corr}\to\text{pol}, \text{struct}\}} \left\{ \mathcal{L}_i \right\} + \mathcal{L}_{\text{corr}} + \mathcal{L}_{\text{DF}} \quad (19)$$

where $\mathcal{L}_{\text{DF}}$ is the $\mathcal{L}_2$ norm between predicted and GT *DF*.

4. Data

We next provide details on our custom camera rig (Sec. 4.1) and CroMo dataset (Sec. 4.2), comprising synchronised image sequences capturing multiple modalities, at video-rate across real-world indoor and outdoor scenes.

4.1. Camera capture rig

Our prototype custom-camera hardware rig is shown in Fig. 5. Our rig is constructed in order to capture synchronised data across multiple modalities including stereo polarisation, i-ToF correlation, structured-light depth and IMU. We rigidly mount two polarisation cameras (Lucid PHX050S-QC) providing a left-right stereo pair, an i-ToF camera (Lucid HLS003S-001) operating at 25Mhz and a camera (RealSense D435i) for active IR stereo capture. All devices are connected with a hardware synchronisation wire resulting in time-aligned video capture at a frame rate of 10fps. The left polarisation camera is the lead camera which generates the *genlock* signal and defines the world reference frame. Accurate synchronisation was validated using a flash-light torch and was further confirmed by the respectable quality observed from stereo Block Matching results [29]. The focus of all sensors is set to infinity, the aperture to maximum, and the exposure is manually fixed at the beginning of each capture sequence. The calibration on all four cameras' extrinsics, intrinsics, and distortion coefficients is done with a graph bundle-adjustment for improved multi-view calibration (see appendix for further details).

4.2. CroMo dataset

We collect a unique dataset comprising multi-modal captures such that each time point pertains to measurement of (1) **Polarisation**: raw stereo polarisation cameras produce $2448{\times}2048$ px stereo images. (2) **i-ToF**: 4 channel $640{\times}480$ px correlation images. (3) **Depth**: a structured-light capture of the scene resulting in a $848{\times}480$ px depth image. In addition to the three main sensors, IMU information is recorded to further enable future research directions. Our dataset consists of more than $20k$ frames, totalling $>80k$ images of indoor and outdoor sequences in challenging conditions, with no constraints on minimum or maximum scene range. We group these sequences into four different scenes which we name: *Kitchen*, *Station*, *Park* and *Facades*. Despite the multitude of senors, operating ranges

are not unlimited and our data collection also does not cover all possible scenarios; we further discuss limitations in our appendix. We report statistics per captured scene in Tab. 1 (lower). These statistics characterise our scene captures and provide useful information, *e.g.*, that the median scene depth differs greatly between indoor (*Kitchen*) and outdoor (*Station*, *Park* and *Facades*) scenes. This is a strong indicator for whether the i-ToF sensor will perform well. Tab. 1 (upper) provides a comparison with other depth datasets showing that CroMo is the first publicly available, modality rich dataset containing a large quantity of image data.

5. Experiments

Our experimental design evaluates (1) the effect of multiple modalities, accessible at training time, for monocular depth estimation and (2) the effect that changing network architecture has on depth quality, under consistent input signal. Our capture setup allows us to employ a standard MVS approach [52] on full temporal sequences of polarisation intensity frames (left-camera), to serve as ground-truth depth for our experimental work. This expensive offline optimisation leverages accordances amongst *all* frames per sequence, affording high quality depth to evaluate our ideas.

Multi-modal training We firstly evaluate combinations of training input signal by changing the number of sensors available to the model. We fix network encoder-decoder backbone components (*i.e.* ResNet50, analogous to [22]) and train models that leverage cues from a maximum of four sensors; left and right polarisation, i-ToF correlation and structured-light. We show predicted depth improvements, attainable by systematic addition of sensors, and quantify where best gains can be made. The training signal components used for our monocular depth estimation experiments are as follow: **Temporal (M)** extracts information from video sequences (3 frames), **Stereo (S)** uses stereo images, **i-ToF (T)** leverages i-ToF correlations via our two interconnected depth branches (see Sec. 3.2). Finally, **Structured-light (L)** incorporates an additional mask into the objective function, derived from information provided by our structured-light sensor. The structured-light signal is utilised *only when the mask improves the projection loss*. We explore alternative strategies to exploit the structured-light signal and discuss details on practical benefits (*e.g.* convergence speed) in the appendix.

Introduced signal components define our set of training experiments. For example, **Stereo and Structure Light (SL)** train the model using self-supervised stereo **(S)** and structured-light **(L)** information. Experiments therefore use differing subsets of the introduced loss terms (see Tab. 3).
Qualitative results are shown in Fig. 6. Unsurprisingly, self-supervised stereo **(S)** is relatively blurry and struggles to capture fine details, such as the thin, metallic arch on the *Facades* sample, or the furniture in the *Kitchen*. Addition of
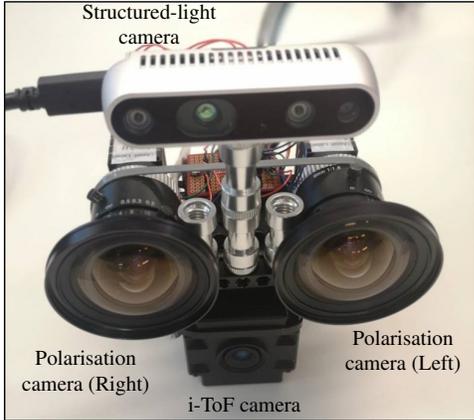
Figure 5. Our multi-modal camera rig (see Sec. 4.1).

Dataset	RGB	Depth	IMU	i-ToF	Polarisation	RAW	Real	Video	Available	Frames
Sturm et al. [56]	✓	✓	✓	-	-	-	✓	✓	✓	>20k
Agresti et al. [4]	-	(✓)	-	✓	-	-	✓	-	✓	113
Guo et al. [24]	-	✓	-	✓	-	✓	-	-	✓	2000
Zhu and Smith [66]	(✓)	-	-	-	✓	✓	✓	-	✓	1
Qiu et al. [48]	✓	-	-	-	✓	✓	✓	-	✓	40
Ba et al. [10]	✓	(✓)	-	-	✓	✓	✓	-	-	300
Kadambi et al. [32]	✓	✓	-	-	✓	✓	✓	-	✓	1
CroMo	(✓)	✓	✓	✓	✓	✓	✓	✓	✓	>20k

Scenes	GT depth statistics (meters)					valid ratio	# of seqs.	# of frames
	mean	*var.*	*min*	*max*	*median*			
Kitchen	3.3	3.6	0.3	15.7	2.9	0.95	3	2859
Station	4.9	14.8	0.3	18.9	3.6	0.86	11	7400
Facades	4.0	8.4	0.3	17.8	3.3	0.86	7	7228
Park	6.1	23.7	0.3	19.7	4.4	0.82	10	5551
Total	4.7	13.6	0.3	18.3	3.6	0.86	31	23038

Table 1. CroMo comparison and dataset statistics.

Models trained with **S**tereo (**S**) input	MP	GMACs	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
ResNet18 architecture [22]	14.36	20.17	1.7928	2.1982	0.3596	0.5061	0.7026	0.8009
ResNet50 architecture [22]	32.55	39.62	1.5037	2.0642	0.3383	0.5324	0.7262	0.8160
p2d [11] (ResNet50 - Stokes)	32.55	39.62	1.5938	2.1291	0.3884	0.4565	0.6632	0.7775
MiDaS architecture [50]	104.21	207.86	1.4021	1.9985	0.3252	0.5409	**0.7901**	**0.8281**
Our architecture (*Stereo (S) input*)	74.40	97.39	**1.3031**	**1.8889**	**0.3233**	**0.5533**	0.7301	0.8213

Table 2. Architectural comparisons under consistent modality sensor input; **S**tereo (**S**). Our proposed architecture improves quantitative results across the majority of metrics whilst remaining competitive in terms of compute and space requirements.

Image sensors	Training strategy	$\mathcal{L}_{stereo}$	$\mathcal{L}_{DF}$	$\mathcal{L}_{corr}$	$\mathcal{L}_{corr \longrightarrow pol}$	$\mathcal{L}_{struct}$	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
2	**S**tereo (**S**) *w/o* DF sampling	✓					1.5037	2.0642	0.3383	0.5324	0.7262	0.8160
	Stereo (**S**)	✓	✓				1.3031	1.8889	0.3233	0.5533	0.7301	0.8213
3	**S**tereo and i-**T**oF (**ST**)	✓	✓	✓	✓		1.2829	1.8573	0.3202	0.5541	0.7308	0.9062
	Stereo and Structured-**L**ight (**SL**)	✓	✓			✓	1.1233	1.7510	0.3168	0.5529	0.7370	0.9251
4	**S**tereo, i-**T**oF, Structured-**L**ight (**STL**)	✓	✓	✓	✓	✓	1.0699	1.6070	0.2891	0.6512	0.7882	**0.9266**
	STL+Temporal (**STLM**)	✓	✓	✓	✓	✓	**1.0031**	**1.4889**	**0.2527**	**0.7061**	**0.8066**	0.9246

Table 3. Model training strategies that differ in terms of available image sensor signals (utilised loss components). Sec. 3 and 4 provide details on loss function components and image sensors, respectively. In spite of having access to only a single, consistent modality during inference, the model benefits from visibility of additional training signals.

i-ToF and structured-light modalities, exclusively at training time, result in (**ST**), (**SL**), (**STL**) and can be observed to improve respective depth quality. Finally, (**STLM**) adds our temporal modality and improves detail recovery (*e.g.* metallic arch and fence). Qualitative results can be observed to corroborate our hypothesis; inclusion of additional modalities at training time afford the model multiple complementary depth cues that can qualitatively improve depth inference. Our experimental work highlights the nature of valuable investigation possible with our unique CroMo dataset.
Quantitative results are reported in Tab. 3. We follow [22], reporting standard evaluation metrics, with focus on the RMSE in our following experiments. Best performance is obtained when all sensors are used together (**STLM**) while self-supervision stereo (**S**) with only polarisation images performs worst. When additional modalities are added to self-supervision (**S**); *i.e.* i-ToF (**ST**) or structure light (**SL**), performance improves in both cases, with larger gains come from the addition of the latter. We conjecture that struc-

tured light information helps more due to the nature of our dataset and current distribution of image content therein *i.e.* ∼85% outdoor imagery, where i-ToF sensors are impaired by ambient light. Combining the i-ToF and structured-light sensors (**STL**), further improves. The best depth prediction utilises the temporal component (**STLM**). RMSE on Tab. 3 displays a clear trend; the availability of additional sensor cues at training time improves monocular inference.

Network architecture We next investigate the effect of network architecture on monocular estimation performance. Of note, we highlight that employing a larger capacity network is not the only way to improve prediction performance. We use our self-supervised stereo (**S**), *i.e.* baseline-modality, training strategy for all experiments that follow in this section. This strategy provided *weakest* performance in our previous investigation of training modality choice (Tab. 3). For this reason, we consider it an appropriate candidate with which to evaluate improvements afforded by changes to network architecture. We report *millions-*
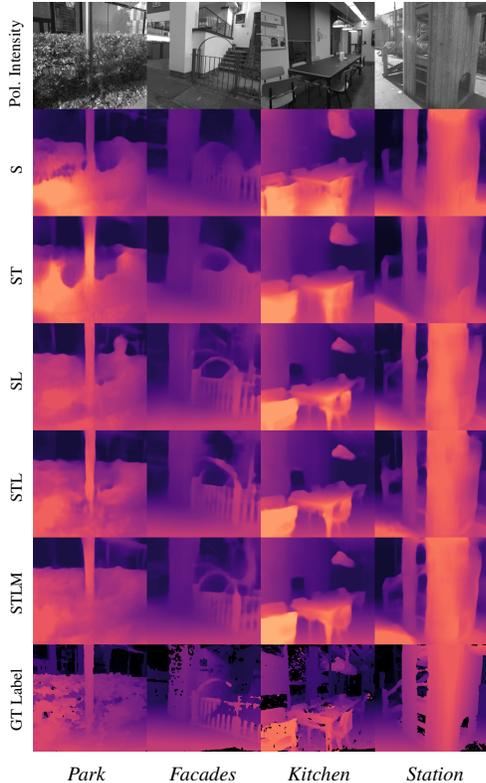
Figure 6. Same ResNet50 architecture as in [22] with different modalities: each new modality closes the gap with GT.

Figure 7. Different architectures, same training strategy **S**tereo (**S**): our new architecture produces the sharpest depth predictions.

of-parameters at inference (**MP**) and the *giga-multiply-accumulates* per second (**GMACs**) in order to evaluate size and compute-cost per architecture. Architectures consist of the **ResNet18** U-Net used in [22] and their supplementary material **ResNet50** variant, the **p2d** architecture [11] using **ResNet50** with a different data representation (Stokes), the **MiDaS** [50] architecture and **Ours** (see Sec. 3.1, Fig. 3a).

Qualitative results are shown in Fig. 7. It may be observed that the **ResNet18** architecture with smallest (**MP**) fails to obtain good background detail of the swing frame structure (*Station* sample) or of the tree (*Park* sample). The **ResNet50** variant slightly improves detail, especially with raw measurements instead of Stokes (**p2d** [11]). Even when increasing network capacity *c.* three-fold with **MiDaS**, results are unsatisfying. Our proposed architecture (**Ours**) requires smaller capacity and computation for a sharper reconstruction of the swing and the tree. We disentangle the benefits of additional sensor modalities from our model contributions, highlighting the advantage of gated convolutions and our DF-based approach towards reducing blur.

Quantitative results are reported in Tab. 2. The smallest architecture **ResNet18** [22] performs worst. The larger U-Net **ResNet50** performs better, and has been generally adopted [11, 21]. Note **p2d** [11] uses a different data representation (Stokes) for polarisation *cf*. **ResNet50**; performance decreases. We believe the Stokes representation, us-
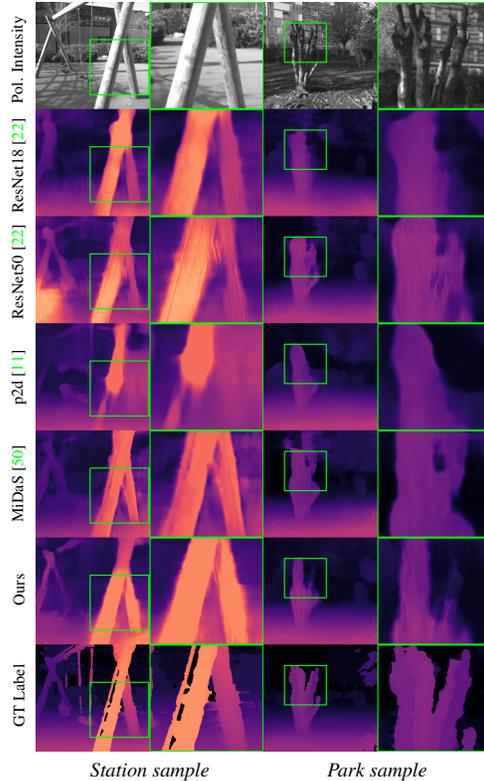
ing angle directly, is more sensitive to noise and not appropriate for an SSIM loss with the self-supervised stereo (**S**) training strategy. MiDaS [50] provides second best performance and yet necessitates roughly ×2 GMACs. Our architecture provides best performance while remaining relatively compact which we largely attribute to gated convolutions and displacement field estimation (see Sec. 3.1).

6. Conclusion

We systematically investigate the effect of using additional information from co-modal sensors at training time, for the task of monocular depth estimation from polarisation imagery. Our exploration is enabled through a unique multimodal video dataset which constitutes synchronized images from binocular polarisation, raw i-ToF and structured-light depth. We quantify the beneficial influence of both *passive* and *active* sensors, leveraging self-supervised and cross-modal learning strategies that lead to the proposal of a new method providing sharper and more accurate depth estimation. This is made possible through two physical models that describe the relationships between polarisation and surface normals on one side and correlation measures and scene depth on the other. We believe that our fundamental investigation of modality combination and the CroMo dataset can accelerate research of both spatial and temporal fusion, towards advancing cross-modal computer vision.

References

[1] Intel realsense depth camera d435i. https://www.intelrealsense.com/depth-camera-d435i/. Accessed: 2021-11-22. 3

[2] Supreeth Achar, Joseph R Bartels, William L'Red' Whittaker, Kiriakos N Kutulakos, and Srinivasa G Narasimhan. Epipolar time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 36(4):1–8, 2017. 3

[3] Gianluca Agresti, Ludovico Minto, Giulio Marin, and Pietro Zanuttigh. Deep learning for confidence information in stereo and tof data fusion. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 3

[4] Gianluca Agresti, Henrik Schaefer, Piergiorgio Sartor, and Pietro Zanuttigh. Unsupervised domain adaptation for tof data denoising with adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 7

[5] Gianluca Agresti and Pietro Zanuttigh. Deep learning for multi-path error removal in tof sensors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 3

[6] Gary A Atkinson. Polarisation photometric stereo. *Computer Vision and Image Understanding*, 160:158–167, 2017. 3

[7] Gary A Atkinson and Edwin R Hancock. Multi-view surface reconstruction using polarization. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 309–316. IEEE, 2005. 3

[8] Gary A Atkinson and Edwin R Hancock. Recovery of surface orientation from diffuse polarization. *IEEE transactions on image processing*, 15(6):1653–1664, 2006. 1, 2

[9] Yunhao Ba, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Physics-based neural networks for shape from polarization. *arXiv preprint arXiv:1903.10210*, 2019. 2

[10] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 554–571, Cham, 2020. Springer International Publishing. 7

[11] Marc Blanchon, Désiré Sidibé, Olivier Morel, Ralph Seulin, Daniel Braun, and Fabrice Meriaudeau. P2d: a self-supervised method for depth estimation from polarimetry. In *25th International Conference on Pattern Recognition (ICPR 2020)*, 2020. 1, 2, 7, 8

[12] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1558–1567, 2017. 3

[13] Rui Dai, Srijan Das, and François Bremond. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13053–13064, October 2021. 4

[14] Adrian A Dorrington, John P Godbaz, Michael J Cree, Andrew D Payne, and Lee V Streeter. Separating true range measurements from multi-path and scattering interference in commercial range cameras. In *Three-Dimensional Imaging, Interaction, and Measurement*, volume 7864, page 786404. International Society for Optics and Photonics, 2011. 3

[15] Sergi Foix, Guillem Alenya, and Carme Torras. Lock-in time-of-flight (tof) cameras: A survey. *IEEE Sensors Journal*, 11(9):1917–1926, 2011. 2

[16] Daniel Freedman, Yoni Smolin, Eyal Krupka, Ido Leichter, and Mirko Schmidt. Sra: Fast removal of general multipath for tof sensors. In *European Conference on Computer Vision*, pages 234–249. Springer, 2014. 3

[17] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 2

[18] N Missael Garcia, Ignacio De Erausquin, Christopher Edmiston, and Viktor Gruev. Surface normal reconstruction using circularly polarized light. *Optics express*, 23(11):14391–14406, 2015. 3

[19] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue, 2016. 2

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 2

[21] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 2, 8

[22] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3838, 2019. 1, 2, 5, 6, 7, 8, 3

[23] John P Godbaz, Michael J Cree, and Adrian A Dorrington. Closed-form inverses for the mixed pixel/multipath interference problem in amcw lidar. In *Computational Imaging X*, volume 8296, page 829618. International Society for Optics and Photonics, 2012. 3

[24] Qi Guo, Iuri Frosio, Orazio Gallo, Todd Zickler, and Jan Kautz. Tackling 3d tof artifacts through learning and the flat dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 368–383, 2018. 3, 7

[25] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks, 2018. 2

[26] Mohit Gupta, Shree K Nayar, Matthias B Hullin, and Jaime Martin. Phasor imaging: A generalization of correlation-based time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 34(5):1–18, 2015. 3

[27] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012. 2

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3, 2

[29] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005. 6

[30] Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clément Ménier. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine Vision and Applications*, 27(7):1005–1020, Jun 2016. 5

[31] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 2017–2025. Curran Associates, Inc., 2015. 5

[32] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Depth sensing using geometrically constrained polarization normals. *International Journal of Computer Vision*, 125(1-3):34–51, 2017. 3, 7

[33] Achuta Kadambi, Refael Whyte, Ayush Bhandari, Lee Streeter, Christopher Barsi, Adrian Dorrington, and Ramesh Raskar. Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. *ACM Transactions on Graphics (ToG)*, 32(6):1–10, 2013. 3

[34] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8602–8611, 2020. 2

[35] Ahmed Kirmani, Arrigo Benedetti, and Philip A Chou. Spumic: Simultaneous phase unwrapping and multipath interference cancellation in time-of-flight cameras using spectral methods. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2013. 3

[36] R. Lange and P. Seitz. Solid-state time-of-flight range camera. *IEEE Journal of Quantum Electronics*, 37(3):390–397, 2001. 5

[37] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1750–1758, 2020. 2

[38] Fotios Logothetis, Roberto Mecca, Fiorella Sgallari, and Roberto Cipolla. A differential approach to shape from polarisation: A level-set characterisation. *International Journal of Computer Vision*, 127(11-12):1680–1693, 2019. 3

[39] Adrian Lopez-Rodriguez, Benjamin Busam, and Krystian Mikolajczyk. Project to adapt: Domain adaptation for depth completion from noisy and sparse sensor data. *arXiv preprint arXiv:2008.01034*, 2020. 3

[40] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3288–3295. IEEE, 2019. 3

[41] Fangchang Mal and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single im-age. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018. 3

[42] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, 126(9):942–960, 2018. 2

[43] Daisuke Miyazaki, Takuya Shigetomi, Masashi Baba, Ryo Furukawa, Shinsaku Hiura, and Naoki Asada. Surface normal estimation of black specular objects from multiview polarization images. *Optical Engineering*, 56(4):041303, 2016. 3

[44] Nikhil Naik, Achuta Kadambi, Christoph Rhemann, Shahram Izadi, Ramesh Raskar, and Sing Bing Kang. A light transport model for mitigating multipath interference in time-of-flight sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–81, 2015. 3

[45] Gregory P Nordin, Jeffrey T Meier, Panfilo C Deguzman, and Michael W Jones. Micropolarizer array for infrared imaging polarimetry. *JOSA A*, 16(5):1168–1174, 1999. 3

[46] Matthew O'Toole, Felix Heide, Lei Xiao, Matthias B Hullin, Wolfgang Heidrich, and Kiriakos N Kutulakos. Temporal frequency probing for 5d transient analysis of global light transport. *ACM Transactions on Graphics (ToG)*, 33(4):1–11, 2014. 3

[47] Can Pu, Runzi Song, Radim Tylecek, Nanbo Li, and Robert B Fisher. Sdf-gan: Semi-supervised depth fusion with multi-scale adversarial networks. *arXiv preprint arXiv:1803.06657*, 2018. 3

[48] Simeng Qiu, Qiang Fu, Congli Wang, and Wolfgang Heidrich. Polarization demosaicking for monochrome and color polarization focal plane arrays. In Hans-Jörg Schulz, Matthias Teschner, and Michael Wimmer, editors, *Vision, Modeling and Visualization*. The Eurographics Association, 2019. 7

[49] M. Ramamonjisoa, Y. Du, and V. Lepetit. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[50] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 7, 8

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 3

[52] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 6

[53] William AP Smith, Ravi Ramamoorthi, and Silvia Tozza. Height-from-polarisation with unknown lighting or albedo. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2875–2888, 2018. 2

10

[54] Kilho Son, Ming-Yu Liu, and Yuichi Taguchi. Learning to remove multipath distortions in time-of-flight range images for a robotic arm setup. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3390–3397. IEEE, 2016. 3

[55] Jaime Spencer, Richard Bowden, and Simon Hadfield. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14402–14413, 2020. 2

[56] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 7

[57] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. Deep end-to-end time-of-flight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6383–6392, 2018. 3

[58] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *Proceedings of the International Conference on 3D Vision (3DV))*, pages 11–20. IEEE, 2017. 3

[59] Jamie Watson, Michael Firman, Gabriel J. Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *The International Conference on Computer Vision (ICCV)*, October 2019. 5, 3

[60] Refael Whyte, Lee Streeter, Michael J Cree, and Adrian A Dorrington. Review of methods for resolving multi-path interference in time-of-flight range cameras. In *SENSORS, 2014 IEEE*, pages 629–632. IEEE, 2014. 3

[61] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks, 2016. 2

[62] Tomonari Yoshida, Vladislav Golyanik, Oliver Wasenmüller, and Didier Stricker. Improving time-of-flight sensor for specular surfaces with shape from polarization. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1558–1562. IEEE, 2018. 3

[63] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-form image inpainting with gated convolution, 2019. 3, 2

[64] Ye Yu, Dizhong Zhu, and William AP Smith. Shape-from-polarisation: a nonlinear least squares approach. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2969–2976, 2017. 2

[65] Pietro Zanuttigh, Giulio Marin, Carlo Dal Mutto, Fabio Dominio, Ludovico Minto, and Guido Maria Cortelazzo. Time-of-flight and structured light depth cameras. *Technology and Applications*, pages 978–3, 2016. 2

[66] Dizhong Zhu and William AP Smith. Depth from a polarisation+ rgb stereo pair. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7586–7595, 2019. 3, 4, 5, 7

7. CroMo: Supplementary Material

We provide additional materials to supplement our main paper. In Sec. 7.1 we provide observations on the properties of light polarisation. Sec. 7.2 states the specifics for our surface normal estimation process. In Sec. 7.3, we provide additional details for our multi-view camera calibration procedure, Sec. 7.4 provides some further modelling details and finally Sec. 7.5 gives supplementary information on our network architectures and learning parameters.

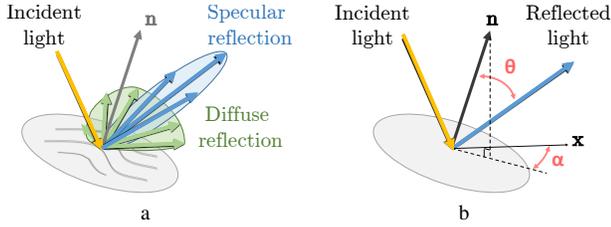## 7.1. Light polarisation parameters

Figure S1. (a) differing types of reflected light and (b) the link between a surface normal **n**, its viewing angle θ and its azimuth angle α (right).

Most natural light sources emit unpolarized light that only becomes polarized if reflected. Hence the type of reflection, illustrated in Fig. S1, either *diffuse* (d) or *specular* (s), influences the characteristics of the reflected polarized light. More specifically, the reflective surface influences the relation between the normals' parameters (θ, α) and the polarisation parameters (ρ, ϕ), defined in Eq. 2 and 3 of the main paper.

7.2. Surface normals

In the main manuscript we estimate polarisation intensity using the varying coordinates of surface normals. Hence, the computation of these normals, derived from network depth prediction, plays an important role in the training process. To increase the robustness of estimated normals, we compute the cross products using four distinct pairs of orthogonal directions as in [S4]:

$$
\begin{cases}
\mathbf{n}_0 & = \partial_x \mathbf{v} \times \partial_y \mathbf{v} \\
\mathbf{n}_1 & = \partial_{-x} \mathbf{v} \times \partial_{-y} \mathbf{v} \\
\mathbf{n}_2 & = \partial_{x+y} \mathbf{v} \times \partial_{-x+y} \mathbf{v} \\
\mathbf{n}_3 & = \partial_{-x-y} \mathbf{v} \times \partial_{x-y} \mathbf{v}
\end{cases}
\tag{S1}
$$

The weighted average of these normals is calculated using weights w_i where:

$$
\begin{cases}
w_0 & = \exp(-0.5\|\partial_x i_{un}\|_1) \cdot \exp(-0.5\|\partial_y i_{un}\|_1) \\
w_1 & = \exp(-0.5\|\partial_{-x} i_{un}\|_1) \cdot \exp(-0.5\|\partial_{-y} i_{un}\|_1) \\
w_2 & = \exp(-0.5\|\partial_{x+y} i_{un}\|_1) \cdot \exp(-0.5\|\partial_{-x+y} i_{un})\|_1) \\
w_3 & = \exp(-0.5\|\partial_{-x-y} i_{un}\|_1) \cdot \exp(-0.5\|\partial_{x-y} i_{un})\|_1)
\end{cases}
\tag{S2}
$$

The final surface normal (unnormalized) is then estimated by their linear combination:

$$
\mathbf{n} = \frac{1}{4} \sum_{i=0}^{3} w_i \cdot \mathbf{n}_i
\tag{S3}
$$

Weights w_i result in neighbouring pixels of i_{un} that contain strong color disparity, to be down-weighted in the normal computation. This follows from the assumption that such pixels are more likely to represent different objects. Conversely, if neighbouring pixels possess similar color, they are more likely to correspond to the same object and their associated partial derivatives are more likely to provide normals that accurately describe the observed object shape.

7.3. Graph-based bundle adjustment

As discussed in Sec. 4.1 of our main paper the calibration of extrinsics, intrinsics and distortion coefficients, for all four capture-rig cameras, is achieved using a graph-based bundle-adjustment [S2] that improves multi-view calibration. We provide here further details of our multi-view calibration procedure.

We start with well established calibration methods [S1] to obtain the intrinsics K_k and distortion coefficients d_k for each camera C_k, where $k \in \{0, 1, 2, 3\}$. We use a standard pinhole camera model and define C_0 as the left polarisation camera, C_1 the right polarisation camera, C_2 the i-ToF camera, and C_3 the structure light camera. We use five parameters for the distortion coefficients and collect n images of a calibration checkerboard, from all cameras synchronously. In practice we move the checkerboard in front of the cameras while keeping the camera rig stationary. We attempt to cover as wide a field-of-view as possible for all four cameras. We find it is more important to thoroughly cover and account for the extremities of the individual images as opposed to attempting to be visible to all cameras simultaneously. Further, we estimate the rigid transformation for each camera pair composed of C_0 (our world reference), and camera C_k in turn, where $k \in \{1, 2, 3\}$. This provides the extrinsics $T_{k \to 0} = [R_{k \to 0}|t_{k \to 0}]$ for camera C_k (with $T_{0 \to 0} = [I|0]$).

These initial intrinsic, extrinsic parameter values and the distortion coefficients are however sub-optimal as they are obtained by solving successive sub-optimisation problems. Towards improving the multi-camera calibration, we define the reprojection error of points X^j on the image I_i for the camera C_k as

$$
\widehat{x_j^i} = \pi \left(T_{k \to 0}, T_0^i, X^j, d_k, K_k \right)
$$

$$
E_k^i = \sum_{j=0}^{\#\text{points}} \mathbb{1}_{\widehat{x_j^i} \in I_i} \cdot \text{dist} \left(x_j^i, \widehat{x_j^i} \right)^2
\tag{S4}
$$

Where T_0^i is the position of camera C_0 for image i, and $\widehat{x_j^i}$ is the distorted 2D point from the projection function $\pi(\cdot)$ which projects a 3D Point X^j visible by the camera C_k at position $T_0^i \cdot T_{k\to 0}$ with distortion coefficients d_k and intrinsic parameters K_k on image I_i. The function dist$(\cdot)$ defines the robustified distance between 2D points, *i.e.* a Huber m-estimator, and x_j^i is the 2D point detected on the checkerboard with a corner detector corresponding to the 3D point X^j in image I_i. The indicator function $\mathbb{1}_{\widehat{x_j^i} \in I_i}$ defines whether the 2D Point $\widehat{x_j^i}$ is visible in image I_i.

Finally, we used a graph-based bundle-adjustment [S2] to model the global problem, for all cameras C_k, jointly as:

$$\min_{T_0^i, T_{k\to 0}, d_k, K_k} \sum_{k=0}^{\#\text{cameras}} \sum_{i=0}^{\#\text{images}} E_k^i, \qquad \text{(S5)}$$

with T_0^0 fixed to $[I|0]$ in order to properly constrain the gauge freedom. All camera calibration parameters are initialised using the values obtained from the original individual calibrations.

This formalism, borrowed from the SLAM community [S3], allows us to optimize all parameters, *i.e.* the intrinsic, the extrinsic and the distortion parameters for all cameras, jointly. We find the global optimisation process is able to improve our calibration RMSE by ∼5–10%.

7.4. Additional modelling details

7.4.1 Reflection ambiguities

A diffuse-specular ambiguity initially exists in our formulation; pertaining to diffuse or specular reflection (see Eq.3, main paper). This ambiguity is addressed during training via the min operator found in Eq.18. We propose to resolve reflection ambiguities (per-pixel) by minimization of the SSIM loss between respective {specular, diffuse} images and the input image, towards consistently providing a valid training signal. Secondly, the azimuthal π-ambiguity is directly accounted for by the formulation of Eq.1; the inherent $\cos(\cdot)$ modulation nullifies ambiguity found in its input (2ϕ component of the argument) and thus supervision is not adversely affected due to ϕ being modulo π.

7.4.2 Wrappings ambiguities

An analytical solution exists for the correlation to depth transform and a wrapping ambiguity remains. However we highlight that a reconstructed depth, although "phase wrapped", is still able to provide reliable surface normals that can be used to produce (1) the degree of linear polarisation and (2) the Angle of Polarisation, for both diffuse and specular surfaces. Once these are projected to the $N2$

referential, this information is used in conjunction with the brightness of the left polarisation image to render valid "Recovered polarisation" images, (see Fig 3b of our main paper).

7.4.3 Polarization intensity recovery

To render the intensity, we require the brightness of each pixel (i_{un} in Eq.1). We obtain the brightness of the polarisation image by channel-wise summing of the left polarisation input pixel values. Two images are rendered following Eq.1; for both the cases of diffuse and specular images. We use a binary mask to select values, *pixel-wise*, from either the specular or diffuse image. The mask selects pixels such that the minimum SSIM loss between the {specular, diffuse} image and the input image are retained.

The two images formed therefore constitute only an *intermediary* step towards producing a final image. We use a binary mask to then select values, *pixel-wise*, from either the specular or diffuse image to form a new image containing the pixels that retain the minimum SSIM loss between the {specular, diffuse} image and the input image (*i.e.* the min in Eq.18 is *per pixel*). We thus form a final image that contains both specular and diffuse components.

7.4.4 Correlation image rendering from depth

Analogous to the Polarization image strategy, we use the input correlation image (obtaining α, β estimates), in addition to depth information, to estimate both the ambient β and reflectance α, for correlation reconstruction.

7.5. Architecture and training details

7.5.1 Architecture

We provide additional description for the network architecture that we propose in order to process the considered input modalities. Instances of this architecture are depicted as edges '$N1$' and '$N2$' in the system design; see Fig 3a of our main paper.

We employ a standard U-Net architecture, similar to our baseline [22], including skip connections. The encoder is based on a 'Resnet' [28] style block, with the original convolutional layer replaced by gated convolution [63]. The size of the input images are $512 \times 544 \times 12$ for polarisation and $640 \times 480 \times 4$ for i-ToF, respectively.

For polarisation, we have the following configuration; layer one: $512 \times 544 \times 64$, layer two: $256 \times 272 \times 128$, layer tree: $128 \times 136 \times 256$, layer four: $64 \times 68 \times 512$. For i-ToF, we have the following configuration; layer one: $640 \times 480 \times 64$, layer two: $320 \times 240 \times 128$, layer tree: $160 \times 120 \times 256$, layer four: $80 \times 60 \times 512$. Both depth and

Displacement Field decoders are a standard cascade of convolutions with layer resizing. Encoder skip connections are concatenated after each resize operation (see Fig. 3a).

7.5.2 Training parameters

To aid reproducibility, we report training parameters and hyperparameters. We use identical training parameters and align with our baseline [22] where possible. We use the Ranger optimiser [S5] and batch sizes of 8, a learning rate of $1e-4$ with an exponential learning rate decay. We train all considered methods for 50 epochs.

7.5.3 Comparison with RGB input

A direct comparison with RGB input forms a relevant and interesting line of enquiry. Our custom capture rig does not currently accommodate this modality directly. However, towards investigating this experimentally, we did transform the polarisation input frame to an RGB frame by considering the polarisation intensity of each RGB channel, individually. We note that this is *not* directly equivalent to an RGB sensor since the bayer pattern differs. We actively decided *not* to include this experimental work in the main paper to avoid misinterpretation and confusion. Preliminary work evaluating our Polarisation input *cf*. the noted "*Polarisation-converted-to-RGB*" showed improvements using Polarisation (RMSE 1.4) over "*Polarisation-converted-to-RGB*" (RMSE 1.53).

7.5.4 Controlling for capture environment

We note that the i-ToF modality excels in indoor environments, however these represent a relatively smaller portion of our dataset. To corroborate this, we report an experiment that considers our various training strategies, tested on only an indoor environment (*Kitchen*). The addition of i-ToF (from (**S**) to (**ST**)), at training time, significantly improves the predicted depth in this restricted setting (see Tab. S1).

Image sensors	Training strategy	Sq Rel	RMSE	RMSE Log
2	(**S**)	0.6202	1.2930	0.2944
3	(**ST**)	0.3001	0.6520	0.2237
4	(**STLM**)	**0.2105**	**0.5431**	**0.180**

Table S1. Test on *Kitchen* scene (780 frames): additional sensors can be observed to improve performance. The largest improvement comes from the addition of the i-ToF, (from (**S**) to (**ST**)), in an exclusively indoor test setting.

7.5.5 Further analysis of *where* additional sensors help

We include preliminary further analysis with respect to investigation of scenarios where additional sensors help. We include an example that highlights two points (see Fig. S2). Due to the concave nature of the scene, the addition of ToF information alone during training (from **S** to **ST**) adversely impacts the depth prediction and we find MPI often detrimental to the ToF sensor in such cases. Additional sensors (from **ST** to **STLM**) do however improve final depth estimation and we show gains achievable by adding orthogonal signal during training, where inference utilises only a single polarisation image in all cases. Additional investigation and rigorous analysis of such scenarios makes for interesting future work.

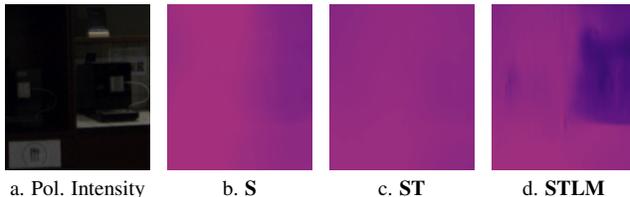| a. Pol. Intensity | b. **S** | c. **ST** | d. **STLM** |

Figure S2. Depth estimation improvements possible from a common input (a). We show gains achievable by adding orthogonal signal during training, where inference utilises only a single polarisation image in all cases. See text for further detail.

7.5.6 Additional structured light experiments

The structured light sensor present in our camera rig offers low-noise signal which we find can also be leveraged in a supervised fashion, directly. For completeness, we compare the resulting depth estimation when supervising directly with structure-light (**D**) and our approach, using unsupervised signals (**STLM**). The structure-light signal, obtained from our Realsense sensor, is claimed reliable up to a 10 meters range according to the constructor [1]. We thus further investigate by evaluating performance over distinct $0 - 10m$ and $0 - 20m$ ranges. Results in Tab. S2 show that the fully supervised method (**D**) can offer similar performance to our approach (**STLM**) in the range $0 - 10m$ yet performance degrades by significant margins when considering the more challenging $0 - 20m$ range. This highlights the benefits of our unsupervised multi-modal strategy (**STLM**); leveraging information from multiple sensor sources and an ability to learn to adapt when a particular sensor results in low quality measurement, due to unsuitable physical conditions (*e.g.* structured light in the $10 - 20m$ range).

7.5.7 Additional details on the $\mathcal{L}_{\mathbf{struct}}$ loss

We select to use a structured light loss similar to the loss proposed in [59]. We find that such indirect supervision of

Image sensors	Training strategy	0-10m			0-20m		
		Sq Rel	RMSE	RMSE Log	Sq Rel	RMSE	RMSE Log
1	(D)	**0.9479**	**1.4246**	**0.2117**	5.447	6.2629	1.6134
4	(STLM)	1.0031	1.4889	0.2527	**1.3994**	**2.9512**	**0.3879**

Table S2. Comparison of training strategies for two depth prediction ranges. Our training strategy (**STLM**) works well in spite of the operational limits of particular sensors.

the structure light signal allows to automatically select the best source of information, particularly in situations where the structure light signal fails or becomes unreliable (as discussed in Sec. 7.5.6). Formally, given a depth from the structured light D_{struct}, the loss reads:

$$\mathcal{L}_{\text{struct}} = E_{\text{pe}}\left(I_l, I_{\text{right}\underset{D_{\text{struct}}}{\longrightarrow}\text{left}}\right) \tag{S6}$$

We make use of an additional $\mathcal{L}_1$ loss between predicted depth and the D_{struct} depth, when $\mathcal{L}_{\text{struct}}$ is minimal (see Eq.19 in the main manuscript).

7.5.8 Limitations and Societal Impact

Limitations We note distinct limitations that relate to our sensor setup. Active sensors have limited range and areas of operativity *e.g.* i-ToF often offers weaker performance outdoors, structure light sensors are of limited range, and polarisation sensors sacrifice spacial sampling resolution for spectral sampling resolution. Our multi-modal ideas attempt to combat these limitations indirectly however we remain bound by the physical laws of light.

Additionally, our current hardware setup is operable by a single person, and yet data capture is currently more cumbersome than *e.g.* use of a modern smartphone. Training data collection, that involves the acquisition of multiple modalities, currently induces a somewhat larger investment of effort over monomodal capture. With the argument being that the cost may then be recouped when assessing monocular inference time performance. Our hardware rig constitutes a research prototype and form factor likely improves as camera evolution results in further reductions in sensor size, weight and cost.

Finally we would note that our current dataset does not yet capture all possible scenarios and represents but a subset of urban scenes where depth estimation can prove valuable. Future capture sessions will look to enrich and widen the recorded capture scenarios, towards increasing the value of the data resource that we provide to the community.

Societal Impact We note that our proposed CroMo dataset was collected by only two human operators in urban environments. While care was taken towards objective scene capture, such collected data may yet reflect the biases of human operators; influencing specific content, scenarios or

capture setups. Efforts towards the reduction of bias, introduced by manual human operators, might suggest mounting of the system on automatic vehicles in future. Additional ideas, toward mitigation of the axis of bias relating to manual data capture, can be considered an interesting future research direction.

References

[S1] J Heikkila and O Silven. A four-step camera calibration procedure with implicit image correction. In *Proceedings of ieee computer society conference on computer vision and pattern recognition*, pages 1106–1112. IEEE, 1997. 1

[S2] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. G2o: A general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation*, pages 3607–3613, 2011. 1, 2

[S3] B Triggs, P F McLauchlan, R I Hartley, and A W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999. 2

[S4] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In *AAAI*, 2018. 1

[S5] Hongwei Yong, Jianqiang Huang, Xiansheng Hua, and Lei Zhang. Gradient centralization: A new optimization technique for deep neural networks. In *European Conference on Computer Vision*, pages 635–652. Springer, 2020. 3