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

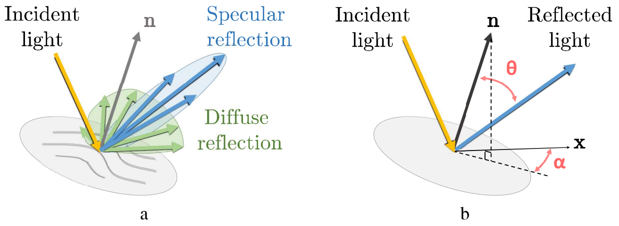

Figure S1. (a) differing types of reflected light and (b) the link between a surface normal **n**, its viewing angle θ and its azimuth angle α (right).

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

$$\begin{aligned} \widehat{x_j^i} &= \pi\left(T_{k\to 0}, T_0^i, X^j, d_k, K_k\right) \\ E_k^i &= \sum_{j=0}^{\#\text{points}} \mathbb{1}_{\widehat{x_j^i} \in I_i} \cdot \text{dist}\left(x_j^i, \widehat{x_j^i}\right)^2 \end{aligned} \tag{S4}$$

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 $\mathrm{dist}(\cdot)$ 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:

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

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 $512\times544\times12$ for polarisation and $640\times480\times4$ for i-ToF, respectively.

For polarisation, we have the following configuration; layer one: $512\times544\times64$, layer two: $256\times272\times128$, layer tree: $128\times136\times256$, layer four: $64\times68\times512$. For i-ToF, we have the following configuration; layer one: $640\times480\times64$, layer two: $320\times240\times128$, layer tree: $160\times120\times256$, layer four: $80\times60\times512$. 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.

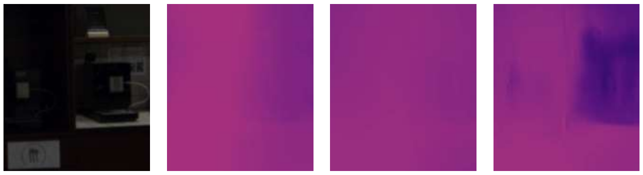
a. Pol. Intensity b. **S** c. **ST** d. **STLM**

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 $\mathcal{L}_{\text{struct}}$ 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:

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

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