Camera Augmentation: Enabling Uncalibrated Stereo Matching of Minimally-Invasive Surgery Images by Training from the Wealth of Public Synthetic Image Datasets

Rasoul Sharifian[1,2*†], Navid Rabbani[1†], Yongcong Zhang[1,2], Adrien Bartoli[1,2]

[1*]DIA2M, DRCI, CHU Clermont-Ferrand, France.
[2]SURGAR, Surgical Augmented Reality, Clermont-Ferrand, France.

*Corresponding author(s). E-mail(s): rasoul.sharifian.cs@gmail.com;
†These authors contributed equally to this work.

Abstract

Purpose. Existing models for stereo matching in Minimally-Invasive Surgery (MIS) require calibrated stereo images. Accurate calibration is however often unavailable intraoperatively. Training an uncalibrated stereo model is thus attractive but challenging owing to the lack of disparity-labelled surgical images.
Methods. We leverage the wealth of non-medical stereo synthetic image datasets. These data were however generated in ideal conditions –rectified and with centred principal points– hence differ from real uncalibrated MIS images. We propose camera augmentation, a new type of image augmentation that augments a dataset by altering the camera's orientation and intrinsic parameters via geometric parameters. We augment the idealised existing datasets, sampling the geometric augmentation parameters from distributions estimated through an in-depth analysis and modelling of stereo laparoscopes. This forms the Camera Augmentation Training Strategy (CATS), with which we retrain RAFTStereo and IGEV++ for zero-shot uncalibrated stereo matching in MIS.
Results. We evaluated using the SCARED, StereoMIS, RIS2017, and an in-house datasets. In the uncalibrated setting on the SCARED dataset, CATS-RAFTStereo and CATS-IGEV++ achieved End-Point-Errors (EPE) of 1.42 and 1.41 pixels. This is a successful result, as the reference pretrained models obtained 1.23 and 1.21 pixels in the calibrated setting and failed in the uncalibrated setting.

Conclusion. Camera augmentation bridges the gap between ideally-conditioned datasets and the real surgical conditions of uncertain or unavailable calibration, enabling the retraining of state-of-the-art architectures. Beyond stereo, the proposed CATS is applicable to tasks sensitive to camera geometry. Code and models will be released publicly.

Keywords: Minimally Invasive Surgery, Stereo Matching, Disparity, Uncalibrated

1 Introduction

Depth estimation in MIS plays a crucial role in downstream applications such as Augmented Reality [1–3] and robotic surgery guidance [3]. Even if depth can be estimated from monocular images acquired by standard laparoscopes, stereo depth estimation benefits from strong triangulation constraints arising from left–right image consistency. This reduces the inaccuracies commonly occurring in monocular depth estimation where the models heavily rely on learnt priors. The use of stereo laparoscopes has been steadily increasing, owing to the rise of robot-assisted MIS systems.

A stereo system comprises two cameras. The stereo calibration process determines each camera's intrinsic parameters and the extrinsic parameters describing their relative rotation and translation. Calibrated stereo methods have standard preprocessing steps: undistortion, rectification, and principal-point alignment (Fig. 1). Disparity estimation is then performed, and the disparity mapped to metric depth. While early works established the fundamental components [4], recent neural architectures such as RAFTStereo [5], IGEV++ [6] and FoundationStereo [7] have shown substantially improved performance and robustness and, importantly, generalisation across datasets. These models require rectified images with centred principal points, which necessitate camera calibration; as expected, their performance directly depends on the accuracy of calibration parameters. Calibrated stereo methods specifically adapted to MIS were proposed, including MSDESIS [8] and MCF-SMSIS [9] which jointly estimate disparity and segmentation with shared features. Beyond the medical domain, recent stereo matching methods have demonstrated promising cross-domain generalisation [5–7]. In practice, however, accurate calibration may not be available throughout surgery, as the parameters may vary with zoom level, focusing and any other digital or mechanical adjustments. Calibration is also impractical to perform routinely in operating room conditions. Consequently, both the intrinsic and extrinsic parameters are often unavailable or unreliable, making the standard stereo pipeline inapplicable. In spite of these known practical challenges in camera calibration, uncalibrated stereo depth estimation –independent of camera calibration parameters, both intrinsics and extrinsics– remains largely unexplored in MIS. We show that, while differences in scene content may not pose a major issue for large learning-based models, discrepancies in camera settings ultimately lead to systematic failures. This issue cannot be resolved by fine-tuning over the limited depth-labelled MIS data.

In the uncalibrated setting (Fig. 1), although the resulting disparity cannot be mapped to metric depth, as this mapping depends on the calibration parameters, it

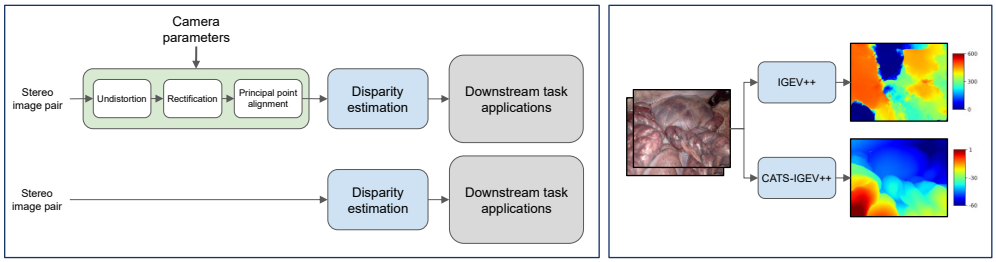

Fig. 1 Left: Pipelines for calibrated (top) and uncalibrated (bottom) stereo matching. Right: disparity estimated by the IGEV++ [6] and the proposed CATS-IGEV++ models in uncalibrated setting.

can still be mapped to *relative* depth. Importantly, many downstream tasks in MIS operate effectively with relative rather than metric depth (*e.g.* collision avoidance [10], instrument tracking [11] and surgical simulation [12]). Augmented reality systems compute the scale of uncalibrated reconstructions concurrently with registration, and visual servoing controls robots effectively from uncalibrated images and reconstructions, highlighting the practical utility of relative depth. We address the problem of uncalibrated stereo in MIS by leveraging the wealth of non-medical stereo synthetic image datasets. A major challenge is that these data were generated in ideal conditions (perfectly rectified, with centred principal point) and published without 3D scene parameters, preventing re-rendering. To overcome these limitations, we introduce the Camera Augmentation Training Strategy (CATS), a framework that enables robust stereo learning under uncalibrated conditions. Concretely, our main contributions are threefold. First, we propose *camera augmentation*, a framework that augments a dataset by altering the camera's orientation and intrinsic parameters. It does not require knowing the camera pose or the depth; yet, it offers geometric control over the augmented data. Second, we use camera augmentation to enhance existing idealised non-medical datasets to simulate uncalibrated yet near-rectified stereo laparoscopes. We sample geometric parameters from distributions derived through a detailed analysis and modelling of stereo laparoscope configurations. Third, we use CATS to retrain RAFTStereo and IGEV++ for zero-shot uncalibrated stereo matching in MIS.

2 Camera Augmentation

We introduce camera augmentation as a generalisation of conventional image augmentation, in which the augmented images are generated by perturbing the camera parameters. Unlike conventional augmentation performed by basic 2D geometric transformations within the image plane, camera augmentation explicitly models the geometric image-formation process. This allows the generation of new, yet geometrically-consistent image samples, thereby enriching datasets that lack sufficient diversity in camera orientation, focal length, principal point, or distortion.

Generating images is challenging, as it typically relies on rendering, which requires at least the lighting, 3D scene surface, reflection function, and camera parameters. This complex process is incompatible with the typical practical conditions in which most

of these parameters are unavailable. In camera augmentation, we bypass this complex procedure by simply keeping the original camera position unchanged, which nonetheless leaves a variety of other geometric controls, subsuming conventional augmentation. These new controls allow the generation of images that preserve 3D-geometric consistency without requiring complex parameters and computations, yet still improve model robustness to changes in camera geometry.

We model the original camera using its intrinsic parameters $\mathbf{K} \in \mathbb{R}^{3\times3}$ and its distortion parameters $\boldsymbol{\theta} \in \mathbb{R}^k$. Choosing as 3D coordinate frame $\mathcal{C}$ the standard coordinate frame related to this camera (origin centred on the projection centre, Z axis coincident with the optical axis, and X, Y axes parallel to the retinal plane axes), the camera pose is the identity rotation $\mathbf{I}$ and zero translation $\mathbf{0}$. We model the augmented image camera using its intrinsics $\mathbf{K'} \in \mathbb{R}^{3\times3}$, its distortion parameters $\boldsymbol{\theta'} \in \mathbb{R}^k$ and its pose in $\mathcal{C}$ as the rotation $\mathbf{R}$. We leave, as already stated, the translation to $\mathbf{0}$ in order to keep the camera position unchanged.

Without distortion, for $\mathbf{P}$ the Euclidean coordinates of a 3D point, and $\tilde{\mathbf{P}}$ its homogeneous coordinates, the projections are $\tilde{\mathbf{p}} \sim \mathbf{K}\begin{bmatrix}\mathbf{I}\ \mathbf{0}\end{bmatrix}\tilde{\mathbf{P}} \sim \mathbf{KP}$ and $\tilde{\mathbf{p}}' \sim \mathbf{K'}\begin{bmatrix}\mathbf{R}\ \mathbf{0}\end{bmatrix}\tilde{\mathbf{P}} \sim \mathbf{K'RP}$, where '$\sim$' means equality up to scale. By inverting the first projection and substituting into the second one, we obtain the well-known homographic relationship $\tilde{\mathbf{p}}' \sim \mathbf{H}\tilde{\mathbf{p}}$, with $\mathbf{H} \sim \mathbf{K'RK}^{-1}$. We can factor the homography as $\mathbf{H} = \mathbf{K}_{\text{rel}}\mathbf{H}_{\text{R}}$ where $\mathbf{K}_{\text{rel}} = \mathbf{K'K}^{-1}$ and $\mathbf{R}$ represent the sought *relative intrinsic* and *relative rotation*, and $\mathbf{H}_{\text{R}} = \mathbf{KRK}^{-1}$. Using the focal length ratios $s_x = f'_x/f_x$ and $s_y = f'_y/f_y$, we have:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix},\ \mathbf{K'} = \begin{bmatrix} f'_x & 0 & c'_x \\ 0 & f'_y & c'_y \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{K}_{\text{rel}} = \begin{bmatrix} s_x & 0 & c'_x - s_x c_x \\ 0 & s_y & c'_y - s_y c_y \\ 0 & 0 & 1 \end{bmatrix}.$$

Therefore, $\mathbf{K}_{\text{rel}}$ is independent of the absolute focal lengths. Conventional image augmentations, including random rotation, scaling, and translation within the image plane, can be interpreted as special cases of the proposed camera augmentation.

We now add lens distortion, for which we define the generic distortion and undistortion functiond $\mathcal{D}_{\mathbf{K},\boldsymbol{\theta}}(\cdot)$ and $\mathcal{D}^{-1}_{\mathbf{K},\boldsymbol{\theta}}(\cdot)$. These functions depend on the intrinsic parameters in $\mathbf{K}$ and the distortion coefficients $\boldsymbol{\theta}$, which typically model radial and tangential distortions. For the perspective function $\Pi([q_1, q_2, q_3]^\top) = [q_1/q_3, q_2/q_3]^\top$ and the homogenisation function $\Gamma([q_1, q_2]^\top) = [q_1, q_2, 1]^\top$, the complete transformation is:

$$\mathbf{p}' \sim \mathcal{D}_{\mathbf{K}',\boldsymbol{\theta}'}\left(\Pi\left(\mathbf{K}_{\text{rel}}\mathbf{H}_{\text{R}}\Gamma\left(\mathcal{D}^{-1}_{\mathbf{K},\boldsymbol{\theta}}\left(\mathbf{p}\right)\right)\right)\right). \tag{1}$$

In performing camera augmentation, $\mathbf{K}_{\text{rel}}$, $\mathbf{R}$ and $\boldsymbol{\theta'}$ are drawn from probabilistic distributions that model realistic variations in camera parameters in the target domain. These distributions act as a bridge between the known camera configuration in the available dataset and the perturbed configuration in the augmented setup, allowing camera augmentation to be performed without re-rendering. The use with stereo laparoscopes is discussed in the next sections, where we measure that lens distortion is negligible for the public datasets at hand. We thus use the following *distortion-free simplification* of Eq. 1:

$$\mathbf{p}' \sim \mathbf{K}_{\text{rel}}\mathbf{H}_{\text{R}}\mathbf{p}. \tag{2}$$

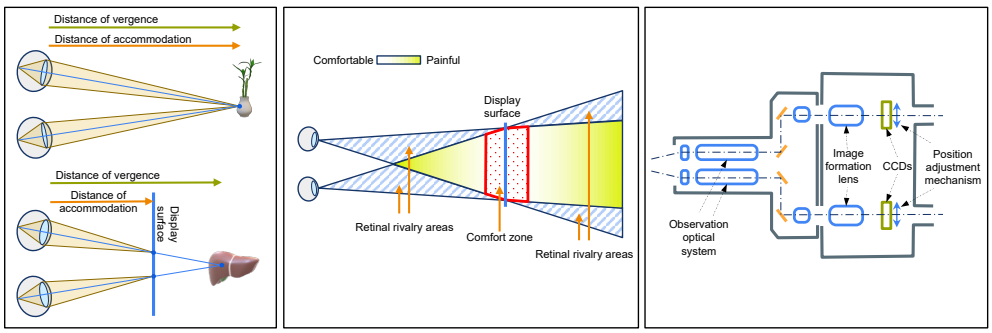

Fig. 2 Left: Vergence–accommodation conflict (VAC); In natural viewing, vergence and accommodation coincide (top). In stereo displays, vergence targets the virtual object while accommodation stays on the display plane (bottom). Middle: Combined effects of binocular rivalry and vergence–accommodation conflict, highlighting the range where both are minimised. Right: CCD position adjustment mechanism in a stereo laparoscope bringing the virtual image into the visual comfort zone.

3 Perception and the Design of Stereo Laparoscopes

We analyse the perceptual constraints and the design of stereo laparoscopes. This establishes a basis for modelling variations in intrinsic and extrinsic parameters. The design choices of stereo laparoscopes given below are general and may not be universally followed by all manufacturers.

Overview

Depth perception in humans may be enabled by displaying stereo images, should some perceptual constraints be satisfied. A stereo laparoscope is a dual-channel imaging system that captures synchronised image pairs for such displays. Its base components, shown in Fig. 2, are two parallel optical pathways with relay lenses, fibre-based illumination, and twin sensors. The sensors are factory-calibrated to provide left and right views for depth perception via a binocular eyepiece or 3D display. The surgeon's visual discomfort, including fatigue and eye strain, and quality of depth perception, directly depend on the extent to which the stereo imaging pipeline's design satisfies the perceptual constraints [13]. These constraints and design choices can be gathered in two groups, which we describe next, along with an experimental verification.

Geometric Alignment

Depth perception requires precise alignment of the left and right images. In an ideal stereo configuration, correspondences exhibit purely horizontal displacements. In computer vision, this alignment is also referred to as rectified images, whose epipolar lines are horizontal. The careful design and factory-based placement adjustment of the imaging lenses and sensors allow modern stereoscopic laparoscopes to produce near-rectified image pairs. We quantified the residual misalignment experimentally from the camera parameters given in the public datasets SCARED [14], StereoMIS [15] and RIS2017 [16], which were acquired using da Vinci systems. We used the extrinsic parameters to estimate three rotations around each camera's optical centre. These

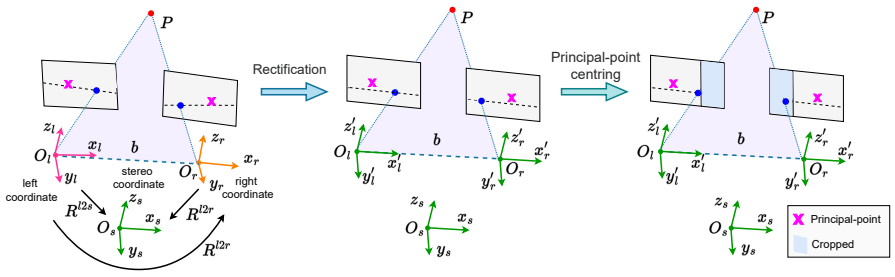

Fig. 3 Geometry of stereo camera pairs in unrectified (left), rectified (middle), and principal-point aligned (right) configurations.The principal point shift and scan-line misalignment are intentionally exaggerated for visualisation purposes. In practice, the applied perturbations are much smaller and preserve sufficient overlap between the two views.

rotations are computed relative to the stereo coordinate system shown in Fig. 3, defined with the X-axis aligned with the baseline and taken as minimal relatively to the left camera. Statistics -mean and standard deviation- for the sampled angles are given in Table 1. Their concentration around zero, all being much lower than 1 degree, and similarity across datasets indicate that different laparoscopes share a comparable small deviation from the rectified design. As indicated by the small Y-axis rotations in Table 1, the display planes follow a planar design without toeing-in. We used the intrinsic parameters to quantify lens distortion, which is known to impair rectification. We measured a residual distortion consistently below half a pixel; given the image resolution of 1280×1024, this confirms a near distortion-free design. Lastly, the principal point for both cameras must be aligned; this is discussed below.

Vergence-Accommodation Conflict and Binocular Rivalry

Vergence rotates the eyes to enable single vision, while accommodation adjusts focus. In natural conditions, both occur at the same distance. In stereoscopic displays however, VAC arises from vergence targetting the virtual object while accommodation remains fixed on the display plane. Binocular rivalry occurs when the left and right images contain inconsistent information, often due to asymmetric fields of view. Its extent increases as the virtual object approaches the viewer. Fig. 2 shows a depth range where the effects of both VAC and binocular rivalry are minimised. Virtual anatomical structures are typically rendered slightly in front of or behind the display plane to optimise depth perception and comfort. The way these two perceptual constraints are coped with in stereo laparoscopes is by shifting the sensors during manufacturer calibration, as documented in patented systems [17–19]. This shift is chosen to place maximal comfort at the typical surgical depth of 5 cm. In terms of the camera parameters, the effect is to offset the principal points horizontally away from the image centre. This creates a camera-induced domain gap between stereo laparoscope images and conventional stereo images, where principal points are centred. A direct effect is that the stereo disparity may be positive and negative. In contrast, state-of-the-art stereo matching methods are typically trained on datasets with only positive disparities, such as SceneFlow [20] and FSD [7], whereas real laparoscopic datasets like SCARED [14] contain both positive and negative disparities. Fig. 4 illustrates this gap.

Table 1 Statistics on camera pose deviations from the ideal rectified configuration for the left (top row) and right (bottom row) cameras, for the three Euler rotation angles, in degrees.

	R_x^{l2s}		R_y^{l2s}		R_z^{l2s}		R_x^{r2s}		R_y^{r2s}		R_z^{r2s}	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
SCARED	0.0000	0.0001	-0.0325	0.0129	0.0904	0.3597	-0.0014	0.0053	0.0013	0.0266	0.0910	0.3602
StereoMIS	-0.0006	0.0006	0.3673	0.0032	-0.1907	0.1812	-0.0020	0.0008	0.3799	0.0187	-0.1887	0.1809
RIS2017	-0.0004	0.0014	-0.3386	0.1652	0.2637	0.4680	0.0002	0.0076	-0.2744	0.1640	0.2639	0.4681
SCARED+StMIS	-0.0001	0.0004	0.0564	0.1666	0.0280	0.3487	-0.0016	0.0047	0.0854	0.1594	0.0289	0.3489

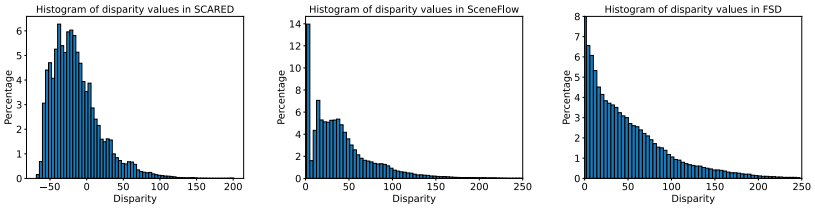

Fig. 4 Disparity distributions (in px) for the SCARED, SceneFlow, and FSD datasets.

4 Camera Augmentation in Stereo Laparoscopes

In classical stereo vision, disparity is defined for rectified image pairs. Surgical laparoscopes produce near-rectified images, as described in Section 3, where vertical displacements are small and horizontal displacements can be used as estimates for disparity. Disparity is related to depth through $d = \frac{fb}{z} + (c_x^L - c_x^R)$, where f is the focal length, b is the baseline, and c_x^L and c_x^R are the x-coordinates of the left and right principal points. This relationship shows that depth in uncalibrated settings is only available *relatively*, up to an unknown scale and shift.

We use camera augmentation from Section 2 to address uncalibrated stereo matching in MIS. The proposed method, termed Camera-Augmentation Training Strategy (CATS), is used to retrain existing stereo matching architectures on camera-augmented datasets that simulate realistic imaging imperfections. We retrained RAFTStereo and IGEV++ using the original training protocol, hyperparameters, and datasets, with as sole modification the use of CATS. Both models are trained on the SceneFlow dataset [20], which contains over 39K stereo pairs with a resolution of 960×540 pixels, rendered from a wide range of 3D scenes. The SceneFlow dataset uses ideal stereo camera geometry, including perfect rectification, centred principal points, and the absence of lens distortion, differing markedly from the real MIS conditions.

We use the analysis of MIS stereo camera imperfections from Section 3 (deviations in rectification and principal point misalignment), to generate camera augmentations. This augmentation strategy encourages the network to learn a stereo matching process adapted to uncalibrated MIS images. Concretely, for each of the left and right cameras, we sample three Euler angles, ϕ, θ, and ψ, from Gaussian distributions reflecting the

measured perturbation statistics, shown in Table 1:

$$\phi_c \sim \mathcal{N}(\mu_{c,\phi}, \sigma^2_{c,\phi}), \quad \theta_c \sim \mathcal{N}(\mu_{c,\theta}, \sigma^2_{c,\theta}), \quad \psi_c \sim \mathcal{N}(\mu_{c,\psi}, \sigma^2_{c,\psi}), \quad c \in \{\text{left}, \text{right}\}. \tag{3}$$

The rotation matrix $\mathbf{R}$ is constructed from the Euler angles (ϕ, θ, ψ) following the Z, Y, X (yaw–pitch–roll) convention as $\mathbf{R}_c = \mathbf{R}_z(\psi_c)\,\mathbf{R}_y(\theta_c)\,\mathbf{R}_x(\phi_c)$, $c \in \{\text{left}, \text{right}\}$. The corresponding homography induced by the rotational perturbation for each camera is computed as:

$$\mathbf{H}_{\text{R},c} = \mathbf{K}\,\mathbf{R}_c\,\mathbf{K}^{-1}, \quad \forall c \in \{\text{left}, \text{right}\}, \tag{4}$$

where $\mathbf{K}$ is the intrinsic calibration matrix from the original dataset. We then sample $\mathbf{K}_{\text{rel}}$ for the left and right cameras. Since random scaling is already included in the standard data augmentation pipelines of the reference methods, additional variation in scaling factors, s_x and s_y, would be redundant, and we use $s_x = s_y = 1$ for both cameras. Furthermore, based on observations from Section 3, we use $c'_y = c_y$ for both cameras, as the vertical principal point displacement was found to be negligible. The dominant parameter affecting $\mathbf{K}_{\text{rel}}$ is the horizontal offset between the principal points of the left and right cameras. Accordingly, we set $c'_{x,\text{left}} = c_{x,\text{left}}$ and sample the right camera offset, $\Delta x_{\text{right}} = c'_{x,\text{right}} - c_{x,\text{right}}$ from a uniform distribution estimated from the empirical distribution given in Fig. 4 as $\Delta x_{\text{right}} \sim \mathcal{U}(-100, 0)$. The uniform distribution is a pragmatic choice ensuring coverage of the observed empirical range and keeping the sampling simple. The use of non-parametric distributions would greatly increase complexity. This distribution strikes a balance between positive and negative disparities during training, improving robustness across varying stereo configurations. We arrive at the following sampled relative intrinsic matrices:

$$\mathbf{K}_{\text{rel,left}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{K}_{\text{rel,right}} = \begin{bmatrix} 1 & 0 & \Delta x_{\text{right}} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

Finally, according to observations from Section 3, the effects of lens distortion are negligible within the analysed MIS datasets; hence, distortion is omitted in CATS. Importantly, we apply the same homography as for the left image to the disparity ground truth and then adjust all disparity values by adding Δx_{right}.

5 Experiments and Results

Reference models. We selected three stereo matching models representing the SoTA: RAFTStereo [5], IGEV++ [6], and FoundationStereo [7] (we used the best performing version for general use, based on ViT-Large), which we used as released, without additional training or fine-tuning. These architectures represent the leading advances in general-purpose stereo matching, independently of our extending them to uncalibrated MIS. For comparison with models specifically designed for MIS, we included MSDESIS [8] and MCF-SMSIS [9], representing the SoTA in MIS stereo matching. We used RAFT [21] as a representative optical flow method, taking the vertical component of the predicted optical flow as an estimate of disparity. Finally, we used DUSt3R [22] as a representative general multi-view reconstruction method.

Table 2 Uncalibrated scenario evaluation over original images from the SCARED dataset. Depth values are computed from predicted disparity using static calibration parameters for reference only. In the DUSt3R method, the calibration parameters are used to find disparity values from the predicted depth.

	Disparity metrics (px)		Depth metrics (mm)			
	EPE ↓	Bad 3.0 ↓	MAE ↓	RMSE ↓	AbsRel ↓	$\delta_{0.25}$ ↑
RAFTStereo	244.37	80.08	43.95	48.69	0.55	0.23
IGEV++	186.49	79.79	39.75	46.03	0.49	0.24
FoundationStereo	331.41	83.60	48.50	52.05	0.62	0.20
MSDESIS	88.25	99.96	38.06	40.42	0.53	0.001
MCF-SMSIS	119.01	97.78	43.38	46.51	0.59	0.03
DUSt3R (GT calib.)	67.81	100	232.78	233.30	4.17	0.00
DUSt3R (DUSt3R calib.)	687.97	100	232.78	233.30	4.17	0.00
RAFT	3.00	27.79	2.85	4.98	0.039	0.81
CATS-RAFTStereo	1.42	6.36	**1.33**	**2.25**	**0.01**	**0.97**
CATS-IGEV++	**1.41**	**6.32**	**1.33**	2.31	**0.01**	**0.97**

CATS models. We retrained RAFTStereo and IGEV++ using the proposed CATS, yielding the CATS-RAFTStereo and CATS-IGEV++ models. We did not retrain FoundationStereo as its training code is not yet publicly available; however CATS would be fully compatible with its pipeline since it operates purely at the data level.

Experimental setup. We evaluated quantitatively on the SCARED dataset and qualitatively on the SCARED, StereoMIS and an in-house datasets. The in-house dataset was collected from hepatectomy procedures with ethical approval IRB00008526-2019-CE58 issued by CPP Sud-Est VI in Clermont-Ferrand, France. In quantitative evaluation, we used keyframes of the SCARED dataset, excluding video sequences that exhibit inconsistencies between the visual data and the kinematic ground truth [14]. In addition, we excluded datasets 4 and 5, as they are explicitly reported in the SCARED paper to contain errors [14]. In the calibrated scenario, we used the available intrinsic and extrinsic parameters to rectify the stereo pairs and centre the principal points before applying the stereo matching models. In the uncalibrated scenario, we applied stereo matching directly on the original stereo image pairs. We evaluated quantitatively using the End-Point Error (EPE) and Bad-3.0 as disparity metrics, and MAE, RMSE, AbsRel, and $\delta_{0.25}$ as depth metrics [23].

Results. Quantitative results for the uncalibrated and calibrated scenarios are given in Tables 2 and 3. Several key observations can be made. First, as shown in Table 2, all reference and MIS-specific models exhibit systematic failures in the uncalibrated setting, with large disparity and depth errors. RAFTStereo, IGEV++, and FoundationStereo yield EPEs exceeding 180 pixels, and MIS-specific models, MSDESIS and MCF-SMSIS, perform only marginally better with EPEs of 88.25 pixels and 119.01 pixels, respectively, still considered failures. DUSt3R exhibits an EPE of 67.81 pixels when using the camera parameters provided by the dataset, and 687.97 pixels when using its own estimated parameters, both of which are far from acceptable. RAFT achieves a substantially lower EPE of 3.00 pixels, which is nonetheless more than twice as large as the CATS methods. The depth metrics confirm the failures, with the lowest MAE value

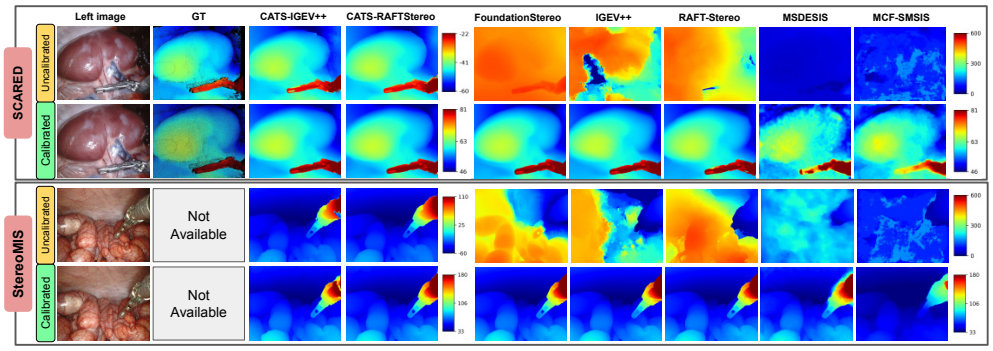

Fig. 5 Qualitative model comparison in the uncalibrated and calibrated scenarios on samples from the SCARED and StereoMIS datasets.

at 38 mm, indicating poor geometric consistency. Second, the CATS-trained models perform successfully on the same uncalibrated scenario, achieving EPEs of 1.42 pixels for CATS-RAFTStereo and 1.41 pixels for CATS-IGEV++. Depth metrics further confirm this improvement with an MAE of about 1.33 mm and an AbsRel of about 0.01, demonstrating that CATS successfully enables reliable stereo matching directly from uncalibrated MIS stereo pairs. Third, the performance of CATS-trained models in the uncalibrated scenario (EPEs of 1.42 and 1.41 pixels for CATS-RAFTStereo and CATS-IGEV++), when compared to the best-performing method in the calibrated setting in Table 3 (EPE of 1.17 pixels for FoundationStereo), underscores the effectiveness of CATS. Remarkably, CATS-trained models achieve nearly equivalent performance despite the absence of camera calibration. Fourth, under the calibrated setting, the same models achieve comparable performance. CATS-RAFTStereo and CATS-IGEV++ reach EPEs of 1.23 and 1.21 pixels, which are on par with their reference counterparts, with RAFTStereo at 1.21 pixels and IGEV++ at 1.23 pixels. This shows that CATS does not compromise performance when calibration is available. Fifth, in the calibrated scenario, CATS-RAFTStereo and CATS-IGEV++ perform nearly as well as FoundationStereo, the best-performing model, with EPEs marginally higher by 0.04 and 0.06 pixels.

Qualitative results in Fig. 5 and 6 illustrate these findings. In the uncalibrated scenario, reference and MIS-specific models produced distorted and inconsistent disparity maps, whereas the proposed CATS models successfully recovered fine structural details and coherent patterns. In the calibrated scenario, all models performed well, including the CATS variants, confirming that CATS remains fully compatible with calibrated conditions and does not degrade performance. Furthermore, the CATS models demonstrated consistent performance across all tested datasets acquired with different da Vinci system models, highlighting their generalisability. Beyond its strong performance on the three public datasets collected on animal models, CATS produced high-quality, structurally coherent disparity maps on the in-house data of human hepatectomy procedures. This demonstrates the method's robustness and translational potential, supporting its use in real-world clinical applications such as computer-assisted interventions and intraoperative guidance.

Table 3 Calibrated scenario evaluation on undistorted, rectified, and principal point–aligned images from the SCARED dataset.

	Disparity metrics (px)		Depth metrics (mm)			
	EPE ↓	Bad 3.0 ↓	MAE ↓	RMSE ↓	AbsRel ↓	$\delta_{0.25}$ ↑
RAFTStereo	<u>1.21</u>	5.67	<u>1.01</u>	<u>2.32</u>	**0.014**	0.979
IGEV++	1.23	5.56	1.03	2.43	<u>0.015</u>	<u>0.980</u>
FoundationStereo	**1.17**	**5.10**	**0.97**	**2.30**	**0.014**	**0.981**
MSDESIS	1.71	9.71	1.39	2.91	0.0200	0.946
MCF-SMSIS	2.08	11.24	1.49	3.16	0.022	0.939
DUSt3R (GT calib.)	71.82	100	231.10	231.64	4.08	0.00
DUSt3R (DUSt3R calib.)	81.63	100	231.10	231.64	4.08	0.00
RAFT	3.68	26.39	2.86	5.60	0.039	0.83
CATS-RAFTStereo	1.23	5.80	1.07	2.59	0.016	0.978
CATS-IGEV++	<u>1.21</u>	<u>5.52</u>	<u>1.01</u>	2.37	<u>0.015</u>	<u>0.980</u>

6 Conclusions

Camera augmentation bridges the gap between ideally conditioned synthetic datasets and the real surgical conditions where calibration is uncertain or unavailable. By introducing controlled geometric perturbations of camera parameters during training, it enables the adaptation of large-scale stereo matching models to the specific challenges and conditions of MIS. The proposed CATS demonstrates that robust and accurate stereo matching can be achieved even in the absence of calibration, while maintaining competitive performance when calibration is available. This capability stems from the in-depth analysis and modelling of stereo laparoscopes, allowing CATS to be rigorously tailored to the specific geometric variability encountered in surgical imaging. Beyond stereo matching, the concept of camera-aware image augmentation opens new perspectives for a broad range of computer vision tasks sensitive to camera geometry. By decoupling model performance from calibration constraints, this approach promotes the development of deep models that are more flexible, generalisable, and clinically deployable. As a direction for future work, we plan to test the CATS framework on a larger set of stereo surgical cameras, and to incorporate optical distortion models.

References

[1] Malhotra, S., Halabi, O., Dakua, S.P., Padhan, J., Paul, S., Palliyali, W.: Augmented reality in surgical navigation: a review of evaluation and validation metrics. Applied Sciences **13**(3), 1629 (2023)

[2] Barcali, E., Iadanza, E., Manetti, L., Francia, P., Nardi, C., Bocchi, L.: Augmented reality in surgery: a scoping review. Applied Sciences **12**(14), 6890 (2022)

[3] Xu, M., Guo, Z., Wang, A., Bai, L., Ren, H.: A review of 3D reconstruction techniques for deformable tissues in robotic surgery. In: MICCAI, pp. 157–167

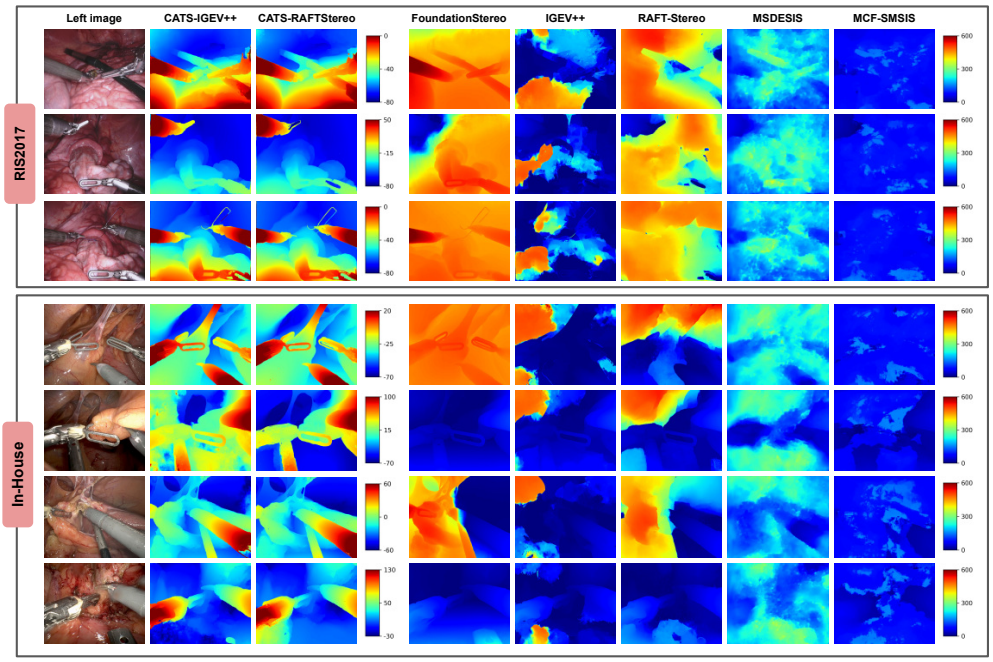

Fig. 6 Qualitative model comparison in the uncalibrated scenario on samples from the RIS2017 and in-house datasets.

(2024)

[4] Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV **47**(1), 7–42 (2002)

[5] Lipson, L., Teed, Z., Deng, J.: RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In: 3DV, pp. 218–227 (2021)

[6] Xu, G., Wang, X., Zhang, Z., Cheng, J., Liao, C., Yang, X.: IGEV++: Iterative multi-range geometry encoding volumes for stereo matching. TPAMI **47**(8), 7108–7122 (2025)

[7] Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: FoundationStereo: Zero-shot stereo matching. In: CVPR, pp. 5249–5260 (2025)

[8] Psychogyios, D., Mazomenos, E., Vasconcelos, F., Stoyanov, D.: MSDESIS: Multitask stereo disparity estimation and surgical instrument segmentation. TMI **41**(11), 3218–3230 (2022)

[9] Wu, R., He, C., Liang, P., Liu, Y., Huang, Y., Liu, W., Shu, B., Xu, P., Chang, Q.: MCF-SMSIS: multi-tasking with complementary functions for stereo matching and surgical instrument segmentation. Computers in Biology and Medicine **179**,

108923 (2024)

[10] Li, L., Li, X., Ouyang, B., Mo, H., Ren, H., Yang, S.: Three-dimensional collision avoidance method for robot-assisted minimally invasive surgery. Cyborg and Bionic Systems **4**, 0042 (2023)

[11] Narasimhan, S., Turkcan, M.K., Ballo, M., Choksi, S., Filicori, F., Kostic, Z.: Monocular 3D tooltip tracking in robotic surgery—building a multi-stage pipeline. Electronics **14**(10), 2075 (2025)

[12] Preetha, C.J., Kloss, J., Wehrtmann, F.S., Sharan, L., Fan, C., Müller-Stich, B.P., Nickel, F., Engelhardt, S.: Towards augmented reality-based suturing in monocular laparoscopic training. In: Medical Imaging, p. 113150 (2020)

[13] Banks, M.S., Read, J.C., Allison, R.S., Watt, S.J.: Stereoscopy and the human visual system. Motion imaging journal **121**(4), 24–43 (2012)

[14] Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)

[15] Hayoz, M., Hahne, C., Gallardo, M., Candinas, D., Kurmann, T., Allan, M., Sznitman, R.: Learning how to robustly estimate camera pose in endoscopic videos. IJCARS **18**(7), 1185–1192 (2023)

[16] Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.-H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)

[17] Akui, N., Honma, S., Kanamori, I., Takahashi, S., Hanzawa, T., Fukaya, T., Karasawa, H., Kubota, T., Hashiguchi, T., Mochida, A.: Three-dimensional vision endoscope with position adjustment means for imaging device and visual field mask. US Patent 5,577,991 (1996)

[18] Zhao, W., Mohr, C., DiMaio, S.: Stereo imaging system with automatic disparity adjustment for displaying close range objects. US Patent 10,178,368 (2019)

[19] Breidenthal, R.S., Forkey, R.E., Smith, J., Volk, B.E.: Stereoscopic endoscope with virtual reality viewing. US Patent 6,139,490 (2000)

[20] Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR, pp. 4040–4048 (2016)

[21] Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: ECCV, pp. 402–419 (2020)

[22] Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric

3d vision made easy. In: CVPR, pp. 20697–20709 (2024)

[23] Sharifian, R., Rabbani, N., Bartoli, A.: The RoDEM benchmark: evaluating the robustness of monocular single-shot depth estimation methods in minimally-invasive surgery. IJCARS **20**(6), 1215–1229 (2025)