Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Generalisable Stereo Depth Estimation with Masked Image Modelling

Samyakh Tukra | **Haozheng Xu** | **Chi Xu** | **Stamatia Giannarou**

[1]Hamlyn Centre of Robotic Surgery, Department of Surgery and Cancer, Imperial College London, London, United Kingdom

Correspondence
Authors can be contacted via
Email: samyakh.tukra17@imperial.ac.uk, haozheng.xu19@imperial.ac.uk, chi.xu20@imperial.ac.uk, stamatia.giannarou@imperial.ac.uk

Present address
Imperial College London, Exhibition Rd, South Kensington Campus, London SW7 2AZ

Funding Information
This research was supported by the Royal Society (UF140290 and RGF\EA\180084) and the NIHR Imperial Biomedical Research Centre (BRC).

Abstract

Generalisable and accurate stereo depth estimation is vital for 3D reconstruction, especially in surgery. Supervised learning methods obtain best performance however, limited ground truth data for surgical scenes limits generalisability. Self-supervised methods don't need ground truth, but suffer from scale ambiguity and incorrect disparity prediction due to inconsistency of photometric loss. In this work, we propose a two-phase training procedure that is generalisable and retains the high performance of supervised methods. It entails: (1) Performing self-supervised representation learning of left and right views via Masked Image Modelling (MIM) to learn generalisable semantic stereo features (2) Utilising the MIM pre-trained model to learn robust depth representation via supervised learning for disparity estimation on synthetic data only. To improve stereo representations learnt via MIM, we introduce perceptual loss terms, which improve the model's stereo representations learnt by explicitly encouraging the learning of higher scene-level features. Qualitative and quantitative performance evaluation on surgical and natural scenes shows that our approach achieves sub-millimeter accuracy and lowest errors respectively, setting a new state-of-the-art. Despite not training on surgical nor natural scene data for disparity estimation.

KEYWORDS
Deep Learning; Depth; 3D; Stereo; Self-Supervised; Supervised; Masked Image Modelling; Vision Transformer

1 | INTRODUCTION

Depth estimation is of paramount importance in Minimally Invasive Surgery (MIS) facilitating, soft tissue 3D reconstruction, surgical robot navigation and augmented reality. A popular method of perceiving depth is with a stereo camera by estimating the horizontal displacement (disparity) from the left image pixels to the corresponding right. End-to-end deep learning stereo reconstruction methods have become state-of-the-art. They are typically categorised as supervised and self-supervised. Supervised learning methods directly predict disparity (regression) by training on ground truth disparity data[1 2 3 4]. They usually undergo training using synthetic data[5], followed by fine-tuning on real-world scenes to overcome the limited availability of ground truth. Despite their popularity, they face several challenges in surgical applications. Discrepancy between training and inference data, limited ground truth data, lower image quality, restricted space, and a multitude of scene variations, are factors that lead to reduced performance.

Self-supervised methods don't require ground truth disparity, making images-only training data abundant. These methods optimise photometric re-projection error through novel view synthesis.[6 7 8 9 10]. Self-supervised methods show potential, but are not as effective as supervised methods. This is because the photometric error can be optimised for a large range of disparity values, resulting in inconsistent geometry. Furthermore, the relative depth information between stereo images is inherently scale ambiguous. Making it challenging to learn robust representation of depth via self-supervision, especially in scenes with occlusions, textureless regions, or repetitive structures, which is prevalent in surgery. Supervised learning is key for robust stereo depth, since training on ground truth provides a strong, unambiguous signal for learning accurate representations. A question

Abbreviations: StereoMAE, Stereo Masked Autoencoder; MIM, Masked Image Modelling; MIS, Minimally Invasive Surgery.

emerges though; can we have the best of both worlds i.e. unambiguous depth that is also generalisable to different scenes?
For achieving generalisability, visual representation learning is key. Learning robust feature representations also allows improved downstream performance as shown in [11 12 13]. These techniques utilise image reconstruction as a preliminary task, relying on the principle that by learning to predict patches in a masked image, useful representations about the context of the scene can be obtained for further tasks. This is proven by the enhanced data label efficiency on various standard benchmark datasets. Hence, in this work we experiment with this masked image modelling (MIM) approach, to train an encoder model to learn stereo feature representations, that can be fine-tuned downstream to encode robust depth feature representations. This endows our model with both generalisable features via MIM and unambigous sharp depth via supervised-learning for disparity estimation.
In this paper we build on [11] and propose *StereoMAE* a two-stage training process which entails (1) training an encoder via MIM, to generate robust feature representations for left and right views, followed by (2) supervised training for disparity estimation. Furthermore, we enhance MIM in (1) by using perceptual similarity learning [14], leading to learned representations that effectively capture the intricacies of the scene and object boundaries without explicit guidance or manually designed inductive biases. Our contributions are the following:

- A novel approach for training stereo depth models by combining self-supervised MIM and supervised stereo depth estimation. To the best of our knowledge, this is the first work to apply MIM for stereo depth estimation;
- We propose a new approach to boost MIM by incorporating perceptual similarity loss term for learning generalisable visual semantic concepts;
- We present a modular model architecture for combining any pre-trained MIM encoder model with any off-the-shelf decoder to enhance depth estimation;
- Our joint MIM-supervised approach enhances performance, yielding a generalisable model with sub-millimeter accuracy in surgical depth estimation

Specific to the field of surgery the lack of datasets with ground truth, restricts state-of-the-art stereo models from reaching sub-mm precision on surgical scenes despite their generalisability. This is due to the need for deep understanding of scene semantics, a task beyond the capabilities of supervised learning alone. Hence, our method pairs Masked Image Modelling (MIM) with StereoMAE model architecture, and perceptual learning to surmount the limited training data issue, achieving high generalisability and sub-mm accuracy in surgery. This blend of generalisability and precision offers a significant advantage, expanding its potential use across diverse medical imaging tasks, not just specific surgical scenarios like Laparoscopic surgery where depth estimation at sub-millimeter accuracy is critical as it helps the localization and perception of surgical instruments, tumour and surrounding healthy tissue (essential in the development of autonomous task execution in Robotic MIS). The performance of StereoMAE has been evaluated on (3) MIS datasets; *SCARED* [15], *Hamlyn* [16] and *SERV-CT* [17], and on (2) non-surgical scenes data; *ETH3D* [18] and *Middlebury* [19].

2 | METHODS

The method is divided into two phases; (1) Pre-training via MIM for enhanced representation learning followed by (2) supervised downstream fine-tuning for stereo depth estimation. The learning framework for (1) is inspired by MAE [11]. Supervised training in (2) is based on RAFT-Stereo [1]. However for (2) any training methodology from the supervised stereo depth literature can be used.

(1) Stereo Masked Image Modelling: MIM involves randomly masking pixels of an input image and training a model to predict the invisible content [11 12 13]. The intuition is that the model learns the representation of the masked pixels by inferring details from the surrounding valid pixels. In the end, if a model is capable of reconstructing the missing content with valid pixels, the model has learnt the context within the image by encoding its semantic features.

We adapt the method in [11], which utilises an auto-encoder model for stereo images as shown in Figure 1. In particular, we construct the auto-encoder model, using vision transformer (ViT) components [20]. The encoder consists of several layers of ViT-Base (ViT-B), a 12 layer transformer model. The custom decoder is composed of 8 transformer layers. Given the left image as input, it is first resized to 224x448, divided into patches of size 16x32 and 75% of them are randomly masked. The encoder is fed only the un-masked patches as input, and it generates corresponding features representing the visible scene parts. These features are then concatenated with the patches that were not fed into the encoder and inputted into the decoder to reconstruct the full left image as output. The model is trained via pixel image reconstruction (photometric error). The same is done for the

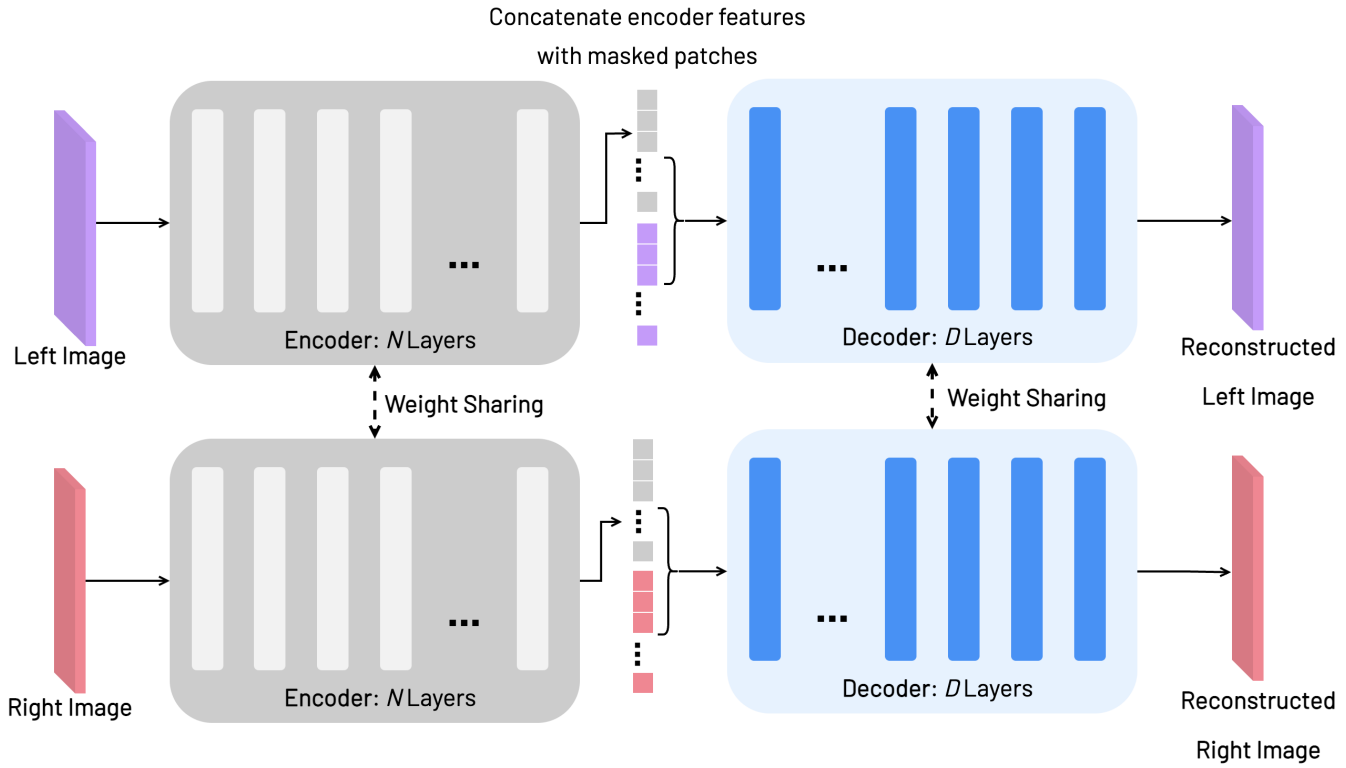

FIGURE 1 Masked Image Modelling (MIM) pipeline for StereoMAE comprising of a weight sharing transformer encoder-decoder model. Where N and D are the total number of transformer layers that make up the corresponding models.

right input image. For the left and right input views, the weights of the encoder and decoder are shared, allowing them to learn joint global stereo representations for reconstructing the scene.

In our work, we do not train the aforementioned model using the mean square error (MSE) loss from the original work[11]. MSE, which compares pixel intensities, is a low-level metric and insufficient for capturing complex structures in an image as it assumes pixel-wise independence. Hence a model's capacity is wasted since high frequency global feature distribution will not be captured. Perceptual similarity, which mimics human visual perception[21], is a better metric for evaluating image similarity and capturing high-level semantic features, essential for model generalization. Hence, we hypothesise, that training MIM with perceptual similarity will enhance the models' output quality and coherence. The focus is on learning high-level relationships between elements in the image, not just low-level details. Thereby reducing the model's vulnerability to masked regions and improving overall semantic content and structure capture.

Therefore, to train StereoMAE we utilise a perceptual loss for MIM for learning the required high-level semantic features that represent the scene context (thereby improving its downstream generalisability). Inspired by[14] our perceptual loss comprises of three loss terms: (i) L1 (also known as absolute error loss, measures the absolute difference between the predicted and actual pixel values, calculated as the sum of the absolute differences between the two values). (ii) Feature matching (compares intermediate features from a pre-trained model like VGG, to encourage learning similar features for the predicted and actual pixel values). Lastly, (iii) style transfer loss (computes correlations between feature maps extracted from a model like VGG, encouraging the predicted and actual image distributions match). L1 loss encourages sparse representations by optimizing absolute intensity values. The feature matching loss compares the feature representations of the reconstructed output and target image, while the style loss measures the difference in statistical distribution of high-level feature activations, capturing image texture and aesthetics.

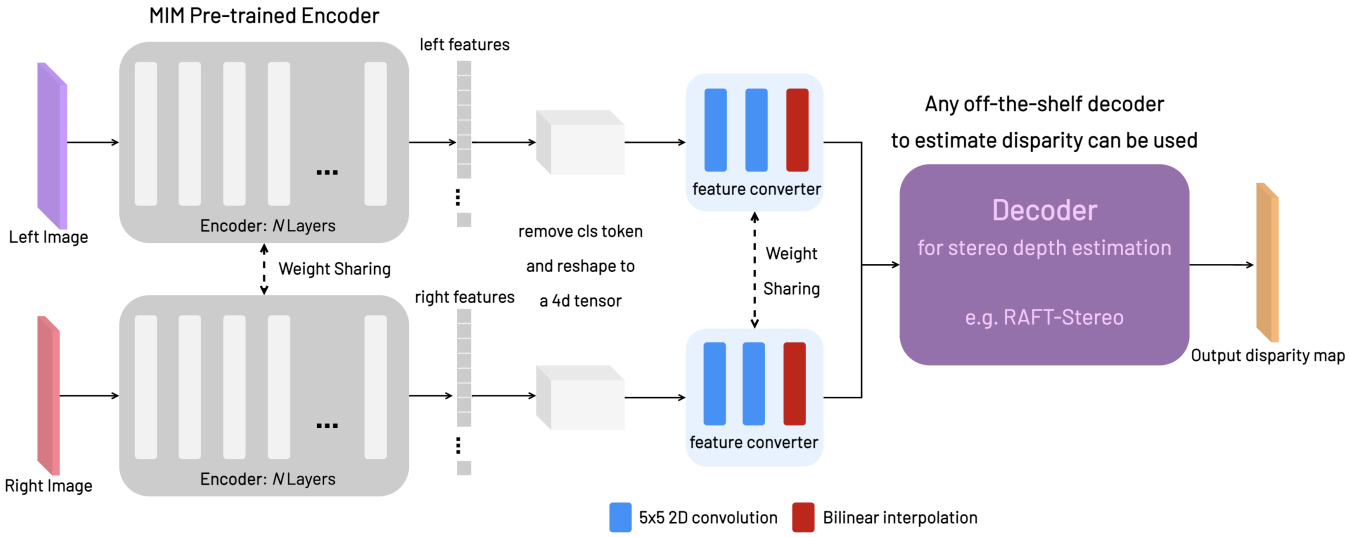

FIGURE 2 The downstream process of supervised stereo depth estimation training. Where the feature extractor encoders pre-trained from MIM, are packaged with an off-the-shelf decoder for depth estimation. Any decoder of user's choice can be used.

To calculate (ii) and (iii) we utilise a pre-trained feature extractor model i.e. VGG16 (though any arbitrary model can be used).

$$L^{I}_{mim} = \|G(I_m) - I\|_1 + \delta_f \sum_{j=1}^{J} \frac{1}{N_j} [\|\phi^j(G(I_m)) - \phi^j(I)\|_1] \\ +\delta_s \sum_{j=1}^{J} \frac{1}{N_j} [\|\Psi(\phi^j(G(I_m))) - \Psi(\phi^j(I))\|_1] \tag{1}$$

where, I is the original input image, I_m is the masked input image, G is the StereoMAE model, ϕ is the feature extractor model and Ψ is the gram matrix. The loss weights δ_f and δ_s are 0.05 and 40.0, respectively. These weightings were selected via experimentation on our dataset (discussed in Implementation Details section). The initial values were inspired by previous research in generative modelling[22,23]. L_{mim} is calculated for both the left and right views, making the full MIM loss for StereoMAE the following:

$$L_{mim} = L^{I_L}_{mim} + L^{I_R}_{mim} \tag{2}$$

where, I_L and I_R are the left and right input views, respectively. MIM training was performed on a mixture of real natural and synthetic scenes (no surgical scenes were used).

(2) Supervised Downstream Fine-tuning: Once the encoders have been pre-trained via MIM in phase (1) we finetune them for disparity estimation. The full model architecture for downstream training is displayed in Figure 2. The architecture is designed to be modular such that the pre-trained encoders can be combined with any off-the-shelf decoder of the user's choice. Thereby enabling a smooth transition from MIM to disparity estimation without further modifications. The pre-trained ViT encoders generate features for the (un-masked) left and right input images, and output a 3D tensor. The features have to be reshaped into a 4D tensor for it to be processed by the decoders which are typically composed of convolution layers. However, once reshaped, the positional encodings of the features change. To ensure the decoder focuses on the necessary elements of the feature map, we parse the feature map through a *Feature Converter* block; its purpose is to learn the transfer of MIM-trained features to stereo-disparity features and select the relevant elements needed for the decoder. It comprises of two 5x5 2D convolution layers and a bilinear interpolation module that resizes the tensor to a scale of choice. The scale depends on the type of decoder used. In our experiments we utilise the RAFT-Stereo decoder[1]. The *Feature Converter* block essentially plays a crucial role in ensuring the appropriate transfer and transformation of features from the pre-trained ViT encoder to the disparity decoder. In our experiments this is vital as the encoder is trained only for the MIM task, hence the features must be adapted for disparity downstream.

To train the final model we utilised a supervised-learning loss comprised of the L1 error between the predicted and ground truth disparity. The RAFT-Stereo decoder generates a sequence of predictions (from its multi-level GRU networks), hence our L1 error is computed over the full sequence of predictions, $[d^1_p, ..., d^N_p]$, with exponentially increasing weights. If a different

decoder that outputs a single disparity map is generated then, only the final single output d_p would have been used for the loss calculation. Considering ground truth disparity is defined as d_{gt}, the loss is defined as

$$L_{sup} = \sum_{i=1}^{Y} \gamma^{Y-i} \| d_{gt} - d_p^i \|_1 \tag{3}$$

where, $\gamma = 0.9$, is the weighting factor for the loss calculated for each scale and Y is the total number of scales. The supervised training for disparity estimation was only conducted on synthetic scenes (no surgical or natural scenes were used).

Implementation Details: For MIM training, we use the ViT-B architecture [11], trained for 150 epochs on a combination of the following datasets [24 25 26 27 28] (a total of 411,942 stereo image pairs). The input patch size is fixed to 16x32 and we mask 75% of input patches during training. The ViT-B encoder, comprises of 12 transformer layers (N), each with 12 self-attention heads and the final hidden dimension output is of size 768. The decoder architecture comprises of 8 transformer layers (D) each with 16 self-attention heads and the hidden dimension output of size 512. The data augmentation strategies include (1) resizing / cropping image to 224x448, (2) adding random distracting shapes like rectangles of varying size and (3) random colour jittering. We train with a batch size of 8 on 4-Nvidia Tesla GPUs. Adam optimizer with a learning rate of 0.00015, and weight decay of 0.05 (cosine strategy), 40 warm-up epochs and the momentum parameters β_1 and β_2 are 0.9 and 0.95 respectively were used. For downstream training, the pretrained ViT-B model is used as feature extractor with the RAFT-Stereo decoder architecture, for disparity estimation. The same training strategy as [1] was utilised, though since our model is modular, it can be combined with any downstream stereo-decoder model and training method. For the downstream training of disparity estimation, only the Sceneflow training split [25] was used (a combination of FlyingThings, Monkaa and Driving). Hence, only synthetic data was used for training. Note, synthetic data is used for perfect, unambiguous ground truth disparity. Synthetic surgical scenes can also be used for training, provided high quality disparity maps are available. Once, trained inference operates at 14 frames per second (fps) on a single Nvidia Tesla GPU (where frames here includes both left and right pairs).

3 | EXPERIMENTAL RESULTS AND DISCUSSION

We evaluated the performance of our model on both surgical and natural scenes, despite not training on them. The following datasets were used for testing; SCARED, Hamlyn, SERV-CT, ETH3D, and Middlebury datasets. To compare our results with other methods in the literature, we calculate the standard metrics for stereo disparity evaluation i.e. average End-Point-Error (EPE) and the percent of pixels with EPE greater than a specified threshold (0.5, 1.0 and 2.0). Specifically for SCARED and SERV-CT, we also calculate the mean absolute depth error in mm on test sets 1 and 2. We also exhibit qualitative performance of our MIM pre-training and downstream disparity estimation in Figures 3 and 4. Furthermore, due to lack of ground truth available in Hamlyn MIS data we show qualitative downstream performance comparison with other SOTA methods in Figure 5.

TABLE 1 Results of generalisation without fine-tuning on ETH3D, Middlebury, and SCARED datasets. Lower is better for all metrics. Method highlighted in bold is best.

Method	ETH3D				Middlebury 2014 (H)				SCARED Small	
	EPE (pixel)	0.5 px (%)	1 px (%)	2 px (%)	EPE (pixel)	0.5 px (%)	1 px (%)	2 px (%)	EPE (pixel)	3 px (%)
LEAStereo [29]	0.63	22.50	8.54	3.94	3.20	49.44	28.51	16.03	22.62	27.07
STTR [30]	1.29	30.37	11.33	4.85	2.22	40.30	22.00	12.21	6.03	9.52
RAFT-Stereo [1]	0.27	8.10	2.33	0.96	1.12	24.28	14.02	8.43	1.16	4.59
StereoMAE	**0.22**	**6.77**	**2.10**	**0.78**	**0.98**	**22.14**	**11.33**	**7.39**	**0.96**	**3.94**

Figure 3 compares the reconstructed outputs of StereoMAE trained via the MSE loss from [11] and the proposed perceptual loss, on unseen datasets. It can be observed that StereoMAE trained with perceptual loss outperforms MSE in all scenes. When using perceptual loss, we observe significant increase in the fidelity of the reconstructed patches with finer textural and structural details. Exhibiting perceptual loss aids the model in learning higher-level feature representations that can generalise to any scene

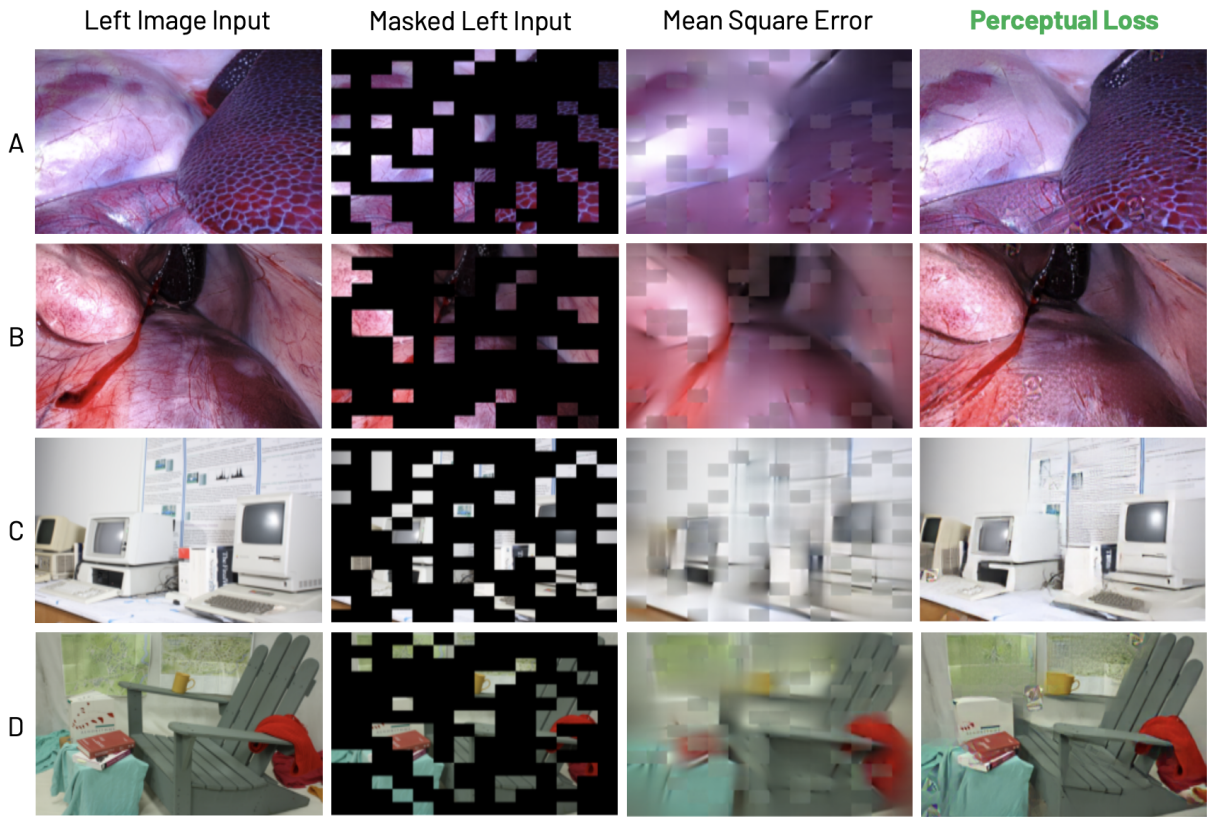

FIGURE 3 Reconstructions of the left images on (A-B) SCARED and (C-D) Middlebury samples. Results of MIM pre-trained StereoMAE via Mean Square Error loss[11] and the proposed perceptual loss shown in 3^{rd} and 4^{th} columns, respectively. The inputs were masked at 75% mask-to-image ratio. StereoMAE was not trained on these datasets.

type despite never having observed surgical scenes. When finetuned for disparity estimation, StereoMAE visibly generates sharper depth maps despite only being trained on synthetic data, as shown in Figure 4. Specifically, in sample A, StereoMAE generates the fine details on the robotic tool, whereas other methods either fail to achieve the correct disparity range or miss fine details around the edges. Similarly in B and C, StereoMAE generates less holes around the edges, as can be seen in the wheels / handle of the bike in B and the structures in the background in C. Hence, by learning a generalisable feature distribution for stereo image representations and disparity estimation, without any training on datasets from Figure 4, StereoMAE creates smooth yet fine disparities. We also conducted an ablation study by creating a new StereoMAE variant. It pairs the StereoMAE encoder with HITNet's decoder[4], fine-tuned on the Scene Flow dataset. It outperforms the original HITNet model in ETH3D, Middlebury and SCARED evaluations, confirming our method's decoder-independence and effectiveness in enhancing any decoder for disparity estimation. Original HITNet achieves EPE of 0.31, 1.62 on ETH3D and Middlebury, respectively and 4.12 mm depth error on SCARED. Whereas our variant achieves 0.25, 1.27 and 2.98 mm error, respectively. Our approach, as demonstrated in Figure 4, is independent of baseline, lighting, image type (like ETH3D's grayscale), or specific scene types. Despite training solely on simulations, it attains sub-mm accuracy (shown in Tables 2 and 3) on surgical scenes, delivering SOTA performance on simulated, natural, and surgical scenes alike. Affirming the significance of learning deep semantic features via MIM in pre-training encoders. This is further reflected in Figure 5 where it is clear samples A, B and E StereoMAE generates sharper details around the background anatomical structures and the edges of the robotic tools. Furthermore StereoMAE in sample D resolves disparity for the anatomical structures and the left-robotic tool even in the presence of smoke. Despite not training on surgical scenes, StereoMAE consistently outperformes competitors by generating sharper disparity maps under challenging lighting conditions and image resolution.

The quantitative evaluation in Tables 1, 2 and 3 also show our model outperforming previous state-of-the-art stereo depth estimation models in Stereo-benchmark datasets. StereoMAE achieves lower EPE across the board on all datasets, despite not being trained on them. In Table 1, SCARED benchmark only shows EPE and 3px error as they are the most commonly reported

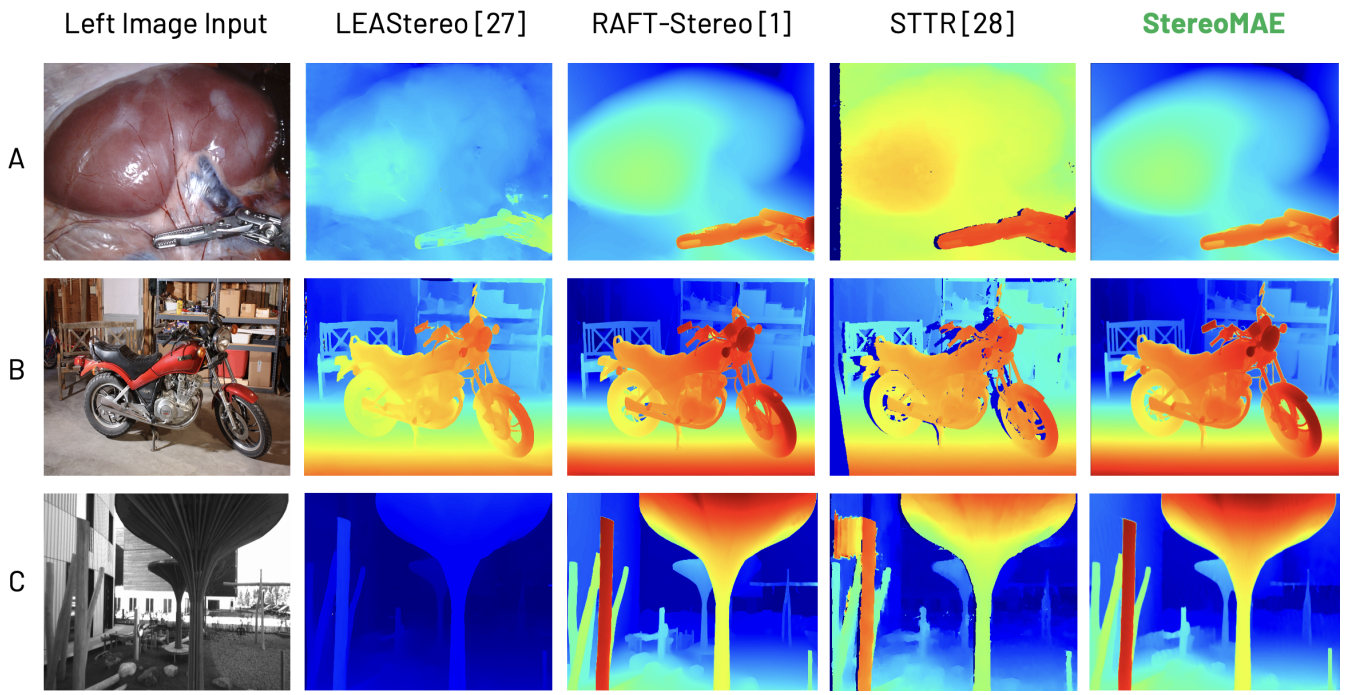

FIGURE 4 Left disparity outputs on (A) SCARED, (B) Middlebury and (C) ETH3D samples. All models were trained only on synthetic data.

by previous methods on this dataset[30]. Furthermore, as shown in Tables 2 and 3, StereoMAE also sets a new record on surgical scenes, achieving sub-millimeter accuracy in depth estimation. This exhibits the importance of visual representation learning for developing generalisable feature distributions, and still retaining the fine details of disparity estimation from supervised learning. Enabling generalisability to surgical images, without training on them.

TABLE 2 Mean absolute depth error (mm) on SCARED. Best results are highlighted in bold. None of the models were fine-tuned on SCARED, inference only.

Methods	SCARED 2019		SCARED Small
	Testdata 1	Testdata 2	Small 19 [30]
	MAE (*mm*)	MAE (*mm*)	MAE (*mm*)
LEAStereo [29]	3.82	4.51	1.48
STTR [30]	4.14	5.91	11.31
RAFT-Stereo [1]	3.74	4.28	1.01
StereoMAE	**1.66**	**2.01**	**0.92**

TABLE 3 Mean absolute depth error (mm) and EPE on SERV-CT data. Best results are highlighted in bold. None of the models were fine-tuned on SERV-CT, inference only.

Methods	SERV-CT			
	Experiment 1		Experiment 2	
	MAE (*mm*)	EPE	MAE (*mm*)	EPE
LEAStereo [29]	3.14	3.05	3.66	5.39
STTR [30]	16.60	7.54	25.69	14.52
RAFT-Stereo [1]	2.03	1.32	2.61	2.27
StereoMAE	**0.99**	**1.06**	**1.19**	**1.68**

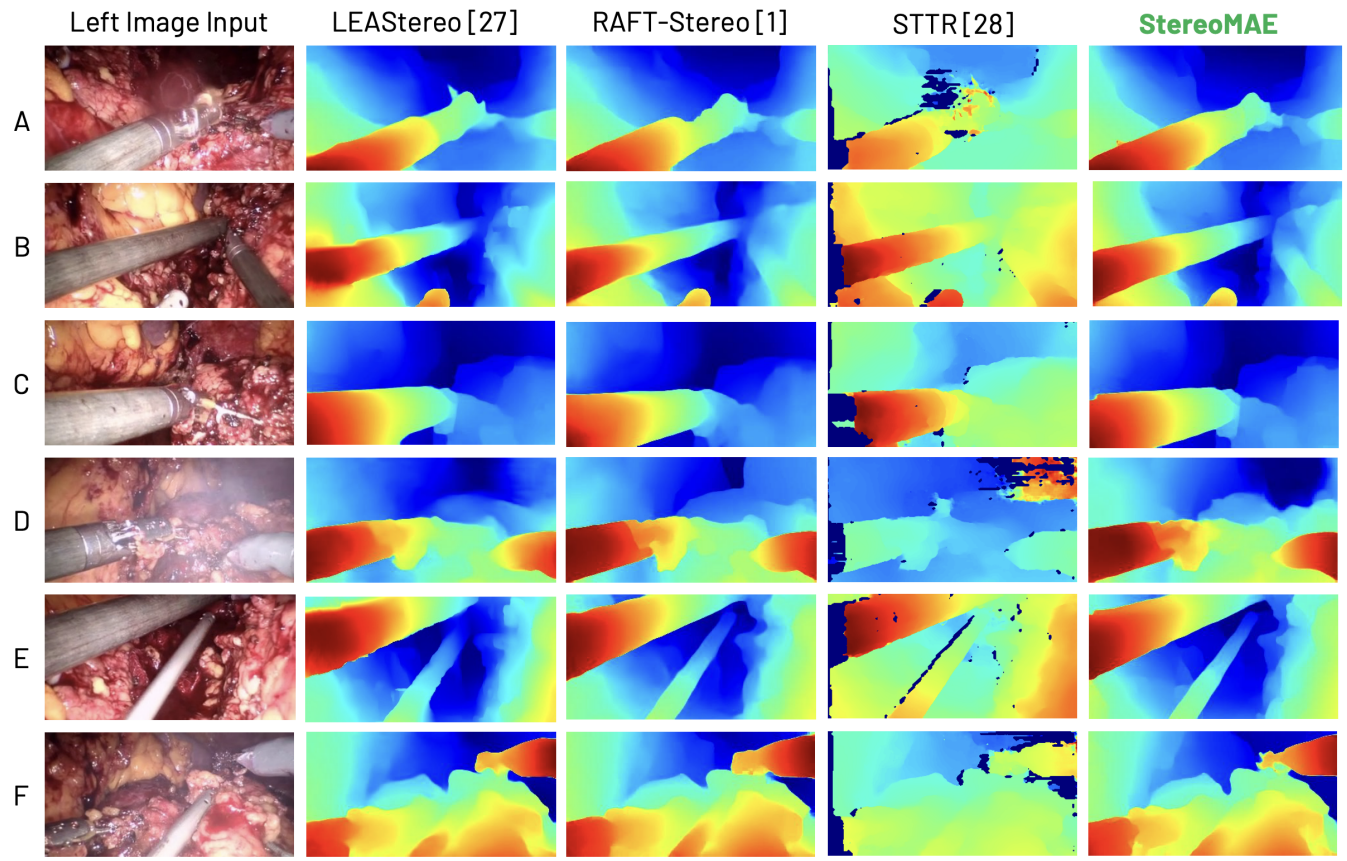

FIGURE 5 Left disparity outputs on the Hamlyn MIS data on different (A-F) samples. All models were trained only on synthetic data.

4 | CONCLUSIONS

Accurate and generalisable stereo depth estimation is crucial for surgical applications. Supervised learning is effective but limited by the scarcity of ground truth data in surgical settings, restricting their generalisability. Self-supervised methods have issues with scale ambiguity and inaccurate disparity prediction. Our StereoMAE approach demonstrates that by integrating MIM for feature representation learning with supervised depth estimation, achieves the benefits of both methods. Despite training only on synthetic data, StereoMAE shows generalisability to surgical and natural scenes, achieving sub-millimeter accuracy in surgical scenes. This is attributed to robust feature representations learned during MIM pre-training, which enhances the model's performance for subsequent tasks. Future work will explore alternative designs and loss terms to enhance MIM and improve stereo depth estimation.

AUTHOR CONTRIBUTIONS

All authors contributed to the conceptualization and design of the study. Samyakh Tukra wrote the basline script for training StereoMAE both pre-training and downstream finetuning. Haozheng Xu and Chi Xu performed the analysis, Haozheng Xu further aided with the interpretation of the results. All authors contributed to manuscript revision, and read and approved the submitted version.

ACKNOWLEDGMENTS

All authors are supported by the Royal Society (UF140290 and RGF\EA\180084) and the NIHR Imperial Biomedical Research Centre (BRC).

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

REFERENCES

1. Lipson L, Teed Z, Deng J. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. *arXiv preprint arXiv:2109.07547.* 2021.
2. Guo X, Yang K, Yang W, Wang X, Li H. Group-wise Correlation Stereo Network. In: 2019:3273–3282.
3. Zhang F, Prisacariu V, Yang R, Torr PH. GA-Net: Guided Aggregation Net for End-to-end Stereo Matching. In: 2019:185–194.
4. Tankovich V, Häne C, Fanello S, Zhang Y, Izadi S, Bouaziz S. HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2021:14357-14367.
5. Mayer N, Ilg E, Häusser P, et al. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In: 2016. arXiv:1512.02134.
6. Tukra S, Giannarou S. Stereo Depth Estimation via Self-supervised Contrastive Representation Learning. In: Springer. 2022:604–614.
7. Pilzer A, Lathuilière S, Xu D, Puscas MM, Ricci E, Sebe N. Progressive Fusion for Unsupervised Binocular Depth Estimation using Cycled Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2019.
8. Wang Y, Wang P, Yang Z, Luo C, Yang Y, Xu W. UnOS: Unified Unsupervised Optical-Flow and Stereo-Depth Estimation by Watching Videos. In: 2019.
9. Zhong Y, Dai Y, Li H. Self-Supervised Learning for Stereo Matching with Self-Improving Ability. *ArXiv.* 2017;abs/1709.00930.
10. Tonioni A, Tosi F, Poggi M, Mattoccia S, Di Stefano L. Real-time self-adaptive deep stereo. In: 2019.
11. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. In: 2022:16000-16009.
12. Bao H, Dong L, Piao S, Wei F. BEiT: BERT Pre-Training of Image Transformers. In: 2022.
13. Chen M, Radford A, Child R, et al. Generative Pretraining From Pixels. In: III HD, Singh A., eds. *ICML.* 119 of *Proceedings of Machine Learning Research*. PMLR 2020:1691–1703.
14. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. In: 2016.
15. Allan M, McLeod AJ, Wang CC, al eJR. Stereo Correspondence and Reconstruction of Endoscopic Data Challenge. *CoRR.* 2021;abs/2101.01133.
16. Giannarou S, Visentini-Scarzanella M, Yang GZ. Probabilistic Tracking of Affine-Invariant Anisotropic Regions. *IEEE Trans. Pattern Anal. Machine Intell.* 2012.
17. Edwards P, Psychogyios D, Speidel S, Maier-Hein L, Stoyanov D. SERV-CT: A disparity dataset from CT for validation of endoscopic 3D reconstruction. *arXiv preprint arXiv:2012.11779.* 2020.
18. Schöps T, Schönberger JL, Galliani S, et al. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In: 2017.
19. Hirschmüller H, Scharstein D. Evaluation of Cost Functions for Stereo Matching. *2007 IEEE Conference on Computer Vision and Pattern Recognition.* 2007:1-8.
20. Kolesnikov A, Dosovitskiy A, Weissenborn D, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: 2021.
21. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: 2018.
22. Tukra S, Marcus HJ, Giannarou S. See-Through Vision With Unsupervised Scene Occlusion Reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2022;44(7):3779-3790. doi: 10.1109/TPAMI.2021.3058410
23. Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In: 2018.
24. Cordts M, Omran M, Ramos S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. In: 2016.
25. Mayer N, Ilg E, Häusser P, et al. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. *CVPR.* 2016:4040-4048.
26. Tremblay J, To T, Birchfield S. Falling Things: A Synthetic Dataset for 3D Object Detection and Pose Estimation. In: 2018.
27. Wang W, Zhu D, Wang X, et al. TartanAir: A Dataset to Push the Limits of Visual SLAM. In: 2020.
28. Butler DJ, Wulff J, Stanley GB, Black MJ. A naturalistic open source movie for optical flow evaluation. In: A. Fitzgibbon et al. (Eds.) ., ed. *European Conf. on Computer Vision (ECCV)*Part IV, LNCS 7577. Springer-Verlag 2012:611–625.
29. Cheng X, Zhong Y, Harandi M, et al. Hierarchical Neural Architecture Search for Deep Stereo Matching. *Advances in Neural Information Processing Systems.* 2020;33.
30. Li Z, Liu X, Drenkow N, et al. Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective With Transformers. In: 2021:6197-6206.