

ARTICLE

Stereo-Endoscopic Disparity Estimation via Pseudo-Label-Pretrained StyleGAN3 and Self-Supervised Test-Time Optimization

Legend Zhang¹, Junmin Lyu¹, Guan Yao², Wei Wei³, Jiawei Tian^{4,*} and Bo Yang^{2,*}

¹Future Tech Institute, Guangzhou Huashang University, Guangzhou, China

²School of Automation, University of Electronic Science and Technology of China, Chengdu, China

³School of Mechanical and Material Engineering, Xi'an University, Xi'an, China

⁴Department of Computer Science and Engineering, Hanyang University, Ansan-si, Republic of Korea

*Corresponding Authors: Jiawei Tian. Email: tianjiawei@hanyang.ac.kr; Bo Yang. Email: boyang@uestc.edu.cn

Received: 02 June 2026; Accepted: 19 August 2026; Published: 28 September 2026

ABSTRACT: Accurate disparity estimation of soft tissue surfaces in endoscopic cardiac imaging is critical for minimally invasive surgical navigation and robotic surgery. Traditional geometric models, such as thin-plate splines, and recent learning-based stereo matching networks often struggle to balance nonlinear modeling capacity, computational cost, and robustness in dynamic surgical environments. We propose TTO-StyleGAN3 (Test-Time Optimization with a simplified StyleGAN3), a hybrid framework combining pseudo-label-supervised prior learning with self-supervised test-time latent optimization. A simplified StyleGAN3 generator is first pre-trained on pseudo-disparity maps generated by a teacher estimator and subsequently serves as a learned prior over plausible cardiac disparity fields. At inference, the generator is frozen, and the latent vector for each stereo pair is optimized without disparity supervision by minimizing photometric reconstruction, structural similarity, smoothness, and region-aware losses. This design avoids training a high-capacity end-to-end disparity predictor on manually annotated dense disparity datasets, although the compact generator prior is learned from teacher-generated pseudo-disparity maps. Experiments on a silicone heart phantom dataset and an *in vivo* dataset from Totally Endoscopic Coronary Artery Bypass (TECAB) procedures show that our method achieves competitive performance under the present pseudo-label and evaluation protocol, while reducing generator parameter count and the cost of training a full end-to-end disparity predictor.

KEYWORDS: StyleGAN3; disparity estimation; surgical navigation; stereo matching

1 Introduction

Disparity estimation is a fundamental task in recovering 3D surface geometry from stereo-endoscopic video, particularly in Minimally Invasive Surgical (MIS) environments [1,2]. Unlike traditional computer vision tasks, MIS scenes pose unique challenges: soft-tissue deformation, illumination variation, specular reflection, occlusion from surgical tools, and a lack of reliable ground truth data. These challenges render traditional supervised stereo matching techniques or dense pixel-wise regression approaches less effective [3–5].

To overcome the inherent limitations of handcrafted models, early methods introduced geometric priors such as Thin-Plate Splines (TPS) [6–8], which can model smooth deformation fields based on sparse control points. Although TPS can represent smooth nonlinear deformations, its fixed radial basis formulation and dependence on a limited number of control points may restrict its ability to capture complex and highly

localized tissue motion. Later improvements incorporated temporal models and high-order deformation [9], yet they still relied on strong assumptions about motion continuity and control point initialization.

With the rise of deep learning, end-to-end stereo matching networks like AANet [10] have shown promise. However, these approaches often require large-scale annotated datasets or handcrafted regularization terms, and struggle with generalization in out-of-distribution scenarios such as real-time intraoperative deployment. Additionally, variational autoencoders (VAEs) and their variants (β -VAE) [11] offer generative capabilities but produce over-smoothed disparity maps lacking fine structural detail.

Recently, Generative Adversarial Networks (GANs) [12–14] and their style-based variants, the StyleGAN series [15–17], have demonstrated exceptional expressiveness and control in image synthesis. In particular, the mapping from a low-dimensional latent vector to high-dimensional image space enables compact and semantically meaningful control over outputs. Generative priors have also been combined with self-supervised consistency constraints in image restoration, where adversarial learning regularizes the output distribution and self-supervision preserves its relationship with the degraded input [18]. While prior works [19,20] demonstrated the feasibility of applying StyleGAN to disparity estimation in stereo endoscopy, challenges remain in model size, overfitting, and convergence speed.

This paper builds on the above efforts and introduces TTO-StyleGAN3 (Test-Time Optimization with a simplified StyleGAN3). Our approach embeds the disparity map in a learned generative latent space and iteratively optimizes a latent vector to minimize photometric consistency, structural similarity, and region-aware losses. This jointly enforces geometric plausibility (via the generator’s learned prior over realistic cardiac shapes) and explicit photometric self-consistency for each stereo pair, yielding disparity estimates that are both anatomically coherent and photometrically faithful, even under challenging surgical conditions. Compared to traditional TPS or VAE approaches, TTO-StyleGAN3 offers greater nonlinear representation power. Compared with end-to-end regression networks, the framework is more annotation-efficient because it does not require manually acquired dense disparity labels; nevertheless, its generator pre-training relies on pseudo-disparity maps generated by a teacher model.

To ensure robustness in real surgical scenes, we further introduce an effective region masking scheme that mitigates the effect of specular highlights and illumination inconsistencies. This scheme enhances both the stability of generator pre-training and the accuracy of the subsequent test-time latent optimization.

We validate our method on two stereo-endoscopic video datasets: (1) Phantom, generated using a silicone heart driven by pneumatic actuation, providing clean motion but simplified texture and lighting conditions; and (2) *In vivo*, a real-world dataset captured from Totally Endoscopic Coronary Artery Bypass (TECAB) procedures using da Vinci robotic systems. The *In vivo* data presents greater challenges due to complex soft tissue deformation, blood presence, fogging, and dynamic illumination, serving as a testbed for disparity estimation models.

Comprehensive experiments demonstrate that TTO-StyleGAN3 achieves competitive performance across Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Root Mean Square Error (RMSE), and photometric consistency metrics, while using fewer parameters and achieving competitive performance on both the Phantom and *In vivo* datasets.

2 Method

The proposed TTO-StyleGAN3 framework comprises two stages with different supervision mechanisms. First, a compact StyleGAN3 generator is pre-trained using teacher-generated pseudo-disparity maps to learn a low-dimensional disparity prior. Second, at test time, the generator is frozen and a scene-specific

latent vector is optimized using only image-derived reconstruction and regularization losses, without using disparity labels for the test sample.

2.1 Generator Architecture (TTO-StyleGAN3)

A closely related style-based GAN approach is the lightweight StyleGAN adaptation proposed by Yang et al. [19], which was tailored for disparity estimation by reducing network depth and parameter count. This work demonstrated that the latent space of a style-based generator can effectively model disparity maps with substantially fewer parameters.

Building upon this idea, we designed a further improved generator, derived from StyleGAN3, which constitutes the generative core of TTO-StyleGAN3. Compared with the full StyleGAN3, our generator reduces the number of convolutional layers and channel dimensions and removes the ToRGB layer, which maps intermediate feature maps to the disparity domain. These changes cut the parameter count by nearly half, mitigating overfitting on small-scale medical datasets and improving computational efficiency. Importantly, the removal of stochastic per-pixel noise makes disparity generation strictly determined, governed solely by the input latent vector. The resulting architecture strikes a practical balance between expressiveness and efficiency, reducing the computational cost of each test-time optimization iteration. However, the complete iterative inference procedure is not yet suitable for real-time surgical deployment. Fig. 1 illustrates the generator structure.

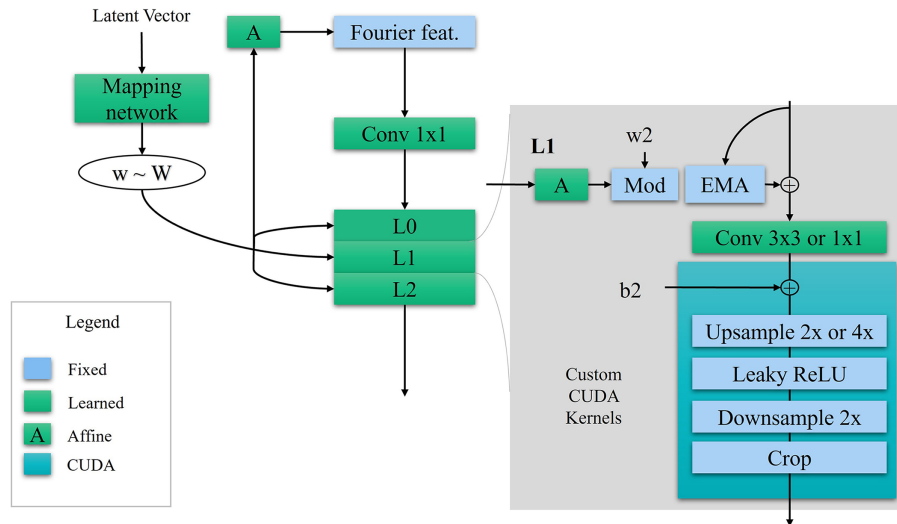


Figure 1: Generator structure of TTO-StyleGAN3.

The generator is pre-trained under pseudo-label supervision using dense disparity maps generated by a previously developed teacher estimator. The teacher model is used only to generate pseudo-disparity maps for the training split and is not used during test-time inference. Therefore, the generator-prior learning stage is not self-supervised. After pre-training, the discriminator is discarded and the frozen generator is used only as a disparity prior during self-supervised test-time optimization.

2.2 Latent Vector Optimization

Instead of directly regressing pixel-wise disparity, the framework recasts disparity estimation as a per-sample latent optimization problem. For each new stereo pair, we search for an optimal latent vector w in the generator's latent space by minimizing an image reconstruction objective, as in Eq. (1):

$$w^* = \arg \min_w \mathcal{L}(I_r, \hat{I}_r, d_w) \quad (1)$$

where d_w denotes the disparity map produced by the frozen generator from the latent vector w , I_r is the original right image, and $\hat{I}_r$ is the reconstructed right image obtained by warping the left image according to d_w .

The loss function $\mathcal{L}(\cdot)$ encapsulates several self-supervised terms that quantify reconstruction fidelity. Because the test-time objective is computed solely from the input stereo pair and does not require a disparity target, this stage is referred to as self-supervised test-time latent optimization.

Through iterative optimization, w converges to a configuration that generates a disparity map capable of accurately warping the left view to match the right view. This formulation transforms the traditional pixel-level regression task into a compact, structured optimization within a low-dimensional latent space, ensuring that the resulting disparity remains both geometrically plausible and photometrically consistent.

2.3 Loss Functions

For a latent vector w , the frozen generator produces a disparity map d_w . The reconstructed right image is obtained by warping the original left image I_l according to d_w as in Eq. (2):

$$\hat{I}_r = \mathcal{W}(I_l, d_w) \quad (2)$$

where $\mathcal{W}(\cdot)$ denotes differentiable backward warping.

To compensate for the global brightness difference between the stereo views, we calculate a constant luminance offset from the masked grayscale means, as shown in Eq. (3):

$$l_{\text{comp}} = \mu_M(I_r^g) - \mu_M(I_l^g), \mu_M(I) = \frac{\sum_{p \in \Omega} M(p)I(p)}{\sum_{p \in \Omega} M(p) + \varepsilon} \quad (3)$$

where I_l^g and I_r^g are the grayscale left and right images, M is the valid-region mask, and ε prevents division by zero. The compensated reconstruction is defined as Eq. (4):

$$\tilde{I}_r(p) = \text{clip}(\hat{I}_r(p) + l_{\text{comp}}, 0, 1) \quad (4)$$

Here, l_{comp} is a luminance correction rather than an additional loss term.

The masked photometric loss is defined as Eq. (5):

$$\mathcal{L}_{\text{pho}} = \frac{\sum_{p \in \Omega} M(p) \|I_r(p) - \tilde{I}_r(p)\|_2^2}{\sum_{p \in \Omega} M(p) + \varepsilon} \quad (5)$$

The appearance matching loss combines L1 and SSIM terms, shown in Eq. (6):

$$\mathcal{L}_{\text{ap}} = \alpha \frac{\sum_p M(p) [1 - \text{SSIM}(I_r, \tilde{I}_r)(p)] / 2}{\sum_p M(p) + \varepsilon} + (1 - \alpha) \frac{\sum_p M(p) |I_r(p) - \tilde{I}_r(p)|}{\sum_p M(p) + \varepsilon} \quad (6)$$

where, $\alpha = 0.85$. The disparity smoothness loss is defined as Eq. (7):

$$\mathcal{L}_{\text{ds}} = \frac{1}{N} \sum_p (|\partial_x d_w(p)| + |\partial_y d_w(p)|) \quad (7)$$

where N is the number of pixels.

The binary mask is generated automatically from the grayscale left image, right image, and reconstructed right image. Pixels with intensity higher than threshold τ are treated as unreliable specular regions. The detected highlight regions in the left image are dilated, combined with the valid regions of I_r and $\hat{I}_r$, and then eroded to remove uncertain boundary pixels, this process is illustrated by Eq. (8):

$$M = \text{erode} \left[\mathbf{1}(I_r^g < \tau) \odot \mathbf{1}(\hat{I}_r^g < \tau) \odot (1 - \text{dilate} [\mathbf{1}(I_l^g \geq \tau)]) \right] \quad (8)$$

No manual annotation or additional segmentation network is required.

The complete objective is shown in Eq. (9):

$$\mathcal{L}(I_r, \hat{I}_r, d_w) = w_{\text{pho}} \mathcal{L}_{\text{pho}} + w_{\text{ap}} \mathcal{L}_{\text{ap}} + w_{\text{ds}} \mathcal{L}_{\text{ds}} \quad (9)$$

where w_{pho} , w_{ap} , and w_{ds} denote the scalar weights of the photometric, appearance-matching, and disparity-smoothness losses, respectively. They were all set to 1.0 in our experiments. The validity mask defined in Eq. (8) is used consistently in the masked loss computation.

3 Experimental Setting

We evaluate the proposed TTO-StyleGAN3 framework on two stereo-endoscopic datasets. All experiments were conducted on a cloud server running Ubuntu 18.04 with PyTorch 1.9.1 and an NVIDIA V100 (32 GB) GPU. StyleGAN3 was configured to use half-precision (FP16) computation, which was adopted for both generator pre-training and latent optimization.

3.1 Datasets

Two calibrated stereo-endoscopic video datasets are used, each containing sequences of 700 frames at a resolution of 288×360 pixels.

Phantom Dataset [21]: Acquired on a pneumatically actuated silicone heart phantom. This dataset provides controlled motion and simplified textures, enabling quantitative evaluation under clean conditions.

***In vivo* Dataset [22,23]:** Recorded during TECAB procedures using a da Vinci robotic system. It contains real surgical challenges including soft-tissue deformation, blood, fogging, specular reflections, and dynamic illumination.

For disparity estimation, a 256×256 -pixel region of interest (RoI) is cropped from each right-view image (yellow box in Fig. 2). For the *In vivo* dataset, the RoI starts at $(u, v) = (32, 14)$, for the Phantom dataset, at $(u, v) = (16, 70)$.

Each sequence is split into 600 frames for training (used to pre-train the generator or to train baseline models) and 100 frames for testing. Because ground-truth disparity is unavailable, we generate teacher-generated pseudo-disparity maps for the training split using a high-fidelity disparity estimator based on our previous StyleGAN-based framework [19] with an extended latent search. The same pseudo-disparity maps were used for all learning-based methods to maintain a consistent supervision source. Nevertheless, because the teacher estimator and the TTO-StyleGAN models belong to related generative-model families, this protocol may favor models with similar inductive biases. Therefore, the comparison does not completely eliminate teacher-model bias.

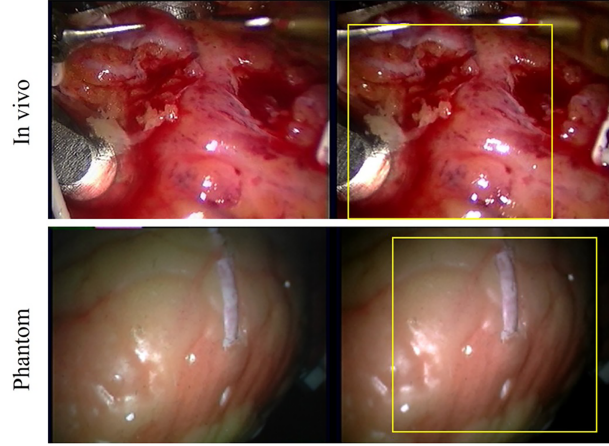


Figure 2: Stereoscopic endoscope video data.

These pseudo-labels serve as supervision for pre-training our TTO-StyleGAN3 generator, and for training the following baseline models: TTO-StyleGAN2, Diff-StyleGAN2 (a StyleGAN2 generator trained under the Diffusion-GAN paradigm [24]), and β -VAE [11]. For fair comparison, all generators operate with a 16-dimensional latent space, ensuring that any performance differences stem from architectural design rather than latent dimensionality. Note that for AANet [10], which is a supervised stereo matching network rather than a generative model, the input consists of the real stereo image pairs and the pseudo disparity maps serve as the training targets.

During testing, we additionally construct a set of sparse feature-matched disparity values to serve as a quantitative reference. For 100 test frames, we extract Oriented FAST and Rotated BRIEF (ORB) features from each stereo pair, perform brute-force matching, and retain the five best matches per frame after rejecting outliers via a disparity consistency threshold. This yields 500 high-confidence matched points across the test set, providing sparse correspondence references for RMSE calculation. These points were selected conservatively to reduce mismatches in texture-poor and specular regions. An example of the ORB feature extraction and stereo matching procedure is shown in Fig. 3.

3.2 Evaluation Metrics

To quantitatively evaluate disparity estimation, four complementary metrics, PSNR, SSIM, RMSE, and photometric consistency error, are employed in Eqs. (10)–(12).

(1) Peak Signal-to-Noise Ratio (PSNR): measures reconstruction fidelity between the original right image I_r and the reconstructed right image $\hat{I}_r$.

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (10)$$

where MAX_I is the maximum possible pixel intensity value and MSE is the mean squared error between I_r and $\hat{I}_r$.

(2) Structural Similarity Index (SSIM): assesses perceptual similarity between I_r and $\hat{I}_r$.

$$\text{SSIM}(I_r, \hat{I}_r) = \frac{(2\mu_{I_r}\mu_{\hat{I}_r} + C_1)(2\sigma_{I_r\hat{I}_r} + C_2)}{(\mu_{I_r}^2 + \mu_{\hat{I}_r}^2 + C_1)(\sigma_{I_r}^2 + \sigma_{\hat{I}_r}^2 + C_2)} \quad (11)$$

where μ and σ represent the mean and variance, and C_1, C_2 are small constants to stabilize the division.

(3) Sparse-correspondence RMSE evaluates local disparity accuracy at the 500 reliably matched ORB feature locations. Let $d_{k,t}$ be the reference disparity of the k -th point at frame t and $\hat{d}_{k,t}$ the estimated disparity at the same location. Then

$$\text{RMSE} = \sqrt{\frac{1}{500} \sum_{t=1}^{100} \sum_{k=1}^5 (d_{k,t} - \hat{d}_{k,t})^2} \quad (12)$$

(4) Photometric Consistency Error: measures the mean squared error between I_r and $\hat{I}_r$ over the valid region, computed identically to the photometric loss in Eq. (5).

Together, these metrics evaluate both the visual quality of the reconstructed views and the numerical accuracy of the estimated disparity.

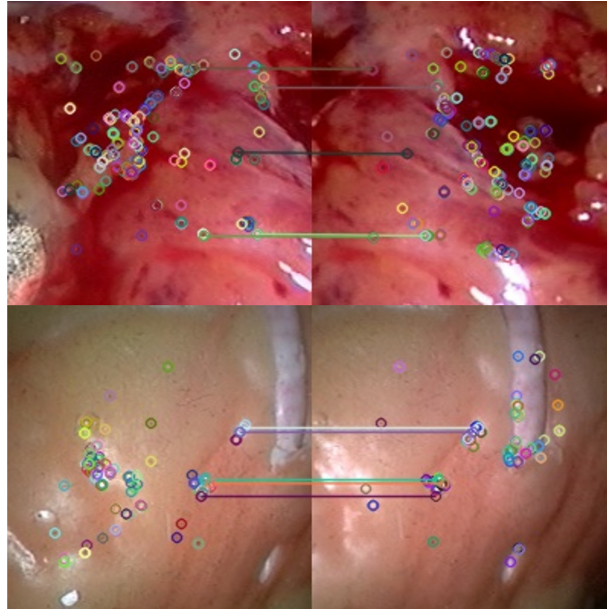


Figure 3: ORB feature extraction and matching.

3.3 Implementation Details

The TTO-StyleGAN generator is pre-trained under pseudo-label supervision using the teacher-generated disparity maps described in Section 3.1. No pseudo-disparity label or teacher-model prediction is used during test-time latent optimization. Following the StyleGAN3 training paradigm, the generator uses a non-saturating logistic loss $\mathcal{L}_G = \log(e^{-D(G(z))} + 1)$, while the discriminator is trained with a logistic loss function and an R1 gradient penalty [25] applied only to real samples, as in Eq. (13):

$$\mathcal{L}_D = \log(e^{D(G(z))} + 1) (e^{-D(x)} + 1) + \frac{\gamma}{2} E_{x \sim p_{\text{real}}} [\|\nabla_x D(x)\|] \quad (13)$$

with $\gamma = 10$. The R2 gradient penalty on generated samples is not used. Training proceeds until the generator has produced 2M samples. The same training protocol is followed for TTO-StyleGAN2 and Diff-StyleGAN2.

β -VAE is trained for 100 epochs using the same pseudo ground truth disparity maps. AANet was optimized using Adam with an initial learning rate of 1×10^{-3} , a batch size of 1, and a maximum disparity of 192 pixels. The maximum number of epochs was initially set to 100. After repeated training runs, the lowest average training loss was consistently observed at approximately epoch 20, and further training produced no meaningful improvement. The checkpoint with the minimum average training loss was therefore selected for evaluation.

At inference, the generator is frozen. For each stereo pair, a latent vector is initialized randomly and optimized for a maximum of 150 iterations using the Adam optimizer with a learning rate of 0.05. The total loss defined in Eq. (9) drives the optimization. Convergence is typically reached within the iteration budget, yielding a per-sample disparity map.

4 Experimental Results and Analysis

4.1 Quantitative Comparison of Performance of Each Model

TTO-StyleGAN2, TTO-StyleGAN3, and Diff-StyleGAN2 were pre-trained using the same pseudo-disparity maps. For the classical geometric baselines, we adopted TPS models with 16 and 25 control points, denoted as TPS-16 and TPS-25, respectively. The control points were arranged on regular 4×4 and 5×5 grids. TPS-16 was selected to match the 16-dimensional latent space used by the generative models, enabling comparison under a similar number of optimizable variables. TPS-25 was included as a higher-capacity geometric baseline to examine whether increasing the number of control points could compensate for the limited nonlinear representation of TPS. These settings were chosen for methodological comparison rather than for a specific clinical rationale. Both TPS variants followed the same test-time optimization procedure as the TTO models, with their control-point disparities optimized for each stereo pair using the same reconstruction objective.

The TTO-StyleGAN2 and TTO-StyleGAN3 generators share a similar synthesis subnetwork structure, consisting of 7 resolution levels, with feature maps progressively growing from 4×4 to 256×256 . Each level contains 32 channels, and all but the final synthesis level employ 32 convolution kernels.

Fig. 4 shows the per-frame photometric loss ($\mathcal{L}_{pho}$) for each model evaluated across 100 test frames from both datasets. To ensure training convergence, the discriminator's architecture mirrors that of the generator in terms of resolution levels and channel counts. All GAN-based models were trained until the generator had processed approximately 2M samples, which took around 2.5 h on our hardware. The distribution of the generated disparity maps after pre-training aligns well with that of the pseudo ground-truth samples.

In the test phase, we evaluated the optimized disparity from each model. For a fair comparison, all latent-optimization-based models (TTO-StyleGAN2, TTO-StyleGAN3, and Diff-StyleGAN2) were initialized using the same latent vector. For each test frame, optimization was performed for a maximum of 150 iterations, and the disparity map corresponding to the minimum was selected as the final output.

The similarity between the original right-view RoI and the one reconstructed by warping the left image with the estimated disparity serves as an indirect yet reliable measure of accuracy. To comprehensively assess the quality of the estimated disparity maps, we averaged several metrics across the 100 test frames. The quantitative results are summarized in Table 1, where $\mathcal{L}_{pho}$ quantifies the per-frame photometric consistency between the original and reconstructed right images, directly reflecting the fidelity of the test-time optimization; PSNR and SSIM measure the pixel-level and perceptual reconstruction quality, respectively, and RMSE evaluates the numerical accuracy of the estimated disparity against the 500 sparse

reference points obtained via ORB feature matching, which provides a direct assessment of geometric precision independent of photometric appearance.

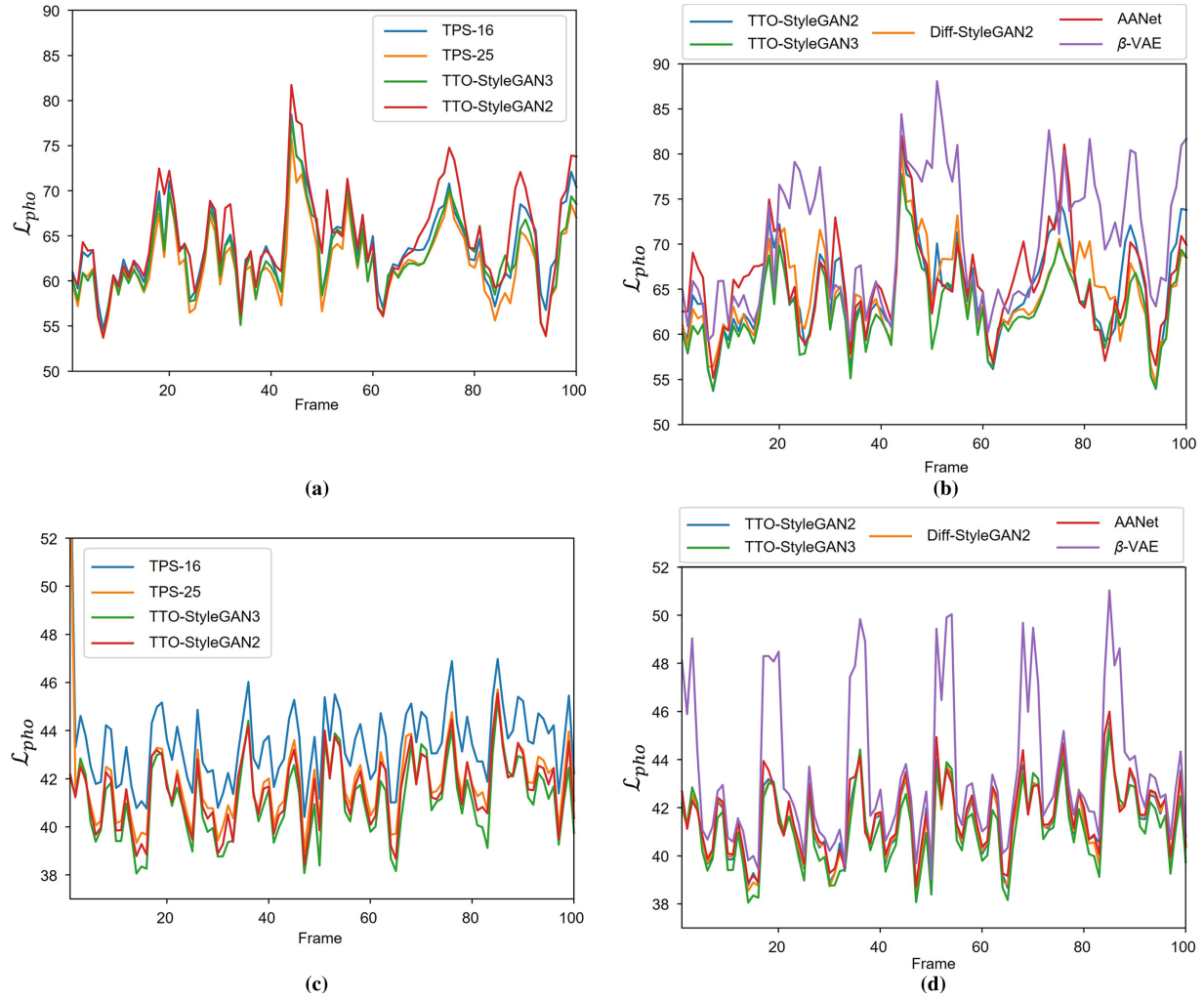


Figure 4: Photometric loss curves for 100 test frames. (a) TPS vs. TTO-StyleGAN series models on the *in vivo* dataset; (b) deep learning models on the *in vivo* dataset; (c) TPS vs. TTO-StyleGAN series models on the phantom dataset; (d) deep learning models on the phantom dataset.

Table 1: Quantitative evaluation of disparity estimation models on the *in vivo*/phantom datasets.

Model	$\mathcal{L}_{pho} \downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	RMSE $\downarrow$
TPS-16	63.779/43.512	30.084/31.745	0.9106/0.8997	2.0882/2.3744
TPS-25	62.008/41.942	30.206/31.904	0.9134/0.9029	1.9272/2.1098
TTO-StyleGAN2	64.844/41.518	30.012/31.948	0.9107/0.9041	2.1759/2.4021
TTO-StyleGAN3	62.901/41.127	30.144/31.989	0.9133/0.9050	1.9355/2.2511
Diff-StyleGAN2	64.520/41.481	30.034/31.952	0.9102/0.9044	2.1629/2.4836
AANet	65.139/41.672	29.992/31.932	0.9079/0.9041	2.2824/2.5001
β -VAE	70.274/43.409	29.663/31.755	0.8984/0.9006	2.6174/2.3834

Note: $\uparrow$ indicates that a higher value is better; $\downarrow$ indicates that a lower value is better.

On the Phantom dataset, the TTO-StyleGAN series and Diff-StyleGAN2 achieve competitive performance under the present evaluation protocol. These methods combine a pseudo-label-pretrained generative prior with self-supervised test-time latent optimization. Specifically, the generator prior is learned offline from teacher-generated pseudo-disparity maps, whereas the latent vector for each test frame is optimized without disparity supervision using image reconstruction and regularization losses. By searching within a compact nonlinear latent space and excluding specular regions from the reconstruction objective, the TTO-based methods obtain photometric losses close to, and in some cases lower than, those of TPS-25.

Compared with the feed-forward AANet, the TTO-based methods can adapt their latent variables to each test frame through iterative reconstruction optimization. This per-frame adaptation may improve reconstruction consistency under distribution shifts, but it also incurs substantially higher inference cost. Therefore, the comparison should be interpreted as an accuracy–adaptability–efficiency trade-off between two different inference paradigms rather than as a strictly equal-computation benchmark. While β -VAE occasionally achieves a low photometric loss on individual frames, its overall performance is less stable. Its average photometric loss is the highest on the *In vivo* dataset, whereas on the Phantom dataset it is slightly lower than that of TPS-16 but higher than those of the other evaluated models.

On the more challenging *In vivo* dataset, TPS-25 achieves the lowest average photometric loss. This result cannot be attributed solely to its larger number of optimized variables. Its denser 5×5 control-point grid provides greater local flexibility, while the globally smooth TPS interpolation may also align well with the predominantly smooth cardiac surface and the photometric objective. TTO-StyleGAN3 achieves comparable performance through a learned nonlinear disparity prior and per-frame latent optimization. AANet obtains lower reconstruction performance under the present training and evaluation settings, possibly because its feed-forward inference does not include sample-specific adaptation. However, this difference should be interpreted cautiously because the methods differ in both parameterization and inference-time computation. The other generative models produce average photometric losses greater than 63 and do not match the performance of TTO-StyleGAN3 under the current protocol.

4.2 Qualitative Comparison of Reconstructed Views and Disparity Maps

To gain deeper insight into the behavior of each model beyond the aggregate metrics reported in Section 4.1, we visually inspect the estimated disparity maps and the resulting 3D surface reconstructions. Figs. 5 and 6 display the disparity maps produced by all models on representative frames from the *In vivo* dataset.

TPS-16 produces overly smooth disparity maps that lack fine structural detail and fail to accurately capture localized soft tissue deformation. In contrast, all generative models and the stereo matching network are capable of preserving higher-frequency geometric detail. The notable exception is β -VAE, whose low-resolution latent representation introduces visible discontinuities and a tiled, blocky appearance in the disparity maps.

Among the generative models, TTO-StyleGAN2 shows signs of overfitting, with noisy artifacts appearing in textureless regions. Diff-StyleGAN2 exhibits the opposite problem: its disparity maps are excessively smooth and lack sharp depth boundaries, resembling the under-representation observed with TPS-16. Although TTO-StyleGAN3 contains approximately half the parameters of TTO-StyleGAN2, its disparity maps are visually on par with those of TPS-25, preserving both fine surface details and clean depth edges. This demonstrates that the simplified architecture does not sacrifice representational quality.

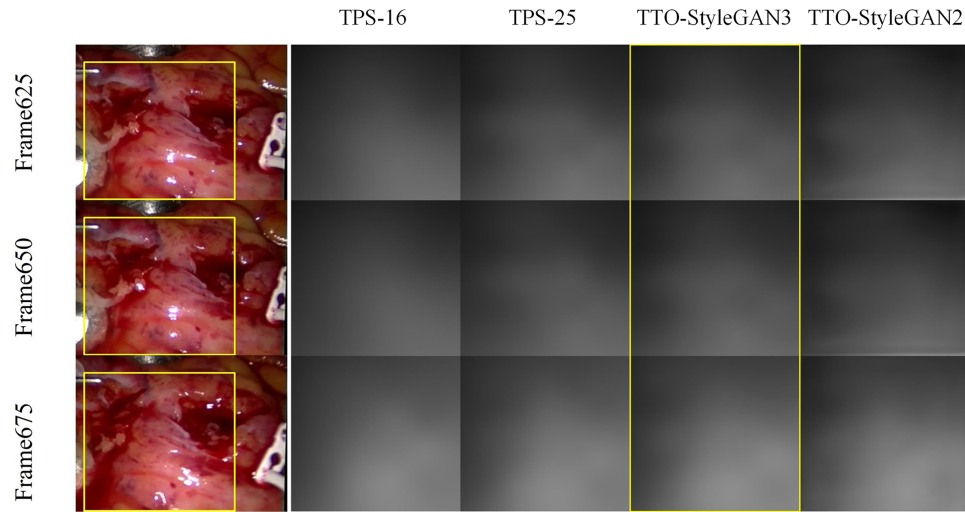


Figure 5: Disparity estimation results of TPS vs. StyleGAN series models on the *in vivo* dataset.



Figure 6: Disparity estimation results of deep learning network on the *in vivo* dataset.

Figs. 7 and 8 show the corresponding reconstructed 3D surfaces for sample frames from both datasets.

All five models (TPS-16, TPS-25, TTO-StyleGAN2, TTO-StyleGAN3, Diff-StyleGAN2) produce globally plausible 3D surfaces. In contrast, the end-to-end network AANet generates disparity maps with poor spatial continuity and numerous small bumps, which are visibly amplified in the 3D reconstructions. While increasing the weight of the disparity smoothness loss in the loss function $\mathcal{L}_{ds}$ can partially suppress these artifacts, it comes at the cost of over-flattening the surface and erasing clinically meaningful geometric detail.

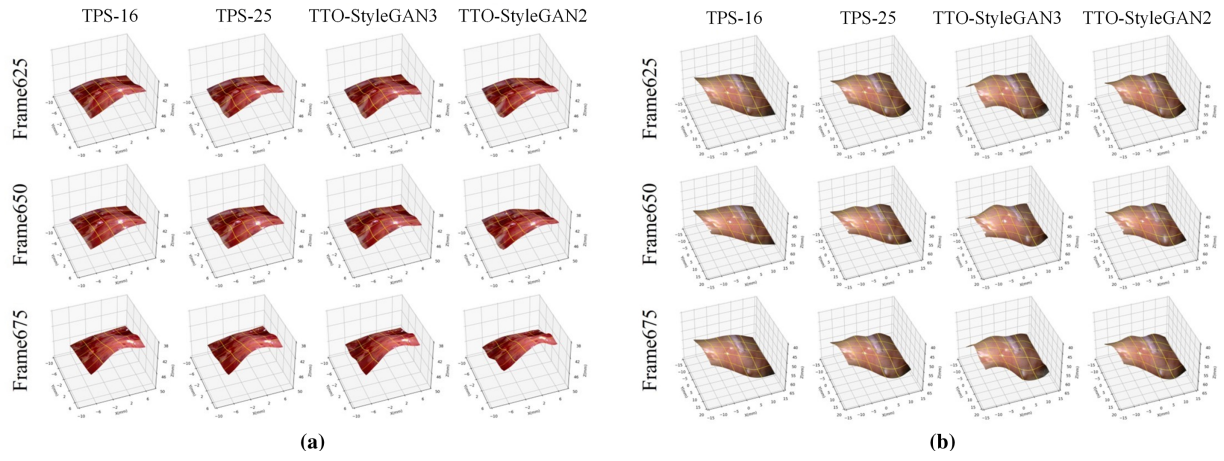


Figure 7: Reconstructed 3D surfaces: TPS vs. TTO-StyleGAN series models. (a) *In vivo* dataset; (b) phantom dataset.

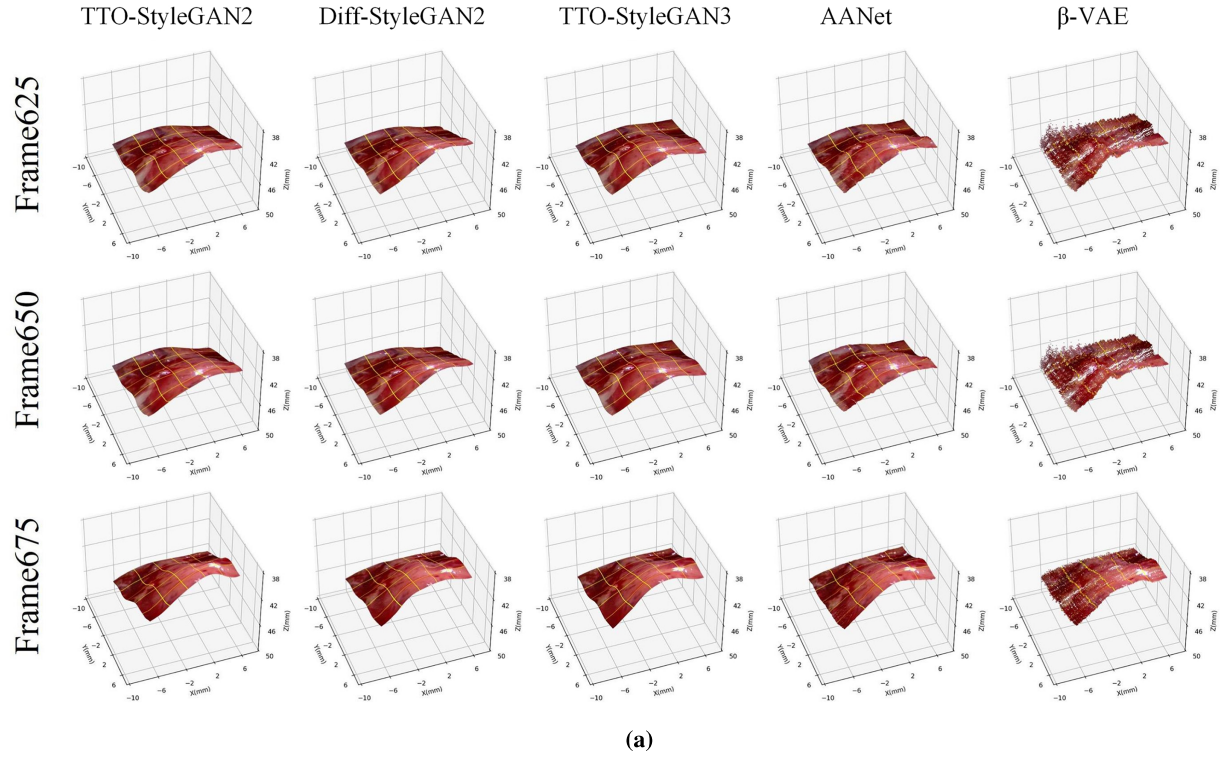


Figure 8: (Continued)

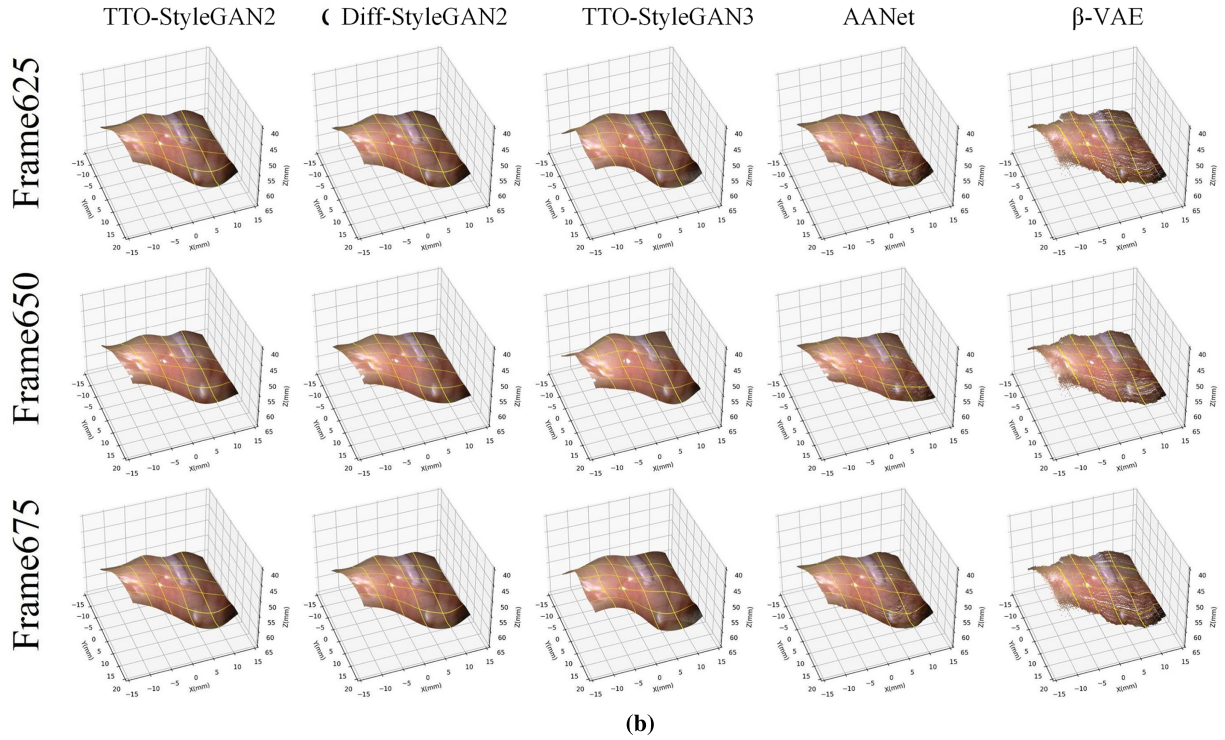


Figure 8: Reconstructed 3D surfaces by deep learning networks. (a) *In vivo* dataset; (b) phantom dataset.

The β -VAE model yields the least coherent surfaces. Even with an increased smoothness weight, its reconstructions remain fragmented and discontinuous, consistent with the discretization artifacts observed in its disparity maps. Overall, β -VAE exhibits poorer surface continuity than the other evaluated models under the present evaluation protocol.

4.3 Robustness to Occlusion and Motion Blur

To evaluate the resilience of the deep learning models under realistic surgical disturbances, we conducted robustness experiments on the *In vivo* dataset, simulating two common artifacts: occlusions and motion blur.

For occlusion, we masked a square region at the center of the RoI with white pixels, simulating the effect of a surgical tool or blood droplet obscuring the tissue surface. Tests were performed at five occlusion scales (e.g., 10×10 , 20×20 , 30×30 , 40×40 , and 50×50 pixels) to assess performance degradation as the occluded area increases. For motion blur, we convolved each frame with Gaussian kernels of varying standard deviations σ from 0 to 9 pixels, where $\sigma = 0$ corresponds to the original sharp image, and larger values produce increasingly severe blur, mimicking rapid endoscope movement or tissue pulsation.

For each perturbation, we computed the averaged photometric loss $\mathcal{L}_{pho}$ over the 100 test frames. Fig. 9 summarizes the performance of each model under these challenging conditions. The results show that all deep models experience performance degradation as occlusion size or blur severity increases, but their relative robustness differs. TTO-StyleGAN3 consistently achieves the lowest photometric loss across all perturbation levels, demonstrating stronger resilience than the other evaluated models, including Diff-StyleGAN2, TTO-StyleGAN2, β -VAE, and AANet.

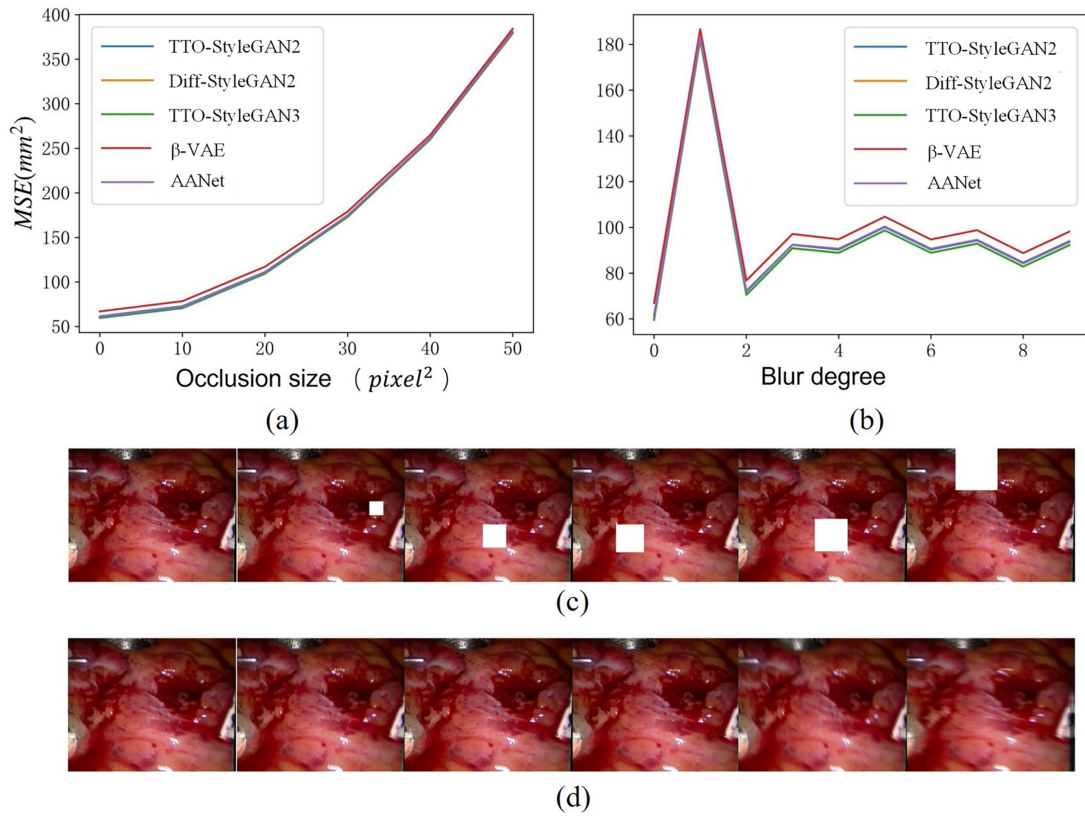


Figure 9: Robustness evaluation results of the compared models under synthetic perturbations. (a) Occlusion experiments; (b) motion blur experiments; (c) examples of synthetically occluded images; (d) examples of synthetically blurred images.

4.4 Computational Efficiency

Table 2 compares the computational complexity of the evaluated models. TTO-StyleGAN3 achieves the lowest floating-point operations (FLOPs) and parameter count among the StyleGAN-based methods, confirming the effectiveness of the simplified generator architecture.

Table 2: Computational complexity and execution time of different disparity estimation models.

Model	FLOPs (G) ↓	Params (M) ↓	Time Per Evaluation (s) ↓
TPS-16	0.021	1.114	–
TPS-25	0.033	1.704	–
TTO-StyleGAN2	3.221	0.120	0.024
TTO-StyleGAN3	1.112	0.056	0.010
Diff-StyleGAN2	3.221	0.120	0.027
AANet	58.984	3.600	0.115
β -VAE	0.405	0.185	0.004

Note: ↓ indicates that a lower value is better. – indicates that the corresponding execution time was not applicable.

The reported execution time for TTO-based methods corresponds to one optimization iteration. Since TTO-StyleGAN3 requires up to 150 iterations, its complete inference time is approximately 1.50 s per frame,

corresponding to 0.67 FPS. In contrast, AANet performs a single forward pass in 0.115 s, corresponding to 8.70 FPS.

5 Discussion

The results indicate that TTO-StyleGAN3 provides a compact framework for stereo-endoscopic disparity estimation by combining a pseudo-label-pretrained generative prior with self-supervised test-time latent optimization. The simplified StyleGAN3 generator maps a low-dimensional latent vector to a dense nonlinear disparity field, while per-frame optimization enforces photometric consistency with the input stereo pair. Compared with TPS-based models, this learned prior offers greater flexibility for representing nonlinear soft-tissue deformation. Compared with feed-forward methods such as AANet, the proposed method provides sample-specific adaptation at test time, which may partially mitigate distribution shifts. However, this adaptability is obtained at the cost of substantially greater inference time, and the comparison should therefore be interpreted as an accuracy–adaptability–efficiency trade-off rather than a strictly equal-computation benchmark.

Beyond stereo reconstruction, reliable disparity estimates may provide useful geometric cues for downstream medical imaging tasks, including depth-aware lesion segmentation, three-dimensional lesion localization, tissue-surface registration, and treatment planning. Recent semi-supervised segmentation research has shown that latent-space representation learning and uncertainty rectification can reduce pseudo-label error accumulation and improve the delineation of complex anatomical boundaries [26]. This suggests that the disparity priors estimated by TTO-StyleGAN3 could potentially be integrated with uncertainty-aware segmentation frameworks to improve spatial consistency and anatomical localization. Nevertheless, the present study does not directly evaluate lesion segmentation, diagnosis, or treatment outcomes, and these potential benefits require task-specific validation.

Several limitations should be acknowledged. First, the generator is pre-trained using pseudo-disparity maps produced by a StyleGAN-based teacher. Although all learning-based methods use the same pseudo-label source, this does not eliminate architecture-dependent bias. TTO-StyleGAN3 is more closely related to the teacher in terms of generative prior and additionally performs per-frame latent optimization, whereas AANet and β -VAE do not receive the same opportunity for online correction. Consequently, the current comparison cannot fully separate the contribution of the proposed architecture from teacher-family similarity and test-time adaptation. Second, the absence of dense measured disparity limits objective validation. The sparse ORB-based RMSE provides only a local reference at reliably matched feature locations and may not reflect performance in textureless, specular, or weakly matched regions. Third, the region-aware mask excludes unreliable highlights but may also remove pixels containing valid depth information, reducing reconstruction completeness. Fourth, the robustness experiments use idealized square occlusions and Gaussian blur. These controlled perturbations do not fully reproduce irregular, semi-transparent, spatially varying, or temporally correlated surgical artifacts such as blood smears, fogging, and complex tissue or instrument occlusions.

Practical deployment also remains challenging. The learned prior may depend on the anatomy, camera calibration, illumination, image resolution, and acquisition conditions represented in the pre-training data; changes in surgical site or imaging setup may therefore require additional pseudo-label generation, adaptation, or retraining. Moreover, although TTO-StyleGAN3 requires only 1.112 GFLOPs and 0.056 million parameters per generator evaluation, its maximum 150-iteration optimization takes approximately 1.50 s per frame, corresponding to 0.67 FPS, whereas AANet requires 0.115 s per frame, corresponding to 8.70 FPS. The current framework should therefore be regarded as a proof-of-concept rather than a clinically deployable real-time system. Future work will investigate independently measured or multi-teacher

disparity supervision, teacher-independent prior learning, cross-domain and multi-anatomy adaptation, denser geometric validation, more realistic surgical artifact modeling, learned latent initialization, temporal warm-starting, adaptive early stopping, and lightweight encoder–optimization hybrids for faster and more stable video reconstruction.

6 Conclusions

This paper proposed TTO-StyleGAN3, a hybrid disparity-estimation framework that combines pseudo-label-guided generative prior learning with self-supervised test-time latent optimization. A simplified StyleGAN3 generator maps a low-dimensional latent vector to a dense disparity field, while the latent vector is optimized for each stereo pair using photometric reconstruction, appearance matching, disparity smoothness, and region-aware masking. Experiments on the Phantom and *In vivo* datasets show that the proposed method achieves competitive reconstruction and disparity-estimation performance under the present pseudo-label and evaluation protocol, while using fewer generator parameters than the other StyleGAN-based models.

The results demonstrate the potential of generative-prior-based test-time optimization for modeling nonlinear soft-tissue deformation without requiring manually annotated dense disparity maps. However, the framework still depends on teacher-generated pseudo-disparity maps for generator pre-training, and the current comparisons may be influenced by teacher-family bias and differences in inference-time adaptation. In addition, the sparse geometric evaluation, dataset specificity, simplified robustness tests, and iterative inference speed limit the strength and generalizability of the current conclusions. With a maximum of 150 optimization iterations, the present implementation does not satisfy real-time surgical requirements and should be regarded as a proof-of-concept rather than a clinically deployable system.

Future work will focus on reducing dependence on a single teacher model, improving cross-domain and multi-anatomy generalization, incorporating denser geometric validation and more realistic surgical artifacts, and accelerating inference through learned latent initialization, temporal warm-starting, adaptive early stopping, and hybrid encoder–optimization strategies.

Acknowledgement: Not applicable.

Funding Statement: Supported by Chengdu Science and Technology Program [2026-YF08-00034-GX].

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Legend Zhang and Bo Yang; methodology, Legend Zhang and Guan Yao; software, Junmin Lyu and Guan Yao; validation, Junmin Lyu and Wei Wei; formal analysis, Jiawei Tian and Junmin Lyu; investigation, Guan Yao and Wei Wei; resources, Jiawei Tian and Bo Yang; data curation, Legend Zhang and Guan Yao; writing—original draft preparation, Legend Zhang; writing—review and editing, Jiawei Tian and Bo Yang; visualization, Legend Zhang and Junmin Lyu; supervision, Jiawei Tian and Bo Yang; project administration, Bo Yang; funding acquisition, Bo Yang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset analyzed in this study is publicly available. The *In vivo* and Phantom datasets can be accessed through <https://hamlyn.doc.ic.ac.uk/vision/>.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sui X, Zhang Y, Zhao X, Tao B. Binocular-based dense 3D reconstruction for robotic assisted minimally invasive laparoscopic surgery. *Int J Intell Robot Appl.* 2024;8(4):866–77. doi:10.1007/s41315-024-00390-7.
2. Tian J, Zhou Y, Chen X, AlQahtani SA, Zheng W, Chen H, et al. A novel self-supervised learning network for binocular disparity estimation. *Comput Model Eng Sci.* 2025;142(1):209–29. doi:10.32604/cmesci.2024.057032.
3. Wang Y, Sun Q, Liu Z, Gu L. Visual detection and tracking algorithms for minimally invasive surgical instruments: a comprehensive review of the state-of-the-art. *Robot Auton Syst.* 2022;149(Suppl 1):103945. doi:10.1016/j.robot.2021.103945.
4. Pedrett R, Mascagni P, Beldi G, Padoy N, Lavanchy JL. Technical skill assessment in minimally invasive surgery using artificial intelligence: a systematic review. *Surg Endosc.* 2023;37(10):7412–24. doi:10.1007/s00464-023-10335-z.
5. Rivas-Blanco I, Pérez-Del-Pulgar CJ, García-Morales I, Muñoz VF. A review on deep learning in minimally invasive surgery. *IEEE Access.* 2021;9:48658–78. doi:10.1109/ACCESS.2021.3068852.
6. Richa R, Poignet P, Liu C. Three-dimensional motion tracking for beating heart surgery using a thin-plate spline deformable model. *Int J Robot Res.* 2010;29(2–3):218–30. doi:10.1177/0278364909356600.
7. Frisken S, Luo M, Machado I, Unadkat P, Juvekar P, Bunevicius A, et al. Preliminary results comparing thin plate splines with finite element methods for modeling brain deformation during neurosurgery using intraoperative ultrasound. *Proc SPIE Int Soc Opt Eng.* 2019;10951(4):1095120. doi:10.1117/12.2512799.
8. Zalzal K, Selladurai S, Rossa C. Real-time simulation of ultrasound image deformation using thin plate spline. In: *Proceedings of the 2025 IEEE Sensors Applications Symposium (SAS); 2025 Jul 8–10; Newcastle, UK.* p. 1–6. doi:10.1109/sas65169.2025.11105189.
9. Wang Z, Li S, Peng J, Tai Y, Yu Z. Viscoelastic cluster-constrained PBD-based soft tissue behavior and interactive media applications for surgical simulation. *IEEE Trans Multimed.* 2025;27(1):2206–20. doi:10.1109/TMM.2024.3521762.
10. Xu H, Zhang J. AANet: adaptive aggregation network for efficient stereo matching. In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA.* p. 1956–65. doi:10.1109/cvpr42600.2020.00203.
11. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. Beta-VAE: learning basic visual concepts with a constrained variational framework. In: *Proceedings of the International Conference on Learning Representations 2017 (ICLR); 2017 Apr 24–26; Toulon, France.*
12. de Souza VLT, Marques BAD, Batagelo HC, Gois JP. A review on generative adversarial networks for image generation. *Comput Graph.* 2023;114(1):13–25. doi:10.1016/j.cag.2023.05.010.
13. Krichen M. Generative adversarial networks. In: *Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT); 2023 Jul 6–8; Delhi, India.* p. 1–7. doi:10.1109/ICCCNT56998.2023.10306417.
14. Chakraborty T, Reddy KSU, Naik SM, Panja M, Manvitha B. Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art. *Mach Learn Sci Technol.* 2024;5(1):011001. doi:10.1088/2632-2153/ad1f77.
15. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA.* p. 4396–405. doi:10.1109/cvpr.2019.00453.
16. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA.* p. 8107–16. doi:10.1109/cvpr42600.2020.00813.
17. Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, et al. Alias-free generative adversarial networks. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems; 2021 Dec 6–14; Online.* p. 852–63.
18. Zhang S, Zhang X, Wan S, Ren W, Zhao L, Shen L. Generative adversarial and self-supervised dehazing network. *IEEE Trans Ind Inform.* 2024;20(3):4187–97. doi:10.1109/TII.2023.3316180.

19. Yang B, Xu S, Yin L, Liu C, Zheng W. Disparity estimation of stereo-endoscopic images using deep generative network. *ICT Express*. 2025;11(1):74–9. doi:10.1016/j.ict.2024.09.017.
20. Xu G, Xu S, Lu S, Liu Y, Yang B, Lyu J, et al. Encoder-guided latent space search based on generative networks for stereo disparity estimation in surgical imaging. *Comput Model Eng Sci*. 2025;145(3):4037–53. doi:10.32604/cmesci.2025.074901.
21. Stoyanov D, Mylonas GP, Deligianni F, Darzi A, Yang GZ. Soft-tissue motion tracking and structure estimation for robotic assisted MIS procedures. In: *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2005*; 2005 Oct 26–29; Palm Springs, CA, USA. p. 139–46. doi:10.1007/11566489_18.
22. Stoyanov D, Scarzanella MV, Pratt P, Yang GZ. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2010*; 2010 Sep 20–24; Beijing, China. p. 275–82. doi:10.1007/978-3-642-15705-9_34.
23. Pratt P, Stoyanov D, Visentini-Scarzanella M, Yang GZ. Dynamic guidance for robotic surgery using image-constrained biomechanical models. In: *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2010*; 2010 Sep 20–24; Beijing, China. p. 77–85. doi:10.1007/978-3-642-15705-9_10.
24. Wang Z, Zheng H, He P, Chen W, Zhou M. Diffusion-GAN: training GANs with diffusion. In: *Proceedings of the International Conference on Learning Representations 2023 (ICLR)*; 2023 May 1–5; Kigali, Rwanda.
25. Mescheder L, Geiger A, Nowozin S. Which training methods for GANs do actually converge?. In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*; 2018 Jul 10–15; Stockholm, Sweden. p. 3478–87.
26. Cheng D, Dong Q, Yang Y, Wang Y, Zhu R, Zheng Y. Semi-supervised medical image segmentation method via dual-view graph contrastive learning and latent space uncertainty rectification. *Eng Appl Artif Intell*. 2026;177:114901. doi:10.1016/j.engappai.2026.114901.